Management Science and Engineering 管理科学与工程, 2025, 14(1), 46-52
Published Online January 2025 in Hans. https://www.hanspub.org/journal/mse
https://doi.org/10.12677/mse.2025.141006

Hans汉斯

基于实体识别的纺织技术主题内容演化研究

胡　莹*，董平军#

东华大学旭日工商管理学院，上海

收稿日期：2024年12月8日；录用日期：2025年1月3日；发布日期：2025年1月10日

摘　要

专利文本是技术创新的核心构建要素，对文本内容进行主题分析有助于厘清技术主题分布及演变趋势。以2018~2022年间知网纺织面料制备技术专利为研究对象，利用命名实体识别进行研究，以提取物体类实体作为专利文本内容分析的依据，按年划分时间窗口，使用困惑度－主体方差得到最优主题数。通过分析技术主题内容演变过程总结得到纺织面料制备的创新模式。通过分析主题内容演变过程，将其归纳为面料原料、面料制备工艺和面料特性三组技术元素，给出进一步面料制备的开发建议。为了克服主题建模中难以准确快速地选定词簇表示主题的难题，利用命名实体识别技术简化技术术语抽取工作，使用ERNIE3.0知识增强预训练模型快速得到具备强概括能力的技术术语集合。

关键词

命名实体识别，技术术语抽取，LDA主题模型，主题内容演化

Research on the Evolution of Textile Technology Theme Content Based on Entity Recognition

Ying Hu*, Pingjun Dong#

Glorious School of Business and Management, Donghua University, Shanghai

Received: Dec. 8th, 2024; accepted: Jan. 3rd, 2025; published: Jan. 10th, 2025

Abstract

Patent text is the core building element of technological innovation, and thematic analysis of text content is helpful to clarify the distribution and evolution trend of technological themes. Taking

*第一作者。
#通讯作者。

文章引用: 胡莹, 董平军. 基于实体识别的纺织技术主题内容演化研究[J]. 管理科学与工程, 2025, 14(1): 46-52.
DOI: 10.12677/mse.2025.141006

CNKI textile fabric preparation technology patents from 2018 to 2022 as the research object, Named-entity recognition was used for research, and object-like entities were extracted as the basis for patent text content analysis. Time windows were divided by year, and the optimal number of topics was obtained using Perplexity subject variance. Summarize the innovative mode of textile fabric preparation by analyzing the evolution process of technical theme content. By analyzing the evolution process of the theme content, it is summarized into three technical elements: fabric raw materials, fabric preparation process, and fabric characteristics, and further development suggestions for fabric preparation are provided. In order to overcome the difficulty of accurately and quickly selecting word clusters to represent topics in topic modeling, Named-entity recognition technology is used to simplify the extraction of technical terms, and ERNIE3.0 knowledge enhancement pre-training model is used to quickly obtain technical term sets with strong generalization ability.

Keywords

Named-Entity Recognition, Technical Terminology Extraction, LDA Topic Model, Theme Content Evolution

1. 引言

伴随着新的科技革命，全球范围内正在加紧以技术预测来抓住先发的优势和占领科技创新制高点。科学有效地开展技术预测工作，可以帮助各国、各公司、各高校准确地把握技术研究热点和技术发展线索，追踪技术发展态势、预测未来的技术发展趋势，尽早找到技术发展的机遇。专利文献是技术创新能力的一个重要体现，包含着巨大的经济价值和技术价值，非常适合进行技术挖掘，但在专利文献激增的情况下，有必要寻找一种快速、精确的技术挖掘分析方法，针对大规模专利文献有效减少数据集、开展技术主题识别及演化分析等将成为一种高效可行的手段。所以，对专利进行技术主题分析能够追踪和预测技术的发展状况，从而提高技术追踪和预测的结果及效率。

在过去十年里，技术主题识别研究一直备受关注，是科学文献文本分析中的一项关键技术。无论使用何种方式进行技术主题识别，技术术语抽取始终是重要子任务，直接影响技术主题识别效果。黄晓斌等指出确认用于指代主题的核心词是较大的难点[1]。随着自然语言处理领域的迅猛发展，为了更加充分地利用文本的语义信息，技术术语抽取技术开始朝着机器学习和深度学习方向发展[2]。技术术语识别任务的目的是识别包含特定领域知识的文本中蕴含技术概念的字符串。蕴含技术概念的字符串也可以被视为一种实体，可以使用命名实体识别方法完成抽取。使用命名实体识别方法抽取命名实体只需进行实体挑选操作，而无需切词分词、建立停用词词典、去停用词、词性标注等机械重复的人工文本预处理操作。

