情报杂志
Journal of Intelligence
ISSN 1002-1965,CN 61-1167/G3

《情报杂志》网络首发论文

题目：基于潜在影响力预测和多源信息融合的新兴技术识别方法
作者：张甜，陈进东，周晓纪，孙胜凯，张永伟
网络首发日期：2025-03-28
引用格式：张甜，陈进东，周晓纪，孙胜凯，张永伟．基于潜在影响力预测和多源信息融合的新兴技术识别方法[J/OL]．情报杂志.
https://link.cnki.net/urlid/61.1167.G3.20250328.1353.002

网络首发时间：2025-03-28 16:38:26　网络首发地址：https://link.cnki.net/urlid/61.1167.G3.20250328.1353.002

基于潜在影响力预测和多源信息融合的新兴技术识别方法*

张　甜[1]　陈进东[1,3]　周晓纪[2]　孙胜凯[2]　张永伟[2]

（1. 北京信息科技大学管理科学与工程学院　北京　102206；
2. 中国航天系统科学与工程研究院　北京　100048；
3. 智能决策与大数据应用北京市国际科技合作基地　北京　102206）

摘　要：[**研究目的**]针对新兴技术识别在前瞻性预测以及单一数据源等方面的不足，提出基于潜在影响力预测和多源信息融合的新兴技术识别方法。[**研究方法**] 首先，从“科学-技术”视角构建影响力评估指标体系，提出基于深度学习模型 Bi-LSTM 的潜在影响力预测方法，识别未来短期、中期、长期具有高影响力的论文和专利；其次，利用 LDA 模型提取研究主题，聚类合并科学主题和技术主题，并基于主题演化网络和主题共现网络识别新兴技术；最后，通过新闻数据验证本文方法的有效性，并结合情感分析挖掘公众诉求。[**研究结果/结论**] 以碳中和领域为例，基于本文提出的新兴技术识别方法，识别得到未来短期、中期、长期新兴技术共 7 项，实验结果验证了潜在影响力预测方法在识别高影响力研究中的有效性，以及融合多源信息的新兴技术识别方法的准确性。

关键词：新兴技术识别；多源数据；潜在影响力预测；主题分析；碳中和

中图分类号：G255.53　　**文献标识码：**A

Emerging Technology Identification Method Based on Potential Impact Prediction and Multi-source Information Fusion

Zhang Tian[1]　Chen Jindong[1,3]　Zhou Xiaoji[2]　Sun Shengkai[2]　Zhang Yongwei[2]

（1. School of Management Science and Engineering, Beijing Information Science and Technology University, Beijing　102206；
2. China Aerospace Academy of Systems Science and Engineering, Beijing　100048；
3. Beijing International Science and Technology Cooperation Base for Intelligent Decision Making and Big Data Application, Beijing　102206）

Abstract: [**Research purpose**] In response to the shortcomings of emerging technology identificationin forward-looking prediction and single data sources, this study proposes an emerging technology identification method based on potential impact prediction and multi-source information fusion. [**Research method**] Firstly, an impact assessment index system is constructed from the "science-technology" perspective, and a potential impact prediction method based on a Bi-LSTM deep learning model is proposed to identify papers and patents with high impact in the short, medium, and long term. Secondly, LDA model is employed to extract research topics, which are then clustered into science-technology topics. Emerging technologies are identified through topic evolution networks and co-occurrence networks.

基金项目：**国家自然科学基金项目“多维视角下的技术预见效用分析与方法优化”**（编号：L2224058）、**国家自然科学基金项目“基于多源数据融合的未来情景生成方法及实证研究”**（编号：L2324224）、**北京市属高等学校优秀青年人才培育计划项目 “中小微企业综合质量智能服务与优化技术研究”**（编号：BPHR202203233）**研究成果**。

作者简介：**张　甜**，女，2000 年生，硕士研究生，研究方向：技术预见；**陈进东**，男，1983 年生，博士，研究员，研究方向：技术预见；**周晓纪**，女，1968 年生，硕士，研究员，研究方向：技术预见；**孙胜凯**，男，1986 年生，硕士，研究员，研究方向：技术预见；**张永伟**，男，1992 年生，硕士，工程师，研究方向：技术预见。

通信作者：**陈进东**

Finally, the proposed method is validated using news data, and public demands are explored through sentiment analysis. [**Research result/conclusion**] Taking the field of carbon neutrality as an example, the proposed emerging technology identification method identified seven emerging technologies in the short, medium, and long term. Experimental results validate the effectiveness of the potential impact prediction method in identifying high-impact research, as well as the accuracy of the emerging technology identification method based on multi-source information.

Key words: emerging technology identification; multi-source data; potential impact prediction; topic analysis; carbon neutrality

新兴技术最早由沃顿商学院提出并定义为基于科学基础的可能建立新行业或改变旧行业的创新[1]，Rotolo 指出其具有新颖性、相对较快的增长、连贯性、突出的影响和不确定性[2]五大特性。新兴技术代表了最新的科学发现和技术突破，具有引领行业创新的潜力，及早识别和评估这些技术有助于提前布局和应对未来的技术变革。另外，随着基础学科新兴技术的涌现以及政府对科技创新活动的政策支持，高效准确地识别新兴技术也成为国家和企业进行科技创新的关键环节。

为了有效识别和追踪新兴技术的发展趋势，常用的识别未来新兴技术方法包括专家经验法、文献计量、网络分析、文本挖掘等方法。但相关方法大多是对过去或当前状况的衡量和评估，而非面向未来重要性预测的研究[3]。此外，大多数研究采用单一数据源进行技术识别，研究视野具有一定局限性，融合多源数据识别新兴技术的方法越来越受到关注。通过融合学术论文、专利和新闻报道等数据，可以更全面、准确地捕捉新兴技术的发展态势和市场潜力[4]。

因此，为克服现有研究在前瞻性预测上的不足以及单一数据源的限制，本文提出一种基于潜在影响力预测和多源信息融合的新兴技术识别方法。首先基于深度学习模型预测并识别出在未来具有高影响力的研究，其次利用主题提取和网络分析识别新兴技术，最后验证本文方法有效性以及分析公众诉求。另外，由于碳中和是我国应对环境资源制约、实现中华民族永续发展的重大战略决策[5]，为抓住科技革命和产业革命的先机，我国必须加强绿色科技创新，所以本文选取碳中和领域作为案例研究对象。

