Statistics and Application 统计学与应用, 2026, 15(4), 24-39
Published Online April 2026 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2026.154068

Hans汉斯

基于专利文本挖掘的健康建筑技术主题识别分析

李　浩

同济大学经济与管理学院，上海

收稿日期：2026年2月27日；录用日期：2026年3月19日；发布日期：2026年4月1日

摘　要

目的/意义：随着“健康中国”战略和建筑高质量发展要求的不断推进，健康建筑已成为建筑领域的重要研究方向。通过对专利摘要文本进行文本挖掘，能够系统地识别健康建筑领域的技术主题，为中国健康建筑技术布局与产业发展提供数据支撑和方法参考。方法/过程：首先，基于《健康建筑评价标准》(T/ASC 02-2021)构建健康建筑专利检索体系，通过国家知识产权局专利检索与分析系统获取专利数据，并采用Jieba分词、停用词表与词典优化方法对筛选后的专利数据进行摘要文本的预处理；其次，在人工标注训练集的基础上构建多分类模型，实现健康建筑技术类别的自动识别；最后，运用LDA主题模型，对健康建筑专利数据进行分类别的技术主题识别，揭示中国健康建筑领域的技术主题分布情况。结果/结论：研究结果表明：(1) 空气与热环境相关技术在健康建筑专利中占据主导地位，反映出建筑的空气质量与热舒适是人们对健康建筑的核心关注点；(2) 监测技术在各类别中广泛存在，体现出针对建筑环境进行持续监测是实现健康建筑的重要技术路径。

关键词

健康建筑，专利分析，文本挖掘，LDA

Topic Identification and Analysis of Healthy Building Technology Based on Patent Text Mining

Hao Li

School of Economics and Management, Tongji University, Shanghai

Received: February 27, 2026; accepted: March 19, 2026; published: April 1, 2026

文章引用: 李浩. 基于专利文本挖掘的健康建筑技术主题识别分析[J]. 统计学与应用, 2026, 15(4): 24-39.
DOI: 10.12677/sa.2026.154068

Abstract

Objective/Significance: With the continuous advancement of the "Healthy China" strategy and the requirements for high-quality development in the construction industry, healthy buildings have become an important research direction in the field of architecture. By conducting text mining on patent abstracts, it is possible to systematically identify the technical themes in the field of healthy buildings, providing data support and methodological references for the technological layout and industrial development of healthy buildings in China. Method/Process: Firstly, a patent search system for healthy buildings was constructed based on the "Standard for Evaluation of Healthy Buildings" (T/ASC 02-2021), and patent data was obtained through the National Intellectual Property Administration's patent search and analysis system. The abstracts of the selected patent data were preprocessed using Jieba word segmentation, stopword lists, and dictionary optimization methods. Secondly, a multi-classification model was built based on a manually labeled training set to automatically identify the technical categories of healthy buildings. Finally, the LDA topic model was applied to classify and identify the technical themes of healthy building patents, revealing the distribution of technical themes in the field of healthy buildings in China. Result/Conclusion: The research results show that: (1) Technologies related to air and thermal environment dominate in healthy building patents, indicating that air quality and thermal comfort in buildings are the core concerns of people regarding healthy buildings; (2) Monitoring technologies are widely present in various categories, demonstrating that continuous monitoring of the building environment is an important technical path to achieving healthy buildings.

Keywords

Healthy Building, Patent Analysis, Text Mining, LDA

1. 引言

在城市化进程不断加快和居民健康需求持续提升的背景下，建筑环境对人体健康的影响日益受到重视。大量研究表明，室内空气质量、水系统和热、声、光环境等，均对居住者的生理与心理健康产生深远影响。传统建筑设计更多关注安全性、功能性与经济性，而对健康因素的系统考虑相对不足。

近年来，我国相继出台多项政策文件推动健康建筑发展。《健康中国 2030 规划纲要》明确提出要将健康理念融入城乡规划、建设和治理全过程。《健康建筑评价标准》(T/ASC 02-2021)的发布，标志着我国健康建筑评价体系逐步走向系统化与规范化[1]。在此背景下，健康建筑相关技术不断涌现，涵盖空气净化、水质保障、噪声控制、采光照明、热环境调节以及智能监测等多个方面。

专利文献作为技术创新活动的集中体现，不仅记录了技术方案的具体实现路径，也反映了技术演进的方向与重点[2]。通过对健康建筑专利的系统分析，可以揭示该领域技术发展的整体格局与内在规律，为学术研究和实践应用提供重要依据。本文将文本挖掘方法引入健康建筑研究领域，基于专利摘要文本进行健康建筑的技术主题识别，拓展了技术主题识别研究的分析视角，系统梳理了健康建筑技术体系，为健康建筑技术布局与产业发展提供了数据支撑和方法参考。

2. 相关研究

现有技术主题识别领域的相关研究主要从方法论和数据源两个维度展开。

在方法论方面早期的技术主题识别研究主要依赖领域专家的知识和经验，采用如德尔菲法、专家咨询法等定性方法[3]。这类方法能够凭借专家的洞察力把握技术发展的整体脉络，但其识别过程与结果不可避免地受到专家主观倾向的影响，导致其在客观性、可重复性以及处理海量数据时的效率方面存在局限[4]。

为克服定性方法的不足，量化分析方法逐渐成为主流。量化研究可进一步划分为基于文献计量学的方法和基于文本挖掘的方法。文献计量学方法主要利用科技文献的外部特征或简单内容特征。其中，引文分析通过构建共被引网络、文献耦合网络或直接引文网络，依据文献间的引用关系来聚类和识别研究主题[5]。该方法能够有效揭示知识单元间的关联与传播路径，但传统引文分析通常忽略引用语境与情感，对文本深层语义信息的挖掘不足[6]。共词分析则通过分析词汇在文献中的共现关系来构建网络并识别主题，其操作相对简便，但依赖于关键词的质量，且对词汇间的语义关系捕捉能力有限[7]。

