基于BERTopic-RAG框架的医学信息学主题演化与知识发现研究

张 琳 任淑敏

（济宁医学院图书馆/济宁医学院学术评价分析与研究中心 济宁 272067）

〔摘要〕 **目的/意义** 构建双层分析框架，全面把握学科结构，识别新兴前沿领域，追踪主题演化。**方法/过程** 检索2016—2025年PubMed、Scopus和Web of Science数据库医学信息学文献，采用BERTopic识别主题，并划分为新兴、稳定、衰退3种演化模式。基于ChromaDB构建检索增强生成系统，通过文档-主题映射实现微观验证与知识关联挖掘。**结果/结论** 医学信息学主题演化呈现研究重心转移、技术融合深化、学科交叉增强3个特征。BERTopic-RAG框架为知识发现提供了新方法。

〔关键词〕 医学信息学；主题建模；BERTopic；检索增强生成；知识演化

〔中图分类号〕R-058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2026.02.005

Study on Topic Evolution and Knowledge Discovery in Medical Informatics Based on BERTopic-RAG Framework

ZHANG Lin, REN Shumin

Library of Jining Medical University, Academic Evaluation Analysis and Research Center, Jining Medical University, Jining 272067, China

〔**Abstract**〕 **Purpose/Significance** To construct a dual-layer analytical framework to comprehensively understand disciplinary structures, identify emerging frontiers, and track topic evolution. **Method/Process** Medical informatics literatures from 2016 to 2025 are retrieved from PubMed, Scopus, and Web of Science databases. The topics are identified by BERTopic and classified into three evolution patterns: emerging, stable and declining. A retrieval-augmented generation (RAG) system is built based on ChromaDB, enabling micro-level validation and knowledge association mining through document-topic mapping. **Result/Conclusion** The evolution of medical informatics topics exhibits three major characteristics: research focus shift, deepening technological integration, and enhanced interdisciplinary convergence. The BERTopic-RAG framework provides a novel methodology for knowledge discovery.

〔**Keywords**〕 medical informatics; topic modeling; BERTopic; retrieval-augmented generation (RAG); knowledge evolution

1 引言

医学信息学是医学与信息技术深度融合的交叉学科，其知识总量在人工智能、大数据和云计算等新兴技术的推动下增长迅猛[1-3]。相关文献数量激增，覆盖领域不断扩展，对研究人员全面把握学科结构、识别新兴前沿和追踪主题演化提出了挑战。共现分析[4-5]、引文分析[6-7]等传统文献计量方法虽然提供了客观性支撑，但其主要基于词频统计，难

〔修回日期〕 2025-11-28

〔作者简介〕 张琳，馆员，发表论文10余篇；通信作者：任淑敏，教授。

以深入理解文献的深层语义内涵[8]。近年来，基于深度学习的主题建模[9-10]为大规模文献的智能分析提供了新路径，但既有研究[11-12]多将其用于宏观领域主题静态呈现，缺乏与语义检索技术的深度融合，制约分析结果的可解释性和应用价值。本研究整合基于深度语义的BERTopic主题建模与检索增强生成（retrieval-augmented generation，RAG）技术，构建面向医学信息学的双层分析框架，通过追踪关键主题的演化轨迹，识别新兴前沿、稳定方向和衰退趋势，为掌握学科发展动态提供实证依据。

2 研究方法与框架

2.1 数据来源与检索策略

在权威学术数据库PubMed、Scopus和Web of Science（WOS）检索医学信息学及其相关子领域研究文献，检索词为“medical informatic*” OR “health informatic*” OR “clinical informatic*” OR “biomedical informatic*” OR “nursing informatic*” OR “public health informatic*”，文献类型限定为Article和Review，时间范围为2016—2025年。各数据库具体检索项及检出文献量，见表1。对检索结果进行合并、去重与数据清洗，最终获得医学信息学领域文献157 844篇。

表1 各数据库检索项及检出文献量

数据库	检索项	检索结果（篇）
PubMed	MeSH、Title/Abstract	187 699
Scopus	Title-Abstract-Keywords	150 334
Web of Science	Topic search	89 317