1 相关研究

1.1 技术识别方法

相较于文献计量、网络分析等其它定量技术识别方法，利用文本挖掘识别新兴技术的方法因其识别的高效率和高准确率，受到广泛应用，主要包括 SAO 结构(Subject-Action-Object)抽取、主题建模和聚类分类等方法。目前使用最为广泛的主题模型是 Blei 等于 2010 年提出的潜在狄利克雷分布(LDA, Latent Dirichlet Allocation)[6]，Chen 等基于 LDA 主题模型挖掘追踪中国 3D 打印技术的发展[7]。在利用机器学习模型识别技术的研究中，已经从传统的机器学习模型逐步向深度学习模型转变。胡泽文等基于 LSTM、BP 神经网络等模型对区块链技术进行主题分类[8]，赵雪峰等组合 LSTM、Word2Vec 及 BERT 模型将技术识别精准度提高至 88.1%[9]。

然而，上述方法大多属于对当前或过往数据的回溯性分析，较少涉及面向未来的预测性研究。即便在预测性研究中，仍存在两个主要问题：一是仅针对单一时间段进行预测，忽略了技术发展的生命周期。例如，冯立杰等利用 Bi-LSTM 模型训练 2017 年至 2022 年间专利指标与技术影响力之间的关系，以预测 2026 年的候选颠覆性技术[10]；二是研究采用的模型存在不足，例如 Li 等采用传统的机器学习算法从学术论文中发现潜在的突破性研究[11]，但是传统的机器学习模型存在准确率低等问题，深度学习模型则在处理复杂数据和发现潜在模式方面表现出显著优势。

鉴于上述局限性，本文提出一种面向未来不同时期的潜在影响力预测方法，将技术识别问题转化为基于深度学习模型的预测问题，根据相关指标预测研究在未来不同时间段的影响力，即采用深度学习模型预测研究在未来 3 年内、5 年内、10 年内的影响力高低类别。

1.2 技术识别数据源

论文和专利是识别新兴技术的重要数据源。使用论文数据可以在研究早期就捕捉到技术发展动态，如 Lobanova 等使用 SciBERT 模型识别论文数据中的创新趋势[12]，但单独依据论文数据识别出的“新兴技术”还不具备市场化的能力，只能视作新兴研究领域[13]。专利相对于论文来说则是从研究过渡到应用，是技术信息的载体。如 Lee 等从韩中日欧等国家专利数据库中提取有前景的 IPC 技术领域[14]。随着科学与技术之间的融合程度不断加深，两者之间呈现出耦合、共生、相互促进和共同发展的特征[15]，因此近年来也有研究将论文和专利代表的“科学-技术”作为整体识别新兴技术，如马亚雪等基于基因工程领域的学术论文与专利，研究专利引证论文的知识特征对颠覆性技术的预见能力[16]。

随着文本挖掘技术的发展，技术识别数据源已经

由“调研数据”向“结构化著录信息”到向“短文本”过渡发展[17],除去技术预见工作中依赖的论文和专利数据,行业报告、网络新闻、社交媒体等数据也受到学界关注。另外,新闻数据中包含了大量公众的情感态度和反应,这些情感态度反映了社会各界对新兴技术的认可度和接受度[18]。通过情感分析技术,可以挖掘出公众对某项技术正面、负面或中立的看法,揭示技术的市场潜力和社会影响。

为了克服单一数据带来的系统偏差,多源数据融合逐渐成为技术识别的研究趋势,有助于研究人员在综合考虑多方情况下识别新兴技术。如谭晓等构建了科学-技术-市场模型识别颠覆性技术[19],苗红等基于论文、专利、科技报告等数据源识别前沿技术[20]。但多源数据存在格式不统一、层次多样等问题,对于不同类型数据融合难度较大,本文提出一种多源数据融合研究思路即利用论文和专利数据识别新兴技术,利用新闻数据对识别结果加以验证。

2 研究方法

本文基于论文、专利和新闻数据,结合深度学习模型与主题建模、网络分析和情感分析等方法,识别并验证未来新兴技术,如图 1 所示。首先,从 WOS 和 incoPat 下载相关领域的论文和专利数据,去除无关及缺漏数据,利用 NLTK 工具对论文和专利数据的标题和摘要进行预处理;从科学性、关联性和技术性三方面构建影响力评估指标体系,训练并评估深度学习模型 Bi-LSTM,识别出在未来短期、中期、长期具有高影响力的研究。其次,利用主题模型分别对潜在高影响力的论文和专利研究提取主题和关键词,再对论文主题与专利主题进行 K-means 聚类合并,用合并后的“科学-技术”主题和关键词字段构建主题演化网络以及主题共现网络,从技术的新颖性、连贯性、影响性出发,通过主题随时间演化关系以及主题的共现强度识别出新兴技术。最后,从 China Daily 网站检索相关领域新闻,提取新闻话题并与新兴技术识别结果进行对比;收集新兴技术识别结果相关新闻数据,通过情感分析和关键词提取挖掘公众诉求。

