基于 BERTopic 模型的学术论坛文本动态主题建模研究

张志祥 [1]
ZHANG Zhixiang

摘 要 针对现有 BERTopic 研究未适配中文学术论坛术语密集与跨域混合特性、缺乏多预训练模型对比选型、无落地化热点预警的研究缺口，为解决学术论坛文本主题分散、动态热点追踪滞后、主题可解释性弱的问题，文章提出一种以预训练模型适配优化 BERTopic 为核心的方法，通过实现学术文本动态主题挖掘与热点追踪，为论坛内容治理及用户需求精准匹配提供技术支撑。采用 jieba 分词结合自定义词库与停用词过滤，对比 5 种预训练模型的语义嵌入效果，经 UMAP 降维、HDBSCAN 聚类及关键词提取，构建含 33 个核心主题的学术主题体系，并开展时序演化分析、可视化与热点预警。实验结果表明，中文专用模型 bge-base-zh 适配 BERTopic 综合性能最优：主题连贯性指标达 -3.180 7，主题多样性（Jaccard 距离）达 0.994 3，较其他预训练模型有显著优势。

关键词 BERTopic；预训练模型；主题建模；降维；聚类；语义嵌入

doi：10.3969/j.issn.1672-9528.2026.06.040

0 引言

学术论坛作为科研交流核心载体，汇聚多领域多元学术文本，其文本具有多领域交叉、语义高密度（含大量中英文专业术语）、动态演化（新兴热点快速迭代）三大特征，对主题挖掘技术的语义捕捉与动态适配能力提出更高要求，传统主题模型（如 LDA、CTM）存在语义理解弱、难以动态追踪热点、主题可解释性差等问题 [1-2]。

BERTopic 作为新兴主题建模框架，融合自注意力网络语义表示能力与 c-TF-IDF 关键词提取优势，通过“文本嵌入 - 降维聚类 - 关键词提取”流程，实现语义理解与主题解释的统一，可为动态主题演化分析提供支撑 [3-5]。与现有研究相比，本文在研究视角与方法上具备明确新颖性：一是填补了学术论坛场景下多预训练模型适配 BERTopic 的研究空白；二是解决了以往主题挖掘仅重建模、无落地应用的共性短板。为此构建了基于 BERTopic 的动态主题挖掘与热点追踪框架，相较于已有研究，本文核心改进与创新点如下：

（1）多模型适配：首次针对学术论坛文本特性，对比中文专用与多语言预训练模型性能，筛选最优语义嵌入方案，弥补现有研究模型适配性不足的问题；

（2）文本预处理优化：结合学术领域自定义词库与停用词表，相比通用预处理方案显著提升专业术语识别精度；

（3）全流程框架构建：区别于传统模块式研究，实现“嵌入、降维、聚类、主题优化、动态追踪”完整闭环；

（4）应用落地：突破现有研究无落地功能的局限，开发主题可视化与热点预警功能，将结果直接转化为内容治理工具。

1 模型概述

BERTopic 是一种融合语义嵌入与无监督聚类的主题建模技术。该模型通过整合 Sentence-BERT、UMAP 降维及 HDBSCAN 聚类，实现从文档嵌入到主题表示的完整流程，依托自注意力网络与 c-TF-IDF 机制，兼顾文档聚类语义密度与主题可解释性 [6]。各阶段技术逻辑如下：

（1）文档嵌入：采用 Sentence-BERT 将清洗后文本映射为高维语义向量，保留深层语义关联并简化特征工程。

（2）文档聚类：通过“UMAP 降维 +HDBSCAN 聚类”的两阶段策略，实现语义相似文档的精准分组：UMAP 降维保留数据结构、降低计算复杂度；HDBSCAN 自动确定簇数、过滤噪声，保障聚类语义一致。

