

基于高质量文献的肿瘤学领域研究前沿识别研究^{*}

朱韶华¹ 李艳红² 张 迅³ 黄海量²

(¹ 上海交通大学医学院附属瑞金医院图书馆 上海 200025 ² 上海财经大学信息管理与工程学院 上海 200433

³ 复旦大学附属肿瘤医院超声科 上海 200032)

〔摘要〕 目的/意义 识别领域研究前沿, 辅助科学研究者有效遴选和追踪重点研究主题, 助力科研管理决策者动态调整政策导向。方法/过程 以 Web of Science 2012—2022 年 37 927 条肿瘤学领域高影响力期刊文献题录和高被引文献题录为数据样本, 运用 BERTop 提取主题, 构建多维指标研究前沿识别模型, 从多维度识别领域内不同类型的研究前沿。结果/结论 所构建模型识别出肿瘤学领域热点研究前沿主题 9 个、新兴研究前沿主题 14 个、潜在研究前沿主题 13 个和衰退研究主题 1 个, 具有有效性。

〔关键词〕 研究前沿; BERTopic; 多维指标; 肿瘤学; 高质量文献

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.07.009

Study on Frontier Identification of Oncology Research Based on High Quality Literature

ZHU Shaohua¹, LI Yanhong², ZHANG Xun³, HUANG Hailiang²

¹ Library, Ruijin Hospital Affiliated to School of Medicine of Shanghai Jiaotong University, Shanghai 200025, China; ² School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China; ³ Department of Ultrasound, Fudan University Shanghai Cancer Center, Shanghai 200032, China

〔Abstract〕 **Purpose/Significance** To identify research frontiers in the field, so as to assist scientific researchers to effectively select and track key research topics, and help research management decision-makers to dynamically adjust policy orientation. **Method/Process** Taking 37 927 pieces of high-impact journal document references and highly cited document references from the WOS database which published in the period of 2012—2022 as the data samples, BERTopic is used to extract topics, a multi-dimensional indicator research frontier recognition model is constructed, and different types of research frontiers in the field are identified from multiple dimensions. **Result/Conclusion** The proposed model identifies 9 hot research frontier topics, 14 emerging research frontier topics, 13 potential research frontier topics and 1 declining research topic in the field of oncology, which is effective.

〔Keywords〕 research frontier; BERTopic; multidimensional indicator; oncology; high quality literature

1 引言

〔修回日期〕 2023-12-11

〔作者简介〕 朱韶华, 馆员, 发表论文 5 篇; 通信作者: 黄海量, 博士。

〔基金项目〕 国家自然科学基金项目 (项目编号: 72271151)。

目前, 每年发表的科技文献数量呈指数增长, 研究者开始聚焦对文献内容层面知识单元的挖掘以提高文献处理效率^[1]。同时恶性肿瘤已成为严重威

胁公众健康的公共卫生问题之一^[2]。2023 年《健康中国行动——癌症防治行动实施方案（2023—2030 年）》发布，其中包括控制肿瘤危险因素和提升防治能力等措施。此外，肿瘤学还是医学创新的重要突破口之一。本研究以肿瘤学研究领域为例，基于国外高质量文献，构建多维指标体系，运用 BERTopic 识别研究前沿，辅助科学研究者有效遴选和追踪重点研究主题，有助于科研管理决策者动态调整政策导向，助力我国在肿瘤学研究领域获得科技发展先机。

2 国内外相关研究

2.1 研究前沿的概念

研究前沿最初由 Price D J^[3] 定义为近期在某一特定科学领域内发表并被广泛引用的文献集。随后这一概念不断演化，先后被界定为“利用当前学科核心论文来描述其活跃研究领域^[4]，与高被引文献集合相互关联的引文文献簇”^[5]，以及“领域内新兴观点、方法和研究结果的不断发展”^[6]等。据此，本研究将研究前沿的概念界定为科学领域中的新发现和新进展，具备高度学术关注度，并持续发展的研究主题。

