基于 LDA 模型的卫生健康媒体数据时间序列主题分析*

吴旭生[1] 查亚东[2] 张冬云[1] 彭祖胜[2] 林 圣[1] 刘宇锋[2] 和晓峰[1]

([1] 深圳市卫生健康发展研究和数据管理中心 深圳 518028 [2] 深圳广播电影电视集团 深圳 518026)

〔摘要〕 **目的/意义** 探索卫生健康领域媒体数据主题及其演化趋势。**方法/过程** 以深圳广电媒资数据库中的 160 549 条卫生健康领域媒体数据为研究对象，采用隐含狄利克雷分布模型结合时间序列进行主题聚类分析，并结合专家经验，进行对比分析。**结果/结论** 得到 25 个与卫生健康领域强相关的主题，根据主题强度演化趋势分为 6 组。主题建模的内容划分和强度变化有效反映了卫生健康领域热点事件的发生及其演进过程。利用隐含狄利克雷分布模型进行主题建模，结合时间序列分析主题分布、解读主题意义，有助于探索媒体数据在卫生健康领域的应用，为卫生健康公共事业赋能。

〔关键词〕 卫生健康媒体数据；隐含狄利克雷分布模型；热点事件；主题演化

〔中图分类号〕R-058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2025.02.010

Thematic Analysis of Time Series of Health Media Data Based on LDA Model

WU Xusheng[1], ZHA Yadong[2], ZHANG Dongyun[1], PENG Zusheng[2], LIN Sheng[1], LIU Yufeng[2], HE Xiaofeng[1]

[1] Shenzhen Health Development Research and Data Management Center, Shenzhen 518028, China; [2] Shenzhen Media Group, Shenzhen 518026, China

〔**Abstract**〕 **Purpose/Significance** To explore the theme and evolution trend of media data in the field of health. **Method/Process** Taking 160 549 pieces of health media data obtained from the Shenzhen Media Group database as the research object, topic clustering analysis is conducted based on the latent Dirichlet allocation (LDA) model combined with time series. Expert experience is used to compare and analyze the themes obtained from the LDA model. **Result/Conclusion** 25 themes are obtained strongly related to the field of health, 6 groups are diveded based on the trend of theme intensity evolution. The content division and intensity change of topic modeling effectively reflect the occurrence and evolution of hot events in the field of health. LDA model is used for theme modeling, combined with time series to analyze theme distribution, interpret theme significance, which is conducive to exploring the application of media data in the field of health and empowering public health undertakings.

〔**Keywords**〕 health media data; latent Dirichlet allocation (LDA) model; hot events; theme evolution

1 引言

通过对卫生健康领域媒体数据的主题建模和分类分析，能够揭示公众的关注焦点、舆论倾向以及社会对细分问题的态度，帮助各层级用户从多维度

〔修回日期〕 2024-07-26

〔作者简介〕 吴旭生，高级工程师，硕士生导师，发表论文 40 余篇；通信作者：和晓峰。

〔基金项目〕 广东省自然科学基金面上项目（项目编号：2022A1515012077）。

全面了解行业动态、趋势、议题及热点，为政府和相关机构提升舆情与智库分析能力、科学制定政策提供参考[1]。由于人工处理此类庞大数据成本高且主观性强，自动化主题建模算法辅助主题分类具有极高的实用价值[2]。目前，卫生健康领域主题分类多应用于在线健康社区和社交网络信息的归集整理[3-4]，医疗机构患者满意度的主题分布和情感分析[5]，以及对医疗信息化政策主题的提取与热点分析[6]等，针对卫生健康新闻的主题建模研究相对较少。世界卫生组织国际家族分类（World Health Organization for the family of international classifications，WHO-FIC）[7]专业性过强，不适用于大众传播及舆情热点分析。总的来说，国内外对卫生健康领域媒体信息的全面主题建模研究尚显不足，现有分类体系难以涵盖行业的各类主题，尤其是涉及政策、改革、基建、技术创新、人才培养等多元维度的精细化多级主题分类。