（3）主题表示：基于 c-TF-IDF 提取关键词，将主题簇文档拼接为“伪单 - 文档”计算每个词汇的重要性得分，形成“主题 - 特征词”的权重分布。

2 研究设计

2.1 研究框架

研究框架技术路线如图 1 所示，涵盖数据处理、主题建模、可视化应用三大模块，整体围绕学术论坛文本主题挖掘与动态追踪目标逐层推进。针对中英文混合文本，通

1. 北京格数致知科技有限公司 北京 100097

过 jieba 分词结合学术自定义词库与停用词过滤保留专业词汇，为精准语义嵌入夯实文本基础；系统对比 5 种预训练模型的语义嵌入向量，筛选适配学术文本的最优嵌入方案；将优选向量输入 BERTopic 模型，依次完成 UMAP 降维、HDBSCAN 聚类、主题关键词提取与主题微调，保障主题挖掘精准有效；最终形成从文本预处理、主题识别到动态分析的全流程技术闭环。

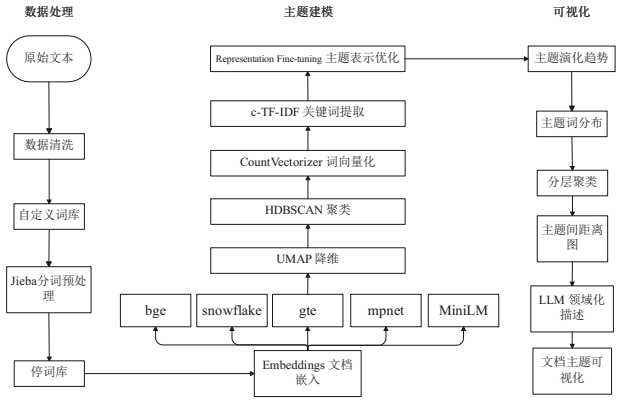

图 1 研究框架技术路径图

2.2 数据来源与预处理

2.2.1 数据来源

研究数据采集于经管之家论坛在 2019 年 1 月 1 日—2024 年 12 月 31 日的发帖数据。包含时间戳（dateline）、标题（subject）与正文（message_text）三类字段。其中时间戳用于动态主题演化分析，标题与正文合并后作为模型输入文本。

2.2.2 数据预处理

为保障主题建模效果，在建模前对原始文本进行预处理。填充标题与正文缺失值，采用 jieba 分词，加载学术领域自定义词库（含数据类“数据分析师”、计算机类“机器学习”等专业术语），再过滤停用词；最后生成模型训练数据集（data）与时间戳列表（timestamps）。预处理后的文本示例如表 1 所示。

表 1 预处理后的部分文本示例

时间戳	处理后文本（标题 + 正文）
2019-01-09 17:05:23	弗里德曼统计学 changyongming ……
2019-01-15 20:37:56	统计中重要的检验：T 检验、F 检验及其统计学意义 ……
……	……
2019-01-26 15:42:30	求助统计学和数据解读的高手：一个关于市场进入决策的小题需要解答……

2.3 模型构建

2.3.1 语义嵌入模型选择与对比

针对学术论坛“中英文混合、多领域术语密集”的文本特性，选取 5 种代表性预训练模型开展语义嵌入对比，含 2 种中文模型（bge-base-zh、gte-base-zh）与 3 种多语言模型 snowflake-arctic-embed-m、paraphrase-multilingual-mpnet-base-v2、paraphrase-multilingual-MiniLM-L12-v2），全面探索不同模型的场景适配潜力。其中，中文专用模型均基于大规模中文语料训练，更适配中文学术文本的专业语义表达；bge-base-zh 依托 RetroMAE 预训练方法与 C-Pack 资源优化，擅长精准捕捉“机器学习”“供应链金融”等专业术语语义，在 C-MTEB 中文嵌入评测中表现突出；gte-base-zh 经过多阶段对比学习优化，在长距离语义依赖与复杂句式理解上更具优势，适配学术论坛多轮深度讨论文本。多语言模型依托 Sentence-BERT 架构优化，可实现跨语言术语对齐，适配中英文混合文本处理需求；其中 paraphrase-multilingual 系列适配多语言场景，snowflake-arctic-embed-m 则兼顾多语言语义捕捉与大规模文本处理效率。本研究系统对比两类模型的适配差异，为 BERTopic 主题建模筛选最优语义嵌入方案，模型选择参考了 Sentence-BERT 与 BERTopic 适配性的相关研究成果。

2.3.2 UMAP 降维

高维语义嵌入存在数据稀疏、计算复杂度高的问题，直接用于聚类易导致效果不佳。UMAP 作为兼顾局部与全局结构的降维算法，能在降低维度的同时保留数据核心语义关联，为后续聚类提供优质输入。本研究基于学术论坛文本“术语密集、语义关联复杂”的特性，对 UMAP 关键参数进行针对性配置，具体参数及设计逻辑如表 2 所示。