为了克服主题建模中难以准确快速地选定词簇表示主题的难题，本文将利用命名实体识别技术简化技术术语抽取工作，快速得到具备强概括能力的技术术语集合。本文还将在此基础上使用 LDA 模型，结合相似度计算等方法来构建不同时序间的主题关联，并以桑基图来直观呈现纺织面料专利主题间的互动演化关系。

2. 相关研究现状述评

2.1. 主题内容演化研究现状

主题内容演化指的是主题内容随时间的发展情况。为了完成主题内容演化分析，需要确定提取主题

的方法。主题提取主要包括两点操作，先从文本中提取技术术语，再将技术术语组合成主题。近年来主题提取实践路径总结如下：

(1) 基于关键词聚类的提取方法。文献[3]利用 RAKE 抽取关键词作为技术术语并通过 word2vec 转换成词向量，然后用 k-means 算法提取主题。

(2) 基于社会网络的共词分析法。文献[4]用 c-value 和 Tf-Idf 抽取关键词作为技术术语并将其表示为共现网络，然后用社区发现算法提取主题。

(3) 基于主题模型的主题提取方法。相比于关键词共现网络和关键词聚类，LDA 模型及其变体关注全文语义信息，通过文档建模对文档进行语义分解，从潜在技术术语集中挑选技术术语，找到隐含主题信息，从而有效表达文档集内部特征。LDA 模型作为无监督学习模型，降低了人工参与所耗费的成本。最优主题数的选取方法不断升级，本文将使用困惑度－主题方差确定最优主题数。

无论使用何种主题提取方法，第一步都是获得技术术语集合。现有研究获取技术术语的方式基本为分词、建立停用词词典或专业领域词典、去停用词、词性标注等机械重复的人工文本预处理操作。相比于上述人工文本预处理操作，命名实体识别作为一项自然语言处理技术，能够更加自动化地完成技术术语提取，但至今仍尚少被应用于主题研究中提取技术术语集合。

2.2. 命名实体识别研究现状

命名实体识别(NER)是自然语言处理的一部分。NER 的主要目标是处理结构化和非结构化数据，并将这些命名实体分类为预定义的类别。一些常见的类别包括姓名、地点、公司、时间、货币价值、事件等。现有命名实体识别任务的解决方案包括基于规则和基于机器学习两种。

基于规则的方式相对于基于机器学习的方式需要使用较多的专家知识来制定规则，规则的设计一般基于句法模式，所以往往只适用于非常有限的文本材料。专利文献的内容特点之一就是创新性高，故基于规则的方法往往不能适应提取新科学出版物所特有的新概念词汇的要求。

现有文献用于完成命名实体识别任务的机器学习模型主要包括三部分：嵌入层、编码层、解码层。嵌入层常用的方法有 Word2vec、Glove 和 CNN，目的是将文本转换到向量空间，将非结构化的数据结构化。嵌入层包括词水平的嵌入和字符水平的嵌入两种嵌入方式。编码层常用的方法有 LSTM、BiLSTM 和 BiGRU。解码层常用的方法有 CRF 和 Softmax。如文献[5]使用 Word2vec 得到字向量，应用 LSTM 抽取文本语义特征完成编码，通过 CRF 完成解码。自 BERT 面世后，因为其具备强大的特征抽取能力，故也常被微调后应用于此任务。无论是作为嵌入层还是编码层，BERT 的加入都能有效提升任务表现。以 BERT 为代表的一系列大语言模型以极多参数和多轮计算将语言文字转化为多维向量，将抽象的文字表达转换为具体的数值计算，对命名实体识别任务产生显著的影响。

在专利技术术语识别垂直领域内，命名实体识别尝试不多而且应用在专利技术术语识别任务中的技术路线与命名实体识别技术的一般技术路线基本一致，并无太多针对专利文本特点的优化方案。文献[6][7]在使用特征迁移思想，使用 BERT 得到字符水平的嵌入后，将字特征迁移至 BiLSTM-CRF 模型。文献[8]使用 Word2vec 得到字和词水平的嵌入向量后放入 BiLSTM 完成编码，编码向量经过字符级的注意力机制和分词词性注意力机制后由 CRF 解码得到最优标签序列。文献[9]在补充特征的同时验证特征迁移思想的有效性，证明加入 BERT 嵌入能提升模型性能。