2.2 技术方法与框架

构建BERTopic-RAG框架，基于多源数据库文献，实现宏观识别、中观演化、微观验证，具体研究框架，见图1。BERTopic模块从海量文献中识别主题并构建宏观知识图谱，结合动态映射从中观层面追踪研究主题演化；RAG模块通过精准语义检索，验证和深化主题分析结果，实现微观层面的知识发现。两模块均利用Python3.10实现。

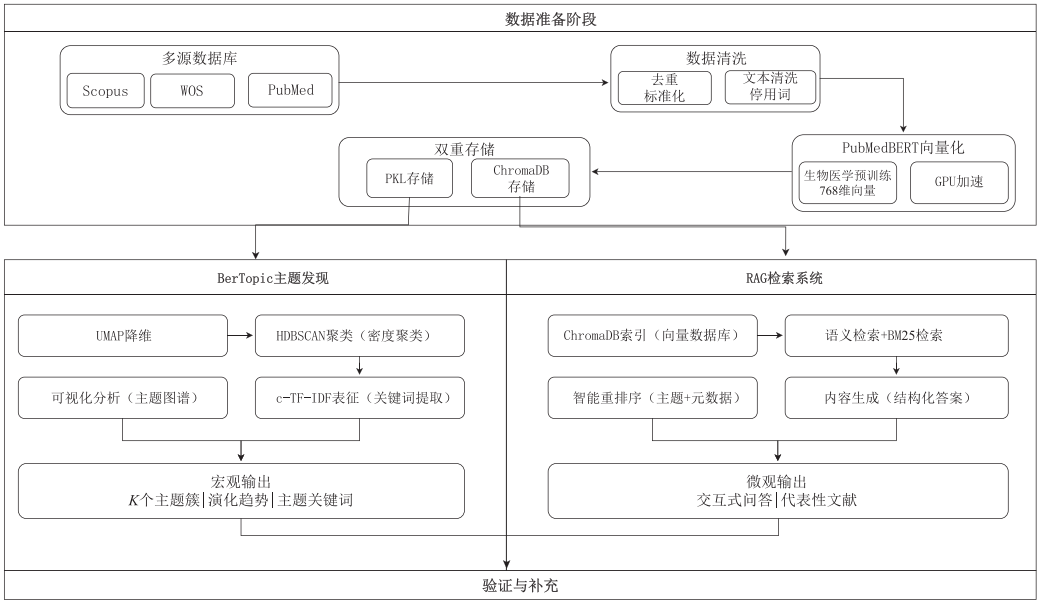


图1 研究框架

与传统隐含狄利克雷分布（latent Dirichlet allocation，LDA）模型相比，BERTopic基于深度学习和语义嵌入，在语义感知、主题质量和自动主题发现方面具有显著优势[13]。本研究BERTopic[14]模块包括主题建模与演化分析两部分。主题建模部分采用PubMedBERT[15]进行文档向量化，利用均匀流形

近似与投影（uniform manifold approximation and projection，UMAP）算法降维，通过层次密度聚类（hierarchical density-based spatial clustering of applications with noise，HDBSCAN）算法识别主题簇，使用基于类别的词频-逆文档频率（class-based term frequency-inverse document frequency，c-TF-IDF）方法提取主题关键词。演化分析部分基于全局主题建模与动态时序映射策略，针对全部157 844篇文献，建立覆盖整个研究时段的全局主题空间；采用变换方法将各时间窗口（每两年）文档映射至全局主题空间，追踪各主题文档频次的时序变化。将5个时间窗口分为前2个和后3个这两段，计算各主题在前后两段各时间窗口的平均文档量变化率。根据变化率及连续性特征，将主题划分为4类，新兴主题（增长超过50%）、衰退主题（减少超过50%）、间歇主题（超过2个时间窗口无文档分布）、稳定主题（增长或减少不足50%且持续出现）。

