基于新媒体的医学舆情平台构建研究*

侯 震 童惟依 邓靖飞 李 扬 王 青

（中国医学科学院/北京协和医学院医学信息研究所 北京 100020）

〔摘要〕 设计并实现基于新媒体的医学舆情平台，分析平台需求，阐述平台构建，包括整体架构、数据层、业务层、应用层等方面，介绍平台应用效果，指出该平台有助于高效完成医学舆情信息收集与分析，为相关领域人员分析、预警医学舆情事件提供有力支持。

〔关键词〕 新媒体；舆情平台构建；医学健康信息

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2021.04.013

Study on the Building of Medical News Events and Public Opinion Platform Based on New Media *HOU Zhen, TONG Weiyi, DENG Jingfei, LI Yang, WANG Qing, Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China*

〔**Abstract**〕 The medical news events and public opinion platform is designed and realized. The paper analyzes the demand of the platform, elaborates the building of the platform, including the overall architecture, data layer, business layer, application layer, etc., introduces the application effect of the platform, and points out that the platform is helpful to efficiently collect and analyze the medical news events and public opinion information, so as to provide strong support for the analysis and early warning of medical news events and public opinion for personnel in related fields.

〔**Keywords**〕 new media; building of public opinion platform; medical and health information

1 引言

根据《中国互联网发展报告（2020）》，截至 2020 年 3 月中国互联网移动用户达 13.19 亿[1]，移动用户已成为互联网信息获取及传播的主要参与者[2]。随着互联网移动用户增加及新媒体平台应用普及，新媒体平台已逐步成为公众信息获取与传播的重要渠道[3]。相较于传统媒体，新媒体平台信息传播及时性、互动性及广域性促进了信息交流，同时也对舆情信息采集、预警及分析带来挑战[4]。自 2020 年初新型冠状病毒肺炎疫情（以下简称新冠肺炎疫情）暴发以来，有超过 10 亿条疫情新闻、防控信息、科普短视频等在新媒体平台中传播[5]，传统舆情采集及分析平台仅通过数据爬虫、文本分词、命名实体识别等手段已无法满足新媒体环境下舆情信息采集与分析，因此本研究通过分布式网络爬虫、新媒体平台数据接口、医学知识组织体系等技术方法，设计并构建基于新媒体平台的医学舆情平台，辅助卫生决策者及研究人员追踪卫生事件源头、分析卫生事件演化进展，及时做出有利于推动社会卫生事件发展的决策。

〔修回日期〕 2020－12－29

〔作者简介〕 侯震，助理研究员，发表论文 19 篇；通讯作者：李扬，副研究员；王青，编审。

〔基金项目〕 中国医学科学院医学与健康科技创新工程项目“医学舆情与预印本平台构建”（项目编号：2020－I2M－2－002）。

2 相关研究现状

2.1 舆情信息采集

邱雅[6]等针对传统网络爬虫技术缺陷，提出设置监控机制监测爬虫收集速度、根据目标网站流量情况调整爬虫策略、调用网络搜索引擎提高爬虫数据收集覆盖率等改进技术方案，为利用网络爬虫技术收集信息数据提供优化方案。袁畅[7]等提出以分布式平台为基础，搭建网络爬虫模块，构建信息收集及算力可拓展的舆情分析架构。刘斌[8]利用 Twitter 平台提供的应用程序接口（Application Programming Interface，API），通过自然语言处理方法进行数据清洗，结合潜在狄利克雷分配（Latent Dirichlet Allocation，LDA）模型，筛选 Twitter 平台中的热点话题。

2.2 舆情事件下公众情感分析

研究人员针对互联网中的文本信息，基于历史数据训练舆情传播分析模型，利用统计学、机器学习方法对舆情事件传播趋势、公众情感取向进行分析。习海旭等[9]通过情感词典及 TF－IDF 算法构建语义情感分析模型，用于判断及分析互联网特定事件下的公众情感。肖倩[10]等针对信息传播中的时序信息，利用卷积神经网络算法，设计热点舆情的识别方法，为新媒体环境下热点舆情发现与分析提供技术支持。