表 2 UMAP 降维算法参数配置

参数名	取值	含义
n_neighbors	15	控制降维时保留的局部结构，值越小主题越细、越大主题越粗，一般在 10~50 之间调整。
n_components	15	降维后维度，设为 15 保留更多语义信息，提升聚类效果。
min_dist	0.1	控制低维空间中点的紧密程度，取值 0.0~0.2，本次设为 0.1 以避免过度聚集。
metric	'cosine'	采用余弦相似度衡量向量语义距离。
random_state	42	固定随机种子，保证可复现
n_jobs	-1	启用全部 CPU 核心并行计算，提升运算效率

2.3.3 HDBSCAN 聚类

HDBSCAN 作为基于密度的层次聚类算法，无需预设聚类数量，可自动识别噪声点并生成紧凑有效的簇结构，在非球形簇数据上的表现优于传统 K-Means、DBSCAN 算法。该

算法适配学术论坛文本语义关联复杂、簇分布不规则的特性，可有效避免聚类碎片化与主题混淆问题；结合 UMAP 降维保留的核心语义信息，为后续主题关键词提取奠定高质量基础。本研究采用 HDBSCAN 对 UMAP 降维后的低维嵌入进行聚类，相关参数如表 3 所示。

表 3 HDBSCAN 聚类算法参数配置

参数名	参数值	含义
min_cluster_size	50	聚类最小样本数，控制主题数量与稳定性。设为 50 确保主题具有足够代表性
metric	‘euclidean’	距离计算方法采用欧氏距离衡量降维后向量间的相似性
prediction_data	TRUE	是否保存预测数据，设为 True 可支持后续对新样本的主题分配
core_dist_n_jobs	-1	计算核心距离的并行核心数，设为 - 1 表示使用所有可用 CPU 核心

2.3.4 主题表示与优化

主题表示旨在提取精准关键词以刻画主题。针对学术论坛文本特点，通过 CountVectorizer 与 ClassTfidfTransformer 实现关键词提取与优化，具体参数配置如表 4 所示。

表 4 主题表示与优化相关参数配置

参数名	参数值	含义
tokenizer	jieba_tokenizer	自定义 jieba 分词器处理学术文本
token_pattern	None	关闭默认分词，保障自定义分词逻辑一致
vocabulary	my_vocab	加载学术自定义词库，提升专业术语与关键词识别精度
stop_words	my_stopword	加载自定义停用词库，过滤无意义词汇、降低噪声
ngram_range	(1, 2)	提取单字与二元组关键词，捕捉学术短语、丰富主题表示
bm25_weighting	TRUE	启用 BM25 加权，优化 c-TF-IDF 关键词权重，提升核心词汇区分度

2.3.5 动态主题追踪

为捕捉主题热度的时间演化特征，采用 BERTopic 模型对象的 topics_over_time 方法构建主题热度时序分析框架，输出各主题在不同时间切片内的文档占比数据，并生成可视化图表量化主题文档占比的时序变化，刻画主题热度的波动规律，直观呈现热点主题的演化轨迹。

3 实验与结果分析

3.1 实验环境及参数配置

实验环境配置：硬件方面，内存 16.0 GB DDR3、GPU 为 NVIDIA RTX 3080 Ti（12288 MB 显存）、CPU 主频约 1.5 GHz；软件方面，NVIDIA 驱动 522.06、CUDA 11.8，基于 Python 3.9，核心依赖库包括 BERTopic 0.17.3、Sentence-Transformers 5.1.1、UMAP-learn 0.5.9.post2、HDBSCAN 0.8.40 等，为模型运行提供软硬件支撑。

3.2 预训练模型对比结果

基于主题连贯性（c_u_mass）与多样性（Jaccard 距离）双维度评价各预训练模型的性能可知：中文专用模型 bge-base-zh 综合表现最优，有效主题数为 33，数量合理且覆盖全面；c_u_mass 均值达 -3.180 7，主题内关键词语义连贯性最强；Jaccard 距离 0.994 3（接近 1）、主题间差异化显著、冗余度极低。其他模型或存在主题数量过多或不足、连贯性较弱等问题。如表 5 所示。