然而，以 BERT 为代表的大规模预训练模型主要依赖纯文本学习，缺乏大规模知识指导学习，模型能力依然存在局限。ERNIE3.0 作为首个包含大规模知识图谱的百亿级预训练模型，在海量无监督文本与大规模知识图谱的平行预训练方法下联合掩码训练结构化和非结构化文本信息，其记忆和推理知识的能力再次得到大幅提升。本文将利用 ERNIE3.0 知识增强预训练模型完成命名实体识别与抽取，获得技术术

语集合。

3. 研究设计

研究思路如图 1 所示。首先，本文基于 CNKI 专利数据库收集专利文本的公开日期和摘要。然后利用 ERNIE3.0 知识增强预训练模型完成命名实体识别与抽取，得到专业术语集合。随后利用 LDA 主题模型对 t 个时间窗口的文献进行划分，结合困惑度－主题方差计算确认最优主题数。最后使用余弦相似度计算主题相似度并进行演化路径分析，以可视化的方式对最终结果进行对比展示。

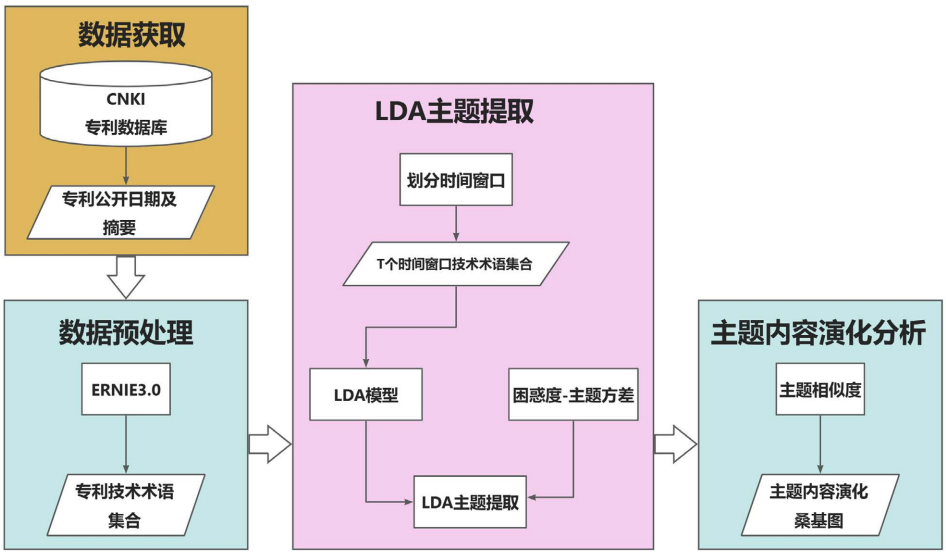


Figure 1. Research ideas for hot topic recognition and evolution analysis
图 1. 热点主题识别及演化分析研究思路图

3.1. LDA 主题模型

LDA 假设一篇文档涵盖了多个主题，且使用词袋模型，即忽略文字在文本中出现的先后顺序。假设语料库中有 M 篇文档。各文档内各主题出现概率被称为文档－主题分布，被用于确认第 m 篇文档中第 n 个位置上的字词属于某一主题的概率 $zmn, zmn \sim Multinomid(\theta m)$。$\theta m$ 是第 m 篇文档属于某一种文档－主题分布的概率，$\theta m \sim Dir(\alpha)$。同理，各主题内各字词出现概率被称为主题－词语分布，被用于确认第 m 篇文档中第 n 个位置上的字词 $wmn, wmn \sim Multinomid(\varphi_{zmn})$。$\varphi_{zmn}$ 是第 m 篇文档中第 n 个位置上的主题属于某一种主题－词语分布的概率，$\varphi_{zmn} \sim Dir(\beta)$。

本文采用 Gibbs 采样算法求解得到文档－主题分布和主题－词语分布。作为无监督机器学习，需要事先确定三个超参数：α、β、最优主题数，α、β 选取默认值[10]。

