



开放科学  
(资源服务)  
标识码  
(OSID)

# 基于 SBERT 的专利前沿主题识别方法研究 ——以我国制氢技术为例

刘晋霞 张志宇

太原科技大学经济与管理学院 太原 030024

**摘要:** [目的/意义] 提出一种基于深度学习的前沿主题识别方法, 使识别结果更加准确客观, 为专利前沿主题的认识工作提供新思路。[方法/过程] 采用专利之星检索系统数据库, 通过 SBERT 预训练模型和聚类算法进行主题抽取工作; 利用主题相似度计算和 LDA 模型对主题进行关联和标识; 并从关注度和质量水平两方面构建前沿性指标, 确定前沿主题。[局限] 在数据源以及指标构建多样性方面尚需进一步研究和构建。[结果/结论] 以我国制氢领域的专利数据作为实证研究对象, 发现电解制氢、重整制氢、光制氢、水解制氢和燃料电池发电 5 个前沿主题, 并通过对比分析验证了主题抽取与指标建立的正确性与有效性。

**关键词:** SBERT; 制氢技术; 主题抽取; 主题识别; 前沿主题

**中图分类号:** G306

## Research on Patent Frontier Topic Recognition Method Based on SBERT —— Case Study of Hydrogen Production Technology in China

LIU Jinxia ZHANG Zhiyu

School of Economics and Management, Taiyuan University of Science and Technology, Taiyuan 030024

**Abstract:** [Purpose/Significance] This paper proposes a method for identifying frontier topics based on deep learning, which makes the recognition results more accurate and objective and provides new ideas for the recognition of patent frontier topics. [Methods/Processes] This paper uses CPRS as the database source. Firstly, topic extraction is based on SBERT pretraining model and clustering algorithm. Secondly, topic similarity calculation and LDA model is used to associate and identify topics. Finally, we construct the index of frontier topics identification from two aspects of attention and quality level. [Limitations] Further research is needed on the diversity of data sources and indicator building. [Results/Conclusions] Taking the patent data in the field of hydrogen production in China as the empirical research object, five frontier topics were found, including electrolytic hydrogen production, reforming hydrogen production, photo-hydrogen production, hydrolysis hydrogen production and fuel cell power generation. The correctness and effectiveness of topic extraction and index establishment are verified by comparative analysis.

**Keywords:** SBERT; hydrogen production technology; topic extraction; topic identification; front topic

**基金项目** 教育部人文社会科学研究青年基金项目“基于区块链的数字版权质押融资模式及其风险管理机制研究”(21YJC630172); 山西省社会科学院(山西人民政府发展中心)项目“山西数字政府建设路径研究”(YWQN202148)。

**作者简介** 刘晋霞(1973-), 博士, 硕士生导师, 研究方向为技术与创新管理、智能决策; 张志宇(1997-), 硕士研究生, 研究方向为大数据驱动的管理决策, E-mail: 15264192112@163.com。

**引用格式** 刘晋霞, 张志宇. 基于 SBERT 的专利前沿主题识别方法研究: 以我国制氢技术为例[J]. 情报工程, 2022, 8(6): 28-45.

## 引言

随着科学技术的不断发展创新,各学科领域的资源数量不断增加,在科技文献和专利文献等方面都产生了庞大的信息数据,科研人员为及时掌握技术发展、辅助技术决策、提高科研效率,力求借助科学计量、信息检索、自然语言处理等方法,快速准确地进行前沿主题识别。

“研究前沿”概念最早由科学计量之父 D. J. Price 提出<sup>[1]</sup>,体现核心文献的高被引程度。近年来,对研究前沿探测的挑战是以科技文献或专利文献为数据信息,通过构建主题识别模型框架<sup>[2]</sup>,更加精细地识别前沿主题。其中,专利文献集技术、法律、经济信息为一体,含有统一的分类体系,是技术创新的载体,相比一般科技文献,更加具有标准化与系统化。目前对“专利前沿主题”的概念并没有明确的阐释,笔者认为专利前沿主题是以专利文献为数据源,构建主题识别模型与指标,识别技术前沿主题。因此,通过对专利前沿主题的识别,有效把握领域技术发展规律,在前沿技术领域实现突破,成为学术界和产业界共同关注的研究热点。目前,研究专利的前沿识别方法大致可以归为定性分析和定量分析两类。定性分析是指专家判断法,专家利用所掌握某一领域的知识和经验,对问题进行主观判断,得出最终结论。定量分析方法主要包括从外部特征分析的引文分析法、从内部特征分析的知识单元分析法、基于主题模型的分析法以及基于深度学习的方法。随着大数据时代的到来,在速度和准确性上,专家判断法难以发挥最佳效能<sup>[3]</sup>;引文分析法没有

深入文献内部进行分析,分析粒度低;知识单元分析法没有很好的解释主题词之间的语义关系;主题模型也没有对文本内容进行深层语义分析,并且在建立前沿指标时,大多数学者对单一指标进行罗列形成多维指标,但会出现设置阈值时主观性较强以及缺乏对指标进行调节等问题;而基于深度学习的方法能更好地抽取主题,进而识别前沿主题。

综上所述,本文利用目前取得较好效果验证的深度预训练 SBERT 模型,以专利文献为数据源,通过深度学习的方法,识别专利文献中的主题;并通过引入调整系数和无需设置阈值的方式,建立关注度和质量水平的前沿性指标,确定前沿主题。在一定程度上提高准确性、降低主观性。

## 1 相关文献

目前对前沿主题识别的定量分析方法主要有四类,分别为引文分析法、知识单元分析方法、基于主题模型的方法以及基于深度学习的方法。

引文分析法中最为常见的是共被引分析和引文耦合分析<sup>[4]</sup>。共被引分析是从被引文献出发,共被引频次越高,越代表前沿性<sup>[5]</sup>,但一篇文献在达到一定引用量时才会得到关注,这会对研究前沿性工作产生时间滞后性。引文耦合分析是从施引文献出发,在其发表后参考文献固定并可以立即获得,一定程度上改进了共被引分析的时间滞后性,但缺少动态变化性。Morris 等<sup>[6]</sup>通过文献耦合方法进行研究前沿的识别,将研究前沿定义为引用一组固定和时间不变的基础文献的集合,将被引文献视为对应

的研究前沿的知识基础。许振亮等<sup>[7]</sup>对引文文献和施引关键词进行剖析, 计量出前沿主题; Huang等<sup>[8]</sup>在探测有机发光二极管领域, 对文献共被引和引文耦合分析方法进行对比, 发现引文耦合在前沿识别的数量和速度上占有优势。Park等<sup>[9]</sup>在太阳能电池领域, 基于文献关系对核心文献进行聚类, 识别出研究前沿。侯剑华等<sup>[10]</sup>通过绘制文献共被引和施引结构的知识图谱, 预测大数据领域的前沿研究与发展趋势。郭伏等<sup>[11]</sup>通过文献的共被引分析与网络可视化, 分析人因工程领域的研究前沿。整体来说, 共被引分析和引文耦合分析两种方法可以在动态性和时效性方面互补, 但它们只从文献外部特征之间的关系进行分析, 没有解决分析粒度问题。