本研究以深圳市传统电视媒体数据作为研究对象，经过多模态数据转换、适于传统电视媒资的数据预处理、卫生健康相关信息抽提，训练隐含狄利克雷分布（latent Dirichlet allocation，LDA）模型，确定主题数、主题标注和主题词；通过时间序列分析追踪深圳市卫生健康相关主题的演化趋势，探索将自动化主题建模与人工专家经验相结合，构建易于理解、传播并适用于卫生健康行业舆情和智库分析的主题分类体系。

2 LDA 模型

2.1 LDA 模型简介

LDA 是一种广泛使用的主题模型，由 Blei D M 等[8]于 2003 年提出。该模型以无监督学习方式对内容数据进行主题聚类或文本分类，通过文本中词与词的共现概率来挖掘隐含主题信息，从而推测文档的主题分布。尽管 LDA 模型本身不直接考虑时间因素，但可以通过将时间因素引入模型来实现动态主题建模，分析主题如何随时间演化。LDA 模型具有很好的可解释性，能够与人工专家经验相结合，为海量数据集的分门别类、相对完整的分类建模提供了可能性。LDA 模型应用广泛，特别是在文本挖掘、信息检索和社交媒体分析等领域[9-10]。其能够发现文本数据中的主题结构，揭示潜在语义信息，并支持更高级的文本分析任务。此外，LDA 模型还可以与其他机器学习方法相结合，如聚类、分类和推荐系统，进一步提升文本数据的处理和理解能力[11-12]。

2.2 LDA 模型评估

主题数的选择直接影响 LDA 主题建模结果的准确性和可解释性。困惑度和主题连贯性可用于确定最佳主题数[13]。困惑度用于评估主题模型的优劣程度（泛化程度）。理论上，困惑度越小，说明模型性能越优，对新文本有更好的预测作用，困惑度曲线的最低点或拐点处对应的主题数通常被视为最佳主题数，但也要考虑主题数过多可能导致的主题概括范围小、语义内容差异小、主题划分困难问题。主题连贯性用于评估主题的可解释性。如果一组陈述或事实相互支持，则称之为连贯的。主题模型学习到的主题结果通常为一组重要的词，主题下词的语义关联性越紧密，一致性越高，模型的可解释性就越好。主题连贯性有多种测量方法，如 C_V、C_UMass、C_UCI 等[14]，这些方法包含不同的测量指标。其中，C_V 测量方法是目前最流行的测量指标，是 Gensim 库 CoherenceModel 类中的默认指标，通过计算每个主题中前 N 个词的共现频率来衡量主题的连贯性，使用词的共现频率创建词的内容向量，然后使用归一化逐点互信息和余弦相似度计算分数。Griffiths T L 等[15]提出主题强度的定义，通过文档的时间属性统计抽取出主题在时间维度的分布情况，以此考查固定时间窗口内每个主题的强度。利用主题强度进行主题挖掘研究，能够直观显示主题随时间的演化进程[16-17]。

3 卫生健康媒体数据主题建模

3.1 数据来源

选取深圳广播电影电视集团（简称深圳广电）1987—2022 年媒体数据作为研究对象，主要包括各类新闻栏目的新闻报道、人物访谈、专题报道等题

材，涵盖文字、视频、音频、图片等多种模态形式。采用语音转文字、光学字符识别等智能处理系统，得到文本数据。处理格式转换过程中产生的各种文字噪音和特殊字符，整合得到用于分析的数据集。通过特征词匹配筛选符合条件的数据，如医院、医术、医疗、医药、医生等，得到与卫生健康领域相关的媒体数据共 160 964 条。统计每年报道数量，见图 1。2000 年之前报道数量过少，因此主要分析 2000—2022 年的 160 549 条数据。可以看出，从 2005 年开始，卫生健康领域的媒体数据逐步增多，到 2010 年呈现井喷式增长。这主要归因于自 2005 年开始，电视台频道和节目快速发展，电视节目制作播出系统的数字化、网络化程度日渐提高，媒体内容数据不断增多。同时，涉及卫生健康领域的报道内容更加精细、专业，报道数量也在持续增长，于 2020 年达到数十年来的最高峰值。