2.3 舆情事件传播和演化特征分析

传染病及其变种模型为模拟舆情信息在新媒体中传播提供理论基础，研究人员通过以上模型对特定舆情事件传播和演化特征进行分析。张连峰等[11]基于 SIR 传染病模型基础，结合超网络理论概念，分析微博平台舆情事件传播关键节点、路径及特征，为微博平台舆情事件传播演化提供技术基础。彭程[12]等基于 SIR 传染病模型与 EGM 灰色预测模型，构建舆情防控及预测模型，通过特定数据训练，挖掘特定舆情事件预测及传播规律。周琦萍[13]等借助 SIS 传染病模型，分析舆情事件随时间的演化过程，挖掘舆情事件暴发临界点，为舆情信息预警与公众疏导提供理论支持。调研以上研究进展为本平台舆情数据采集、舆情信息预警、公众情感分析、舆情信息传播路径追踪提供设计及技术支持。

3 医学舆情平台需求分析

传统舆情信息收集及分析平台通常采用浏览器/服务器（Browser/Server，B/S）架构，由数据层完成数据采集及封装处理，业务层实现平台各功能逻辑，应用层用于响应用户请求及平台功能呈现[14]。随着政府、研究机构、商业、媒体等机构对互联网热点信息、公众舆论分析需求日益增加，国内厂商陆续推出各类舆情分析产品，如百度舆情、人民网舆情、中青华云舆情、慧科舆情等，为用户提供领域热点信息分析、舆情信息预警、特定事件分析等服务。由于医学名词专业性及复杂性、呈现方式的多元化以及新媒体环境下医学健康信息数据不断激增，基于传统文本采集及分析的舆情系统已无法满足各领域用户对医学健康事件的挖掘、追踪和分析需求，影响用户对医学健康舆情事件的分析和演化判断。

4 医学舆情平台设计与构建

4.1 架构设计（图 1）

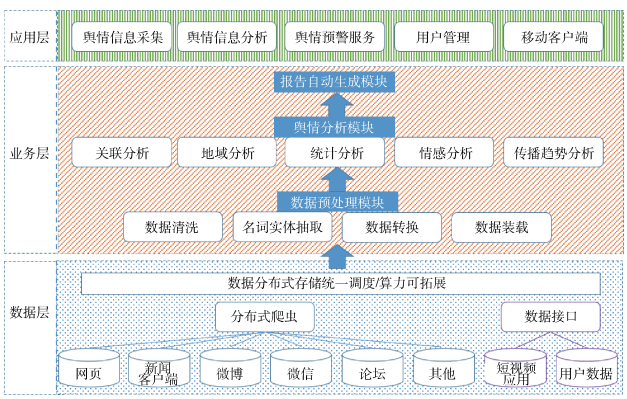

图 1 医学舆情平台架构设计

4.2 数据层建设机制

4.2.1 概述 医学舆情数据采集的广泛性与准确

性是舆情信息分析及趋势判断的关键因素，为保证医学舆情数据采集全面性、及时性、可靠性，采取数据爬虫及数据接口两种模式获取相关数据，通过采集网页、新媒体、客户端等渠道信息为本平台提供底层数据支持。

4.2.2 数据爬虫模式 利用数据爬虫技术，根据采集任务需求实时抓取网页、新闻客户端、微博微信及新媒体、贴吧论坛等渠道全网数据。平台通过协作式爬虫技术对各渠道信息数据进行分工式爬取，同时部署针对各渠道内容分类列表、关键词内容的多线程数据爬取方案，以满足医学舆情数据采集的及时性和高效性。在数据爬虫虚拟地址部署方面，采取代理式 IP 技术，通过代理 IP 池信息及时更换，完成数据不间断采集，以保证平台信息采集稳定性。同时为保证数据采集能力拓展性，针对不同规模数据采集需求，采用私有云爬虫部署机制，通过云技术协调和部署数据采集规模，提升平台数据采集拓展能力。