3.2. 主题数计算方法

LDA 主题模型被广泛引用于主题建模，但主题数的确认方法始终没有出现绝对合理的方案。困惑度常被作为确定主题数的指标，但使用困惑度确定的主题数较大，提取的主题干扰较多，相似性较大[11]。为了优化 LDA 主题提取效果，文献[12]在使用 LDA 主题模型时补充使用专利共现网络，以解决主题区分度不高的问题。文献[13]则验证了主题方差能够有效衡量潜在主题空间的整体差异性和稳定性。故本文将使用困惑度－主题方差(Perplexity-Var)指标确定最优主题数，准确有效地避免主题冗余。

困惑度计算公式[14]如下：

$$Perplexity(D)=\exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\}$$

D 表示测试集，共 M 篇文档，N_d 表示文档 d 中的单词数，w_d 表示文档 d 中的词，$p(w_d)$ 即文档中词 w_d 产生的概率。

主题方差计算公式如下：

$$Var(T)=\frac{\sum_{i=1}^{K}\left[D_{JS}(T_i,\varphi)\right]^2}{K}$$

T_i 表示 LDA 主题结果中的第 i 个主题，K 表示主题总数目，D_{JS} 表示 JS 散度。φ 表示主题－词概率分布的均值。当主题数为 1 时，JS 散度为 0，主题方差也为 0。

困惑度－主题方差计算公式如下：

$$Perplexity-Var_{test}=\frac{Perplexity(D)_{test}}{Var(T)_{test}}$$

当 *Perplexity-Var* 指标最小时，对应的 LDA 主题模型最优。

4. 案例分析

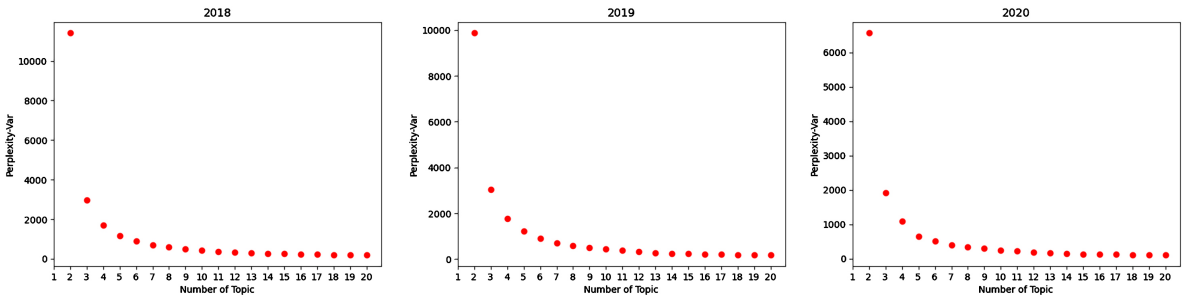
4.1. 数据获取及预处理

知网专利检索式为 TI = “面料”，并以公开日期为时间戳，查询公开日期介于 2018 年 1 月 1 日至 2022 年 12 月 31 日的相关专利，得到共 62,319 条专利信息。实际检索日期为 2023 年 6 月 3 日。经过专家筛选，先去除全部外观设计型专利，再在实用新型和发明专利中筛去用于加工纺织面料的纺织机械相关的专利数据，最终保留 27,225 份与纺织面料材料相关的专利信息。为了更好地进行主题演化路径的分析，本文将获取到的文献记录按照时间顺序进行窗口划分，一年为一个时间窗口。

本文在每个时间窗口下利用 ERNIE3.0 知识增强预训练模型完成命名实体识别与抽取，考虑到纺织面料专利的内容特征，结合专家意见，确认最终拟获得的实体类型标签为“物体类”，获得技术术语集合。

4.2. 主题提取

将依照专利的时间戳信息将文档集合划分到不同时间窗口，分别对每个时间窗口下的文本计算困惑度－主题方差得到最优主题数。困惑度－主题方差随主题数量变化的散点图如图 2 所示。当 Perplexity-Var 指标最小时，对应的 LDA 主题模型最优。设定 LDA 模型每轮迭代次数为 1500 次，共迭代 10 轮。虽然高迭代次数对算力要求较高，但模型收敛效果也更好。最终将 Perplexity-Var 的下降幅度低于 100 设定为最优主题数标志点，确定纺织面料专利在 2018、2019、2020、2021、2020 年的最优主题数分别为 9、8、7、6、7。