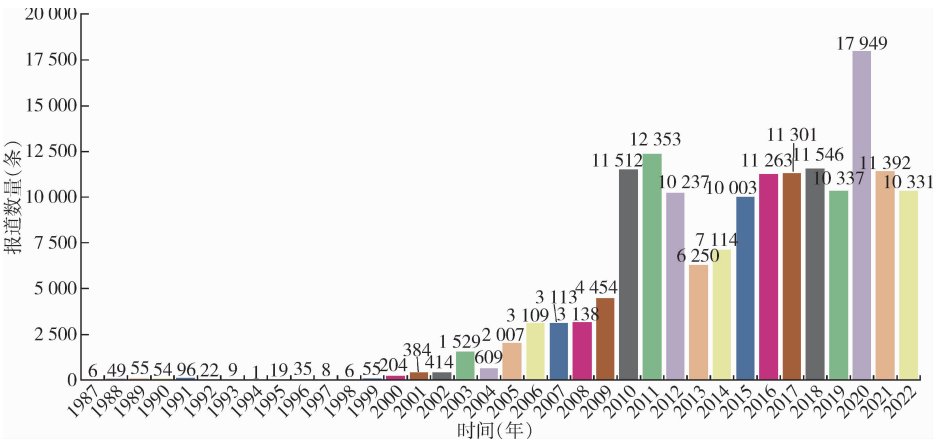

图 1　1987—2022 年深圳卫生健康领域每年报道数量

3.2　数据预处理

数据预处理阶段包括拼接、清洗和分词 3 个步骤。原始媒体数据中主要包含标题和正文两个字段，将二者拼接得到一个长文本，作为原始文档。使用 Python 语言编写程序，完成对纳入数据的清洗、去停用词和分词等处理。针对涉及广播电视系统自带的特殊字符标记，如“【正文】”“【导语】”“【标题】”“【同期声】”等，以及涉及既往敏感人物和敏感事件的报道等，使用正则清洗。使用 jionlp 库（https：//github. com/dongrixinyu/JioNLP）进行文本清洗，去除文本中的时间、日期、异常字符、冗余字符、HTML 标签、括号信息、URL、E-mail、电话号码，将全角字母、数字转换为半角，去除长度小于 2 的词。引用百度停用词表，去除“主张”“举行”等无实际意义的词。使用 jieba 库（https：//github. com/fxsjy/jieba）进行分词，并结合自定义词表将文档切分为词语[18]。

3.3　LDA 模型训练参数

使用 gensim 库建立 LDA 主题模型[19]。将清洗后的文档导入 gensim 库的 Dictionary 类中，得到相应词典，利用该词典将数据集转化为词袋模型向量，保存为 corpus。为了充分利用服务器多核心优势加速模型训练，选择 gensim 库的 LdaMulticore 类实现 LDA 模型，设置参数为 passes = 20、iterations = 100、eta = ‘auto’、eval_every = None，固定 random_ state，主题数量 num_topics 为变量 k，词典和 corpus 使用当前数据集对应的数据，其余参数使用默认值。在分析每条数据的主题分布时，使用系统默认的概率阈值 minimum_probability = 0. 01，每条数据可被划分到多个主题。LDA 训练完成后，使用 log_perplexity 函数计算困惑度，使用 CoherenceModel 函数计算主题连贯性。LDA 主题模型得到的结果使用 pyLDAvis 库（https：//github. com/bmabey/pyLDAvis）进行可视化，以辅助确定最优主题数。

4 基于时间序列的主题分析

4.1 全量数据分析

4.1.1 主题数实验分析 针对上述数据预处理后得到的约 16 万条全量数据，进行主题分类建模。主题模型的困惑度、主题连贯性与主题数目间的对应关系，见图 2。综合考虑困惑度与连贯性的评价结果，将主题数确定为 43，作为最优解。