表 5 多预训练模型性能对比

模型名称	有效主题	连贯性 (c_u_mass)	多样性 (Jaccard 距离)
bge-base-zh	33	-3.180 7	0.994 3
snowflake-arctic-embed-m	23	-3.204 4	0.980 3
gte-base-zh	20	-3.474 1	0.989 2
paraphrase-multilingual-mpnet-base-v2	43	-3.493 7	0.995 3
paraphrase-multilingual-MiniLM-L12-v2	30	-3.643 4	0.991 4

3.3 主题体系构建结果

基于 bge-base-zh 适配的 BERTopic 模型，识别出包含噪声主题在内的全部主题集合。其中噪声主题（标记为 -1）由离群点及文档数量低于阈值的主题组成，通过采用基于语义相似度的 K 近邻分配策略，将噪声文档重新分配至语义最相近的核心主题，最终确定 33 个核心主题。进一步对模型输出的“主题 - 特征词”及主题分布情况进行系统整理，按领域属性将 33 个核心主题聚类为 6 大类别，核心主题及占比如表 6 所示。

表 6 学术主题体系分类及占比

主题类别	包含核心主题编号	文档数	占比 /%
经管与产业研究	1、6、9、13、29	4 631	28.01
计算机技术应用	2、3、5、8、14、22、25、26、31、32	4 405	26.65
科研方法与工具	0、4、11、23、24、28	6 231	37.69
学术出版与评价	12、20、30	268	1.62
学术资源与交流	7、15、16、18、21	613	3.71
跨学科交叉研究	10、17、19、27	385	2.32

经管与产业研究（28.01%），聚焦全球市场、区块链金融等，体现经管领域产学研融合；计算机技术应用（26.65%），涵盖人工智能、大数据等，凸显技术对学术研究的渗透；科研方法与工具（37.69%，占比最高），以统计检验、工具应用为核心，反映学者对规范方法与实用工具的迫切需求；学术出版与评价（1.62%）、学术资源与交流（3.71%）、跨学科交叉研究（2.32%）占比较低，分别对应学术成果评价、知识共享与前沿交叉趋势。

3.4 动态主题演化分析

通过 BERTopic 内置的 visualize_topics_over_time 函数生成主题热度时序演化图（如图 2 所示），该图清晰地呈现 2019—2024 年各主题的时间演化特征：

2019—2021 年：主要围绕传统统计方法、科研数据支撑类主题，占比超 60%，以传统研究范式为主。

2022 年：数据驱动、开放科学相关主题的热度显著上升，占比达到 25%，数据化与开放科学受到关注。

2023—2024 年：全球市场与产业类主题的热度大幅攀升，机器人教育、生成式 AI 等技术类主题也同步快速增长，整体以学术研究向技术驱动全球化产业视野转型。

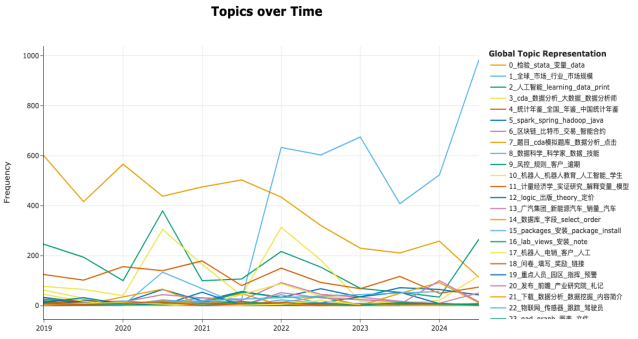

图 2 主题热度时序演化图

4 应用实践

4.1 数据可视化展示

（1）主题热度时序图（如图 2 所示）：以时间为横轴、主题频率为纵轴，动态呈现了 2019—2024 年热点主题的演化轨迹，可为主题动态趋势判断提供数据支撑。

（2）主题关键词柱状图（如图 3 所示）：展示 33 个核心主题的关键词，快速定位主题核心研究的内容。

（3）主题层级聚类图（如图 4 所示）：借助 BERTopic 内置的 visualize_hierarchy 函数生成，以树形结构展示不同主题间的语义关联与层次聚类关系。