随着自然语言处理技术的发展，文本挖掘方法成为技术主题识别研究的重要方向，其核心优势在于能够深入挖掘文本本身的语义信息。以 LDA (Latent Dirichlet Allocation)为代表的概率主题模型被广泛应用，它能够从大量非结构化文本中自动发现潜在的主题结构[8]。然而，传统 LDA 模型基于词袋假设，难以捕捉词汇的上下文语义和顺序信息，导致生成的主题可能存在连贯性差、可解释性不强等问题[9]。为提升语义理解能力，研究引入了词嵌入技术(如 Word2Vec)与主题模型结合，通过词的向量化表示来捕获语义关联，从而优化主题建模效果[10]。近年来，以 BERT 为代表的预训练语言模型推动了主题识别技术的进一步发展。BERTopic 等新型神经主题模型利用深度语义表示进行文档聚类和主题提取，在多项研究中展现出优于传统方法的性能[11]。然而，在面向专利等具有特定领域术语和法律语言的专用文本时，使用通用领域的预训练模型可能无法获得最优的文本向量表示，从而影响主题识别的准确性[12]。

从数据源视角看，技术主题识别研究主要基于论文、专利、政府基金项目等单一或多源数据展开。学术论文数据侧重理论基础与前沿探索，有助于发现新兴概念[13]；专利数据则紧密关联技术应用与产业化，蕴含丰富的技术细节与商业价值[14]；而基金项目数据能够反映研发投入与政策导向。越来越多的研究倡导融合多源数据(如论文与专利结合)，以期更全面、立体地刻画技术全景，弥合基础研究与应用研究之间的鸿沟[15]。

综上所述，技术主题识别研究在方法上正从依赖外部特征的计量分析向深度语义挖掘的文本分析演进，在数据上从单一数据源向多源融合分析发展。然而，现有研究大多聚焦于技术研发的基础研究与应用研究阶段，对开发研究阶段的关注相对不足；同时，如何进一步提升模型在跨领域、跨场景下的语义理解精度与自动化水平，减少对人工先验知识的依赖，仍是未来需要深入探索的方向[16]。

3. 研究框架与方法

首先，基于《健康建筑评价标准》(T/ASC 02-2021)构建健康建筑专利检索体系，通过国家知识产权局专利检索与分析系统获取专利数据，并采用 Jieba 分词、停用词表与词典优化方法对筛选后的专利数据进行摘要文本的预处理；其次，在人工标注训练集的基础上构建多分类模型，实现健康建筑技术类别的自动识别；最后，运用 LDA 主题模型，对健康建筑专利数据进行分类别的技术主题识别，揭示中国健康建筑领域的技术主题分布情况。本文研究框架见图 1。

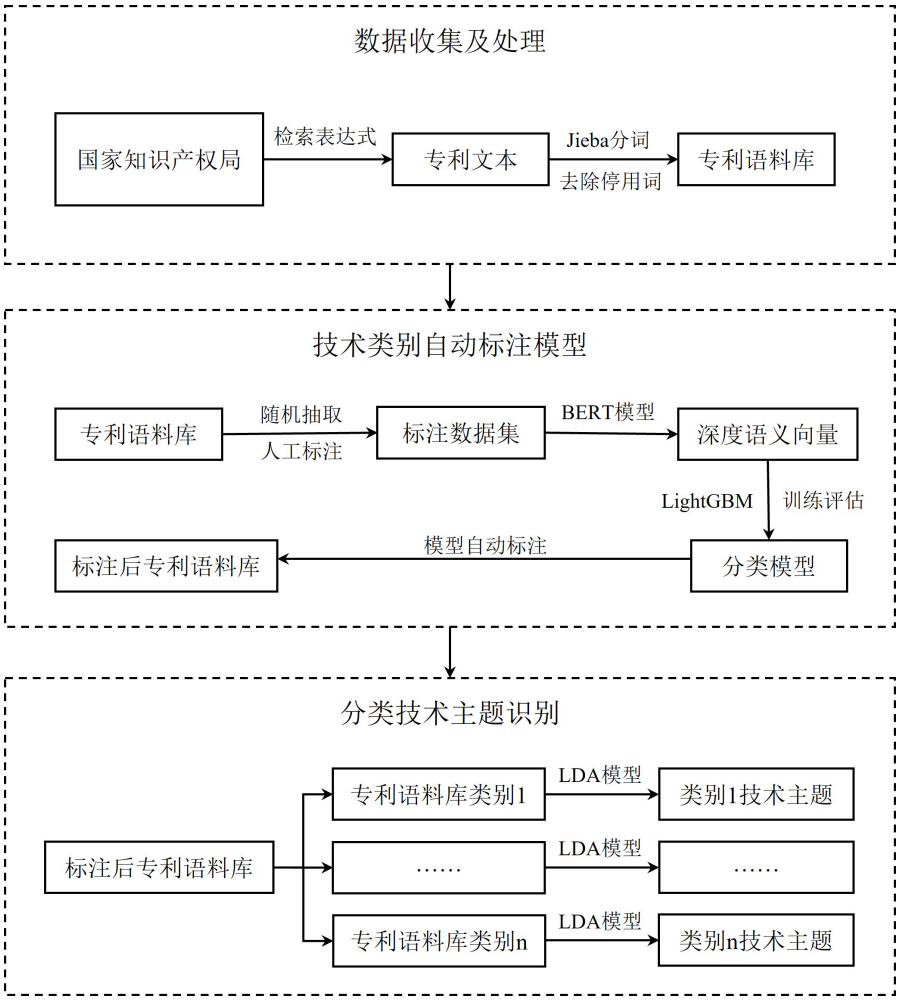

Figure 1. Diagram of research framework
图 1. 研究框架图

3.1. 数据筛选和清理

初始检索获得的专利数据中，仍包含部分与健康建筑关联度较低的专利。为提高数据质量，本文采用“自动筛选–人工校验”相结合的方式进行多轮清洗：(1) 筛选申请人所在国为中国的专利；(2) 剔除摘要内容缺失或信息严重不足的专利；(3) 基于程序规则再次匹配检索条件，确保专利文本严格符合健康建筑定义；(4) 对边界模糊样本进行人工复核，对施工工程、农业养殖、食品加工等明显偏离研究主题的内容进行排除。通过上述步骤，最终构建健康建筑专利分析数据库。

专利摘要文本属于非结构化中文文本，需进行系统预处理。本文主要包括以下步骤：(1) 中文分词：采用 Jieba 分词工具，并结合健康建筑领域特征，对分词词典进行人工补充与优化；(2) 停用词处理：引入哈尔滨工业大学停用词表，并根据研究需要手动补充领域无关高频词；(3) 低频与高频词过滤：删除出现频次过低或过高、区分度不足的词语；(4) 构建健康建筑领域的词袋模型与词典，为后续主题模型与分类模型提供输入。