RAG技术通过整合外部知识库与生成模型，可缓解“幻觉”问题，并提高内容可信度。基于ChromaDB构建RAG[16]检索系统，对BERTopic主题演化结果进行微观验证。先通过BERTopic构建向量数据库，生成文档ID与主题ID映射表，记录每篇文档的主题编号。再采用混合检索策略，稠密检索基于PubMedBERT语义向量（权重0.6）捕捉深层语义关联，稀疏检索基于BM25算法（权重0.4）精确匹配关键术语。从向量数据库检索候选文献，通过文档-主题映射表筛选有明确主题标注的文献，按混合检索得分排序，返回最具代表性的10篇文献。通过分析检索结果的主题分布和时间演化趋势，验证BERTopic主题划分的合理性。

3 医学信息学主题演化

3.1 主题分布总体特征

医学信息学领域呈现多元化、高度交叉的分布格局，BERTopic识别出的88个核心主题，可归纳为4大核心主题群，见表2。生物医学计算是领域发展的主导方向，聚焦基因表达、癌症机制、蛋白结构预测等微观层面的计算模拟与机制挖掘，是医学信息学技术的典型应用场景。临床信息应用直接面向医疗实践，涵盖医疗信息系统、疫苗设计等主题，是连接基础研究和临床转化阶段的信息学应用。公共卫生信息关注肠道微生物组、流行病传播等传统健康管理和宏观公共卫生问题。技术方法创新持续为前沿研究提供可视化、深度学习等新型方法与工具，该主题群主题数量多、文档占比低，表明新兴技术仍处于探索阶段，文档分布分散。上述主题分布特征反映出医学信息学是一门以计算方法为核心、衔接基础研究与临床公共卫生应用的综合性学科。

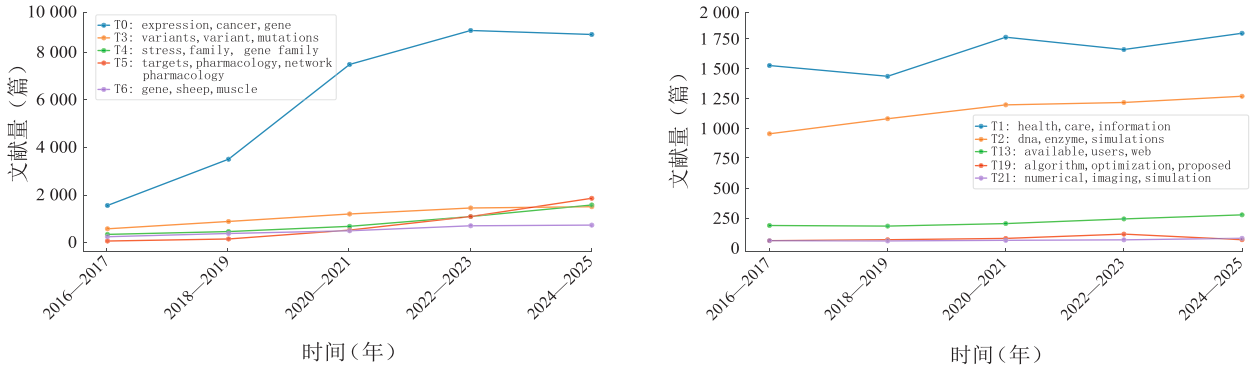
表2 医学信息学主题分布（2016—2025年）

主题群	主题数量（个）	文档占比（%）	代表主题与关键词
生物医学计算	24	41.7	T0 基因表达与癌症（expression，cancer，gene）；T2 DNA酶模拟（dna，enzyme，simulations）；T45 蛋白结构预测（structures，predictions）
临床信息应用	19	28.3	T1 医疗信息系统（health，care，information）；T9 疫苗设计（vaccine，epitopes）；T33 NGS检测（ngs，sequencing）
公共卫生信息	17	18.2	T10 肠道微生物组（microbiota，gut）；T18 膳食微生物（diet，rumen）；T55 流行病传播（transmission，epidemic）
技术方法创新	28	11.8	T13 可视化工具（visualization，tool）；T20 深度学习分类（classification，accuracy）；T83 并行序列比对（alignment，parallel）

3.2 主题演化趋势

3.2.1 主题演化模式 以两年为时间窗口追踪主题演化，将88个主题归类为新兴、稳定、衰退3种模式。文档数量最多的5个新兴主题演化趋势，见图2a。新兴主题呈现显著增长趋势：T0基因表达与癌症增长最为迅速，文献量从2016—2017年的约