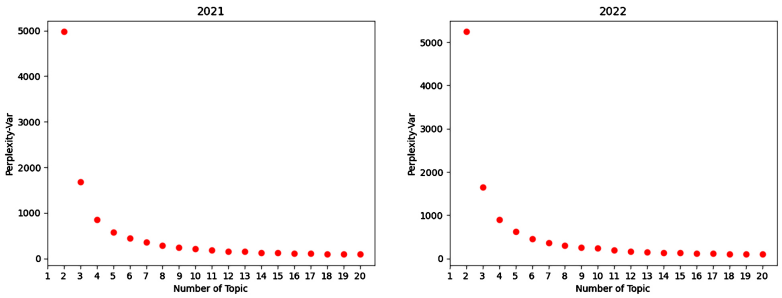


Figure 2. Perplexity-Var scatter chart with number of topics
图 2. Perplexity-Var 随主题数变化散点图

为了全面而准确地解读主题语义，本文选择在参考中图分类法的同时辅助专家经验，得到各主题标签如表 1 所示。两两各时间窗口下计算主题相似度得到主题内容演化桑基图如图 3 所示。

Table 1. Topic label
表 1. 主题标签

时间窗口	主题数	主题标签
2018	9	防水面料、抗菌面料、棉－聚酯纤维混纺的保暖面料、化纤针织面料、羊毛－聚酰胺纤维－聚酯纤维－阻燃剂混纺的阻燃面料、聚氨酯纤维为主的复合面料、涂层面料、棉－纤维混纺面料、网眼面料
2019	8	棉－聚酰胺纤维混纺的防水面料、混纺面料、石墨烯面料、棉－橡胶复合面料、保暖面料、网眼面料、聚酯纤维为主的复合面料、涂层面料
2020	7	无纺布面料、纤维复合面料、防水面料、棉为主要材质的保温面料、涂层面料、棉－纤维混纺面料、聚酯纤维－聚酰胺纤维－聚氨酯纤维
2021	6	保暖面料、防水面料、涂层面料、抗菌面料、棉为主的复合面料、聚酯纤维为主的复合面料
2022	7	棉为主的复合面料、防水面料、含石墨烯和保温层的复合面料、聚酰胺为主的复合面料、聚酯纤维－棉混纺面料、保暖面料、透气面料

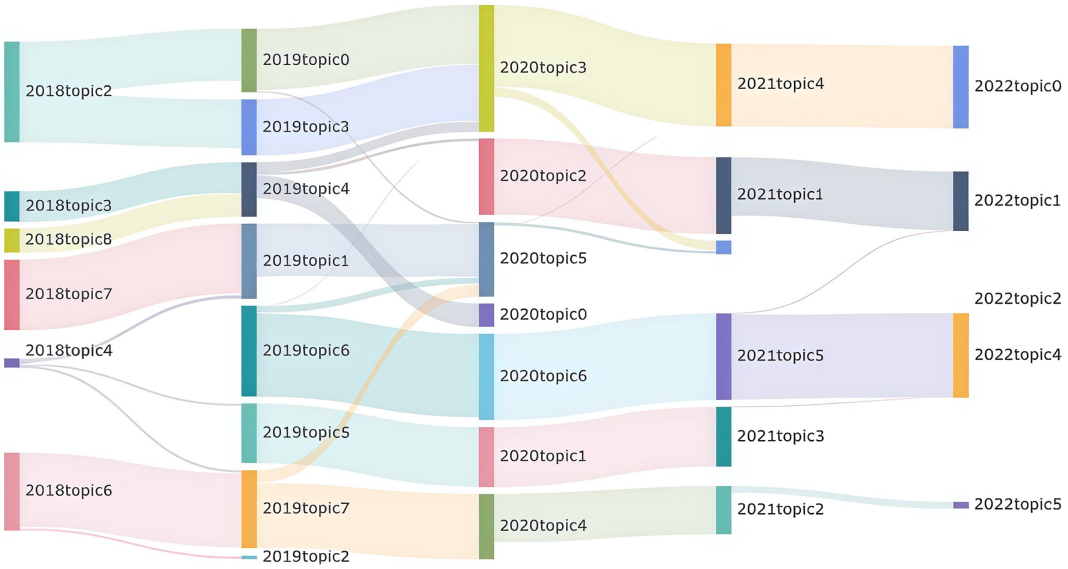


Figure 3. Theme content evolution Sankey
图 3. 主题内容演化桑基图

通过观察桑基图，可以看到如 2018 年主题“棉–聚酯纤维混纺的保暖面料”分化得到 2019 年主题“棉–聚酰胺纤维混纺的防水面料”“棉–橡胶复合面料”。然后与 2019 年主题“保温面料”一起被 2020 年主题“棉为主要材质的保温面料”继承。然后分化为 2021 年主题“保暖面料”和“棉为主的复合面料”，其中“棉为主的复合面料”延续出现到 2022 年。