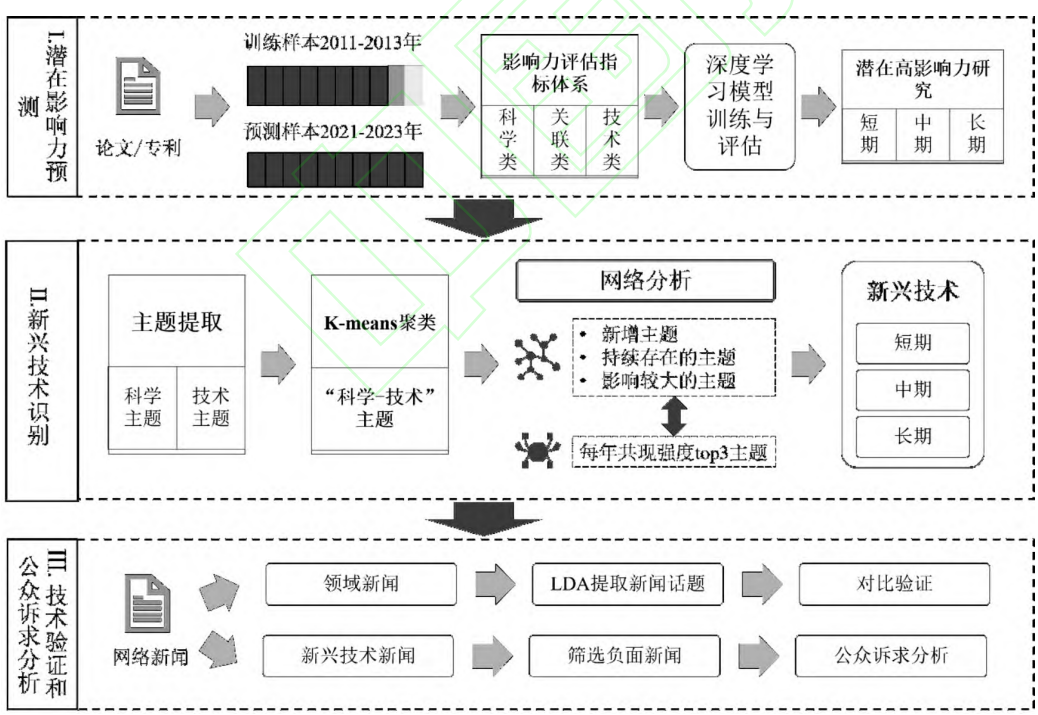


图 1 研究框架图

2.1 潜在影响力预测

研究的潜在影响力预测流程如图 2 所示。首先从科学、技术以及科学与技术的关联性三个方面构建指标体系,对训练样本(2011-2013 年)和预测样本(2021-2023 年)分别提取并计算相关指标值。同时,计算训练样本在未来短期、中期和长期的被引频次及其上四分位数,将训练样本划分为高被引研究和低被引研究两类。随后将这些提取的指标值输入至 Bi-LSTM 模型进行训练,模型训练完成后,基于预测样本的指标值,模型可以输出其影响力高低类别预测结果。未来短期、中期和长期的潜在高影响力研究用于后续技术识别。

2.1.1 指标体系构建

前期研究大多基于单一数据源(论文或者专利)

属性来识别新兴技术，较少有研究同时采用论文和专利数据并考虑到论文与专利属性之间的关联关系，本文从科学、科学关联技术、技术三方面十一个指标构建影响力评估指标体系，如表 1 所示。基于表 1 中的科学类指标，提取论文数据相关指标作为输入，以论文未来影响力高低为输出。同理，基于“科学–技术”和技术类指标提取专利数据相关指标作为输入，以专利未来影响力高低为输出。

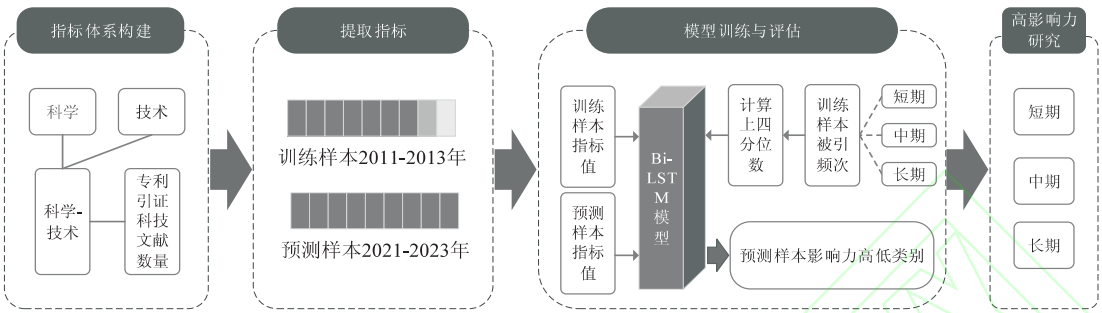


图 2　潜在影响力预测流程图

研究表明，被引频次与研究的影响力显著正相关[21]，因此本文采用论文和专利的被引频次来衡量其影响力。但是研究的实际被引频次可能受到多种因素影响，鉴于本文构建的指标体系未必考虑到所有因素，所以本文并不关注研究在未来的具体被引频次，而是根据研究的被引频次将其划分为高影响力研究和低影响力研究两类。通过类别划分可以有效减小原始数据中可能存在的较大波动和噪声影响，模型复杂度的简化也有利于提高模型的稳定性和预测效果。具体而言，本文将被引频次大于上四分位数的研究视为高影响力研究，其余研究视为低影响力研究，也就将预测问题转换为二分类问题[22]。此外，本文分别预测了研究在 3 年、5 年和 10 年后的影响力高低类别，以此识别未来短期、中期和长期内的潜在高影响力研究。

表 1　影响力评估指标体系

指标维度	指标名称
科学类指标	合著者数量
	180 天使用量
	引证次数
	所属领域类别数量
	有无受到资助
“科学–技术”关联类指标	专利引证科技文献量
技术类指标	所属 IPC 类别数
	引证次数
	权利要求数量
	发明人数量
	简单同族个数

a. 科学类指标

科学类指标从论文作者和论文本身两方面考虑，选用合著者数量、180 天使用量、引证次数、所属领域类别数量、有无受到资助 5 个指标进行衡量。合著者数量是研究力量与强度的体现，合著者数量越多说明研究受到的关注越多，越容易促进知识流动和产生高水平的创新性研究成果。引证次数越高，表明该论文在学术界的认可度和传播度越广，反映了其学术价值和影响力[23]。复合影响因子越大以及 180 天使用量越高，说明论文影响力越大。所属领域类别越多，意味着该研究的跨学科性和广泛应用前景，能够在更多领域产生影响。受到资助的论文通常有更好的资源支持，研究质量和影响力可能更高，与未受资助的论文相比，更有可能产生重要的科研突破。

b. “科学–技术”关联类指标

在技术识别相关研究中，常用论文表示科学知识，用专利表示技术知识，用专利与论文在作者、主题、引用之间的关系来表示科学和技术之间的关联，以反映科学知识流向技术知识，由于本文同时采用了论文和专利数据，因此在指标选取时还考虑了科学与技术之间的关联关系[24]。技术创新与科学和技术之间的紧密结合相关，利用专利的“科学–技术”知识关联特征可以区分高影响性研究与一般研究[25]。专利引证科技文献量越多，说明该技术与科学联系越紧密，就越有可能开发出具有创新性和影响性的新兴技术。

c. 技术类指标

所属 IPC 类别数、权利要求数量和简单同族个数都反映专利涉及的范围，所属 IPC 类别数越多说明专利涉及范围越广[26]；权利要求数量越多说明专利保护的技术越多，专利质量越高，创新性越高；简单同族个数越多说明该专利越重要且具有较高的技术价值和市场竞争力。引证次数越多则说明该专利越具有较高的技术质量。发明人数量反映了专利发明的合作情况，发明人数量越多越有可能研制出新兴技术。

2.1.2　模型训练与预测

首先，潜在影响力预测模型构建。由于 2011 至 2013 年这三年的论文和专利数据在未来三年、五年、十年的被引频次均为已知，选用这三年的数据训练影响力评估指标体系与被引频次高低类别之间的关系。对于输入特征，根据前文构建的指标体系，从 2011—

2013 数据集中提取并计算相关指标值。对于类别标签,根据数据集中每条数据的被引数据发表年份,人工计算每条数据在未来短期(三年内)、中期(五年内)和长期(十年内)的被引频次,根据被引频次上四分位数将其划分为高影响力研究和低影响力研究两类。