2.2 研究前沿的识别方法

研究前沿的识别以定量识别方法为主，包括基于引文、关键词和主题模型的分析方法。一是引文分析法，以科技文献之间的引用关系为基础，构建引文网络识别前沿主题。如卢超^[7]基于生物医学领域、田理等^[8]基于基因编辑领域的文献耦合网络发现新兴研究话题。二是关键词分析方法，以文献中的主题词汇为对象，采用关键词和共词分析等技术发现领域前沿方向。如姜鑫等^[9]结合词频分析法和共词分析法，通过揭示“数据科学”领域近 10 年的核心词、突变词和新生词发现前沿主题。三是主题模型方法，以隐含狄利克雷分布（latent Dirichlet allocation, LDA）模型为主，采用无监督学习对全文本进行语义结构和聚类分析，提取有价值的主题及关键词分布。如罗恺等^[10]用 LDA 和社会网络方

法构建关键词主题共现矩阵，通过“关节机器人”研究领域的专利和技术融合识别其技术前沿。BERTopic 是基于深度学习的主题建模方式，结合双向编码器表征（bidirectional encoder representations from transformers, BERT）的上下文词嵌入和狄利克雷过程，能够提供更准确的主题表示。如王秀红等^[11]将 BERT 与 LDA 相结合识别农业机器人研究领域关键技术。

综上，目前领域研究前沿的识别方法应用已取得一定进展。但是，基于引文的方法依赖文献之间的引用关系且不能反映研究领域动态变化；基于关键词的方法注重词汇频次和共现关系，对于一些主题的复杂表达和深度理解有一定局限性；以 LDA 为主的主题模型方法需要预先设置主题数、高维向量计算复杂且数据稀疏影响聚类效果，无法获得文本的深层语义理解。而 BERTopic 模型使用浅层神经网络模型获取语义信息，将单词表示为密集向量的词嵌入，可以从深层挖掘语义并更好地解释主题。

2.3 多维指标评价模型研究

单一评价指标可用于比较主题的某个方面，但全面分析研究前沿要考虑选择多个角度、可量化或调整后可量化的指标数据。如彭国超等^[12]构建多维突变词检测指标体系，预测学科领域的研究前沿和热点。多维指标在研究前沿识别中需要设定阈值和权重，但由于不同研究者对关键指标的认识存在差异，阈值筛选缺乏统一标准，影响了评价指标体系的客观性和系统性。此外，多维指标计算复杂，难以形成统一量化体系。因此，有必要综合考虑多维指标的不同特征，简化指标计算，更准确地反映研究前沿本质。

3 研究前沿识别模型构建

3.1 多维研究前沿识别指标体系构建

结合研究前沿的属性，借鉴国内外期刊文献特点，在主题识别的基础上，引入主题新颖性^[13]、主题强度^[14]和主题研究热度^[15]3 项指标。同时，借鉴中国科学院等机构发布的研究前沿相关报告^[16]中用

• 51 •

日,文献类型为 Article 或 Review,总计 36 545 篇文献;二是来自 Web of Science (WoS) 核心集数据库,检索类别为“Oncology”,时间范围同上的高被引论文(在过去 10 年内被引用次数位于该学科前 1% 的论文),共计 7 001 篇文献。将两部分数据合并、去重,获得 38 030 条文献数据。删除缺失摘要字段的 103 条数据,最终得到 37 927 条研究数据,其年度分布情况,见图 1。