3.2. 技术类别自动识别模型

本研究旨在构建一个高效、准确的专利文本自动分类系统，主要分为模型训练与批量标注两大阶段，

核心是结合 BERT (Bidirectional Encoder Representations from Transformers)模型获取深度语义特征，并利用 LightGBM (Light Gradient Boosting Machine)分类器完成分类任务。

首先，对已经进行筛选和预处理的专利摘要文本实行人工标注。随机抽取部分专利作为数据集，人工对其进行类别标注(空气、水、声、光、热等)，并将文本类别(字符串形式)转化为模型可处理的数值标签，使用 LabelEncoder 进行标签编码。

其次，采用 BERT 模型作为特征提取器，进行专利摘要的深度语义特征提取[17]。本研究选择中文预训练模型“bert-base-chinese”作为基础模型，该模型在海量中文语料上训练，能有效理解中文语义。将处理后的文本输入 BERT 模型，取模型输出的[CLS]标记对应的隐藏状态向量作为整个句子的语义表示。该向量是一个高维度的稠密向量，蕴含了文本的综合性语义信息。

再次，将 BERT 模型提取的语义特征向量作为输入特征，进行分类模型的构建、训练与保存。本研究选用 LightGBM 梯度提升决策树模型作为分类器。该模型具有训练效率高、内存消耗低、且能有效处理表格型数据的优点，尤其适合处理由 BERT 生成的高维特征[18]。将数据集按 8:2 的比例随机划分为训练集与测试集，并确保划分时保持各类别样本的比例(分层抽样)。使用训练集特征训练 LightGBM 模型，并在测试集上评估性能，输出包括精确率、召回率、F1 分数在内的详细分类报告以评估模型效果，通过对模型参数进行迭代优化，提高技术类别识别的稳定性与准确性。训练完成的 LightGBM 模型以及对应的标签编码器(LabelEncoder)将被序列化保存(joblib.dump)，供后续的批量标注任务直接加载调用。

最后，为将训练好的模型应用于大规模未标注数据，设计并实现了自动化的批量标注流程。预测完成后，系统将预测的类别标签(经标签编码器逆转换回原始类别名称)添加至原数据表中，并输出为结构化的 Excel 文件，完成整个自动化标注任务。

3.3. LDA 主题模型

本研究采用潜在狄利克雷分配(LDA)模型对健康建筑专利文本进行主题识别。LDA 是一种生成式概率图模型，其核心思想在于假设所有文档共享 K 个潜在主题，而每篇文档则表现为这些主题的混合[19]。具体地，模型假设文档的主题分布 θ 和主题的词语分布 φ 均服从狄利克雷(Dirichlet)先验分布。对于文档 m 中的第 n 个词语 $\omega_{m,n}$ 的生成过程，LDA 遵循以下概率步骤：(1) 从以参数 α 为超参数的狄利克雷分布中，抽样生成文档 m 的主题分布 θ_m。(2) 从多项式分布 Multinomial (θ_m)中，抽样生成词语 $\omega_{m,n}$ 所对应的主题编号 $z_{m,n}$。(3) 根据被抽中的主题 $z_{m,n}$，从其对应的词语多项式分布 $\Phi_{z_{m,n}}$ (其本身服从以 β 为超参数的狄利克雷分布)中，最终抽样生成观测到的词语 $\omega_{m,n}$ [20]。

因此，文档中任意一个词语 ω_i 出现的概率，可由所有主题的贡献求和得到，其数学表达为：

$$P(\omega_i|d)=\sum_{j=1}^{K}P(\omega_i|z_i=j)\cdot P(z_i=j|d)$$

其中，$P(\omega_i|z_i=j)$ 表示词语 ω_i 在主题 j 下的概率，$P(z_i=j|d)$ 表示主题 j 在文档 d 中的概率[21]。通过吉布斯采样或变分推断等算法，可以从观测到的文档集合中逆向估计出文档–主题分布(θ)和主题–词语分布(Φ)，从而实现对隐藏主题结构的挖掘。

LDA 模型需要预先指定主题数量 K，而 K 值的选取直接影响模型性能与主题的可解释性。为确定最优主题数，一般采用困惑度(Perplexity)或主题一致性(Topic Coherence)进行评估与筛选[22]，困惑度用于衡量模型对未知数据的预测能力，其值越低表明模型的泛化性能越好，主题一致性用于评估单个主题内部词语之间的语义一致性，其值越高意味着该主题越清晰、具有可解释性。本研究采用广泛使用的 c_v 一致性度量，它通过计算主题中高频词对之间的点互信息(PMI)来量化语义连贯性[23]。计算公式可简示为：

$$Coherence(V)=\frac{1}{N}\sum_{i<j}score(v_i,v_j)$$

其中，V 是描述该主题的 top-N 个高频词列表，score 是基于语料库计算的词语间语义关联度。

本研究通过遍历一个合理的 K 值范围(从 2 到 20)，分别训练多个 LDA 模型，并计算每个 K 值对应的平均主题一致性。最终，选取主题一致性较高的拐点区域作为最优主题数 K 的候选区间，并进一步结合对生成主题词的人工审阅，选定在统计意义和领域解释性上最为平衡的 K 值[22]。基于此最优 K 值重新训练最终的主题模型，并对生成的主题及其代表性词语进行人工凝练与命名，从而完成技术主题的提取。

4. 实证研究

4.1. 数据获取与处理

本文研究数据来源于国家知识产权局专利检索与分析系统(CNIPA)。选取该平台作为数据来源，主要基于其覆盖范围全面、数据权威性强以及对中国专利信息收录完整等优势。研究对象限定为申请人所在国为中国(CN)的专利文献，以保证研究结论对我国健康建筑技术发展的现实解释力。时间跨度设定为从 2000 年 1 月 1 日至 2025 年 12 月 31 日。

为系统识别健康建筑相关专利，本文以《健康建筑评价标准》(T/ASC 02-2021)为理论依据，从“健康要素”和“建筑对象”两个维度构建检索逻辑。在健康要素维度，重点涵盖空气、水、声、光、热等方面，具体包括空气质量监测与净化、水质保障、隔声降噪、采光照明、热舒适调节、无障碍设计、健身设施以及心理与安全相关技术。在建筑对象维度，涵盖住宅建筑、公共建筑及其细分类型，如住宅、公寓、宿舍、办公建筑、科研建筑、商业建筑、医疗建筑、文化建筑、交通建筑等。