2018 年主题“聚氨酯纤维为主的复合面料”和主题“棉–纤维混纺面料”共同被 2019 年主题“混纺面料”继承，然后与 2019 年主题“涂层面料”一起被 2020 年“棉–纤维混纺面料”继承。

2018 年主题“聚氨酯纤维为主的复合面料”分化为 2019 年主题“网眼面料”后在 2020 年演化成“无纺布面料”。

2018 年主题“聚氨酯纤维为主的复合面料”和主题“涂层面料”一起被 2019 年主题“涂层面料”继承并延续出现到 2021 年，在 2022 年分化为“保暖面料”。

2018 年主题“化纤针织面料”和“网眼面料”在 2019 年被主题“保暖面料”继承后，在 2020 年分化为主题“无纺布面料”。

5. 纺织面料专利创新模式总结

棉、聚酯纤维、聚酰胺纤维、聚氨酯纤维、橡胶是近五年的重要面料原料，涂层和网眼是近五年的重要面料制备工艺，保暖和防水是近五年热度较高的面料特性。以上热点要素的交叉组合可以作为未来面料专利开发时的重点指导方向，但新专利的创新程度可能会被影响，需要注意避免重复造轮。

此外，无纺布工艺、抗菌功能以及石墨烯纤维原料作为间断出现但尚未形成演化趋势的技术术语，可能是未来面料专利开发时的新兴指导方向，在合理评估可实现性后可以作为重点开发方向。

参考文献

[1] 黄晓斌, 吴高. 学科领域研究前沿探测方法研究述评[J]. 情报学报, 2019, 38(8): 872-880.

[2] 邱科达, 马建玲. 机器学习在术语抽取研究中的文献计量分析[J]. 图书情报工作, 2020, 64(14): 94-103.

[3] 靳嘉林, 王曰芬, 巴志超, 等. 基金项目研究的主题挖掘与动态演化分析——以美国 NSF 数据中 AI 领域为例[J]. 情报学报, 2022, 41(9): 967-979.

[4] 邢晓昭, 任亮, 雷孝平, 等. 基于专利主题演化的颠覆性技术识别研究——以类脑智能领域为例[J]. 情报科学, 2023, 41(3): 81-88.

[5] 高佳奕, 杨涛, 董海艳, 史话跃, 胡孔法. 基于 LSTM-CRF 的中医医案症状命名实体抽取研究[J]. 中国中医药信息杂志, 2021, 28(5): 20-24.

[6] 李建, 靖富营, 刘军. 基于改进 BERT 算法的专利实体抽取研究——以石墨烯为例[J]. 电子科技大学学报, 2020, 49(6): 883-890.

[7] 傅源坤, 柳先辉, 赵卫东. 基于 BERT 的智能制造装备命名实体识别方法[J]. 制造业自动化, 2022, 44(9): 120-124.

[8] 杨佳鑫, 杜军平, 邵鋆侠, 李昂, 奚军庆. 面向知识产权的科技资源画像构建方法[J]. 软件学报, 2022, 33(4): 1439-1450.

[9] 罗艺雄, 吕学强, 游新冬. 融合多特征的专利功效短语识别[J]. 中文信息学报, 2022, 36(12): 139-148.

[10] Wei, X. and Croft, W.B. (2006) LDA-Based Document Models for Ad-Hoc Retrieval. *SIGIR* 2006: *Proceedings of the* 29*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, 6-11 August 2006, 178-185.

[11] 杨洋, 江开忠, 原明君, 等. 新闻话题识别中 LDA 最优主题数选取研究[J]. 数据分析与知识发现, 2022, 6(11): 72-78.

[12] 单晓红, 韩晟熙, 刘晓燕. 基于技术主题演化的颠覆性技术识别研究[J]. 情报理论与实践, 2023, 46(8): 113-123.

[13] 关鹏, 王曰芬. 科技情报分析中 LDA 主题模型最优主题数确定方法研究[J]. 现代图书情报技术, 2016(9): 42-50.

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