知识单元分析方法包括词频分析法和共词分析法。两者从主题词的角度入手。词频分析法主要利用文献计量学, 统计主题词的词频或词频变化率来识别前沿主题<sup>[4]</sup>, 但这种方法中得到的是孤立、没有连接的词语。与词频分析法相比, 共词分析法通过构建共词网络知识图谱与聚类分析, 建立了词语之间的关系, 在一定程度上展现知识结构变化。Liu等<sup>[12]</sup>利用数据库和可视化工具创建了被引出版物的共被引网络, 通过纳入这些刊物中频率较高的关键词, 分析研究前沿与发展趋势。Peters等<sup>[13]</sup>利用共词分析并结合多维尺度分析绘制了化学工程领域的知识图谱, 梳理了该领域的研究进展和重点问题。潘黎等<sup>[14]</sup>通过应用文献计量学可视化软件绘制文献关键词的知识图谱, 探测前沿主题。张丽华<sup>[15]</sup>对主题演化情况, 通过建立主题演化指数、主题演化率、主题演化强度、前沿

特征演化指数的前沿指标, 进行详细指标量化。武建鑫<sup>[16]</sup>运用文献计量学和 Citespace 工具绘制关键词共现知识图谱, 揭示其研究主题和前沿趋势。郑彦宁等<sup>[17]</sup>利用关键词共现的方法, 建立研究主题年龄、研究主题关注作者数量指标, 研究前沿识别方法。范少萍等<sup>[8]</sup>基于文献计量学, 建立时效性、创新性、应用性、风险性与学科交叉性指标。整体来说, 知识单元分析法对文献内部间的主题词进行分析, 与引文分析法相比, 分析粒度更细。但该类方法主题词以及主题关联主题词的确定较为复杂, 设置阈值时有一定的主观性; 另外, 没有较好的表达和挖掘文本语义信息。

基于主题模型的识别方法, 首先利用主题模型挖掘文本主题, 其次建立指标识别研究前沿主题。目前的研究中, 主题模型以潜在狄利克雷分布 LDA (Latent Dirichlet Allocation)<sup>[19]</sup>模型应用广泛, 但由于其词袋模型、不考虑时间动态等问题, 其改进的模型也具有广泛应用。如引入时间动态概念的 DTM (Dynamic Topic Models) 模型<sup>[20]</sup>、将词的顺序作为影响主题确定的因素的 BTM (Biterm Topic Model) 算法<sup>[21]</sup>、基于 Gibbs sampling 并采用全局同步的思想的 AD-LDA (Approximation Distribution Latent Dirichlet Allocation) 模型<sup>[22]</sup>和提高 AD-LDA 分析效率的 PLDA (Parallel Latent Dirichlet Allocation) 模型<sup>[23]</sup>等。范云满等<sup>[24]</sup>基于 LDA 主题模型, 通过构建主题新颖度、作者发文章量和文章被引量的指标, 对新兴主题进行探测。冯佳等<sup>[25]</sup>利用 LDA 模型抽取主题, 建立主题强度和主题新颖度指标识别研究前沿主题, 并基于领域本体挖掘前沿主题语义类型。王效岳等<sup>[26]</sup>

利用 PLDA 模型识别研究主题后,建立主题的资助时间、资助金额和中心性指标,进行研究前沿主题识别。王菲菲等<sup>[27]</sup>通过引入 Altmetrics 数据的前沿探测方法,构建用于前沿探测的即时性、增长性、影响力、关注度、交叉性五项评价指标,通过 LDA 算法提取研究主题,计算出研究主题的前沿评价指标得分。吴一平等<sup>[28]</sup>运用 PLDA 模型进行主题识别,并建立主题热度和主题新颖度等指标识别前沿主题。整体来说,这类方法首先基于主题模型,通过对大规模文本集合进行聚类来发现隐含的语义结构,很大程度上弥补了引文分析法和知识单元分析法对文本语义内容和语义关系的忽略,但它只停留在对于文本语义的浅层理解上,没有获得文本的深层语义;其次,考虑到单一指标的片面性,如今大量学者利用多维度指标进行前沿主题的识别,但会出现建立的多数指标普适性不强、综合前沿性指标设定阈值的方式具有一定的主观性,从而导致结果不同的问题。

基于深度学习的识别方法,可以获取文本的深层语义信息,解决主题模型在文本主题挖掘时的缺点。陈虹枢等<sup>[29]</sup>运用概率主题模型 LDA 进行主题提取并结合词嵌入 Word2vec 的方法进行向量化,克服了语义表达上存在盲点等问题,并构建新颖性、突变性、影响力以及学科交叉性的指标体系对突破性创新主题进行识别。通过引入 Word2vec 方法,可以深入文本语义,将每个词转化为一个向量,但这种模型忽略了上下文语境问题,而 BERT (Bidirectional Encoder Representation from Transformers) 模型能够很好地体现语义和语法方面的复杂性,对表达词汇在不同语境的含义更有帮助。李松繁

等<sup>[30]</sup>以 BERT 模型构建文本的句嵌入集合为基础,对农业领域前沿研究进行主题识别;王秀红等<sup>[31]</sup>提出基于 BERT-LDA 模型的关键技术识别方法,通过结合 BERT 与 LDA 相结合,弥补了单一使用 LDA 主题模型缺乏上下文语义信息的缺陷,提高了主题聚类的连贯性及细粒度划分的精准度。但 BERT 模型中向量空间的各项异性,使得词向量的空间分布受频率影响,所以词向量之间的距离不能很好的表示词之间的相关性;在句向量方面,是对词向量的平均池化,因此 BERT 模型生成的句向量效果也并不理想,不适合文本相似度 (STS) 搜索和聚类等无监督任务。

综上所述,为改进前沿主题识别研究的不足,本文使用可以解决 BERT 问题的 SBERT<sup>[32]</sup> (Sentence-BERT) 深度预训练模型,提出基于 SBERT 的专利前沿主题识别方法。一方面以专利文献为数据源,通过获取句表征向量、聚类主题、主题关联、主题词提取和主题标识,深度挖掘数据内部语义信息,识别专利主题;另一方面,建立关注度和质量水平的前沿主题识别指标,通过对前向引用量和主题文本量的调整,增强前沿识别结果的准确性和客观性。

## 2 方法构建

基于 SBERT 模型的专利前沿主题识别方法流程如下:(1)数据获取与文本建立。从专利数据检索系统中收集专利文献数据,构建不同字段样本集。(2)主题抽取。对不同字段样本集使用 SBERT 获取各自的句表征向量;以年为单位划分时间窗口,通过对句向量进行聚类,获取样本每年的若干聚类主题与对应的主题中