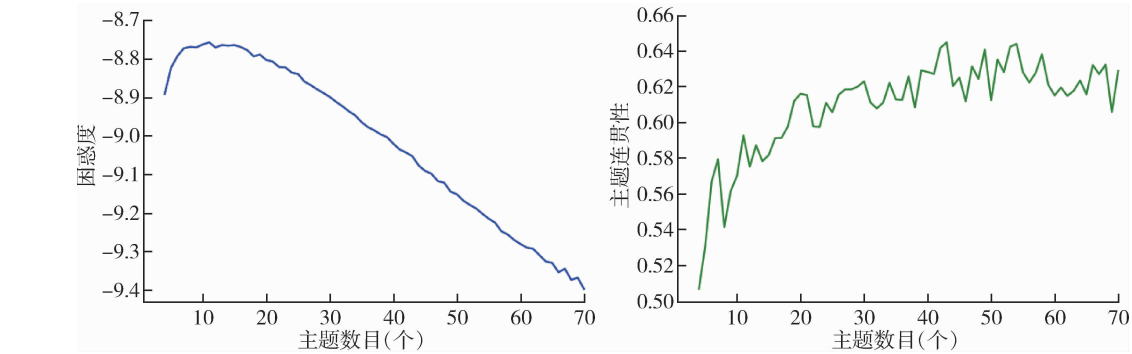

图 2 困惑度和连贯性分数曲线

4.1.2 主题词及主题标注分析 邀请来自医学临床、医院宣传科和信息科、广电部门以及卫生健康智库机构的多位具有高级职称的专家作为主题审议人员。针对 LDA 主题模型得到的 43 个主题，每个主题挑选出现频次最高的 10 个主题词。结合专家经验，参照实际业务中常见的卫生健康领域主题类型，进行主题摘要总结和标注解读。经过过滤与卫生健康相关性不高的主题，最终得到 25 个卫生健康领域强相关主题，见表 1。主题强度值越大，表明该主题在所有文本中越突出。可以看出，这些强相关主题基本覆盖了卫生健康领域大众传播和舆情热点的主要方面，其中，数量最多的主题依次是“警情处置及救治”“手术治疗”和“新生儿救治”，均是大众持续关注的舆情热点；而数量最少的主题依次是“无偿献血”“国际突发公共卫生事件”“传染病防治”，已被公众熟知或具有时段性特征，因此总报道数量相对较少。被过滤的与卫生健康领域无关或低相关的主题多为数据筛选阶段引入的噪音数据，这与新闻报道中同时报道多个人物或事件的特点有关，同时也揭示了 LDA 模型在处理海量内容数据时的另一个重要功效——从数据集中剔除一部分噪音数据。

表 1 LDA 主题标注和主题强度分析

主题编号	主题词	主题标注	数量(次)	占比(%)	主题强度
T01	医院、老人、派出所、现场、监控、发生、调查、受伤、救人、送往	警情处置及救治	42 089	25.98	0.026 39
T02	医院、患者、医生、手术、治疗、患者、病情、护士、抢救、检查	手术治疗	41 228	25.45	0.035 81
T03	孩子、妈妈、医院、父母、出生、孕妇、照顾、婴儿、治疗、新生儿	新生儿救治	39 013	24.08	0.026 93
T04	公司、律师、负责人、赔偿、责任、事件、调查、法律、医院、鉴定	受伤治疗赔偿纠纷	36 158	22.32	0.029 31
T05	医院、工人、受伤、伤者、骨折、现场、事发、工地、意外、烧伤	工伤意外	32 612	20.13	0.020 09
T06	捐献、希望、生活、故事、器官、生命、全国、精神、遗体、英雄	器官捐赠	31 802	19.63	0.024 06
T07	健康、医生、疾病、治疗、运动、糖尿病、近视、睡眠、预防、高血压	预防保健	31 755	19.60	0.022 29
T08	社区、工作、街道、居民、人员、服务、工作人员、现场、保障、物资	社区医疗保障	30 609	18.89	0.042 65
T09	医院、医疗、社康、患者、门诊、预约、诊疗、看病、挂号、就医	预约挂号	30 468	18.81	0.013 21
T10	医保、政策、参保、社保、老人、医疗保险、家庭、保障、费用、标准	医保事务	28 715	17.72	0.025 46
T11	活动、培训、服务、社区、健康、公益、爱心、知识、急救、义工	社区健康宣传	28 710	17.72	0.027 15
T12	国际、科技、创新、研究、团队、发展、人才、实验室、生物、基因	科技创新	28 523	17.61	0.031 09
T13	事故、现场、爆炸、受伤、医院、原因、伤者、火灾、救治、抢救无效	事故救援	27 804	17.16	0.026 05
T14	症状、医院、食物、中毒、细菌、呕吐、腹泻、感染、过敏、不适	食物中毒	27 765	17.14	0.039 64
T15	孩子、家长、儿童、医院、幼儿园、小朋友、医生、手足口、儿科、牙齿	儿童疾病	27 564	17.01	0.004 46