LSTM(Long Short-Term Memory)模型是一种特殊的递归神经网络(RNN),通过"门控"机制捕捉时间序列数据中的长短期依赖,避免信息在长序列中逐渐丢失的问题[27]。Bi-LSTM(Bidirectional Long Short-Term Memory)在 LSTM 的基础上,通过双向传播的方式同时处理数据的前后依赖关系,使得模型能够同时分析序列中的过去和未来信息,进一步提升对研究的影响力高低类别预测能力[28]。

模型训练公式如下:

$$h_t = BiLSTM(x_t) \tag{1}$$

其中,x_t 为输入的影响力评估指标值,h_t 为模型输出的状态表示,经过双向传播后,模型能够捕捉输入特征和类别标签之间的复杂关联关系。在模型训练时,目标是最小化损失函数 L,即:

$$L = \sum_{t=1}^{T} loss(h_t, y_t) \tag{2}$$

y_t 为训练样本的影响力高低类别标签。采用准确率、精确率、召回率和 F1 值指标评估模型的训练效果,并将结果与传统机器学习模型(如支持向量机、逻辑回归和随机森林)以及深度学习模型(如人工神经网络)的运行效果进行对比。

其次,依据上述训练模型预测研究在未来的影响力高低类别。2021 年至 2023 年这三年的论文和专利数据在未来三年、五年、十年的被引频次均为未知,所以预测这三年数据在未来短期、中期、长期的被引频次高低类别。提取 2021—2023 年数据集中影响力评估指标,将这些指标值输入训练完成的 Bi-LSTM 模型中,输出研究在未来短期、中期和长期被引频次所属高低类别,将未来高被引研究视为潜在高影响力研究,用于后续研究。

2.2 新兴技术识别

本文利用主题提取、聚类和网络分析的方法识别新兴技术,具体流程如图 3 所示。

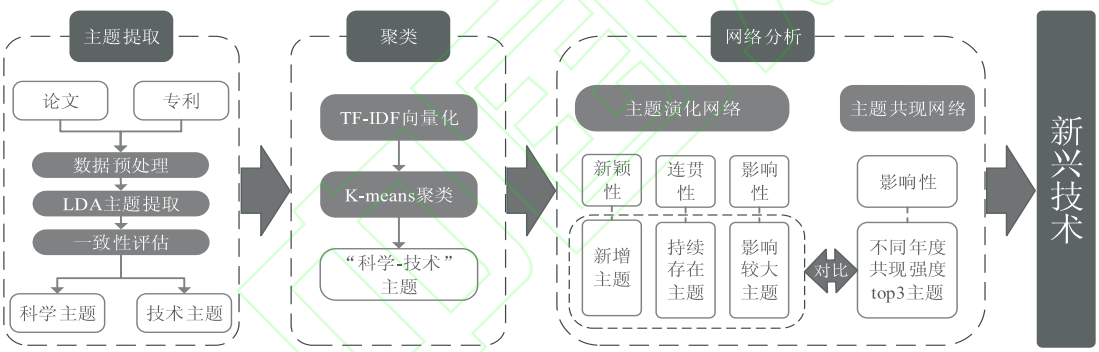


图 3 新兴技术识别流程图

2.2.1 主题提取和合并

如图 3 所示,针对论文和专利两种异构数据,分别对它们进行主题提取。在正式处理数据前首先合并标题和摘要创建语料库用作后续研究。利用 NLTK 工具对语料库进行分词,同时采用英文停用词表和自定义词表,剔除高频但无实际意义词语,加载用户自定义词典防止目标领域的部分专有名词被切割,保证数据有效性。预处理后,使用 LDA 模型提取文本主题。然而,LDA 模型无法自动确定最优主题数,常用的主题模型性能评价指标包括一致性和困惑度。主题一致性评价主题模型中单词的语义一致性,一致性越高说明模型主题越具有可解释性,因此本文使用主题一致性来确定 LDA 模型的最优主题数。

为了整合学术研究与技术创新的成果以及避免后续重复研究,将论文主题和专利主题进行聚类合并。对论文与专利主题关键词进行 TF-IDF 向量化,以便提取文本数据中的重要特征。采用 K-means 聚类算法对向量化处理后的主题进行聚类分析,得到合并后的"科学-技术"主题用于后续研究。

2.2.2 基于网络分析识别新兴技术

如图 3 所示,主题演化网络考虑了时间因素,从动态视角识别新兴技术,能够反映主题随时间推移的演化进程。对合并后的"科学-技术"主题构建主题演化网络,研究主题随时间产生和消亡的演化趋势,利用新兴技术的新颖性(Novelty, N)、连贯性(Coherence, C)和影响性(Impact, I),筛选出新增主题、持续存在的主题以及影响较大的主题作为候选新兴技术[29]。其中节点为主题,连线为主题之间的相似度,相似度值越大则边越粗,节点的大小反应其与其相邻时间窗口节点的关联,关联越大则节点越大。

从新兴技术的"新颖性"特征出发,将未出现在主题演化网络中的主题视为新增主题(首年除外)。定义是否为新增主题的公式如下:

是否新增主题(NT) =

$$\begin{cases}1, & \text{if } T(t)_i \notin \{T(t-1)_j\} \text{ and } t > t_0 \\ 0, & otherwise\end{cases} \quad (3)$$

其中$T(t)_i$为第t年出现的主题，$\{T(t-1)_j\}$为第$t-1$年出现的主题，t_0为主题演化网络的起始年份，未继承上一年的主题且首次出现的主题$T(t)_i$则被视为新增主题，可能反映新的研究趋势和技术发展。

从新兴技术的“连贯性”特征出发，如果某些主题在多年的演变中一直出现，则表明其受到持续关注。可以通过以下公式衡量：

$$连贯性(C) = \frac{N}{L} \quad (4)$$

其中$C(T_i) \in [0,1]$，表示主题T_i的连贯性，N为T_i在整个时间窗口内出现的年份数，L为时间窗口总长度，本文将C为1的主题视作持续受到关注的主题。

从新兴技术的“影响性”特征出发，如果某些年份的主题在连接线上显示出较大的流量，则表示这些主题具有相对较高的影响力，指标定义如下：

$$影响力(I) = \sum_{j=1}^{n_i} 流量(T_i, T_j) \quad (5)$$

其中T_i为目标主题，T_j为与其关联的其他主题，流量(T_i, T_j)为T_i与T_j之间的相似度值，n_i为T_i的关联主题数量。本文将影响力排名前五的主题视作影响力较大的主题。