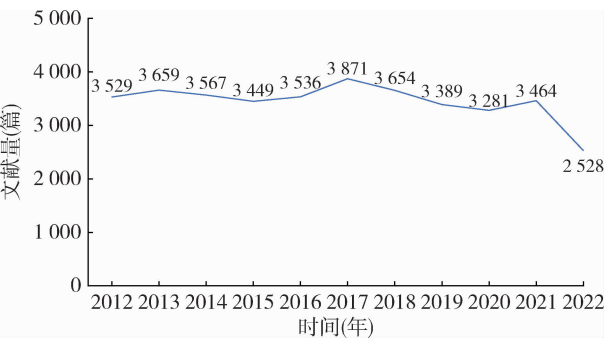


图 1 2012—2022 年肿瘤学领域发表文献年度分布情况

4.2 主题抽取

选择 BERTopic 的“all - MiniLM - L6 - v2”模型获取文本的句嵌入表征结果。通过 HDBSCAN 算法获得文本数据集的相似文档集合^[19];将相似度参数 Ngram_Range 值设置为 (1, 1),将文档中单个单词视为聚类单位;根据最小聚类数量自动设置主题数。比较互视图的聚类结果后,将最小聚类大小设置为 60。排除标签为“-1”的未识别主题后生成 54 个主题,将 54 作为 LDA 的预设主题数。对比 LDA 和 BERTopic 模型结果。一方面,基于 BERTopic 模型识别的关键词之间关联度更高,更易解

释。另一方面,从评估主题连贯性的两个指标分析,LDA 模型 C_v 为 0.466,BERTopic 模型为 0.669;LDA 模型 C_{UNASS} 为 -4.286,BERTopic 模型为 -0.556。BERTopic 模型在主题连贯性方面表现优于 LDA 模型。

4.3 基于多维指标的研究前沿识别

4.3.1 研究前沿识别多维指标计算 计算每个主题的主题强度、主题新颖性、主题研究热度、篇均被引频次及高被引文献数。利用 CRITIC 法计算 5 个指标的权重值,见表 1。

表 1 CRITIC 法权重计算结果

指标名称	赋值结果
主题强度	0.154 1
主题新颖性	0.228 9
主题研究热度	0.197 9
篇均被引频次	0.270 3
高被引文献数	0.148 8

4.3.2 研究前沿主题发现 根据前文对研究前沿及其分类的定义,参考白如江^[20]的筛选方法,综合考虑主题集综合指数及识别指标,将综合指数高于 50% 的主题标记为“综合指数高”,其余为“综合指数低”;将新颖性高于 80% 的主题标记为“新颖性高”,其余为“新颖性低”;将研究热度高于 50% 的主题标记为“研究热度高”,其余为“研究热度低”;对于交叉分类的主题,归类的优先顺序为热点研究前沿主题>新兴研究前沿主题>潜在研究前沿主题。主题集筛选结果,见表 2。

表 2 研究前沿主题发现

主题集	主题编号	主题名
热点研究前沿主题 (前 5 位)	T2	脑胶质瘤的诊断和治疗
	T4	急性性白血病的诊疗策略
	T5	治疗 EGFR 突变的非小细胞肺癌
	T6	他莫昔芬和内分泌疗法治疗激素受体阳性乳腺癌
	T9	基于前列腺特异性膜抗原抑制剂的靶向治疗前列腺癌
新兴研究前沿主题 (前 5 位)	T1	免疫治疗中 PD - 1 和 PD - L1 的临床应用
	T3	癌症相关成纤维细胞的生物标志物
	T7	线粒体代谢对肿瘤治疗的影响
	T11	曲妥珠单抗治疗 HER2 阳性乳腺癌
	T14	靶向 DNA 修复途径治疗肿瘤的机制及临床

续表 2

主题集	主题编号	主题名
潜在研究前沿主题（前 5 位）	T18	测序法应用于基因突变检测
	T25	KRAS 基因突变的研究
	T36	调节 KEAP1 – NRF2 通路的分子机制
	T37	骨肉瘤的诊断标准和治疗策略
	T38	纳米颗粒和药物输送的应用
衰退研究主题	T51	狄诺塞麦与唑来膦酸在骨髓瘤治疗中的比较研究