（4）主题相似度矩阵图（如图 5 所示）：通过热力图来量化展示主题间语义相似度，数值越接近 1 主题的关联性越强，可为主题合并和优化提供依据。

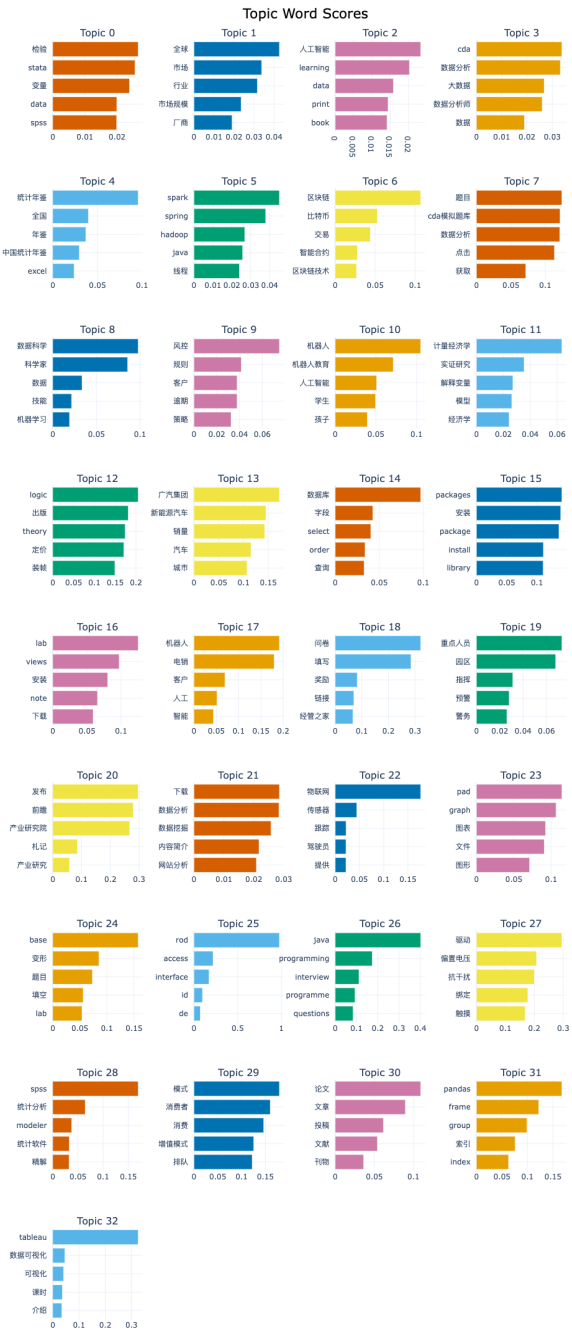

图 3 主题关键词分布图

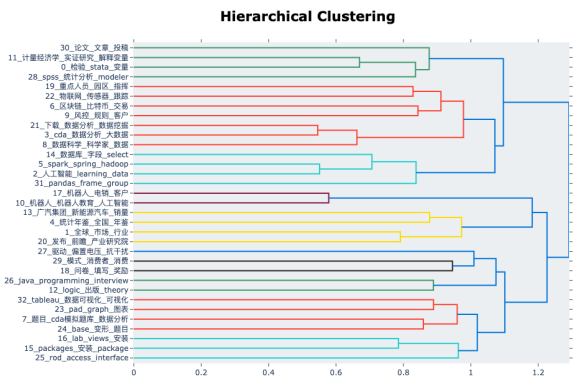

图 4 主题层级聚类图

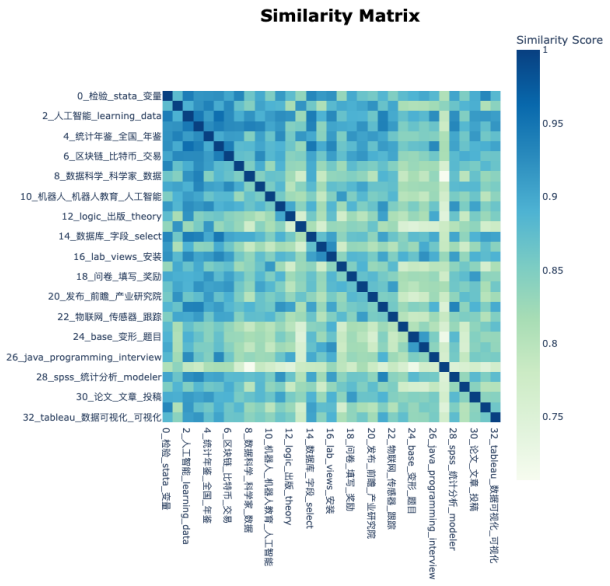

图 5 主题相似度矩阵图

（5）文档 - 主题分布可视化（如图 6 所示）：以二维散点图展示文档在主题空间的分布，区分不同主题类别，可直观评估聚类效果与映射合理性。

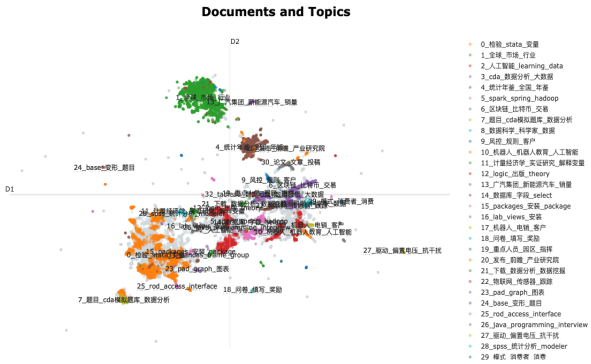

图 6 文档 - 主题分布图

4.2 热点动态预警

（1）预警阈值设定

以单时间切片内主题热度的增幅大于 100% 为预警阈值，把 2019—2024 年划分为 12 个均等的时间区间，实时计算热度变化率，为新兴热点识别提供量化标准。

（2）预警主题识别与输出

当主题热度达到预设预警阈值时，系统自动识别生成式 AI 学术应用等爆发增长型热点，并同步推送预警到论坛管理端，为学术论坛趋势引导工作提供参考。

5 结论

本文围绕学术论坛的主题挖掘与动态治理需求，构建了一套从主题识别、模型适配到可视化应用与预警机制的完整方案，主要结论如下：

（1）BERTopic 框架有效性：构建的主题挖掘框架有效解决了学术论坛“主题分散、热点追踪滞后”问题，识别出 33 个核心主题，覆盖多领域学术需求，为主题治理与资源调配提供语义支撑。

（2）中文专用模型 bge-base-zh 的适配性优势：在多模型对比实验中，bge-base-zh 表现出最优适配性，其主题连贯性指标 c_u_mass 达 -3.180 7，多样性指标 Jaccard 距离达 0.994 3，显著优于 gte-base-zh 等模型，验证了中文学术论坛场景专用模型适配的有效性，填补了相关研究缺口。