心向量；对比不同样本正确率，选择结果最优字段样本的主题结果做进一步研究。（3）主题关联与主题标识。通过计算主题相似度，进行

主题关联；运用 LDA 模型提取主题词，进行主题关联效果验证与主题标识。（4）前沿主题识别。构建前沿识别指标，识别前沿主题。如图 1 所示。

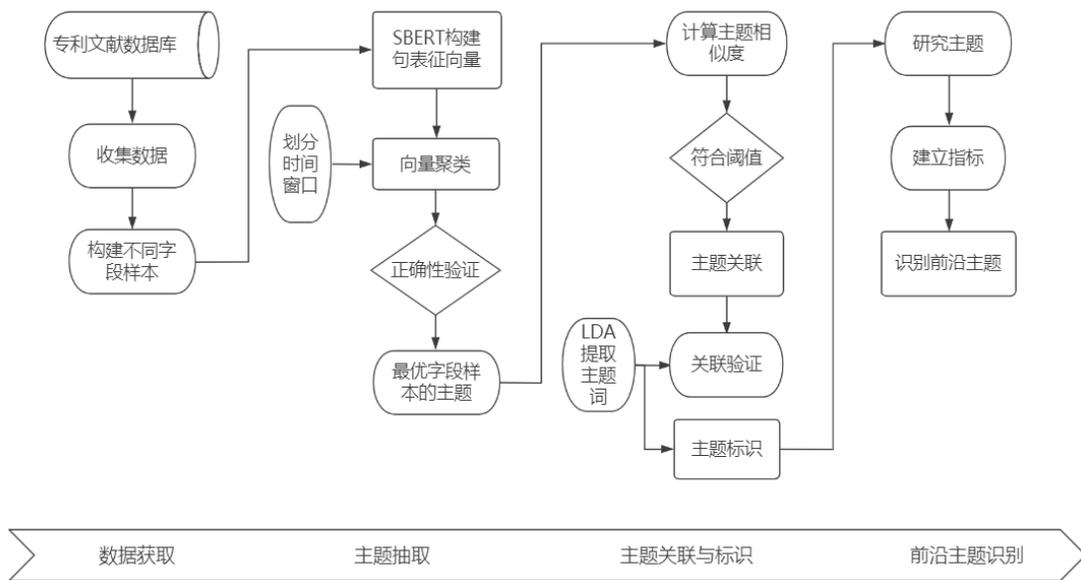


图 1 方法流程图

## 2.1 数据获取与文本建立模块

本模块分为两个步骤：

（1）数据获取。在专利平台中对专利类型进行选取，并且提取每个专利中的摘要、标题、申请时间、分类号等字段。

（2）文本建立。考虑到 SBERT 输入样本对输出句表征向量的影响，本文建立三种不同字段的样本，分别为标题、摘要和分句。

## 2.2 主题抽取模块

### 2.2.1 获取句表征向量

当前许多学者对 BERT 的向量空间的各项异性问题提供了相应的解决方案，其中包括双塔模型 SBERT 和对比学习 SimCSE<sup>[33]</sup>、ConSERT<sup>[34]</sup> 模型。两类模型的选用可以依据计算损

失函数时的输入数据，Cosine Similar Loss 方法计算损失函数时输入样本为句子对和相似度，Multiple Negatives Ranking Loss 方法计算损失函数时的输入样本为句子。本文利用 Cosine Similar Loss 方法计算损失函数，而 SBERT 的架构可以通过输入句子对计算相似度得分。因此，本文利用双塔模型 SBERT 对样本进行训练，获取句子的表征向量。

SBERT 模型解决 BERT 模型在 STS 计算任务和无监督任务中的不足，其本质是对原生 BERT 模型进行微调。通过在微调阶段加入计算相似度的任务，生成具有语义信息的句子表征向量，并使得语义相似的句子表征向量空间距离更加接近，减少寻找相似向量的时间、提高准确率，适用于分类与回归任务。在分类任

务中, SBERT 模型借鉴孪生网络 (Siamese Network) 结构, 使两个句子分别经过共享参数的两个 BERT 神经网络; 在 BERT 的输出中加入池化操作, 推导出固定大小的句子嵌入  $u$  和  $v$ ; 拼接  $u$ 、 $v$  以及它们两者的差值  $|u - v|$ , 并乘以可训练的权重句子  $W_i \in R^{3n \times k}$ , 输入全连接层做分类任务, 使用 Cosine Similar Loss 函数作为优化的目标函数 (如公式 (1) 所示) 更新权重值。

$$o = \text{soft max}(W_i(u, v, |u - v|)) \quad (1)$$

### 2.2.2 聚类主题

首先以获取的句表征向量为依据, 利用聚类算法得到主题; 其次对主题聚类结果进行正确性对比验证。

#### (1) 主题聚类

以年为划分单位进行主题聚类。先从文本挖掘角度和向量角度, 分别运用最小困惑度 (如公式 (2) 所示) 和肘部法则共同确定最佳簇的数量  $K$ ; 再使用二分 K-Means 算法选择质心, 将其得到的质心作为主题中心向量。二分 K-Means 算法是 K-Means 的变种算法, 克服 K-Means 算法容易收敛于局部极小值问题, 在每次划分时最大限度降低 SSE (Sum of Squared Error 误差平方和), 弱化随机初始值质心的影响。

$$\text{perplexity} = e^{\frac{-\sum \log(p(w_j|d_j))}{N}} \quad (2)$$

其中,  $p(w_j|d_i)$  代表文档集中词语的概率信息,  $N$  代表文档集中出现的所有词。

#### (2) 正确性对比验证

在获取句表征向量和聚类后得到各年份专利的所属主题。提取各个主题的专利, 依据每个主题中 IPC 号的相同率构建正确性验证指标。IPC 号来源于国际专利分类系统体系, 它是我

国进行专利分类的常用体系, 按照不同的技术对专利进行划分。数据库中每项专利所分配的 IPC 号, 体现其专利技术的所属类型。因此, 本文以 IPC 号作为正确性验证标准, 如公式 (3) 所示,  $A_k$  的结果值越大, 说明一个主题下 IPC 号相同率高, 相应地聚类正确性好。

$$A_k = \frac{H_k}{N_k} \quad (3)$$

其中,  $H_k$  是  $k$  主题含有相同专利号的专利数,  $N_k$  是  $k$  主题下的专利总数。

由于不同的样本集或不同的模型得到每年的最优主题数不同, 因此本文对整年的聚类正确性进行比较。以年为单位, 对利用  $A_k$  计算出的每个主题的正确性进行加和平均, 得到各样本集或模型中每年聚类的正确性结果  $A_y$ , 第  $i$  年的正确性如公式 (4) 所示, 其中  $N$  表示第  $i$  年的最优主题数。

$$A_{y_i} = \frac{1}{N} \sum_{k=0}^N A_k \quad (4)$$

## 2.3 主题关联与主题标识模块

选择正确率最高的样本, 从向量角度进行主题关联, 并从文本挖掘角度进行对比验证主题关联的准确性与一级主题、二级主题的标识。

### 2.3.1 主题关联

首先运用皮尔逊 (Pearson) 相关系数来计算不同时间窗口之间主题中心向量的相似度, 其公式为 (5), 相关系数越大, 说明两个主题间的相似性越高。其次设置相似度上限阈值与下限阈值, 确定主题关联。

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)} \sqrt{E(Y^2) - E^2(Y)}} \quad (5)$$

其中,  $X$ 、 $Y$  表示两个主题的中心向量。

以提取数据的初始年份为基础, 计算前两年所有主题之间的相关系数, 当主题间相关系数高于上限阈值时, 视为主题关联; 当主题间低于下限阈值时, 视为新主题。对于其它年份, 先计算其与上一年主题的相关系数, 若都低于下限阈值, 再往前计算一年, 寻找相关主题, 以此类推, 若与之前年份的主题都低于下限阈值, 视为新兴主题。