4.2.3 数据接口模式 考虑到新媒体平台已逐渐成为医学舆情信息的首发源，针对新媒体平台数据传播速度快、影响广泛的特点，根据医学舆情任务采集需求，通过数据加密接口间断性获取微博、问答网站、短视频应用等新媒体舆情热点信息、信息传播流量、用户个人数据、用户评论信息、信息传播路径等数据。为平台舆情传播趋势分析、传播路径追踪、参与用户分析提供全面数据支撑。

4.3 业务层建设机制

4.3.1 概述 业务层由数据预处理、舆情分析、报告自动生成模块构成。可实现底层数据清洗及结构化处理，匹配名词标准库、医学主题词表、行业知识库等权威知识库信息，开展舆情信息关联分析、数据地域分析、统计分析、舆情传播趋势分析。

4.3.2 数据预处理模块 （1）数据清洗。将所有原始数据分别进行噪声过滤、重复度检验。平台基于人工神经网络算法提取站点中的有效标签内容，通过人工纠正不断提升标签内容提取正确率，从而过滤站点中的广告、链接信息等噪音数据。系统将过滤后的数据通过 hash 算法进行重复度检验，重复数据只保留原始链接地址，用于医学舆情信息溯源及路径追踪。（2）数据关联标注及结构化转换。为追踪医学舆情信息传播路径，辅助医学舆情信息分析，平台基于医学主题词表、行业知识库、舆情信息库等本体库实现原始数据重新组织及信息关联。系统将清洗后的每条医学舆情信息按内容及类别存储在结构化文档中，并形成信息的层级关联，见图 2。

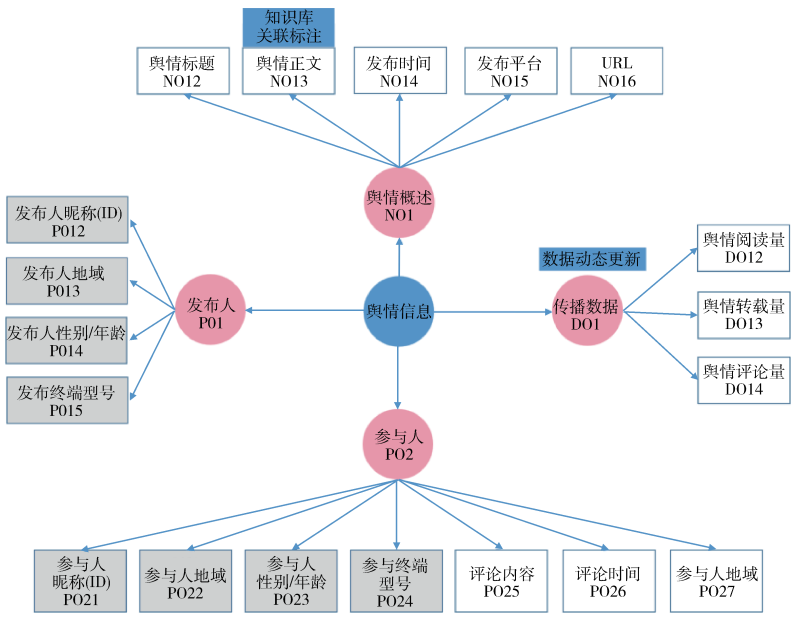

图 2 舆情信息关联结构

图中红色节点代表构成舆情信息的重要分类，如舆情概述、传播数据、发布人、参与者，蓝色节点表示该分类的主要元素信息，灰色节点表示基于新媒体平台接口的隐私数据，箭头表示各层级间的关系，系统同时赋予关联网络每个节点相应数字作为唯一标识。系统根据医学主题词表信息自动标注舆情内容中的疾病名称与别名、疾病及症状、药物成分及商品名称，实现医学舆情信息内容关联。