（3）可视化工具与热点预警机制为主题分析提供直观支撑，有效提升了内容治理效率与需求匹配精度，为学术论坛智能化治理提供可行路径。

参考文献：

[1] 牛振东，和晓峰，刘晓琦，等 . 基于 BERTopic 模型的医学人工智能研究主题挖掘及演化特征分析 [J]. 医学信息学杂志，2025, 46 (2): 45-54.

[2]Yuan Ye, Idaya A M K Y, Noorhidawati A, et al. Unveiling the dynamics of research data management: insights from bibliometric and BERTopic analysis[J]. The Journal of Academic Librarianship, 2025(5): 103126.

[3] 高道斌 . 基于 BERTopic 的国内智慧图书馆“科学 - 技术”主题研究 [J]. 文献与数据学报，2025, 7(1): 114-128.

[4] 李文科，钟子琪，丛挺 . 基于 BERTopic 模型的我国学术出版研究主题与趋势展望 [J]. 出版与印刷，2025(1): 57-71.

[5] 宋俊杰，尹裴，邓诗语，等 . BERTopic 在医疗领域文章主题挖掘中的应用与分析 [J]. 软件工程，2025, 28 (4): 62-66.

[6]Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Brussels: ACL, 2019: 3982-3992.

【作者简介】

张志祥（1994—），男，甘肃平凉人，硕士研究生，研究方向：用户行为数据挖掘、文本挖掘，email:3344382514@qq.com。

（收稿日期：2026-04-09 修回日期：2026-06-04）