1 600 篇增至 2022—2023 年的 9 000 余篇，之后保持在 9 000 篇左右，显示该领域研究热度持续高涨；T3 基因变异、T4 基因家族功能分析、T5 药理学网络虽增长相对平缓，但均保持稳定上升，文献量从数百篇增至 1 500~1 800 篇。文档数量最多的 5 个稳定主题演化趋势，见图 2b。稳定主题呈现成熟研究领域的平稳演化特征：T1 医疗信息系统文献量维持在 1 400 篇左右；T2 DNA 酶分子动力学、T13 可视化分析工具和 T19 算法优化缓慢增长，反映了基础性研究方向的持续性和稳定性。衰退主题共 4 个，代表影响力逐渐减弱或已被整合的方向，如 T12 补充信息与数据可用性，文献量自 2016 年逐渐下降。随着数据平台标准化的推进，此类方法性研究逐渐减少。



a. 新兴主题演化趋势　　b. 稳定主题演化趋势

图 2　主要新兴和稳定主题演化趋势

3.2.2　总体演化特征　医学信息学研究主题演化总体呈现 3 个特征。一是研究重心转移。文献数量最多的 30 个主题时间窗口分布，见图 3。

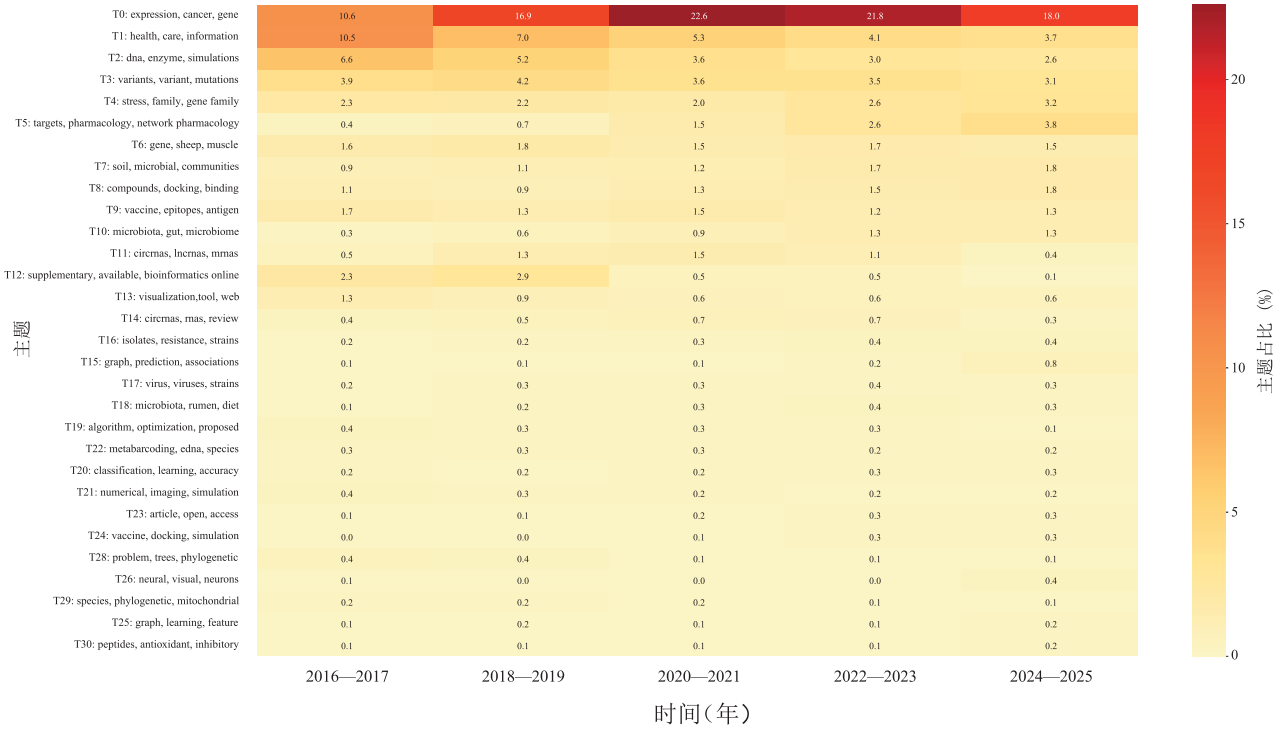


图 3　文献数量最多的 30 个主题时间窗口分布

T0基因表达与癌症在所有时间窗口均呈现最深红色，研究热度从2016—2017年的10.6%持续增长至2024—2025年的18.0%，标志着生物医学计算已成为领域核心。与之对比，T1医疗信息系统虽保持稳定产出，但下降趋势明显。T3基因变异等计算生物学主题同样展现由橙向红的渐进升温特征，表明领域重心已从信息管理转向数据驱动的生命机制解析。二是技术融合加速深化。多个主题在2020年后骤然升温，如T5网络药理学，体现了系统生物学方法与人工智能的持续融合。T15图预测、T26神经网络等深度学习主题的研究热度从0.1%分别上升至0.8%、0.4%，表明新兴计算技术正快速渗透至传统研究领域。三是学科交叉纵深拓展。本研究识别出多个新兴跨学科主题，如T72骨再生与生物支架、T57空间转录组学、T86铁死亡与代谢等，虽然当前文献量较低，但其融合信息学与生物材料学、空间生物学、代谢生物学等领域，体现了医学信息学向多学科交叉纵深发展的态势。