4.3.3 医学舆情分析模块 （1）可视化分析。基于清洗后的结构化数据，根据数据渠道来源、舆情传播量、医学热词数据等维度，通过饼状图、折线图、散点图、词云等可视化分析图像动态展示舆情信息传播趋势，直观呈现舆情信息传播数据。（2）信息传播路径分析。基于SIR传染病模型及社会网络模型，以用户转发及评论、用户关系、用户个人数据为基础，结合舆情信息传播量，分析医学舆情事件传播关键用户节点，筛选用户评论意见领袖，通过力导图呈现传播路径，为分析舆情事件暴发节点、跟踪信息传播路径、分析事件演化过程提供有力支撑。（3）评论信息情感分析。基于自然语言处理及语言情感判断技术，通过拆分医学舆情事件中每个用户评论语句，进行情感分析及加权计算，经过数据汇总获得公众对该舆情事件的情感取向，为判断公众对舆情事件情感态度、意见取向提供适当参考。

4.3.4 报告自动生成模块 为便于用户撰写医学舆情分析报告，平台设置报告自动生成模块。报告模板预设由事件标题、事件概述、事件传播趋势、事件传播路径、公众热点评论5部分构成。用户可根据个人需求选择性插入舆情分析图表、特殊案例等信息。平台可导出带有页眉页脚标签、个性化水印的PDF或DOCX文档，同时支持事件传播数据导出功能，便于用户对舆情事件多类型数据分析。

4.4 应用层功能集成与服务

4.4.1 概述 应用层提供舆情信息采集、分析、预警服务等功能，通过对医学舆情事件数据整合分析、挖掘利用实现事件分析、信息预警、数据大屏展示等多种服务方式，为用户追踪健康事件源头、分析健康事件演化进展提供支持，辅助用户及时做出事件处理决策，根据用户需求可个性化集成各类型服务。

4.4.2 事件分析服务 用户通过“与”、“或”、“非”关系设置分析关键词的关联逻辑，引导平台数据采集模块完成目标区域数据收集。由于不同身份特征用户对医学分析关键词表述具有差异性（如不同用户将阿尔茨海默症表述为AD、老年痴呆症、失智症等），平台在医学舆情事件数据采集、内容关联、事件数据回溯等分析服务模块中引入医学主题词表及医学拓展词表，结合舆情事件传播路径追踪、数据可视化等模块辅助分析医学健康舆情事件传播过程。（1）舆情事件传播路径分析。平台通过采集舆情事件首发时间、首发人、转发人、评论信息、传播层级等数据绘制舆情事件传播路径，并对促进事件传播的节点人物、关键评论进行高亮显示。有助于用户梳理舆情事件传播发展脉络，挖掘舆情传播关键节点，为分析舆情事件演化提供参考。（2）舆情事件数据可视化分析。平台基于用户设置的关键词，采集全网新闻、文章等信息，并关联网站链接、发帖时间等数据。利用数据可视化插件，分别按时间、传播热词等维度提供舆情事件传播趋势、热词概览等可视化服务。

4.4.3 信息预警服务 根据用户提交的预警关键词，设置监测地理区域、媒体范围、时间段及预警信息推送方式、推送频率等关键指标。平台提供国内网页、纸媒及新媒体在内的全网媒体24小时不间断监测预警。平台根据用户设置的预警关键词，自动匹配并推荐相应医学关联词、事件关联词及地域信息，辅助用户对关键词事件的预警设置，同时平台实时根据新媒体客户端热点榜数据接口、文章传播趋势数据，提供健康、时政等领域的分类舆情热点预警。平台提供短信、微信、邮箱等不同渠道

的预警信息推送功能。满足用户对预警推送的各类需求。

4.4.4 数据大屏展示 基于数据可视化分析及展示模块，自适应不同尺寸大屏展示需求，用户可通过拖拽方式设置舆情传播数据、预警信息等各类型舆情数据可视化模块展示位置。同时基于用户物理坐标及地理信息系统（Geographic Information System，GIS）定位，提供舆情监测区域动态数据呈现功能。