### 2.3.2 主题标识

运用 LDA 模型提取主题词, 并依据主题词提取结果进行主题关联的效果验证以及标识主题。LDA 模型可以将语料库中的所有词语进行分主题聚类, 实现无监督地从数据集中抽取每项文档所属主题, 以及文档中的每个词语所属主题。但其主题数的确定需要困惑度的辅助, 并且本身是一个词袋模型, 不考虑文本之间的上下文逻辑关系, 无法深入挖掘语义。因此本文在使用 LDA 模型前, 通过 SBERT 预训练模型深度挖掘语义信息, 并利用困惑度、肘部法则确定主题数的方法, 消除和缓解 LDA 模型的问题。

具体步骤如下:

(1) 主题词提取: 按照 SBERT 和二分 K-Means 聚类得到的每个主题结果, 选取其文本项, 并对文本项进行预处理得到语料库。将每个主题的语料库作为 LDA 模型的输入, 分别进行主题词的挖掘。

(2) 主题关联效果验证: 对挖掘出的主题词, 通过正点互信息 (公式 (6)) 进行混淆效果计算, PPMI 的值越大说明主题关联性强。

$$PPMI(X_{t_k}, X_{t_j}) = \max(\log_2 \frac{P(X_{t_k}, X_{t_j})}{P(X_{t_k})P(X_{t_j})}, 0) \quad (6)$$

其中,  $t_k, t_j$  分别表示  $k$  主题和  $j$  主题,  $X_{t_k}, X_{t_j}$  表示两个主题之间相同词的个数。若依据主题中心向量计算出有关联的主题, 其通过文本挖掘得到的 PPMI 值也大, 视为主题关联效果好。

(3) 标识主题: 依据主题关联结果以及利用 LDA 模型提取出的主题高频词, 标识主题一级方向和主题二级方向。

## 2.4 前沿主题识别研究

本文构建了关注度和质量水平两个方面的指标进行前沿主题 FT (Frontier Topics) 识别。其中关注度指标是从主题中专利的前向引用次数角度出发, 通过调整来平衡时间维度对关注度的影响, 分析主题的被关注程度。质量水平指标是从主题文本量的角度出发, 通过引入主题中含有前向引用专利的数量, 分析每个主题的质量水平。

### 2.4.1 关注度指标

专利的前向引用数 (施引数) 是使用最广泛的技术进步性评估方法, 它反映出专利所代表的技术对之后技术发展的贡献程度。一项技术被越频繁、越广泛地引用到未来技术之中, 说明其被关注的程度越高<sup>[35-38]</sup>, 越具有前沿性。然而, 一个主题中专利的前向引用次数会受时间维度的影响, 因此本文在前向引用数据的基础上进行标准化调整, 以平衡时间因素对关注度的影响。

首先, 由于引用频率的影响, 前向引用次数会随着时间积累, 使得近期发表的专利前向引用次数相对低于前期发表的专利。它应该通

过专利的申请日期对前向引用数进行调整,以提高近期申请专利的相对关注度;其次由于近因性的影响,如果一项专利被其它最近的专利引用,会被认为是一种最近可获得的技术,更加具有前沿性。它应该通过施引专利的申请时间对前向引用数进行调整,以提高被近期引用的专利相对关注度。因此,本文建立的关注度指标  $AD$  (Attention Degree),是在前向引用数的基础上通过引用频率和近因性两个相关系数进行标准化调整。 $i$  主题的关注度公式为(7)。

$$AD_i = \sum_t [F_{i,t} \times \frac{1}{EY-t+1} \times \frac{1}{\sum_j^{F_{i,t}} (EY-AY_j+1)_{i,t}}] \quad (7)$$

其中,  $F_{i,t}$  表示第  $t$  年  $i$  主题专利的前向引用次数,  $EY$  是数据集中的结束年,  $\frac{1}{EY-t+1}$  是对第  $t$  年引用频率的调整系数,  $t$  值越大其值也越大;  $AY_j$  是第  $j$  个施引专利的申请年,  $\sum_j^{F_{i,t}} (EY-AY_j+1)_{i,t}$  值越小代表第  $t$  年该主题的专利被最近施引的次数越多,  $\frac{1}{\sum_j^{F_{i,t}} (EY-AY_j+1)_{i,t}}$  是对第  $t$  年近因性的调整系数;  $F_{i,t} \times \frac{1}{EY-t+1} \times \frac{1}{\sum_j^{F_{i,t}} (EY-AY_j+1)_{i,t}}$  表示第  $t$  年  $i$  主题的关注度。

#### 2.4.2 质量水平指标

专利主题文本量可以反映该技术的研究热度。但依靠主题总数量的计算,会因只关注主题研究数量忽略研究质量,而质量高的主题往往更能引领前沿发展。一项专利被其他专利引用是专利质量高的体现,一个主题中含有这种专利的数量越多代表其质量相对较高。因此,质量水平指标  $QL$  (Quality Level) 是在主题文本量的基础上,引入主题中含有前向引用的专利数量,通过对两者数量比的计算,反映主题质量水平,公式为(8)。

$$QL_i = \frac{PN_i}{TP_i} \quad (8)$$

其中,  $PN_i$  是  $i$  主题中含有前向引用专利的数量,  $TP_i$  是  $i$  主题的专利总数。

前沿主题  $FT$  是对两个指标值归一加和后的综合对比。为消除两个指标之间数据范围的差距,将两个指标的数据先分别进行归一化处理,后进行加和计算。主题  $i$  的前沿指标值  $FI_i$  (Frontier Index) 的计算公式为(9)

$$FI_i = \text{sum}(\frac{AD_i}{\sum_j^n AD_j}, \frac{QL_i}{\sum_j^n QL_j}) \quad (9)$$

按照年份划分计算集与验证集,依据这两个指标对计算集的进行指标计算,选取出具有前沿性的主题,并通过验证集数据等信息验证分析指标建立的有效性。

## 3 实证研究

### 3.1 数据获取与文本建立

#### 3.1.1 数据获取

本文研究对象为我国制氢技术领域的专利文献,检索平台选择“专利之星检索系统(CPRS)”,使用表格检索,检索式为“TX=制氢”(TX表示关键词)、“AD=20110101>20201231”(AD表示申请日、20110101>20201231表示时间跨度为2011—2020年),共检索得到9243篇专利文献。其中专利类型为发明的专利6753篇、实用新型专利2403篇、外观设计专利87篇。由于发明专利具备突出的实质性特点、显著性进步、新颖性和创造性水平更高、保护年限长、保护产品方法与技术的优点,因此选择发明专利为研究对象。发明专利中的法律状态包括有效、审中

和失效专利，其中有效专利 2344 篇、失效专利 2974 篇（失效但有过授权的专利为 249 篇）。由于专利会随着时间失效，但有过专利的授权就说明该技术曾为有效专利，创新研究被认可，故选取法律状态为有效和已失效但曾有授权时间的专利。根据以上选取条件，共检索得到 2693 篇专利文献。

### 3.1.1 文本建立

将 2693 篇专利文献导出后对专利权人、发明人名称规范与消歧<sup>[38]</sup>以及对重复文本剔除，经过筛选，共有 1968 条专利记录。每项记录中包括发明人、标题、分类号、摘要、法律状态、专利类型等 20 项内容。

本文建立三种字段样本，分别为标题、摘要和分句，其中分句是在摘要中提取出包含标题并体现领域、重要技术方法或材料所组合出的句子。

## 3.2 主题抽取

### 3.2.1 获取句表征向量

本文实验使用 Python3.8 作为开发平台，安装 sentence-transformers 和 transformers 库，

在 Sentence-Transformer 预训练模型中，选用 STS 任务 STSb performance 值高且使用率高的 'sentence-transformers/all-MiniLM-L6-v2' 中文模型，对该模型基础进行微调后分别训练三种字段样本，输出句表征向量。