续表 1

主题编号	主题词	主题标注	数量(次)	占比(%)	主题强度
T16	司机、事故、车辆、交警、现场、医院、肇事、受伤、伤者、交通事故	车祸救治	27 292	16. 85	0. 014 82
T17	市场、执法、检查、食品、餐厅、卫生、监管、食品安全、非法、超市	市场卫生监管	27 249	16. 82	0. 029 05
T18	药品、检测、销售、药店、消费者、成分、市场、合格、质量、超标	药品监管	25 457	15. 71	0. 014 31
T19	信息、网络、广告、保健品、诈骗、宣传、网上、老人、虚假、养生	保健养生	25 322	15. 63	0. 025 93
T20	医生、医院、美容、手术、医疗、整形、注射、资质、美容院、整容	美容整形	23 964	14. 79	0. 024 81
T21	救援、应急、地震、被困、紧急、灾区、搜救、群众、伤员、救灾	灾害救援	22 878	14. 12	0. 038 78
T22	新增、病例、隔离、报告、输入、防控、本土、措施、境外、风险	本土突发公共卫生事件	22 237	13. 73	0. 010 71
T23	疫苗、接种、感染、病毒、流感、病例、预防、患者、传播、症状	传染病防治	21 864	13. 50	0. 026 39
T24	美国、英国、印度、全球、死亡、口罩、组织、确诊、世卫、抗疫	国际突发公共卫生事件	18 747	11. 57	0. 018 42
T25	血液、献血、志愿者、无偿献血、爱心、干细胞、捐血、毫升、用血	无偿献血	14 776	9. 12	0. 034 88

4.1.3　主题强度变化分析　主题强度值越大，表明该主题在所有文本中越突出，本研究中最突出的主题是“社区医疗保障”“食物中毒”“灾害救援”等。以年为单位划分时间窗口，分别计算每个主题在 2000—2022 年之间每年的主题强度，并绘制主题强度随时间演变的折线图。根据演化趋势，将 25 个卫生健康主题分为 6 组，分别为上升型、下降型、上升－下降型、下降－上升型、波动型和平稳型，见图 3。