5 平台应用与实践

医学舆情平台基于面向服务架构（Service Oriented Architecture，SOA），灵活地将各功能模块通过接口及协议的方式整合在一起，实现平台功能更新的快速部署及新模块整合。平台在开发过程中采用分布式处理及算力、存储拓展技术，保证平台面向多任务处理的拓展性、稳定性。平台前端利用 Echarts 插件实现数据可视化呈现，用户可根据个人需求设置图表颜色及样式。同时平台通过采用相应数据安全等级保护、用户动态登录认证、用户行为安全认证等多种数据安全保证措施，保证数据存储、利用及用户各项操作的安全性。新冠肺炎疫情期间，应用该平台进行“国内各地区新增确诊新冠病例”、“秋冬季流感病毒”、“丹麦变异新冠病毒”、“新冠疫苗研发”等事件的全网舆情信息收集与分析，共收集舆情信息 1 071 153 条、公众评论 12 035 条，撰写舆情报告 23 期，部署专题医学舆情预警 4 期，高效完成医学舆情信息收集、预警与分析工作。医学舆情事件案例监控页面，见图 3。

图 3 医学舆情事件案例监控页面

6 结语

新媒体环境下医学健康事件传播的复杂性和广泛性为舆情信息分析带来挑战，本研究采用分布式数据爬虫技术及新媒体平台数据接口方式，广泛采集舆情数据，利用人工智能算法、医学知识组织方法、数据可视化技术等从大量数据中梳理舆情事件发展脉络、分析公众舆论焦点、挖掘事件传播关键节点，为相关领域人员分析、预警医学舆情事件提供平台保障。面对新媒体环境下各类应用不断增加的短视频信息，目前平台通过协议方式依托数据接口获取短视频数据，过程繁琐且被动，易受到短视频应用平台限制和制约，下一步将继续开展视频爬虫及视频内容识别技术的探索与应用，不断完善平台在新媒体信息收集、存储及分析的功能模块，为新媒体环境下医学舆情信息收集与分析提供有力保障。

（下转第 78 页）

gy and workflow process for providing translational research informatics support [J]. J Biomed Inform, 2009, 42 (2): 377 – 381.
7 Cantor Z, Plint AC, Kamil Y, et al. REDCap: a pediatric ED experience [J]. Am J Emerg Med, 2016, 34 (10): 2048 – 2049.
8 Harvey LA. REDCap: web – based software for all types of data storage and collection [J]. Spinal Cord, 2018, 56 (7): 625.
9 Dai W, Zhang Y, Feng W, et al. Using Patient – reported Outcomes to Manage Postoperative Symptoms in Patients with Lung Cancer: protocol for a multicentre, randomised controlled trial [J]. BMJ Open, 2019, 9 (8): e030041.
10 耿辉, 贺海蓉, 曾宪涛, 等. 多中心临床数据采集系统 REDCap 系统应用及架设 [J]. 中国循证心血管医学杂志, 2017, 9 (9): 1025 – 1028.
11 贺海蓉, 郑婕, 耿辉, 等. 如何利用 REDCap 数据采集系统实现病例报告表的设计 [J]. 中国循证心血管医学杂志, 2018, 10 (2): 133 – 137.
12 Tomko RL, Gray KM, Oppenheimer SR, et al. Using REDCap for Ambulatory Assessment: implementation in a clinical trial for smoking cessation to augment in – person data collection [J]. Am J Drug Alcohol Abuse, 2019, 45 (1): 26 – 41.
13 李蒙, 田国祥, 贺海蓉, 等. 如何利用 REDCap 实现重复测量数据的录入及管理 [J]. 中国循证心血管医学杂志, 2018, 10 (6): 647 – 650.
14 侯倩, 江真, 刘卿, 等. REDCap 数据库在嗓音外科临床研究中的初步应用 [J]. 中国耳鼻咽喉颅底外科杂志, 2017, 23 (4): 338 – 341.
15 贺海蓉, 田国祥, 任晓东, 等. 如何利用 REDCap 实现临床试验的随机化操作 [J]. 中国循证心血管医学杂志, 2018, 10 (9): 1039 – 1042.
16 耿辉, 贺海蓉, 郑婕, 等. 多中心临床数据采集系统 REDCap 权限设置与管理 [J]. 中国循证心血管医学杂志, 2018, 10 (1): 11 – 13.
17 高凡, 田国祥, 贺海蓉, 等. REDCap 实现多中心研究数据管理的方法 [J]. 中国循证心血管医学杂志, 2018, 10 (4): 392 – 395.
18 耿辉, 田国祥, 贺海蓉, 等. 如何安全实现 REDCap 系统的数据录入、存储 [J]. 中国循证心血管医学杂志, 2018, 10 (3): 262 – 264.
19 Basch E, Deal AM, Kris MG, et al. Symptom Monitoring With Patient – reported Outcomes during Routine Cancer Treatment: a randomized controlled trial [J]. J Clin Oncol, 2016, 34 (6): 557 – 565.
20 Campion TR, Sholle ET, Davila MA. Generalizable Middleware to Support Use of REDCap Dynamic Data Pull for Integrating Clinical and Research Data [J]. AMIA Jt Summits Transl Sci Proc, 2017 (2017): 76 – 81.