### 3.2.2 聚类主题

(1) 以年为单位，通过二分 K-Means 算法聚类句表征向量，选取最佳主题数 K 值与主题中心向量，得到聚类结果。

(2) 对其聚类结果和主题模型 LDA、BTM 以及深度学习模型 BERT 结果进行正确性对比。如图 2 所示。通过验证发现，三个不同字段样本聚类结果的正确性高于 LDA 模型与 BTM 模型，并且在大多数年份高于 BERT 模型，说明利用本文方法进行主题聚类的有效性。其中字段样本为分句时结果最优，在各个年份的正确性效果均高于其它字段样本和模型，说明选取文本中有效信息作为 SBERT 模型的输入，会提高训练结果。因此以分句计算出的聚类结果为依据，对主题进行关联和前沿识别。其中每年的最佳簇 K 值分别为  $K_{2011-2020}=7, 8, 9, 10, 11, 11, 10, 10, 8, 6$ 。

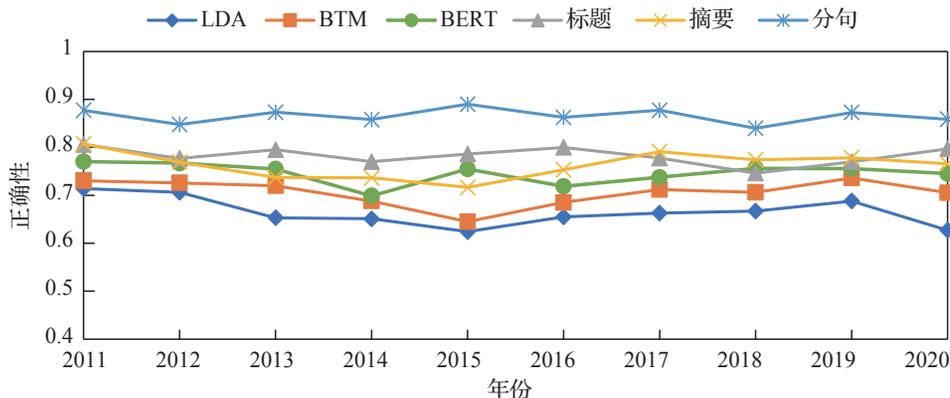


图 2 各年度正确性对比验证

### 3.3 主题关联与主题标识

通过计算聚类所输出的主题中心向量相似度进行主题关联，并利用 LDA 模型输出主题词，依据主题词进行主题关联的比较验证和标识主题。

#### 3.3.1 主题关联

以 2011 年的所有主题为基础，通过计算主题间相关系数设置阈值。上限阈值的设定依据计算年与下一年份各个主题之间的相关系数，

确保其相关系数高的主题中有 1~2 个关联主题，并使关联结果保持清晰，故将其设置为 0.94；下限阈值的设定依据计算年与之前年份之间的相关系数，使其相关系数都小的主题成为新兴主题，故将其设置为 0.5。在 2011 年 6 个主题的基础上（2011\_00 和 2011\_02 归为同一主题），2012 年产生 2 个新兴主题，2013 年产生 2 个新兴主题，整体可归为 10 个主题。关联结果图以 2011-2018 年主题关联为例，如图 3 所示。

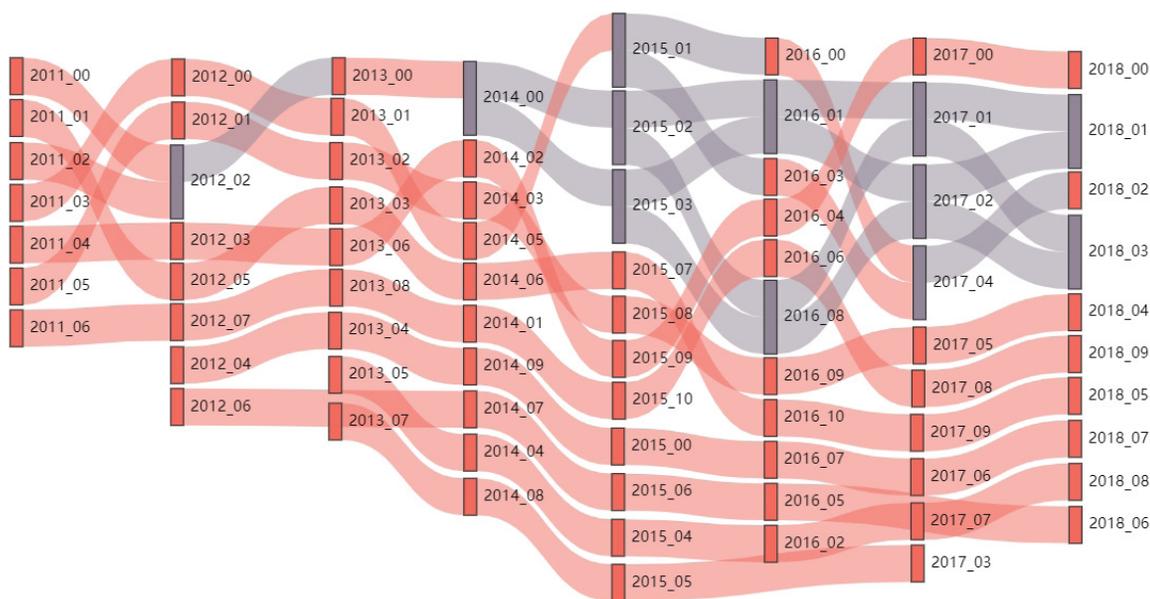


图 3 2011-2018 主题关联效果图

#### 3.3.2 主题标识

##### (1) 主题词提取：

对分句文本进行中文停用词表的清洗、建立自定义词典、使用中文分词工具 jieba 和人工方法进行分词得到语料库。

将每个主题语料库结果分别输入到 Scikit-Learn 中的 LDA 主题模型，设置主题数  $K=1$ ，文档迭代次数为 600 次，输出每个主题

的主题词。

##### (2) 主题关联效果验证

对每个主题选取 LDA 模型输出的前 20 个主题词，进行 PPMI 的计算，如图 4，以 2011 年的 7 个主题与 2012 年的 8 个主题为例。

通过图 3 中 2011-2012 年间主题关联结果与图 4 的结果对比发现，根据主题中心向量计算出有主题关联关系的主题，它们之间的

PPMI 值也高，并且 2012 年中 Topic4 和 Topic6 与 2011 年各主题之间的 PPMI 值都低，可以将其作为新生主题。因此证明主题关联效果好。