结合上述三个指标，可以筛选出新兴技术，但利用主题演化网络直接确定新兴技术仍有一定偏差，如上述将网络中未出现的主题视作的新增主题，也可能是因为与其他主题关联程度较小而未出现在网络中。新兴技术的影响性是其显著特征之一，因此将通过主题演化网络识别出的技术视作候选新兴技术，结合主题共现网络分析节点的共现强度，进一步评估技术的影响力，以此确定最终的新兴技术。主题共现网络以主题为节点，节点的大小为中心度，边的粗细为共现频次大小，以主题相应关键词的共现关系来表示主题的共现关系。采用固定时间段即以年为单位划分“科学-技术”主题，构建基于时序的主题共现网络，筛选出每年共现强度排名前三的技术，与主题演化网络识别出的候选新兴技术进行对比，确定最终未来的新兴技术。

2.3 技术验证和公共诉求分析

首先，基于新闻数据的新兴技术识别结果验证。过往的研究在识别出新兴技术后，往往缺乏对这些识别结果的有效验证[3]。本文提出通过对比前文识别出的新兴技术与新闻话题，评估和验证所提出研究方法的有效性。在 China Daily 新闻网站检索领域相关新闻，利用 LDA 主题模型对这些新闻数据进行主题提取，识别出其中主要话题，这些话题代表了公众和媒体对某个技术领域的关注热点。将这些提取出的新闻话题与前文通过论文和专利数据挖掘出的“科学-技术”主题进行对比分析，从而验证本文方法的有效性。

其次，基于新闻数据的公众诉求分析。网页新闻既包含公众对新兴技术的实时反应，也包含他们对新兴技术未来发展的期待和诉求。在新闻网站检索上述研究确定的新兴技术并对每个新兴技术相关新闻进行情感分析和关键词提取，以此挖掘公众诉求。以上述识别出的未来新兴技术为关键词，在 China Daily 中检索技术相关新闻。

BERT（Bidirectional Encoder Representations from Transformers）模型是一种基于 Transformer 的预训练语言模型，通过双向编码器机制，能够从输入文本中捕捉前后文信息[30]。BERT 在大量无监督文本数据上预训练得到丰富的上下文表示，因此在文本分类、情感分析等下游任务中具有良好的泛化能力。结合 BERT 的语义理解能力，本文利用微调后的 BERT 模型计算每个技术的情感得分，以揭示公众对不同新兴技术的情感态度。具体而言，对于给定的文本输入x，微调后的 BERT 模型通过特征提取生成情感向量表示，经过全连接层及激活函数计算情感得分s：

$$s = Sigmoid(MLP(BERT(x))) \quad (6)$$

其中，$s \in [0,1]$表示情感得分，当s接近 1 时表示正面情感，接近 0 时表示负面情感。鉴于负面新闻往往更能揭示相关事件或问题的深层次原因，本文更关注负面新闻内容。为进一步分析负面新闻内容，本文采用词频逆向文件频率（Term Frequency-Inverse Document Frequency，TF-IDF）算法提取负面新闻中的关键词，TF-IDF 算法能够有效地突出负面新闻中重要的关键词，标识出公众诉求所在。

3 案例研究-以碳中和领域为例

本文选取碳中和领域进行实证分析，详细解释方法流程并验证其有效性，识别碳中和领域的新兴技术，助力我国实现双碳目标和绿色可持续发展。

3.1 碳中和领域数据收集及预处理

WOS 是全球最大的外文数据库之一，具有信息含量大、覆盖范围广等优点，因此本文在 WOS 检索相关领域论文数据，收集数据字段包括标题、摘要、作者、180 天使用量、被引次数、所属领域等。incoPat 是一家全球知识产权服务商，涵盖全球专利，在 incoPat 中下载专利及所属 IPC 类别数、引证次数、权利要求数量、发明人数量、简单同族个数等指标。

分别在 WOS 和 incoPat 中检索 2011—2013 年以及 2021—2023 年间主题为“carbon neutral *”的相关论文和专利研究，将 2011—2013 数据集用作模型训

练、验证与测试，将时间跨度在2021—2023数据集用作预测。在WOS核心数据库中下载论文数据，去除缺漏及无关数据共收集14842条有效数据；划分论文数据集，论文2011—2013数据集含有384条论文，论文2021—2023数据集含有14458条论文，再爬取论文2011—2013数据集的被引文献，共收集4431条。在incoPat专利库中下载专利数据，去除缺漏及无关数据共收集5560条有效数据；划分专利数据集，专利2011—2013数据集含有1516条专利，专利2021—2023数据集含有4044条专利，再爬取专利2011—2013数据集的被引专利，共收集9620条。不同年份的论文和专利数据数量如表2所示。

表2　不同年份数据量　　单位：条

年份	论文数据量	专利数据量
2011	126	483
2012	109	548
2013	149	485
2021	2655	1422
2022	5362	1548
2023	6441	1074

影响力评估指标体系中技术类指标、“科学-技术”关联类指标可通过incoPat数据库直接下载得到，科学类指标中的180天使用量、引证次数指标值可通过WOS数据库直接下载得到，而合著者数量、所属领域类别数量、有无受到资助指标以及短期、中期、长期被引频次指标值则通过人工计算相关数据获取。收集完所有指标数值后对各个指标值进行标准化处理。

3.2　碳中和领域研究的潜在影响力预测结果

计算2011—2013数据集中被引频次的上四分位数并视为阈值，不低于阈值的数据视为高影响力研究，低于阈值的数据视为低影响力研究。论文和专利数据在未来三年、五年、十年的高影响力类别与低影响力类别数量如表3所示。由表可知在所有时间段内，论文和专利的低影响力类别数量远高于高影响力类别数量，表明大多数研究成果和技术创新在后续时间内并未获得广泛关注或产生较大影响。利用贝叶斯优化调整Bi-LSTM模型参数，表4展示的是预测论文和专利在未来五年影响力高低类别的Bi-LSTM模型参数设置。

表3　训练集分类结果

	论　文		专　利	
	高	低	高	低
未来三年	84	300	359	1157
未来五年	83	301	311	1205
未来十年	89	295	338	1178

表4　中期预测Bi-LSTM模型参数设置

论文		专利	
参数	值	参数	值
Units	52	units	127
dropout_rate	0.1334	dropout_rate	0.405
learning_rate	0.0032	learning_rate	0.0001
num_layers	3	num_layers	1
Epochs	65	Epochs	54
batch_size	243	batch_size	227