3.2.3 演化驱动因素分析 综合演化趋势与时间窗口分析，医学信息学主题演化主要由技术突破、公共卫生事件和学科交叉3类因素驱动，呈现多元驱动、分层作用的特征。首先，关键技术的成熟是主题演化的主要推动力。单细胞测序技术的普及推动T0在2020—2021年达到峰值，空间转录组[17-18]技术的突破催生新兴主题T57，AlphaFold蛋白结构预测技术[19]促进T45结构生物学主题增长。其次，公共卫生事件推动T9疫苗设计、T17病毒传播等主题达到峰值，并于2022—2023年后快速消退，表明公共卫生事件驱动具有短期特征。最后，学科交叉持续融合，如气候变化与粮食安全推动T4增长，精准医疗与免疫治疗维持T0高位产出，药物发现效率提升促进T5网络药理学发展。从作用机制看，技术能力跃迁决定了研究的可行性边界，是领域演化的根本动力；公共卫生事件在短期内重塑资源配置，推动应急类议题的集中涌现；学科交叉持续拓展研究边界，孵化新兴前沿方向。3类驱动因素在不同时间尺度和主题类型上协同作用，塑造了医学信息学核心深化、应急响应、边界拓展的演化格局。

4 RAG驱动的动态验证与知识发现

4.1 案例1：成熟概念主题聚焦验证

案例1 查询“precision medicine”，检索结果呈显著主题聚焦性，见表3。T1医疗信息系统占90%，揭示了精准医疗作为医学信息学领域成熟核心概念的特征。时间分布显示，30%文献发表于2020年后，整体跨度为2016—2023年，且持续聚焦于T1主题，验证了BERTopic对成熟研究范式的稳定识别能力。

表3 案例1查询结果

主题ID	主题名称	文献量（篇）	关键词
T1	医疗信息系统	9	health，care，information，informatics，healthcare
T80	无创诊断与基因测序	1	liquid，biopsy，review，sequencing，diagnosis

4.2 案例2：跨学科场域知识发现

案例2查询“machine learning diagnosis”，检索结果呈现更高的主题多样性，涉及3个不同研究主题且均匀分散，表明机器学习诊断具有跨学科特性，是医学信息学、深度学习技术、图像分类方法论3个知识域交汇的研究，见表4。

表4 案例2查询结果

主题ID	主题名称	文献量（篇）	关键词
T1	医疗信息系统	6	health，care，information，informatics，healthcare
T37	深度学习应用	2	learning，deep learning，deep，applications，review
T20	深度学习分类与精度优化	2	classification，learning，accuracy，proposed，images

时间-主题交叉分析显示，2017—2019年文献以主题T1为基础，主题T37出现于2018—2019年，主题T20出现于2020、2022年，主题切换揭示了研究范式从技术可行性探索向工程实现与精度优化的