通过布尔逻辑运算，将“建筑类关键词”与“健康类关键词”进行组合，形成“建筑 AND 健康”的复合检索式。最终确定检索式如下：“((建筑、室内、楼宇、楼房、房屋、别墅、住宅、公寓、宿舍、办公楼、行政楼、写字楼、实验楼、实验室、教学楼、教室、研究所、剧院、KTV、网吧、电影院、图书馆、博物馆、档案馆、文化馆、展览馆、音乐厅、礼堂、美术馆、商店、商场、超市、便利店、菜市场、旅馆、宾馆、餐店、餐厅、食堂、银行、酒店、、酒吧、邮局、体育场、体育馆、游泳馆、医院、康复中心、急救中心、疗养院、诊所、客运站、旅客站、航站楼、地铁站、法院、看守所、监狱、游乐场、景点建筑、大楼、大厦、商住楼、商务中心) AND (健康、OR (空气污染、空气质量、空气净化、空气监测、新风系统、气密性) OR ((浓度、AND、甲醛、颗粒物、PM2.5、PM10、苯、TVOC) VOCs、霉菌、防霉)、OR (水质、直饮水、(饮用水、AND 硬度、菌落、污染、杀菌、抑菌))、OR ((管道、AND、渗漏、防结露) (卫生间、AND、防干涸、同层排水、无接触、无障碍))、OR (声环境、隔声、隔音、降噪、低噪) OR (光环境、天然光、日照标准、采光、照度、照明控制、照明亮度、室内亮度) OR (热环境、热舒适、热感觉、热湿、遮阳、通风、(空气、AND、湿度)、暖通空调)、)))、NOT (施工、工程、工人)”。

4.2. 健康建筑技术类别体系构建

本文在健康建筑评价标准与既有研究基础上，将健康建筑技术划分为若干核心类别，包括空气、水、声、光、热等。该分类体系既能够覆盖健康建筑的主要技术方向，又具备较好的可操作性与可解释性。根据该分类体系，本研究构建人工标注训练集 1750 条，经模型训练、评估与优化后，最终得到健康建筑自动标注模型，该模型评估结果见表 1。

Table 1. Evaluation results of automatic annotation model for healthy buildings
表 1. 健康建筑自动标注模型评估结果

	precision	recall	f1-score	support
空气	0.72	0.84	0.78	82
光	0.91	0.72	0.82	48
声	0.74	0.76	0.75	65
水	0.83	0.86	0.85	57
热	0.69	0.85	0.77	98
accuracy			0.77	350
macro avg	0.78	0.81	0.79	350
weighted avg	0.76	0.81	0.79	350

在人工标注训练集的基础上构建的多分类模型，对健康建筑专利技术类别具有较好的识别能力。通过对测试集的分类结果进行评估，模型在总体准确率、宏平均与加权平均指标上均取得较为稳定的表现，表明该模型能够在多类别、不均衡样本条件下实现有效分类。

4.3. 健康建筑技术类别总体分布特征

基于自动标注模型对健康建筑专利进行技术类别识别后，可得到专利数量分布情况，见图 2。

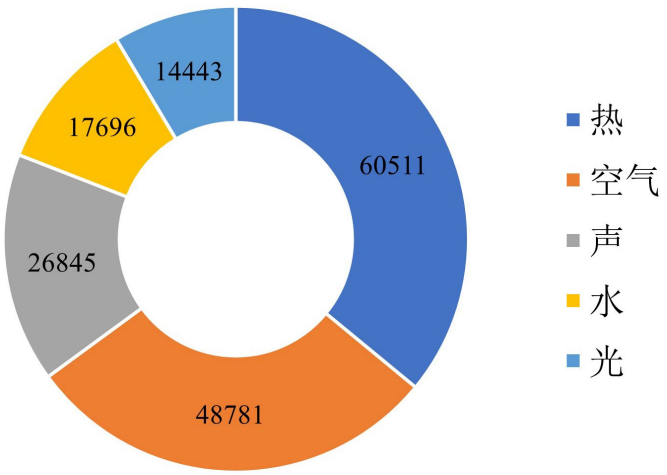

Figure 2. Distribution map of health building patents in China by category
图 2. 中国健康建筑专利类别分布

整体来看，健康建筑技术呈现出明显的结构性差异，不同技术类别在专利数量和发展活跃度方面存在显著不均衡。从总体分布看，空气与热环境相关技术专利数量明显高于其他类别，构成健康建筑技术体系中的核心部分。这一结果表明，在建筑健康性能提升过程中，室内空气质量控制与热舒适调节仍是技术创新的主要发力点。相比之下，声环境、光环境、水系统等技术类别处于第二梯队。

这种分布特征在一定程度上反映了健康建筑技术研发的现实需求结构。一方面，空气污染和热环境问题对人体健康影响直接且显著，易形成明确的技术需求与市场空间；另一方面，在建筑隔音，建筑采光，水系统等间接作用于人体健康的领域，未来有望进一步在建筑相关技术的研发中受到更多的重视。

4.4. 基于 LDA 主题模型的分类别主题识别

4.4.1. 空气类技术主题

通过综合考虑主题一致性分数折线图(见图 3)和领域解释性，最终选定空气类技术主题数量为 7。

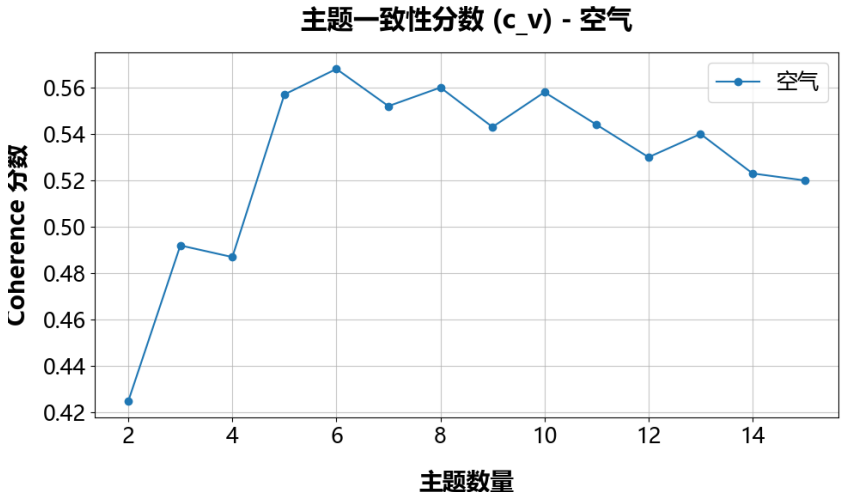

Figure 3. Line chart of air-related topic consistency scores
图 3. 空气类主题一致性分数折线图

Figure 4. Word cloud of air-related topic
图 4. 空气类主题词云图

在设定主题数量下经过 LDA 聚类得到空气类别技术主题分布及其 TOP 10 关键词，见表 2。

Table 2. Distribution of air-related technology topics and their TOP 10 keywords
表 2. 空气类技术主题分布及其 TOP 10 关键词