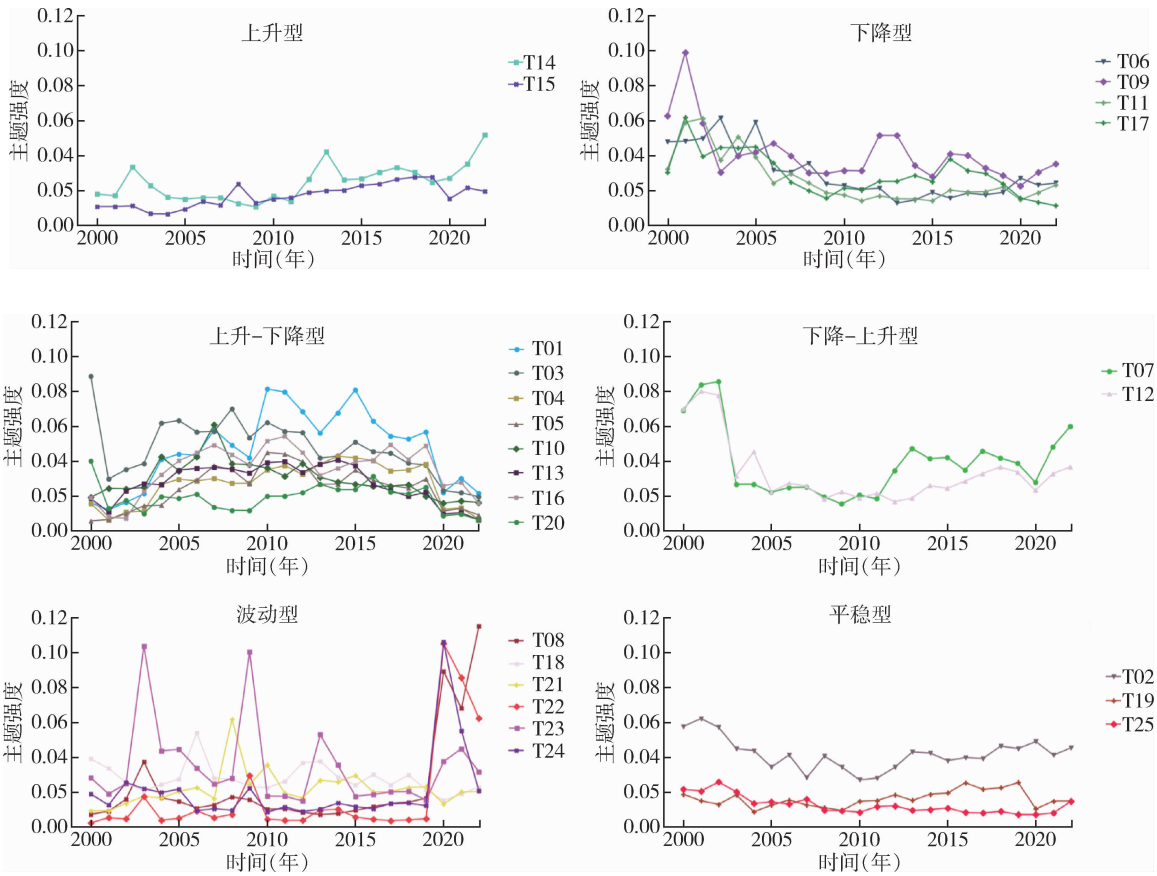

图 3　主题强度随时间演变

通过分组分析，可以了解大众媒体对卫生健康领域各类主题关注程度的变化趋势。例如，随着医疗信息化水平的不断提高，“预约挂号”主题报道呈下降型趋势，而近年来“儿童疾病”的关注度不断上升，“保健养生”“无偿献血”等主题报道则常年较为稳定。“灾害救援”“传染病防治”“突发公

共卫生事件”“社区医疗保障”等具有典型时间阶段性特征的主题，其强度变化呈现波动型规律。主题强度随时间变化的趋势，可作为舆情监测评估与政策反馈的重要依据。

4.2 年度数据分析

采用上述相同方法，对各年数据分别进行 LDA 主题建模和主题总结标注。需要注意的是，主题构建的结果可能是二级甚至三级主题，需结合实际进行分析。逐年进行主题建模和标注后，除了上述经常出现的主题外，还发现一些独特的主题，如 2008 年北京奥运会和汶川大地震、2009 年 H1N1 甲型流感、2014 年 H7N9 禽流感和埃博拉病毒疫情、2015 年中东呼吸综合征等。这些主题都是具有深刻影响的事件，具有较强的时效性，很好地反映了在一些重大时事中涉及卫生健康主题的内容。

5 结语

本研究以深圳市权威媒体数据库数十年积累的新闻内容为研究对象。首先，进行数据预处理和统计分析，包括数据清洗、从多模态媒体数据中筛选并整理出一份可用于 LDA 主题建模的文档数据集。在 LDA 主题建模分析中，基于主题聚类的连贯性、困惑度指标共同确定模型主题数量的合理选择，并由人工专家根据经验进行主题筛选与标注。其次，通过主题强度来判断并解读主题随时间的变化趋势。然后，针对时间切片的数据集进行 LDA 迭代计算，逐级细化主题聚类分析。最终，提出融合 LDA 主题建模和人工总结标注的方法，用于对大规模媒体数据进行主题分类建模，建立了一套适用于卫生健康领域行业资讯内容的分类主题标签体系。该体系可以应用于行业舆情热点分析和智库分析等领域，具有很大的实用价值。