转变。而主题T1占比60%，贯穿2017—2023年，体现了数据标准化、电子病历整合等基础设施的重要性，揭示了机器学习诊断临床部署对信息化基础设施的高度依赖。基于RAG的知识发现，聚焦了BERTopic全局视角下的局部特征。

4.3 双层框架的互补性分析

BERTopic与RAG在医学信息学文献分析中形成宏观-微观、静态-动态互补关系。BERTopic构建全局主题结构，识别出88个主题的层级关系与分布特征，为领域知识体系提供系统化视图。然而，其基于全量数据的静态建模无法针对特定查询揭示隐含的知识关联与研究机会。RAG则通过语义检索实现局部知识发现，案例1的高主题集中度从微观验证了BERTopic对“precision medicine”等成熟概念的有效识别，案例2通过主题分布对比揭示了跨学科场域特征，通过时间-主题交叉分析发现方法论成熟度转变，通过主题持续性分析揭示隐性基础设施需求。结合两种方法，利用BERTopic提供全局主题结构与演化趋势，利用RAG提供微观验证与动态知识发现，能够实现从宏观到微观、从静态到动态的完整分析。

5 讨论

5.1 领域知识结构呈现多元化格局

基于BERTopic-RAG框架识别出的88个核心主题勾勒出医学信息学的多元化知识版图，相关研究正从医疗信息系统、基因表达分析等传统核心，向单细胞空间转录组学、人工智能应用、铁死亡机制[20]等前沿方向拓展。4大主题群的分布特征表明，医学信息学已发展为以计算方法为核心、衔接基础研究与临床应用的成熟交叉学科。

5.2 主题演化揭示前沿发展路径

通过主题演化分析划分新兴、稳定和衰退3类主题。新兴主题如空间转录组学、网络药理学、铁死亡机制等，展现快速增长态势。研究重心从信息管理向数据驱动的生命机制解析转移，技术融合深化，学科交叉增强，推动领域向精准化、智能化方向演进。

5.3 RAG增强主题模型的知识发现能力

RAG检索验证了BERTopic主题建模的有效性并增强了其知识发现能力。案例1验证了BERTopic对成熟概念的有效识别。案例2通过主题分布、时间-主题交叉、主题持续性分析，揭示了跨学科场域特征、方法论成熟度转变、基础设施依赖等隐含知识。RAG有效补充了主题模型在动态追踪和深层关联挖掘方面的局限。

6 结语

本研究构建BERTopic-RAG框架，实现医学信息学知识体系解析的宏观结构化与微观精准化有机结合。BERTopic提供领域全景与主题分类依据，RAG实现按需知识验证与隐含关联挖掘，形成全局建模、动态演化、局部验证与发现的完整链路。本研究仍存在一定局限：一是验证案例有限，二是跨学科术语的语义理解精度有待提升，未来可整合引用网络、作者合作等多维数据，结合动态主题模型实现领域热点实时追踪，拓展知识扩散路径解析、学术社群识别，为科研决策和资源配置提供支持。

作者贡献：张琳负责数据分析、论文撰写；任淑敏负责提供指导、论文修订。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 GOMATHI DR R M, SIVASANGARI DR A, AJITHA D P, et al. Health informatics and medical data analysis [EB/OL]. [2025-09-29]. https://www.magesticts.com/book/Health%20Informatics%20Book%20Final.pdf.

2 WEN S. Medical informatization and emerging technologies: artificial intelligence, big data, and the internet of things [J]. Journal of innovations in medical research, 2024, 3(4): 28-35.