技术主题	TOP 10 关键词										数量
空气加湿与调节装置	加湿	水箱	加热	除尘	雾化	腔室	加湿器	机箱	集尘	喷头	4692
新风空调系统	新风	空调	风机	空调器	排风	进风	新风系统	送风	风道	调节	8209

续表

空气过滤装置	过滤	通风	过滤网	净化器	滤网	电机	灰尘	管道	风机	进气	12939
空气质量监测技术	检测	空气质量	传感器	监测	浓度	数据	控制器	调节	采集	信息	8316
甲醛吸附净化技术	甲醛	吸附	材料	涂料	纳米	活性炭	装饰	光触媒	复合	抗菌	6390
空气消毒与抗菌技术	净化器	消毒	风机	杀菌	负离子	过滤器	发生器	过滤	紫外线	臭氧	8235

根据词云图(见图 4)和关键词表(见表 2)可以看出，空气类技术是健康建筑中专利布局最密集、创新最活跃的领域，核心技术围绕净化处理、环境调节与智能监控三大体系协同作用。净化处理是核心，包含物理过滤(过滤装置)、化学与物理消杀(消毒抗菌技术)以及针对气态污染物的吸附分解(甲醛净化技术)，构成多层次防御。环境调节体系通过新风空调系统实现空气的循环与温湿度的基础调控，并辅以专用的加湿装置精细控制湿度。所有这些系统的运行都依赖于智能监控体系，即通过高密度的传感器网络实时监测空气质量，并基于数据驱动实现系统的自动调节与联动控制。这标志着健康建筑的空气管理正从单一设备向全屋、主动、智能的整体解决方案飞速演进。

净化处理是研发最密集的领域，这直接反映了 COVID-19 疫情对技术路线的深刻影响，催生了大量针对病毒消杀的技术创新。在政策方面，《健康中国 2030》战略及配套的《健康环境促进行动实施方案(2025~2030 年)》将室内空气质量提升为国家行动目标，《健康建筑评价标准》将室内空气质量控制纳入评分体系，这些政策促进了监测与净化联动系统的技术集成。

4.4.2. 水类技术主题

综合考虑主题一致性分数折线图(见图 5)和领域解释性，最终选定水类技术主题数量为 7。

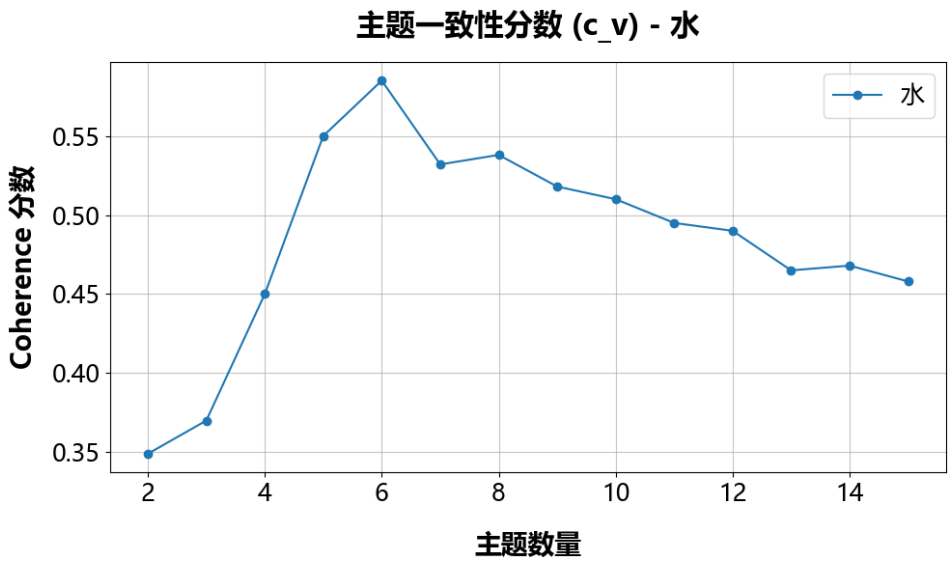

Figure 5. Line chart of water-related topic consistency scores
图 5. 水类主题一致性分数折线图

Figure 6. Word cloud of water-related topic
图 6. 水类主题词云图

在设定主题数量下经过 LDA 聚类得到水类别技术主题分布及其 TOP 10 关键词，见表 3。

Table 3. Distribution of water-related technology topics and their TOP 10 keywords
表 3. 水类技术主题分布及其 TOP 10 关键词

技术主题	TOP 10 关键词										数量
水质监测技术	检测	水质	监测	数据	传感器	取样	分析	采集	样品	采样	2778
防水防渗技术	防水	混凝土	密封	墙体	清洗	水泥	保温	废气	减压	干燥	2671
排水系统技术	管道	排水	排水管	地漏	支管	卫生间	水封	排水管道	密封	接头	2683
供水与热水系统	水箱	供水	加热	水泵	储水	水管	太阳能	水质	热水	空调	2725
水处理与净化技术	过滤	净化	水质	污水	废水	污水处理	消毒	过滤器	出水	沉淀	4014
雨水收集利用技术	雨水	收集	通风	除尘	水箱	回收	垃圾	粉尘	喷淋	风机	2825