本研究仍存在以下不足。一是数据预处理采用目前流行的分词工具 jieba，未来研究应对比新兴分词工具[20]后选用最优工具，通过提高分词效果优化 LDA 模型的计算结果。二是建模方法方面，LDA 模型需经历多次计算后由人工确定主题数量，这需要更多的辅助手段和评价指标来确定最佳主题数，未来可考虑采用基于深度学习的 BERTopic 主题模型进行分析[21]。一般而言，面对海量数据时，可先采用 LDA 算法持续深化挖掘多级分类，再结合人工专家经验，快速形成达到一定质量水平的多级分类主题标签及对应的数据集，可为后续采用深度学习模型进行训练提供有力的数据支撑。

作者贡献：吴旭生负责研究设计、项目管理、论文修订；查亚东负责实验指导；张冬云负责主题标注；彭祖胜负责编程与模型训练；林圣负责数据分析、论文撰写；刘宇锋负责数据收集与预处理；和晓峰负责研究设计。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 汤紫薇，文庭孝，张冬云. 2000—2022 年我国卫生健康政策演进分析［J］. 中国现代医生，2023，61（9）：89－92.
2 LIU L，TANG L，DONG W，et al. An overview of topic modeling and its current applications in bioinformatics［J］. Springerplus，2016，5（1）：1608.
3 高慧颖，刘嘉唯，杨淑昕. 基于改进 LDA 的在线医疗评论主题挖掘［J］. 北京理工大学学报，2019，39（4）：427－434.
4 吴胜男，田若楠，蒲虹君，等. 基于社交媒体的医药领域关联主题预测方法研究［J］. 数据分析与知识发现，2021，5（12）：98－109.
5 张瑶，夏晨曦，马敬东. 某医院患者投诉信息中服务体验主题建模与情感分析［J］. 中华医院管理杂志，2019，35（12）：1037－1041.
6 蔡琼，吴荣飞，李瑞锋，等. 基于 LDA 模型的医疗信息化政策主题提取与热点分析［J］. 中国数字医学，2023，18（4）：112－120.
7 JAKOB R，USTÜN B，MADDEN R，et al. The WHO family of international classifications［J］. Bundesgesundheitsblatt－Gesundheitsforschung－Gesundheitsschutz，2007（7）：924－931.
8 BLEI D M，NG A Y，JORDAN M. Latent Dirichlet allocation［J］. Journal of machine learning research，2003，3（4/5）：993－1022.
9 JELODAR H，WANG Y，YUAN C，et al. Latent Dirichlet allocation（LDA）and topic modeling：models，applications，a survey［J］. Multimedia tools & applications，2019，78（11）：15169－15211.