4.3.3 结果验证 采取专家访谈和文献支持两种方式对研究结果进行合理性和有效性验证。(1) 专家访谈。邀请上海市消化外科研究所研究员，以及肿瘤超声诊断、肿瘤外科治疗、肿瘤内科治疗和肿瘤病理分析等科室临床医师、医学院图书馆肿瘤学科馆员，共 6 名专家，进行访谈调查。使用预先制定的访谈提纲，采用面谈、微信访谈、电话访谈等形式，专家基于研究经验，以及通过文献查找国内外肿瘤研究领域基础研究、肿瘤检查及诊断方法、肿瘤治疗和内科治疗方法前沿热点，一致认为本研究有效地揭示了肿瘤学研究领域当前研究热点和潜在研究前沿。(2) 文献支持。《2023 年癌症统计报告》^[21] 概括肿瘤领域 6 个研究热点与本研究识别出的热点研究前沿、新兴研究前沿所涉及主题相符，如肿瘤药物研究方面（曲妥珠单抗治疗 HER2 阳性乳腺癌）、肿瘤靶点分析（靶向 DNA 修复途径治疗肿瘤的机制研究）、肿瘤信号通路研究（circRNA 相关 WNT 信号通路研究）、肿瘤药物联合（他莫昔芬和内分泌疗法治疗激素受体阳性乳腺癌）、肿瘤药物耐药性研究（复发性去势抵抗性前列腺癌治疗策略）、肿瘤微环境研究（肠道微生物群和肿瘤发生与肿瘤治疗）。同时，《细胞》杂志^[22]2023 年列举肿瘤学领域新兴研究主题也包括肿瘤细胞在肿瘤微环境中的作用，以及非癌症细胞如癌症相关成纤维细胞在肿瘤微环境中的重要性。Chhikara B S^[23] 和 Smith S M^[24] 发表的两项关键性肿瘤年度进展报告中提出纳米系统和纳米结合物应用于药物输送系统、免疫疗法应用、生物标志物用于肿瘤早筛等是未来肿瘤学领域的研究重点。本研究发现与以上文献观点基本吻合，从侧面印证了肿瘤学研究前沿主题识别结果较准确，其他新发现的内容也具有可信性。

5 结语

通过对肿瘤学领域高质量文献的深入分析，识别出该领域研究前沿，主要聚焦于以下关键方向。一是热点研究主题主要集中在探索各种高发肿瘤的新型诊疗方法，还包括免疫疗法个体化策略，如免疫检查点抑制剂、CAR – T 细胞疗法等。二是新兴研究前沿主题主要涉及肿瘤早筛、信号传导和转移机制等领域。Cescon D W^[25] 指出利用 ctDNA 和小囊泡中的生物标志物等进行“动态监测”可以全面了解肿瘤发展进程、检测肿瘤患者的耐药基因，有助于及时调整治疗方案，是肿瘤学未来发展方向之一。三是潜在研究前沿主题主要集中在肿瘤相关基因研究、新型抗癌药物研发以及药物再利用等领域。如寻找更具选择性的靶点或将特定病毒攻击癌细胞的机制用于新型抗癌药物研发；利用大数据识别药物再利用机会；设计纳米递药系统提高递送效率和增强抗肿瘤治疗效果。这些技术被视为潜在的癌症治疗策略，有望克服部分传统治疗难题。

本研究基于近 10 年 Web of Science 核心集数据库中肿瘤学领域的高质量期刊论文和高被引文献数据，运用 BERTopic 提取主题，结合主题强度、主题新颖性、主题研究热度、篇均被引频次、高被引文献数 5 个指标构建多维指标研究前沿识别模型。采用专家访谈和文献支持法验证研究结果的可靠性。本研究仅采用期刊数据，可能在一定程度上影响研究的全面性和深入性。未来研究可引入其他多源数据，进一步深入分析和解读主题兴起和衰退的原因，以更全面地理解肿瘤学领域的发展动态。

作者贡献：朱韶华负责数据收集与分析、论文撰写；李艳红负责研究设计、数据解释、论文撰写及

修订; 张迅负责研究设计和论文修订、黄海量负责研究设计、项目管理和论文修订。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