与传统的机器学习模型和其他深度学习模型进行对比，分别计算不同模型的准确率、精确率、召回率、F1值，论文和专利数据的预测效果如表5和表6所示。

表5　论文数据不同模型预测效果

预测时间	模型	Accuracy /%	Precision /%	Recall /%	F1 Score /%
三年	SVM	87.18	43.59	50.00	46.58
	LR	84.21	71.67	77.08	73.73
	RF	76.92	52.27	52.65	52.37
	ANN	84.62	61.11	57.06	58.21
	Bi-LSTM	**92.11**	**86.97**	**81.77**	**84.06**
五年	SVM	76.32	68.89	64.64	65.9
	LR	84.62	61.11	57.06	58.21
	RF	79.49	60.61	60.61	60.61
	ANN	82.05	59.72	55.30	56.04
	Bi-LSTM	**89.47**	**80.21**	**80.21**	**80.21**
十年	SVM	81.58	68.77	75.52	70.93
	LR	82.05	59.72	55.30	56.04
	RF	79.49	53.79	55.36	54.12
	ANN	84.62	61.11	57.06	58.21
	Bi-LSTM	**86.84**	**75.35**	**78.65**	**76.80**

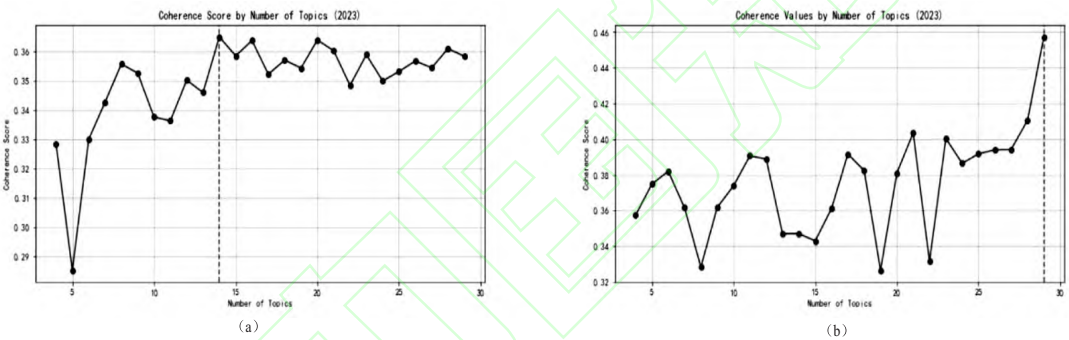
表6　专利数据不同模型预测效果

预测时间	模型	Accuracy /%	Precision /%	Recall /%	F1 Score /%
三年	SVM	69.54	45.62	49.31	43.03
	LR	66.89	58.46	57.96	58.15
	RF	67.76	53.57	51.73	49.75
	ANN	71.52	35.76	50.00	41.70
	Bi-LSTM	**74.17**	**67.62**	**64.45**	**65.41**
五年	SVM	71.52	35.76	50.00	41.70
	LR	70.20	48.21	49.77	43.30
	RF	69.08	51.20	50.66	49.26
	ANN	69.54	57.16	53.31	52.16
	Bi-LSTM	**71.52**	**62.80**	**58.40**	**58.76**
十年	SVM	68.87	59.69	57.95	58.31
	LR	68.21	56.90	54.68	54.45
	RF	69.74	56.90	52.44	49.65
	ANN	68.87	48.16	49.55	44.56
	Bi-LSTM	**70.86**	**62.53**	**60.03**	**60.62**

由表 5 和表 6 可知，从不同时间段的预测效果来看，Bi-LSTM 模型在短期和中期预测中表现优异，尤其是在短期能够提供相对较高的准确率和 F1 值。在长期预测中模型性能略低于短期和中期，表明模型在长期预测中的平衡性需要进一步优化。从不同模型运行效果对比来看，相较于 SVM、LR、RF、ANN 等模型，Bi-LSTM 模型在处理论文和专利数据的预测任务中表现突出，证明其在复杂数据分析中的有效性和优越性。利用 Bi-LSTM 模型对 2021—2023 数据集进行预测，论文和专利数据高影响力与低影响力分类结果如表 7 所示。

表 7　分类预测结果

	论　文		专　利	
	高	低	高	低
未来三年	4058	11231	432	3612
未来五年	731	14558	311	1205
未来十年	739	14550	338	338



（a）　（b）

图 4　论文（a）与专利（b）最佳主题数

对提取的论文和专利主题进行 TF-IDF 向量化处理，并通过计算轮廓系数确定最佳聚类数，利用 K-means 聚类实现论文主题和专利主题的合并。以 2023 年论文和专利数据为例，通过 LDA 模型一共提取出 43 个主题，如图 5 所示当合并后的“科学-技术”主题聚为 16 类时，轮廓系数最高即聚类效果最好。根据合并后的“科学-技术”主题关键词内容命名技术。

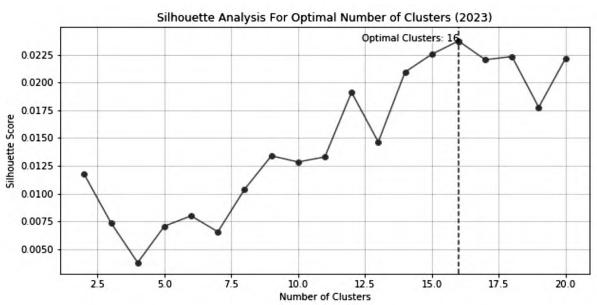


图 5　主题数与轮廓系数

对其他年份的数据进行同样操作。不同年份论文和专利提取出的主题数量以及合并后的主题数量如表 8 所示，对不同年份“科学-技术”合并主题命名如表 9 所示。

3.3　碳中和领域新兴技术识别

3.3.1　主题提取和合并

未来三年、五年和十年预测分别代表短期、中期、长期预测，因为不同时期数据处理过程相同，下文以中期为例解释数据具体的处理流程。分别对 2021—2023 数据集具体时间段中每年潜在高影响力研究提取主题，以 2023 年为例对主题提取和聚类部分进行说明。在提取主题前先对数据进行预处理，利用 NLTK 工具进行分词、去停用词和提取词干，通过主题一致性确定最佳主题数，利用 LDA 模型提取潜在高影响力研究中主题及对应关键词。图 4 中的图（a）和图（b）分别展示了提取论文数据和专利数据时，LDA 模型一致性随不同主题数量变化的情况。当论文主题数为 14、专利主题数为 29 时主题一致性最高，即此时 LDA 主题聚类效果最佳。