### (3) 标识主题

对主题关联后形成的 10 个主题进行主题标识作为主题一级方向，依据一级方向主题下每个年份提取的主题词，选择有代表性的高频关键词作为主题的二级方向。结果如表 1。

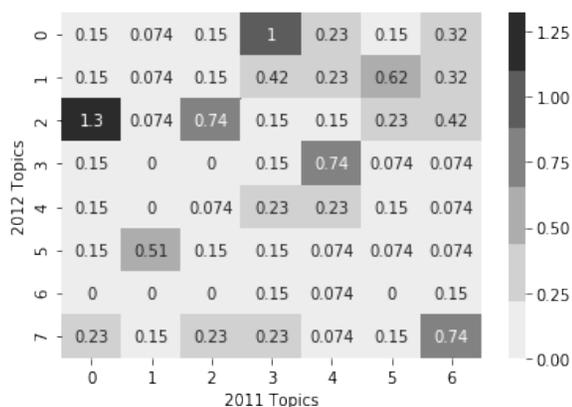


图 4 2011 年与 2012 年主题词热度图

表 1 主题一级、二级方向

主题一级方向	主题二级方向
光制氢	光催化分解水、光解水制氢、光伏、光电化学、光热、光催化剂
化学链制氢	化学链循环制氢、热化学、化学链气化
重整制氢	甲醇重整、乙酸自热重整、甲烷重整
电解制氢	电解水、电催化、车载、电化学、生物电化学
生物制氢	生物质制氢、微生物、混合菌源
化石燃料制氢	煤、天然气、石油化工
燃料电池发电	发动机、甲醇重整、氢燃料电池
工业副产品制氢	焦炉煤气、尾气、煤沥青、地沟油
水解制氢	硼氢化钠水解、三元合金、富镁
金属材料	合金、非贵金属、过渡金属

### 3.4 前沿主题识别

通过 2011-2018 年的数据作为计算集，计算关注度和质量水平两个指标，并将两个指标归一加和计算前沿指标值，来确定前沿主题。

#### (1) 关注度 $AD_i$ 指标计算

以光制氢主题的关注度  $AD_i$  指标中各参数结果为例，如表 2 所示。起始年  $EY=2011$ ，结束年  $FY=2018$ ，各年度的前向引用次数、引用频率的调整系数、施引专利的申请年和近因性的调整系数分别计算列出。

依据公式 (7) 对各参数进行整合，计算前向引用强度  $IL_i$  指标，各主题的计算结果如表 3， $F_i$  为主题  $i$  中专利的前向引用次数。由此可以得到电解制氢、重整制氢、光制氢、燃料电池发电和水解制氢主题的  $IL_i$  指标相对较高。

#### (2) 质量水平 $QL_i$ 指标计算

依据公式 (8) 计算质量水平  $QL_i$  指标，各主题的计算结果值如表 4。由此可以得到重整制氢、电解制氢、光制氢、水解制氢、化石燃料制氢和燃料电池发电的  $AI_i$  指标值相对较高。

表 2 光制氢主题的指标参数结果

光制氢	2011	2012	2013	2014
$F_t$	5	8	8	1
$\frac{1}{EY-t+1}$	1/8	1/7	1/6	1/5
$AY_j$	2012 2012 2013 2014 2015	2013 2014 2014 2014 2014 2014 2016 2018	2014 2014 2015 2015 2015 2015 2016 2016	2018
$\frac{1}{\sum_j^{F_t} (EY - AY_j + 1)_t}$	1/29	1/35	1/32	1
光制氢	2015	2016	2017	2018
$F_t$	0	2	0	0
$\frac{1}{EY-t+1}$	1/4	1/3	1/2	1
$AY_j$	--	2017 2017	--	--
$\frac{1}{\sum_j^{F_t} (EY - AY_j + 1)_t}$	--	1/4	--	--

表 3 各主题的关注度指标

$i$	光制氢	化学链制氢	重整制氢	电解制氢	生物制氢	化石燃料制氢	燃料电池发电	工业副产品制氢	水解制氢	金属材料
$FCI_i$	1.463	0.100	2.168	2.220	0.115	0.324	1.088	0.167	1.080	0.250

表 4 各主题的质量水平指标

$i$	光制氢	化学链制氢	重整制氢	电解制氢	生物制氢	化石燃料制氢	燃料电池发电	工业副产品制氢	水解制氢	金属材料
$AI_i$	0.051	0.010	0.079	0.076	0.023	0.037	0.030	0.009	0.042	0.012

### (3) 前沿指标值计算与前沿主题确定

对关注度和质量水平指标结果值进行归一加和，得到每个主题前沿指标值  $FI_i$  的计算结果如图 5 所示。电解制氢和重整制氢两个主题前沿指标值远高于其它主题；光制氢、水解制氢和燃料电池发电也具有较高前沿指标性值，生物制氢、化石燃料制氢、工业副产品制氢、化学链制氢和金属材料五个主题前沿性低。本文将具有高前沿指标值与较高前沿指标值的 5 个主题作为前沿主题。

### 3.5 有效性验证

由于本文建立的前沿性指标是对前向引用数和文本量的调节，因此首先利用 2011-2018 年的数据，对仅使用前向引用数和文本量的主题结果，与本文建立两个指标的主题结果进行对比分析。其次，通过 2019-2020 年的主题文本量数据进行分析，验证指标建立的有效性。

2011-2018 年各主题的前向引用数和文本量如表 5。仅从主题前向引用数分析时，化石燃料制氢高于燃料电池发电与水解制氢，而如今，

随着“双碳”目标的提出，我国能源结构调整的步伐加快，以绿色零排放方式制取“绿氢”将成为制氢研究的重点，而化石燃料制氢将被逐步取代。仅从文本量角度分析时，生物制氢和化石燃料制氢高于水解制氢，但生物制氢的方法占地面积大，不适合大规模制取；并且，当今寻找运氢和储氢方面的有利方法与材料，来解决储运成本、氢能利用率等瓶颈问题迫在

眉睫，在储运方面水解制氢方法优于生物制氢方法。

2019-2020年的各主题占比量如图6所示，其中电解制氢、重整制氢、光制氢、燃料电池发电和水解制氢五个前沿主题在2019年和2020年均有所发展，并且相比其它非前沿主题，文本占比量大。因此验证两个指标所选的前沿主题更加准确，具有一定的有效性。

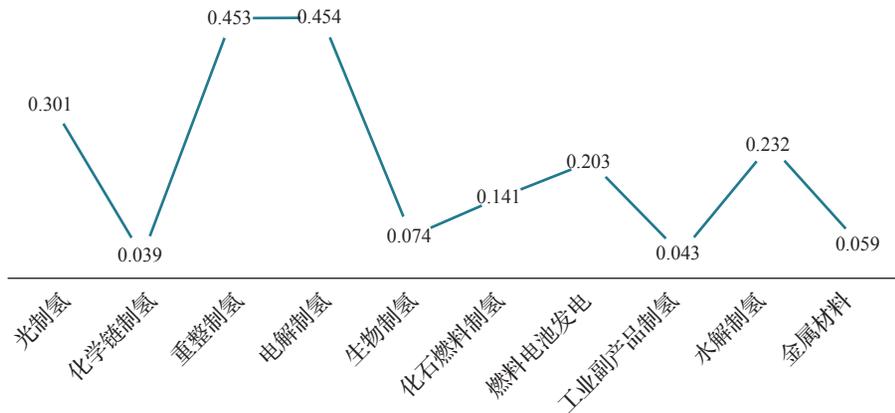


图5 各主题前沿指标值

表5 各主题前向引用数和文本量

$i$	光制氢	化学链制氢	重整制氢	电解制氢	生物制氢	化石燃料制氢	燃料电池发电	工业副产品制氢	水解制氢	金属材料
$F_i$	24	1	43	19	3	9	6	1	6	1
$TP_i$	467	96	267	249	133	189	197	109	120	86