根据词云图(见图 6)和关键词表(见表 3)可以看出，水类技术是健康建筑专利中的重要组成部分，相关技术主要围绕水质监测、水系统保障与水质净化三大方向展开。水质监测通过传感器、采样与分析设备实现对建筑用水的实时监控与数据采集，是智能水管理的基础。水系统保障涵盖多个具体技术环节：从建筑结构层面的防水、密封，到管道层面的排水、防堵，再到生活所需的稳定供水与节能热水(如太阳能加热)。水质净化则聚焦于通过过滤、沉淀、消毒等方式处理生活污水与废水，并结合雨水收集、回收与再利用技术，实现水资源的循环与节约。以上技术主题共同体现了社会对于建筑用水安全、系统可靠与资源效率的高度重视，着力于构建从源头到排放的全流程健康水环境。

从专利数据来看，政策与标准直接塑造了技术方向：“水十条”与环保督察：推动了污水处理、废水净化技术的密集创新，以满足日益严格的排放标准。“海绵城市”试点与评价标准：显著促进了雨水收集利用技术的发展，关键词中“回收”、“水箱”体现了对蓄水与资源化的关注。《绿色建筑评价标准》与《健康建筑评价标准》将节水器具、非传统水源利用、供水水质安全纳入评分体系，这不仅驱动了供水系统的节能与水质保障技术，也使得水质在线监测成为实现标准认证与健康承诺的基础支撑技术。未来，在“双碳”目标与建筑高质量发展背景下，研发将更侧重于智慧水务(监测、处理、输送的全系统智能联动)和水资源在建筑内部的闭环循环与高效利用。

4.4.3. 声类技术主题

综合考虑主题一致性分数折线图(见图 7)和领域解释性，最终选定声类技术主题数量为 4。

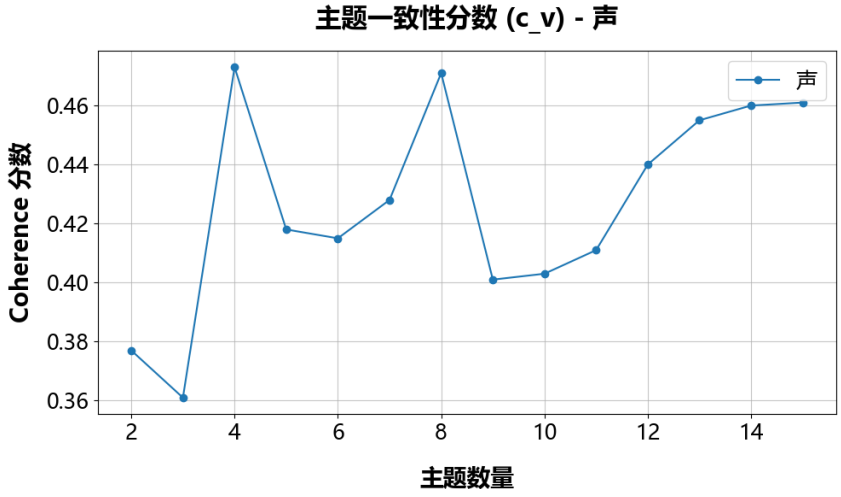

Figure 7. Line chart of sound-related topic consistency scores
图 7. 声类主题一致性分数折线图

Figure 8. Word cloud of sound-related topic
图 8. 声类主题词云图

在设定主题数量下经过 LDA 聚类得到声类别技术主题分布及其 TOP 10 关键词，见表 4。

Table 4. Distribution of sound-related technology topics and their TOP 10 keywords
表 4. 声类技术主题分布及其 TOP 10 关键词

技术主题	TOP 10 关键词										数量
建筑结构消声与减震技术	消音	幕墙	降噪	减震	通风	弹簧	缓冲	管道	隔音板	安装板	5893
复合保温材料	保温	材料	复合	防火	装饰	板材	隔热	阻燃	防水	轻质	8219
装配式墙体隔声技术	墙体	墙板	保温	隔音板	龙骨	装配式	混凝土	凹槽	装饰	保温层	7550
门窗降噪与密封技术	玻璃	降噪	隔热	型材	门窗	密封	铝合金	隔声	窗框	吸声	5183

根据词云图(见图 8)和关键词表(见表 4)可以看出，声类技术专注于为健康建筑打造低干扰、高舒适的声学环境，其专利布局体现了从材料基础、构造设计到系统集成的全方位噪声控制策略。材料基础层面，研发重点在于具备保温、防火、防水等多功能一体化的轻质复合材料，这些材料本身也是良好的隔声体。构造设计层面，则通过装配式墙体、集成龙骨与隔音板的墙体系统，以及采用密封型材、隔热玻璃的门窗系统，在建筑围护结构上形成高效的声屏障。系统集成层面，针对通风管道、设备安装等特定噪声源，发展出专门的消音、减震与缓冲技术。这些技术共同构建了一个“材料－构件－系统”三层级的综合噪声防控体系，以系统性思维应对建筑内外的各类噪声挑战。

根据声类别专利数据，复合保温材料是研发最集中的领域，这体现了节能与隔声的一体化融合趋势，直接响应了国家“双碳”目标下对建筑围护结构高性能化的要求。装配式建筑相关政策强力推动了装配式墙体隔声技术的快速发展，关键词“墙板”、“龙骨”、“凹槽”反映了与工业化生产、现场干法施工紧密结合的构造创新。《绿色建筑评价标准》对室内噪声级有强制要求，《健康建筑评价标准》更强调声舒适度。这促使技术从单一“隔声”向系统化“降噪”升级，推动了建筑结构消声与减震和门窗密封降噪技术的精细化发展，以应对交通、设备等不同噪声源。在提升居住品质和建筑工业化的双重背景下，下一步的研发预计将更侧重于材料的声学－热工－力学性能多功能集成，以及面向学校、医院等敏感建筑的智能主动降噪与声环境个性化调节技术。

4.4.4. 光类技术主题

综合考虑主题一致性分数折线图(见图 9)和领域解释性，最终选定光类技术主题数量为 4。

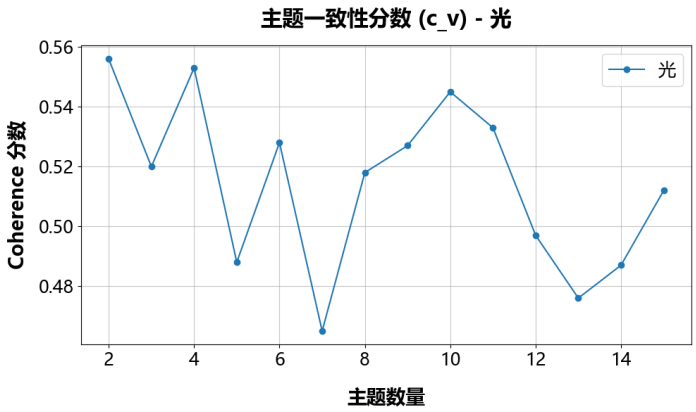

Figure 9. Line chart of light -related topic consistency scores
图 9. 光类主题一致性分数折线图