表 8　不同年份主题数量

年份	论文主题数	专利主题数	合并主题数	聚类主题数
2021	29	4	33	18
2022	4	13	17	17 *
2023	14	29	43	16

注：由于 2022 年主题数较少，所以不进行聚类处理。

表 9　不同年份合并后的“科学-技术”主题

年份	技术名称
2021	土壤改良技术、电催化技术、电解水技术、碳循环技术、碳排放监测技术、生物质能技术、碳管理技术、环境认证技术、农业碳管理技术、污水处理技术、碳足迹评估技术、温室气体监测技术、微生物碳循环技术、光伏技术、催化剂开发技术、可再生能源技术、废物管理技术、植物碳管理技术
2022	可再生能源技术、电催化技术、碳管理与环境技术、碳排放建模技术、碳吸附技术、废水处理技术、气体处理技术、气体分离与回收技术、生物质材料技术、能源与材料处理技术、复合材料处理技术、废渣回收技术、电化学材料技术、碳管理数据技术、催化剂准备技术、液体废物处理技术、污泥处理技术
2023	环境监测技术、电催化技术、材料处理技术、水处理技术、催化材料技术、电解质技术、纳米催化技术、生态保护技术、废物处理技术、碳捕集与存储技术、经济评估技术、智能监测技术、环境测量技术、碳排放计算技术、液体处理技术、生态修复技术

3.3.2 基于网络分析的新兴技术识别结果

本文使用余弦相似度计算相邻时间窗的主题向量相似度,利用 Python 中的 sankeny 库绘制主题随时间演化网络图。研究发现设定相似度阈值为 0.5 时,主题较为完整且连贯。根据碳中和领域从 2021—2023 年不同技术的演变,可以发现 2022 年新增可再生能源技术、废水处理技术、气体处理技术、气体分离与回收技术、生物质材料技术、能源与材料处理技术、催化剂技术,2023 年新增碳排放计算技术、生态修复技术。从 2021 年到 2023 年中,电催化技术、废水处理技术、复合材料技术持续存在,即这些技术一直在碳中和领域受到较多关注。废水处理技术、环境监测技术、材料处理技术、电解质技术、废物处理技术与其他技术关联较大,影响性较强。

然后根据 2021—2023 年每年的主题和关键词绘制主题共现网络图,并识别共现强度排名前三的技术,其中 2021 年共现强度排名前三的技术分别是生物质能技术、污水处理技术、电催化技术,2022 年共现强度排名前三的技术分别是废水处理技术、碳吸附技术、电化学材料技术,2023 年共现强度排名前三的技术分别是环境监测技术、环境测量技术、电解质技术。结合主题演化网络及主题共现网络分析,可以发现近三年内一直受到持续关注的电催化技术、废水处理技术,影响较大的环境监测技术、储能技术与不同年份共现强度排名前三的技术重合,因此将这四个技术视为未来中期内的新兴技术。

分别对短期和长期数据进行同样处理,可以得到未来短期内的新兴技术为废物处理与回收技术、电催化技术、废水处理技术、生物质能制备与利用技术,未来长期内的新兴技术为绿色催化技术、废水处理技术、电化学转化技术、废物处理与回收技术。值得注意的是绿色催化技术包括电催化技术,短期、中期、长期内都将废水处理技术视为新兴技术。与碳中和技术分类体系进行对比,其中大部分技术属于公认的对碳中和起到促进作用的技术[31],说明了本文所提方法的有效性。不同的是本文还识别出废水处理技术和环境监测技术为新兴技术。废水处理技术不仅可以减少水环境中的污染物排放,还可以回收利用废水中的有用物质,有助于实现碳中和目标。环境监测技术则能够实时监测和管理碳排放,确保环境质量的持续改善,并支持碳中和政策的有效实施。

3.4 碳中和领域技术验证与公众诉求分析

在 China Daily 中检索“carbon neutral”相关网页新闻,检索时间为 2024 年 5 月 29 日,去除重复和缺漏数据一共获得 2311 条有效新闻,时间跨度为 2020—2024 年。对数据经过预处理后利用 LDA 模型提取主题,一共识别出 21 个新闻话题,将这 21 个话题与前文识别的 7 项新兴技术对比,发现这 7 项技术均涵盖在公众关注话题中,因此验证了本文研究框架的有效性,也说明下文检索这 7 项新兴技术相关新闻并挖掘公众诉求是可行的,根据主题文档分布标识出这 7 项新兴技术相关新闻如表 10 所示。

表 10 新兴技术相应新闻

新兴技术	标志性新闻
废水处理技术	ADB offers China 300 mln USD loan for rural wastewater treatment
环境监测技术	Ministry to further strengthen environmental monitoring data accuracy
储能技术	China to boostbattery storage technology for clean energy
废物处理与回收技术	China makes progress in solid waste treatment
生物质能制备与利用技术	Biomass technology shows huge growth
绿色催化技术	Chinese scientists develop low – cost, eco – friendly propylene catalyst
电化学转化技术	China contributesto global green development with low–carbon tech innovation

根据前文确定的新兴技术,在 China Daily 网站中检索每个新兴技术,各个技术收集到的有效新闻数据量如表 11 所示。使用微调后的 BERT 模型计算每条新闻的情感得分,将低于情感分数均值的新闻视为负面新闻,统计不同新兴技术负面新闻比例如表 11 所示。

表 11 不同技术对应新闻数量

新兴技术	新闻数量/条	消极新闻占比/%
废水处理技术	1973	60.21
环境监测技术	1305	62.30
储能技术	**958**	**63.99**
废物处理与回收技术	413	58.84
生物质能制备与利用技术	1036	46.81
绿色催化技术	**92**	**36.96**
电化学转化技术	92	41.30

从表 11 可知,储能技术的消极新闻比例最高,达到 63.99%,而绿色催化技术的消极新闻比例最低,仅为 36.96%。这一差异可能源于储能技术的广泛应用及其潜在风险,容易引发更多消极报道;相对而言,绿色催化技术因其环保优势和较低的直接社会影响,受到更为正面的评价。

为了挖掘公众诉求,利用 TF–IDF 算法提取新兴技术负面新闻中的高频关键词。结果显示,首先,"energy" 在所有新兴技术负面新闻中都具有较高的权重,与能源相关的词语如"power"、"renewable"、"wind"、"solar"、"clean" 等词也频繁出现,表明在碳中和的讨论中,清洁和可再生能源是重要的关注点。其次,"technology" 和"innovation" 也是高频词,显示出技术进步

和创新在实现碳中和目标中的关键作用。相关词语如"digital"、"data"、"transformation"等也出现较多,表明数字化和技术变革是推动碳中和的重要手段。另外,"environmental"、"protection"和"quality"等词语反映了环境保护和提升环境质量在碳中和战略中的重要性。"company"、"industry"和"market"等词语的频繁出现,表明商业和市场力量在碳中和进程中的重要性。此外,"government"、"national"和"policy"等词汇的出现频率也很高,强调了政府和政策在实现碳中和目标中的领导作用。综上可以发现在这些负面新闻中,公众比较关注碳中和领域的清洁与可再生能源、技术创新、环境质量、企业参与、政府支持等方面,聚焦这些方面能更好地理解当前碳中和研究和实践的重点与挑战,早日实现碳中和目标。