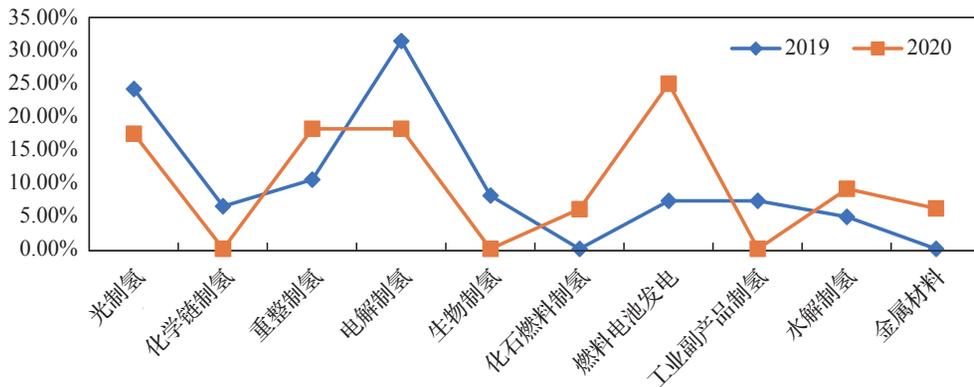


图6 2019-2020年的各主题文本量占比



近年来,在国家大力发展的低碳经济背景下,利用可再生能源进行电解制氢是当前制氢方案中碳排放最低且应用较广的工艺<sup>[39]</sup>。电解制氢主要包括电解水制氢和电催化制氢,相比化石燃料制氢,电解制氢工艺简单、无污染并且所得氢气纯度高。以往电解制氢的贵金属等催化材料价格昂贵,储量稀缺。当今电解制氢方面的专利研究,不断向寻找储量高且价格低廉的催化剂或方法的方向发展,如采用非金属材料,与再生能源风、光、太阳能相结合形成光电制氢等方法,使电解制氢方法的技术得到进一步提高。

重整制氢主要包括甲醇重整制氢和乙醇自热重整制氢,尤以甲醇重整制氢发展迅速。澳大利亚国家工程院外籍院士、南方科技大学创新创业学院院长刘科认为,甲醇是目前最好的制氢材料<sup>[40]</sup>。我国有丰富的甲醇资源、生产技术成熟、产业链完善、价格低廉,并且甲醇本身是液态含氢元素比例最高的清洁能源、安全性高、使用和运输方便。因此,重整制氢研究工作不断向甲醇重整与燃料电池系统相结合的方向发展,以降低氢能压缩、存储与运输的成本,有效解决氢能推广的瓶颈问题。

行业分析机构标普全球普氏援引 Philippe Malbranche 称:目前可再生能源电解水制氢项目成本的 2/3 都来自于电解水所需的能耗,另 1/3 则来自于电解槽。这意味着,降低绿电成本是降低绿氢成本的关键<sup>[41]</sup>。近年来我国可再生能源发展迅速,光解水制氢、光催化分解水制氢、光伏发电制氢、光电化学制氢、太阳能制氢等方法可直接利用一次能源,通过光伏发电、太阳能、光解水、光催化等方式获取廉价的电能,

但目前弃电率较大。因此,在利用光制氢方面,将可再生能源释放的电能应用于电解水制氢或转化为氢能进行存储,这种方法的结合可以降低电成本与弃电率,是未来清洁能源的终极解决方案之一。

水解制氢主要包括硼氢化物水解制氢、金属基水解制氢和氢化物水解制氢,制备技术安全可靠,在热力学趋势上很容易与水发生反应产生氢气,能够即时制氢和即时供氢,有效解决氢能在使用和运输中的瓶颈,降低成本。另外,董仕节指出水解制氢后的产物都有很高的利用和经济价值,通过新能源汽车的使用回收后,可解决节能和环保的问题<sup>[42]</sup>。

氢能运用的途径首选是燃料电池,氢燃料电池技术的突破带动了氢的市场需求<sup>[37]</sup>。国家发改委、国家能源局发布的《氢能产业发展中长期规划(2021-2035年)》对氢能燃料电池汽车发展提出规划。氢气作为燃料电池的必要燃料,成为绿色能源转型的需求。目前,我国虽然在化石燃料制氢中发展成熟,但它所制得的氢气纯度不高并且会造成环境污染,违背绿色需求。因此,在考虑燃料电池所使用的氢能来源时,依靠较为环保的电解和重整方法,会对氢能发展产生巨大助力。

## 4 结论

针对当前主题抽取以及建立专利前沿主题指标的局限性,提出一种基于 SBERT 的专利前沿主题识别方法研究。首先获取专利文献数据源,其次对数据进行主题抽取工作:利用 SBERT 预训练模型获取句表征向量;二分

K-Means 对句表征向量聚类, 输出聚类主题结果和主题中心向量; 并以主题中专利分类号的相同率为标准进行正确率对比验证。然后进行主题关联与主题标识工作: 通过计算主题中心向量的相似度来关联主题; 利用 LDA 模型提取的主题词, 计算正点互信息值对主题关联效果进行验证, 并标识主题一级方向与二级方向。最后, 进行前沿主题识别工作: 建立关注度和质量水平的前沿识别指标。通过实证研究, 发现电解制氢、重整制氢、光制氢、水解制氢和燃料电池发电五个前沿主题, 同时验证了指标建立的有效性。

本文在主题抽取、主题关联与主题标识、前沿主题识别工作中, 较好地提高了主题抽取的准确性、解决构建指标主观性较强的问题, 为专利前沿主题的识别研究工作提供了新方法、新思路。在主题抽取方面, 首先利用 SBERT 模型对专利文本数据进行深度学习, 弥补了获取句表征向量时忽略上下文语境、向量空间各向异性等问题。其次在对表征句向量进行聚类时, 通过文本角度计算困惑度和向量角度利用肘部法则, 两者共同确定最佳簇的数量 K, 提高 K 取值的准确性。并且利用二分 K-Means 输出聚类主题以及主题中心向量, 以弱化随机初始值质心的影响, 加速执行速度并保证每一步误差最小。最后在聚类正确率对比验证时, 引入国际专利分类系统体系中的 IPC 号, 以其相同率对每个聚类主题的正确率分别验证, 增强验证结果的客观性。在主题关联与主题标识方面, 利用 LDA 模型提取主题词进行主题关联效果验证时, 以主题抽取工作中的主题聚类结果作为文本输入, 消除和缓解 LDA 模型无法自动确定

最优主题数和无法深入挖掘语义的问题。在前沿主题识别方面, 以每个主题的关注度与质量水平作为识别专利前沿主题的指标, 改进了前向引用次数、主题文本量作为前沿主题的识别指标, 并且无需设置阈值, 一定程度上提高了准确性和客观性。

然而, 该方法仍然存在一定的局限性。在数据源方面, 本文只使用了我国专利数据作为数据源, 今后可考虑加入其它国家或其它数据进行对比研究。在建立指标体系方面, 主要考虑专利前向引用与主题文本量的相关指标, 未来可以引入学科交叉性等指标, 从而形成更综合的识别方法。