Figure 10. Word cloud of light -related topic
图 10. 光类主题词云图

在设定主题数量下经过 LDA 聚类得到光类别技术主题分布及其 TOP 10 关键词，见表 5。

Table 5. Distribution of light-related technology topics and their TOP 10 keywords
表 5. 光类技术主题分布及其 TOP 10 关键词

技术主题	TOP 10 关键词										数量
自然光采集与调节系统	采光	调节	照明	光线	电机	支架	采光板	阳光	光管	太阳光	3048
光伏建筑一体化技术	玻璃	采光	太阳能	光伏	窗框	通风	幕墙	保温	窗户	门窗	5090
智能照明检测与控制系统	照明	传感器	检测	控制器	数据	信号	电路	采集	控制系统	信息	3499
LED 照明技术	LED	光源	照明	灯具	发光	照度	散热	模型	电源	灯罩	2138

根据词云图(见图 10)和关键词表(见表 5)可以看出，光类技术致力于为健康建筑创造舒适、节能且智能的光环境。相关专利主要围绕自然光的最大化利用与精准调控、人工光的健康化与智能化两大核心展开。自然光利用是重点，通过可调节的采光装置、以及将光伏发电与建筑围护结构(如幕墙、窗户)深度集成的一体化设计，在引入阳光的同时实现能源生产。人工光营造则包含两个层面：一是通过传感器和控制系统实现照明环境的智能监测与按需调节；二是针对 LED 光源、灯具本身在发光效率、散热等方面的持续优化。这些技术共同构建了一个从自然采光到智能补光，兼顾视觉健康、心理舒适与能源可持续的综合性解决方案。

从政策方面看，“双碳”目标与能源政策：直接推动了光伏建筑一体化技术的爆发式发展。关键词“玻璃”、“幕墙”、“窗户”显示，研发重点在于将光伏发电功能无缝集成到建筑外围护结构中，使建筑从能源消耗者转变为生产者；《绿色建筑评价标准》对节能与可再生能源利用有明确要求；《健康建筑评价标准》则强调光健康与视觉舒适度。这共同引导了自然光采集与调节系统的发展，以实现节能与健康的平衡，并推动了智能照明控制系统向基于传感器和数据的个性化、自适应调节演进。智慧城市与物联网等概念的发展则为智能照明检测与控制技术提供了底层支撑，使其从单一开关控制升级为可监测、可采集、可联网的智慧环境子系统。未来，相关领域的研发将更聚焦于光伏构件与建筑美学、结构安全的一体化融合，以及遵循人体节律的健康光环境智能调控系统。

4.4.5. 热类技术主题

综合考虑主题一致性分数折线图(见图 11)和领域解释性，最终选定热类技术主题数量为 7。

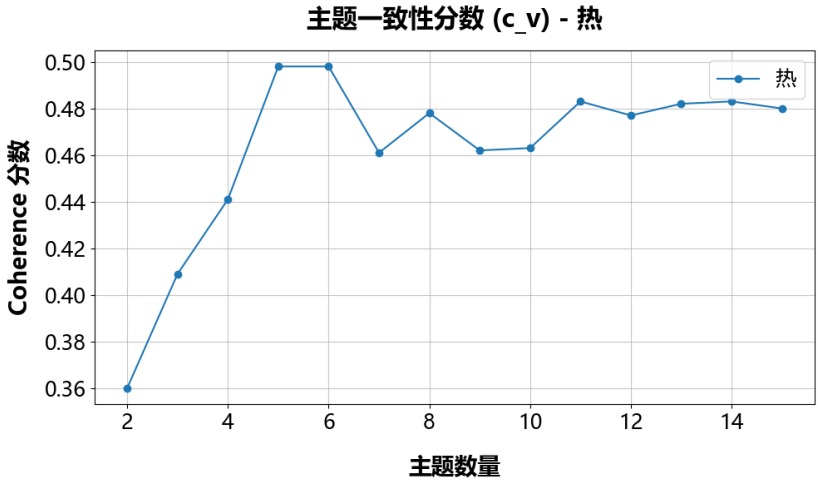

Figure 11. Line chart of heat-related topic consistency scores
图 11. 热类主题一致性分数折线图

Figure 12. Word cloud of heat-related topic
图 12. 热类主题词云图

在设定主题数量下经过 LDA 聚类得到热类别技术主题分布及其 TOP 10 关键词，见表 6。

Table 6. Distribution of heat-related technology topics and their TOP 10 keywords
表 6. 热类技术主题分布及其 TOP 10 关键词

技术主题	TOP 10 关键词										数量
通风散热技术	通风	风机	空气	散热	进风	过滤	管道	排风	电机	通风管	17877
太阳能光伏发电技术	太阳能	通风	墙体	光伏	保温	幕墙	屋顶	材料	温度	发电	6183
空调温控技术	空调	温度	空调器	导风	出风	风机	调节	送风	环境温度	设定	6495
通风遮阳装置	通风	遮阳	烘干	玻璃	调节	窗框	电机	叶片	窗户	百叶	12825
温度监测与调控技术	温度	数据	调节	传感器	检测	控制器	湿度	模型	调控	采集	7119
新风换热系统	新风	空调	空气	换热器	风道	换热	风机	除湿	送风	机组	10012

根据词云图(见图 12)和关键词表(见表 6)可以看出，热类技术致力于为健康建筑营造舒适、稳定且节能的室内热环境，其专利围绕温度调节、通风散热、能源利用与智能控制四大方向紧密布局。温度调节是核心目标，通过空调系统、新风与热回收系统实现空气的加热、冷却、除湿与能量高效交换。通风散热是基础手段，既包含强制性的机械通风与空气循环，也涵盖通过可调节遮阳、窗户等实现的自然通风与被动散热。在能源利用层面，太阳能光伏与建筑围护结构的一体化设计，为实现低能耗的温度调节提供了可能。整体系统的运行则依赖于智能控制，即通过传感器网络实时监测温湿度，并利用数据模型驱动通风、空调、遮阳等系统协同工作。这标志着一个从被动适应到主动优化、从单一设备到系统联动、兼顾舒适性与能效的综合性热环境管理体系的形成。