- 1 王茜, 谢秀芳, 王蕾, 等. 数据驱动背景下美国国立医学图书馆研究布局分析 [J]. 医学信息学杂志, 2023, 44 (1): 79-86, 97.
- 2 ZHENG R, ZHANG S, ZENG H, et al. Cancer incidence and mortality in China, 2016 [J]. Journal of the national cancer center, 2022 (1): 1-9.
- 3 PRICE D J. Networks of scientific papers [J]. Science, 1965 (149): 510-515.
- 4 SMALL H. Co-citation in the scientific literature: a new measure of the relationship between two documents [J]. Journal of the American society for information, 1973 (4): 265-269.
- 5 PERSSON O. The intellectual base and research fronts of JASIS 1986-1990 [J]. Journal of the American society for information science, 1994 (41): 3-7.
- 6 BOYACK K W. Co-citation analysis, bibliographic coupling, and direct citation: which citation approach represents the research front most accurately [J]. Journal of the American society for information science and technology, 2010 (12): 2389-2404.
- 7 卢超. 学科领域中新兴研究话题的内容特征识别及其实证研究 [D]. 南京: 南京理工大学, 2020.
- 8 田理, 郝继英. 基于引文网络的前沿科技热点演进分析框架 [J]. 中华医学图书情报杂志, 2021, 30 (6): 1-10.
- 9 姜鑫, 王德庄, 马海群. 关键词词频变化视角下我国“科学数据”领域研究主题演化分析 [J]. 现代情报, 2018, 38 (1): 141-146, 161.
- 10 罗恺, 袁晓东. 基于 LDA 主题模型与社会网络的专利技术融合趋势研究——以关节机器人为例 [J]. 情报杂志, 2021, 40 (3): 89-97.
- 11 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.
- 12 彭国超, 孔泳欣, 王玉文. 多维指标融合的主题突变检测研究 [J]. 情报学报, 2022, 41 (6): 584-593.
- 13 李军莲. 基于共被引耦合与 LDA 模型的研究前沿识别方法优化研究 [D]. 北京: 中国科学院大学, 2020.
- 14 刘自强, 岳丽欣, 朱承宁. 基金项目 and 论文主题扩散演化路径识别及其可视化研究 [J]. 现代情报, 2022, (11): 76-86.
- 15 钱莉, 朱恒民, 魏静. 话题演化研究综述 [J]. 数字图书馆论坛, 2021 (11): 57-64.
- 16 中国科学院发布《2023 研究前沿》报告和《2023 研究前沿热度指数》报告 [EB/OL]. [2023-11-28]. http://www.casisd.cn/mtsm/202311/t20231130_6939216.html.
- 17 王磊, 高茂庭. 基于改进 CRITIC 权的灰色关联评价模型及其应用 [J]. 现代计算机 (专业版), 2016, (23): 7-12.
- 18 王琳. 复合系统的耦合协调性及其对产业结构的影响分析 [D]. 长春: 吉林财经大学, 2021.
- 19 GROOTENDORST M. BERTopic: neural topic modeling with a class-based TF-IDF procedure [EB/OL]. [2023-11-28]. <https://arxiv.org/abs/2203.05794>.
- 20 白如江, 冷伏海, 廖君华. 一种基于多数据源主题对比的科学研究前沿识别方法 [J]. 情报理论与实践, 2017 (8): 43-48, 36.
- 21 SIEGEL R L, MILLER K D, WAGLE N S, et al. Cancer statistics 2023 [J]. CA: a cancer journal for clinicians, 2023, 73 (1): 17-48.
- 22 Cell Editorial Team. Cancer: untangling complexity together [J]. Cell, 2023, 186 (8): 1513-1514.
- 23 CHHIKARA B S, PARANG K. Global cancer statistics 2022: the trends projection analysis [J]. Chemical biology letters, 2023, 10 (1): 451-451.
- 24 SMITH S M, WACHTER K, BURRIS III H A, et al. Clinical cancer advances 2021: ASCO's report on progress against cancer [J]. Journal of clinical oncology, 2021, 39 (10): 1165-1184.
- 25 CESCOON D W, BRATMAN S V, CHAN S M, et al. Circulating tumor DNA and liquid biopsy in oncology [J]. Nature cancer, 2020, 1 (3): 276-290.