4 结 语

本文构建了一种基于潜在影响力预测和多源信息融合的新兴技术识别研究框架,并在碳中和领域进行案例研究,对技术预见领域、碳中和领域都有一定参考价值。本文主要有以下两点贡献:首先,本文利用深度学习模型系统地拟合了影响力评估指标体系与被引频次高低类别之间的关系,属于对未来的前瞻性预测,可以为政策制定者提供未来战略支持。其次,本文综合考虑了多种数据来源,具体而言,既考虑了论文与专利数据之间的关联,又利用新闻数据进行了结果验证和公众诉求分析。

但本文研究仍然存在以下局限性:首先本文选用的预测指标为被引频次,但论文和专利的被引频次受到时间影响、以及可能并非被本领域相关研究引用,后续可以考虑选用其他能够衡量数据影响力的指标;其次本文虽然采用了多源数据进行分析,后续仍然可以考虑融入更多数据如社交媒体数据、行业报告等进行技术识别。

参 考 文 献

[1] Geroge S D, Paul J H S. 沃顿论新兴技术管理[M]. 石莹,等译. 北京:华夏出版社, 2002. 01.

[2] Daniele R, Diana H, Ben R M. What is an emerging technology? [J]. Research Policy,2015,44(10):1827-1843.

[3] 刘小玲,谭宗颖. 新兴技术主题识别方法研究进展[J]. 图书情报工作,2020,64(11):145-152.

[4] 杨思洛,江 曼. 新兴技术内涵特征和识别方法研究进展[J]. 情报科学,2023,41(5):181-190.

[5] 袁晓玲,金中国,李朝鹏. 中国实现碳中和:进程评估与实践困境[J]. 北京工业大学学报(社会科学版),2024,24(4):90-106.

[6] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research,2003,3: 993-1022.

[7] Chen W, Lin C R, Li C A Y, et al. Tracing the evolution of 3-d printing technology in China using lda-based patent abstract mining[J]. IEEE Transactions on Engineering Management, 2022,69(4):1135-1145.

[8] 胡泽文,王梦雅,韩雅蓉. 基于机器学习的中国区块链专利技术主题识别与自动分类研究[J]. 数字图书馆论坛,2023,19(12):32-43.

[9] 赵雪峰,吴德林,吴伟伟,等. 基于深度学习与多分类轮询机制的高质量"卡脖子"技术专利识别模型——以专利申请文件为研究主体[J]. 数据分析与知识发现,2023,7(8):30-45.

[10] 冯立杰,秦 浩,王金凤,等. 融合专利数据与社交媒体数据的潜在颠覆性技术识别——基于深度学习模型[J]. 情报学报,2024,43(2):181-197.

[11] Li X, Wen Y, Jiang J J, 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.

[12] Lobanova P, Bakhtin P, Sergienko Y. Identifying and visualizing trends in science, technology, and innovation using SciBERT[J]. IEEE Transactions on Engineering Management,2024, 71:11898-11906.

[13] 王 玏,吴新年. 新兴技术识别方法研究综述[J]. 图书情报工作,2020,64(4):125-135.

[14] Lee Y J, Han Y J, KimS S, et al. Patent data analytics for technology forecasting of the railway main transformer[J]. Sustainability,2023,15(1):1-15.

[15] 曹 琨,吴新年,白光祖,等. 基于"科学-技术"复杂网络的关键核心技术识别研究——以数控机床领域为例[J/OL]. 数据分析与知识发现,1-18[2024-12-23]. http://kns.cnki.net/kcms/detail/10.1478.G2.20240506.1342.002.html.

[16] 马亚雪,王嘉杰,巴志超,等. 颠覆性技术的后向科学引文知识特征识别研究——以基因工程领域为例[J]. 图书情报工作,2024,68(1):116-126.

[17] 张 硕,汪雪锋,乔亚丽,等. 技术预测研究现状、趋势及未来思考:数据分析视角[J]. 图书情报工作,2022,66(10):4-18.

[18] Li X, Xie Q, Huang L C. Identifying the development trends of emerging technologies using patent analysis and web news data mining: The case of perovskite solar cell technology[J]. IEEE Transactions on Engineering Management,2022,69(6),2603-2618.

[19] 谭 晓,西桂权,苏 娜,等. 科学—技术—项目联动视角下颠覆性技术识别研究[J]. 情报杂志,2023,42(2):82-91.

[20] 苗 红,王浩桐,李伟伟,等. 面向应用场景的前沿技术识别方法[J/OL]. 情报杂志,1-9[2024-12-23]. http://kns.cnki.net/kcms/detail/61.1167.G3.20240722.1713.004.html.

[21] 鲍玉芳,马建霞. 科学论文被引频次预测的现状分析与研究[J]. 情报杂志,2015,34(5):66-71.

[22] Lee C, Kwon O, Kim M, et al. Early identification of emerging technologies: A machine learning approach using multiple patent indicators[J]. Technological Forecasting and Social Change, 2018,127:291-303.

[23] 陶文倩,潘云涛,王海燕. 基于主题演化动态情境的高被引论文影响力形成模式探索[J]. 现代情报,2024,44(4):114-

126,153.

[24] 梁镇涛,毛　进,李纲. 融合“科学-技术”知识关联的高颠覆性专利预测方法[J]. 情报学报,2023,42(6):649-662.

[25] Liu X W, Wang X Z, Lyu L C, et al. Identifying disruptive technologies by integrating multi-source data[J]. Scientometrics,2022,127(9):5325-5351.

[26] Zhou Y, Dong F, Liu Y, et al. A deep learning framework to early identify emerging technologies in large-scale outlier patents: An empirical study of CNC machine tool[J]. Scientometrics,2021,126(2):969-994。

[27] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation,1997,9(8):1735-1780.

[28] Alex G, Jürgen S. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks,2005,18(5/6):602-610.

[29] 慎金花,王　薇,张更平,等. 基于动态主题网络的新兴技术主题识别——以氢燃料电池领域为例[J]. 情报杂志,2024,43(9):92-100.

[30] Devlin J, Chang M W, Lee K, et. al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2019,1:4171-4186.

[31] 周启星,王　辉,欧阳少虎. 基于碳中和新技术的美丽中国建设[J]. 中国环境科学,2024,44(4):1777-1787.