## 参 考 文 献

- [1] Price D J. Networks of scientific papers[J]. Science, 1965, 149(3683): 510-515.
- [2] 崔宇红, 王飒, 高晓巍, 等. 基于全域微观模型的研究前沿主题探测和特征分析 [J]. 图书情报工作, 2018, 62(15): 75-82.
- [3] 王效岳, 白如江. 海量网络学术文献自分类技术研究 [M]. 北京: 人民出版社, 2015: 40-42.
- [4] 罗瑞, 许海云, 董坤. 领域前沿识别方法综述 [J]. 图书情报工作, 2018, 62(23): 119-131.
- [5] Garfield E. Research Fronts[J]. Current Contents, 1994, (41): 3-7.
- [6] Morriss A S, Yen G, Wu Z, et al. Time line visualization of re-search fronts[J]. Journal of the Association for Information Science & Technology, 2003, 54(5): 413-422.
- [7] 许振亮, 郭晓川. 国际技术创新研究前沿的科学计量学分析 [J]. 图书情报工作, 2011, 55(8): 49-53.
- [8] Huang M H, Chang C P. A comparative study on detecting research fronts in the organic light-emitting diode (OLED) field using bibliographic coupling and co-citation[J]. Scientometrics, 2014.
- [9] Park I, Lee K, Yoon B. Exploring promising research frontiers based on knowledge maps in the solar cell

- technology field[J]. Sustainability, 2015, 7(10): 13660-13689
- [10] 侯剑华, 李莲姬, 杨秀财. 基于引文网络结构变换的大数据研究前沿预测 [J]. 情报科学, 2018, 36(6): 142-148+168.
- [11] 郭伏, 胡名彩, 李明明. 文献共被引视角下的国际人因工程研究热点分析 [J]. 工业工程与管理, 2018, 23(3): 1-8.
- [12] Liu G Y, Jiang R, Jin Y. Sciatic nerve injury repair: a visualized analysis of research fronts and development trends[J]. Neural regeneration research, 2014, 9(18): 1716-1722.
- [13] Petrrs H P F, Raan A F J. Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling[J]. North-Holland, 1993, 22(1): 23-45.
- [14] 潘黎, 王素. 近十年来教育研究的热点领域和前沿主题——基于八种教育学期刊 2000-2009 年刊载文献关键词共现知识图谱的计量分析 [J]. 教育研究, 2011, 32(2): 47-53.
- [15] 张丽华. 研究前沿探测及其演化分析方法与实证研究 [D]. 北京: 中国科学院大学, 2015.
- [16] 武建鑫. 世界一流大学研究的主题演进与前沿探讨——基于 CNKI 数据库关键词共现知识图谱的分析 [J]. 湖南师范大学教育科学学报, 2016, 15(3): 92-99.
- [17] 郑彦宁, 许晓阳, 刘志辉. 基于关键词共现的研究前沿识别方法研究 [J]. 图书情报工作, 2016, 60(4): 85-92.
- [18] 范少萍. 基于文献计量学的研究前沿指标体系及其对我国科学基金评审的借鉴意义 [J]. 中国科学基金, 2016(2): 171-178.
- [19] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
- [20] Blei D M, Lafferty J D. Dynamic topic models [C]// International Conference on Machine Learning (ICML). 2006: 113-120.
- [21] yan x, guo j, lan y, et al. A biterm topic model for short texts[C]//International Conference on World Wide Web, 2013: 1445-1455.
- [22] Newman D, Asuncion A U, Smyth P, et al. Distributed inference for latent allocation [C]// Advances in neural information processing systems 20. New York: Curran Associates, Inc., 2007: 1-6.
- [23] Wang Y, Bai H J, Stanton M, et al. PLDA: Parallel latent Dirichlet allocation for large -scale applications[C]//Algorithmic aspects in information and management. Berlin: AAIM, 2009: 301-314
- [24] 范云满, 马建霞. 基于 LDA 与新兴主题特征分析的新兴主题探测研究 [J]. 情报学报, 2014, 33(7): 698-711.
- [25] 冯佳, 张云秋. 基于 LDA 和本体的科学前沿识别与分析方法研究 [J]. 情报理论与实践, 2017, 40(8): 49-54.
- [26] 王效岳, 刘自强, 白如江, 等. 基于基金项目数据的研究前沿主题探测方法 [J]. 图书情报工作, 2017, 61(13): 87-98.
- [27] 王菲菲, 刘明. Altmetrics 视角下的交叉学科研究前沿探测——以医学信息学领域为例 [J]. 情报学报, 2020, 39(10): 1011-1020.
- [28] 吴一平, 于纯良, 曲佳彬, 等. 文本主题视域下的高校论文研究前沿领域及演化发展趋势研究 [J]. 情报科学, 2021, 39(5): 156-162.
- [29] 陈虹枢, 宋亚慧, 金茜茜, 等. 动态主题网络视角下的突破性创新主题识别: 以区块链领域为例 [J]. 图书情报工作 2022, 66(10): 45-58.
- [30] 李松繁, 黄永, 杨金庆. 基于 BERT 的农业领域前沿研究主题识别方法研究 [J]. 情报工程, 2021, 7(5): 100-114.
- [31] 王秀红, 高敏. 基于 BERT-LDA 的关键技术识别方法及其实证研究——以农业机器人为例 [J]. 图书情报工作, 2021, 65(22): 114-125.
- [32] Nils R, Iryna G. Sentence-BERT: Sentence embeddings using siamese BERT-networks [C]// Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 2019: 3982-3992
- [33] Gao Tianyu, Yao Xingcheng, CHEN Danqi. SimCSE: simple contrastive learning of sentence embeddings[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for

- Computational Linguistics, 2021:6894-6910.
- [34] Yuanmeng Yan, Rumei Li, Sirui Wang, et al. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer[C]. //59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Part 8: 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021), and 11th International Joint Conference on Natural Language Processing (IJCNLP 2021), 2021, 1:5065-5075.
- [35] Geum Y, Kim C, Lee S, et al. Technological convergence of IT and BT: evidence from patent analysis[J]. Etri journal, 2012, 34(3): 439-449
- [36] Harhoff D, Narin F, Scherer F M, et al. Citation frequency and the value of patented inventions [J]. Review of economics and statistics, 1999, 81(3): 511-515.
- [37] 孔德婧, 董放, 陈子婧, 等. 离群专利视角下的新兴技术预测——基于 BERT 模型和深度神经网络[J]. 图书情报工作, 2021, 65(17): 131-141.
- [38] 田鹏伟, 张娴. 基于异构信息网络融合的专利技术主题识别研究[J]. 情报杂志, 2021, 40(8): 45-52.
- [39] 俞红梅, 邵志刚, 侯明, 等. 电解水制氢技术研究进展与发展建议[J]. 中国工程科学, 2021, 23(2): 146-152.
- [40] 龙跃梅, 叶青. 刘科院士谈甲醇经济与氢燃料电池: 以甲醇为原料水氢机拉开产业化序幕 [EB/OL]. (2019-07-03) [http://www.china-nengyuan.com/exhibition/exhibition\\_news\\_141640.html](http://www.china-nengyuan.com/exhibition/exhibition_news_141640.html).
- [41] 佚名. 光伏制氢的成本竞争力展望[J]. 新能源科技, 2022(1): 20-21.
- [42] 李阳, 雷燕超, 张熙廷. 四问“水氢车” 研发团队回应: 铝合金水解制氢 [EB/OL]. (2019-05-26) [http://k.sina.com.cn/article\\_1644114654\\_61ff32de02000m6gq.html](http://k.sina.com.cn/article_1644114654_61ff32de02000m6gq.html)