(上接第 69 页)

参考文献

1 唐维红, 唐胜宏, 廖灿亮. 跨入 5G 时代的中国移动互联网—《中国移动互联网发展报告 (2020)》发布 [J]. 中国报业, 2020 (17): 32 – 35.
2 张鹏, 赵动员, 梅蕾. 移动社交网络信息传播研究述评与展望 [J]. 情报科学, 2020, 38 (2): 170 – 176.
3 赵加奎, 陈虹仪, 庄颖, 等. 三种呈现方式在新媒体健康传播中的效果评价 [J]. 健康教育与健康促进, 2020 (4): 411 – 413, 421.
4 张明, 刘嘉浩, 陈晨. 疫情防控期间新媒体传播的功能及特征研究 [J]. 新闻研究导刊, 2020, 11 (20): 106 – 107.
5 潘维华, 林静, 孙秀芸. 从新冠肺炎疫情看当下网络舆情应对新变革 [J]. 新闻研究导刊, 2020, 11 (19): 140 – 141.
6 邱雅, 王华彬. 网络舆情分析中智能爬虫的设计 [J]. 电脑知识与技术, 2011, 7 (33): 8301 – 8302.
7 袁畅, 郝志峰, 蔡瑞初, 等. 一种分布式的舆情分析系统架构 [J]. 电信科学, 2013 (7): 66 – 71.
8 刘斌. 基于 Twitter 大数据处理的境外舆情分析系统设计与实现 [J]. 电脑知识与技术, 2020, 16 (27): 30 – 33, 42.
9 习海旭, 蒋红芬, 程志凡, 等. 特定事件下网络舆情的情感分析与可视化方法 [J]. 情报理论与实践, 2020, 43 (9): 132 – 136, 143.
10 肖倩, 谢海涛, 刘平平. 一种融合 LDA 与 CNN 的社交媒体中热点舆情识别方法 [J]. 情报科学, 2019, 37 (11): 27 – 33.
11 张连峰, 周红磊, 王丹, 等. 基于超网络理论的微博舆情关键节点挖掘 [J]. 情报学报, 2019, 38 (12): 1286 – 1296.
12 彭程, 祁凯, 黎冰雪. 基于 SIR – EGM 模型的复杂网络舆情传播与预警机制研究 [J]. 情报科学, 2020, 38 (3): 145 – 153.
13 周琦萍, 杨芳. 基于 SIS 模型的网络舆情无监督预警机制研究 [J]. 情报科学, 2019, 37 (8): 51 – 55.
14 谌志华. 基于大数据的网络舆情分析系统 [J]. 现代电子技术, 2017, 40 (24): 15 – 17.