3 PENTEADO B E, FORNAZIN M, CASTRO L. The evolution of artificial intelligence in medical informatics: a bibliometric analysis [M]. Cham: Springer International Publish-

ing, 2021.
4 曹树金, 吴育冰, 韦景竹, 等. 知识图谱研究的脉络、流派与趋势——基于SSCI与CSSCI期刊论文的计量与可视化[J]. 中国图书馆学报, 2015, 41(5): 16-34.
5 罗依宁, 崔雷. 医学相关信息学子学科研究重点与关系探究[J]. 医学信息学杂志, 2025, 46(5): 50-55, 66.
6 梁镇涛, 巴志超, 徐健. 基于引文的跨学科领域发展路径分析——以眼动追踪领域为例[J]. 图书情报工作, 2019, 63(23): 65-78.
7 李点, 文庭孝, 许林勇. 引文施引双角度医学信息学交叉测度分析[J]. 医学信息学杂志, 2024, 45(5): 46-52, 64.
8 TAMINE L, GOEURIOT L. Semantic information retrieval on medical texts[J]. ACM computing surveys, 2021, 54(7): 1-38.
9 WU X, NGUYEN T, LUU A T. A survey on neural topic models: methods, applications, and challenges[J]. Artificial intelligence review, 2024, 57(2): 1-30.
10 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, 72.
11 YU D, XIANG B. Discovering topics and trends in the field of artificial intelligence: using LDA topic modeling[J]. Expert systems with applications, 2023, 225(9): 120114.
12 夏苏迪, 谢靖, 是沁, 等. 基于BERTopic模型的我国中医药老年康养政策量化研究[J]. 医学信息学杂志, 2025, 46(6): 43-49.
13 ROMAN E, JOANNE Y. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts[J]. Frontiers in sociology, 2022(7): 886498.
14 MAARTEN G. BERTopic: neural topic modeling with a class-based TF-IDF procedure[EB/OL]. [2025-09-29]. https://arxiv.org/abs/2203.05794.
15 YU G, ROBERT T, HAO C, et al. Domain-specific language model pretraining for biomedical natural language processing[J]. ACM transactions on computing for healthcare, 2022, 3(1): 1-23.
16 GARG M, WANG L, GHANCHI B, et al. Biomedical literature QA system using retrieval-augmented generation (RAG)[EB/OL]. [2025-09-29]. https://arxiv.org/abs/2509.05505.
17 PATRIK L S, FREDRIK S, SANJA V, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics[J]. Science, 2016, 353(6294): 78-82.
18 CHENGLONG X, JEAN F, GEORGE E, et al. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression[J]. Proceedings of the national academy of sciences of the United States of America, 2019, 116(39): 19490-19499.
19 JOHN J, RICHARD E, ALEXANDER P, et al. Highly accurate protein structure prediction with alphafold[J]. Nature, 2021, 596(7873): 583-589.
20 ZEYE L, HONG J, FENGWEN Z, et al. The examination of expression patterns, underlying mechanisms, diagnostic accuracy, and potential AI-driven drug development approaches for ferroptosis-related genes in heart failure via single-cell and bulk RNA sequencing analyses[EB/OL]. [2025-09-29]. https://doi.org/10.1161/circ.150.suppl_1.4140707.

（上接第29页）

层级医疗数据协同的创新探索[J]. 医学信息学杂志, 2025, 46(11): 28-34, 41.
10 王彦霞. 新时代医院电子病历档案共享机制构建研究[J]. 兰台内外, 2025(17): 1-4.
11 MAURER M M, PFITZNER B, VAN DE WATER R P, et al. Privacy preserving federated learning for 90-day mortality prediction in colorectal surgery: a multicenter retrospective development and comparison study[J]. International journal of surgery, 2025, 111(12): 9065-9074.
12 王钰涵, 孙燕杰. 区块链隐私计算赋能智慧医疗数据共享[J]. 中国科技信息, 2025(21): 131-134.
13 LI J, WANG D, QI G, et al. Alliance chain-based simulation on a new clinical research data pricing model[J]. Annals of translational medicine, 2022, 10(15): 836.
14 李哲成, 张波. 基于区块链的轻量级工业物联网跨域认证与数据共享方案[J]. 计算机研究与发展, 2025, 62(10): 2416-2427.
15 刘鲲, 王羽赫. 数字政府背景下基于数据可信计算沙箱技术的应用[J]. 网络安全和信息化, 2023, (11): 113-116.
16 胡业飞, 陈美欣, 张怡梦. 价值共创与数据安全的兼顾: 基于联邦学习的政府数据授权运营模式研究[J]. 电子政务, 2022(10): 2-19.