从政策方面分析，“双碳”目标与建筑节能设计标准所代表的强制性的节能要求，推动了新风换热系统的广泛应用，以实现能量回收；同时激励了太阳能光伏发电技术与建筑本体的结合。《健康建筑评价标准》中对室内热舒适与自然通风的强调，促进了温度监测与调控技术向基于数据模型的精细化、智能化发展。未来，相关领域技术研发预计将更侧重于通风、遮阳、散热与能源回收的集成化系统设计，以及利用人工智能算法实现的个性化、预测性热环境调控，以同时满足节能、健康与舒适的综合目标。

5. 总结与展望

本文围绕健康建筑技术这一交叉研究对象，构建了融合文本挖掘与主题建模的方法框架。具体而言，

通过引入多类别文本分类模型，实现了健康建筑技术类别的自动识别；通过与 LDA 主题模型的结合，系统刻画了健康建筑领域的技术主题分布情况。研究结果表明：(1) 空气与热环境相关技术在健康建筑专利中占据主导地位，反映出建筑的空气质量与热舒适是人们对健康建筑的核心关注点；(2) 监测技术在各类别中广泛存在，体现出针对建筑环境进行持续监测是实现健康建筑的重要技术路径。

尽管本文在方法与实证分析方面进行了系统探索，但仍存在一定不足。首先，专利文本分析难以全面反映技术实际应用效果。其次，未对动态演化过程进行深入刻画。未来研究可在以下方面进一步拓展：一是结合标准、规范与工程案例数据，对健康建筑技术应用效果进行多源验证；二是引入动态网络分析方法，刻画技术时间演化特征；三是结合语义嵌入与因果推断方法，深化对健康建筑技术演化机制的理解。

参考文献

[1] 中国建筑学会. 健康建筑评价标准: T/ASC 02-2021 [S]. 北京: 中国建筑工业出版社, 2021.

[2] 杨铁军. 专利信息利用导引[M]. 北京: 知识产权出版社, 2011.

[3] 万校基, 李海林, 何雨晴, 等. 热度演化视角下新兴主题识别分析研究[J]. 图书情报工作, 2024, 68(22): 126-138.

[4] 许佳琪, 汪雪锋, 陈虹枢, 等. 跨领域颠覆性技术主题识别研究: 以脑科学技术为例[J]. 图书情报工作, 2024, 68(15): 44-57.

[5] Kleminski, R., Kazienko, P. and Kajdanowicz, T. (2020) Analysis of Direct Citation, Co-Citation and Bibliographic Coupling in Scientific Topic Identification. *Journal of Information Science*, **48**, 349-373. https://doi.org/10.1177/0165551520962775

[6] 柴文越, 刘小平, 梁爽. 新兴主题识别方法研究综述[J]. 现代情报, 2023, 43(12): 164-177.

[7] Wang, X., He, J., Huang, H. and Wang, H. (2022) Matrixsim: A New Method for Detecting the Evolution Paths of Research Topics. *Journal of Informetrics*, **16**, Article ID: 101343. https://doi.org/10.1016/j.joi.2022.101343

[8] Suominen, A., Toivanen, H. and Seppänen, M. (2017) Firms' Knowledge Profiles: Mapping Patent Data with Unsupervised Learning. *Technological Forecasting and Social Change*, **115**, 131-142. https://doi.org/10.1016/j.techfore.2016.09.028

[9] 王晨, 廖启明. 基于改进的 LDA 模型的文献主题挖掘与演化趋势研究——以个人隐私信息保护领域为例[J]. 情报科学, 2023, 41(10): 112-120.

[10] Ma, J., Wang, L., Zhang, Y., Yuan, W. and Guo, W. (2023) An Integrated Latent Dirichlet Allocation and Word2vec Method for Generating the Topic Evolution of Mental Models from Global to Local. *Expert Systems with Applications*, **212**, Article ID: 118695. https://doi.org/10.1016/j.eswa.2022.118695

[11] Rejeb, A., Rejeb, K., Simske, S. and Süle, E. (2025) Industry 5.0 Research: An Approach Using Co-Word Analysis and BERTopic Modeling. *Discover Sustainability*, **6**, Article No. 402. https://doi.org/10.1007/s43621-025-01252-3

[12] 薛航, 施国良, 陈挺. 基于对比学习的高价值发明专利识别: 以无线通信网络领域为例[J]. 情报杂志, 2024, 43(9): 179-187.

[13] 王桂芳, 何涛, 马廷灿, 等. 基于科技文献的生物核磁领域技术机会识别[J]. 科技管理研究, 2016, 36(10): 142-147.

[14] 杨恒, 王曰芬, 张露. 基于核心专利技术主题识别与演化分析的技术预测[J]. 情报杂志, 2022, 41(7): 49-56.

[15] 国家市场监督管理总局, 国家标准化管理委员会. 科学技术研究项目评价通则: GB/T 22900-2022 [S]. 北京: 中国标准出版社, 2022: 4.

[16] 马铭, 王超, 周勇, 等. 基于语义信息的核心技术主题识别与演化趋势分析方法研究[J]. 情报理论与实践, 2021, 44(9): 106-113.

[17] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.

[18] Ke, G., Meng, Q., Finley, T., *et al*. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 3146-3154.

[19] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993-1022.

[20] 白如江, 刘博文, 冷伏海. 基于多维指标的未来新兴科学研究前沿识别研究[J]. 情报学报, 2020, 39(7): 747-760.

[21] Steyvers, M. and Griffiths, T. (2007) Probabilistic Topic Models. In: *Handbook of Latent Semantic Analysis*, Routledge, 424-440.

[22] 余厚强, 王玥, 吴婷婷, 等. 基于政策文献计量的我国新时期科技评价体系改革进程研究[J]. 情报科学, 2022, 40(8): 20-28.

[23] Stevens, K., Kegelmeyer, P., Andrzejewski, D., *et al*. (2012) Exploring Topic Coherence over Many Models and Many Topics. *Proceedings of the* 2012 *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, July 2012, 952-961.