（下转第 75 页）

综述［J］. 河北科技大学学报, 2021, 42（1）: 48 – 59.
2 邹然, 柳杨, 李聪, 等. 图表示学习综述［J］. 北京师范大学学报（自然科学版）, 2023, 59（5）: 716 – 724.
3 WU Y, LAN W, FAN X, et al. Bipartite network influence analysis of a two – mode network［J］. Journal of econometrics, 2024, 239（2）: 105562.
4 张佳慧, 张婷, 吕来水, 等. 基于加权投影的二分网络的链路预测［J］. 计算机应用与软件, 2021, 38（3）: 264 – 268, 297.
5 张斌, 马费成. 科学知识网络中的链路预测研究述评［J］. 中国图书馆学报, 2015, 41（3）: 99 – 113.
6 贺圣平, 王会军, 李华, 等. 机器学习的原理及其在气候预测中的潜在应用［J］. 大气科学学报, 2021, 44（1）: 26 – 38.
7 HU F, ZHANG Y, YAN X Y, et al. An improved heterogeneous graph convolutional network for inter – relational medicine representation learning［J］. IEEE multimedia, 2023, 30（1）: 52 – 61.
8 余黄樱子, 董庆兴, 张斌. 基于网络表示学习的疾病知识关联挖掘与预测方法研究［J］. 情报理论与实践, 2019, 42（12）: 156 – 162.
9 DRUCE M, ROCKALL A, GROSSMAN A B. Fibrosis and carcinoid syndrome: from causation to future therapy［J］. Nature reviews endocrinology, 2009, 5（5）: 276 – 283.
10 RAM P, PENALVER J L, LO K B U, et al. Carcinoid heart disease: review of current knowledge［J］. Texas heart institute journal, 2019, 46（1）: 21 – 27.
11 丁炎波, 陈炳芳, 庄耘, 等. 上消化道类癌的诊断和治疗［J］. 中国医药指南, 2013, 11（8）: 592 – 594.
12 EDINOFF A N, RAVEENDRAN K, COLON M A, et al. Selective serotonin reuptake inhibitors and associated bleeding risks: a narrative and clinical review［J］. Health psychology research, 2022, 10（4）: 39580.
13 RORSTAD O. Prognostic indicators for carcinoid neuroendocrine tumors of the gastrointestinal tract［J］. Journal of surgical oncology, 2005, 89（3）: 151 – 160.
14 KAKIMOTO K, INOUE T, TOSHINA K, et al. Multiple mesenteric panniculitis as a complication of sjögren's syndrome leading to ileus［J］. Internal medicine, 2016, 55（2）: 131 – 134.
15 SANKPAL U T, GOODISON S, JONES – PAULEY M, et al. Tolfenamic acid – induced alterations in genes and pathways in pancreatic cancer cells［J］. Oncotarget, 2017, 8（9）: 14593 – 14603.
16 YANG B, XIE X, WU Z, et al. DNA damage – mediated cellular senescence promotes hand – foot syndrome that can be relieved by thymidine prodrug［J］. Genes & disease, 2023, 10（6）: 2557 – 2571.

（上接第 67 页）

10 周梅佳佳, 朱庆华. 基于 LDA 主题模型的智慧养老研究主题分析［J］. 医学信息学杂志, 2024, 45（3）: 8 – 15.
11 徐文秀. 基于 LSTM 和 LDA 的可再生能源领域主题分类研究［D］. 济南: 山东大学, 2020.
12 李散散. 基于用户行为分析和 LDA 模型的数字媒体推荐系统的设计与实现［J］. 现代电子技术, 2020, 43（7）: 146 – 149.
13 张东鑫, 张敏. 图情领域 LDA 主题模型应用研究进展述评［J］. 图书情报知识, 2022, 39（6）: 143 – 157.
14 RÖDER M, BOTH A, HINNEBURG A. Exploring the space of topic coherence measures［EB/OL］.［2024 – 06 – 25］. https: //dl. acm. org/doi/10. 1145/2684822. 2685324.
15 GRIFFITHS T L, STEYVERS M. Finding scientific topics［J］. Proceedings of the national academy of sciences, 2004, 101（S1）: 5228 – 5235.
16 陶胜阳, 许新华, 余亚烽, 等. 基于 LDA 模型的教育技术学研究主题挖掘及演化趋势分析［J］. 现代信息科技, 2023, 7（6）: 176 – 180.
17 袁永旭, 李黛, 王思源, 等. 基于 LDA 模型的我国医保政策主题建模与演化分析［J］. 预防医学情报杂志, 2023, 39（6）: 710 – 716.
18 江锐鹏, 钟广玲. 中文分词神器 Jieba 分词库的应用［J］. 电脑编程技巧与维护, 2023（9）: 87 – 89.
19 ŘEHŮŘEK R, SOJKA P. Software framework for topic modelling with large corpora［C］. Valletta: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010.
20 倪维健, 孙浩浩, 刘彤, 等. 面向领域文献的无监督中文分词自动优化方法［J］. 数据分析与知识发现, 2018, 2（2）: 96 – 104.
21 GROOTENDORST M. BERTopic: neural topic modeling with a class – based TF – IDF procedure［EB/OL］.［2024 – 06 – 25］. https: //arxiv. org/pdf/2203. 05794.