

小数据导向的医院优势学科信息资源建设研究^{*}

赖 昕 范美玉

(广州中医药大学第二附属医院 广州 510120)

〔摘要〕 介绍小数据发展背景、概念、研究现状,分析小数据在医院优势学科信息资源建设中的意义,阐述优势学科小数据的分类、获取、管理、分析方法以及小数据导向的优势学科信息资源建设策略。

〔关键词〕 小数据;科学数据管理;学科建设;信息资源建设

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2022.09.017

Study on Information Resource Construction of Hospital Preponderant Disciplines Oriented by Small Data LAI Xin, FAN Meiyu, The Second Affiliated Hospital of Guangzhou University of Traditional Chinese Medicine, Guangzhou 510120, China

〔Abstract〕 The paper introduces the development background, concept and research status of small data, analyzes the significance of small data in the information resource construction of hospital preponderant disciplines, and expounds the classification, acquisition, management and analysis methods of small data of preponderant disciplines and the construction strategy of small data-oriented information resources of preponderant disciplines.

〔Keywords〕 small data; scientific data management; discipline construction; information resource construction

1 引言

近年来国家、地方一直大力支持中医药学科建设和中医药信息化建设,发布《国家中医药管理局关于加强中医药重点学科建设的指导意见》《国务院关于印发中医药发展战略规划纲要(2016-2030年)的通知》等文件。广州中医药大学第二附属医

院是全国中医系统拥有重点专科/学科最多的医院之一^[1],也是广东省高水平医院建设“登峰计划”的首批重点医院,研究型医院的定位明确,学科建设的资源管理与利用需求日益增多。在数据科学时代,信息呈现爆炸式增长态势,碎片化信息越来越多,对传统的信息收集、处理、利用方法提出挑战^[2]。目前大数据技术应用存在数据价值密度低、数据欺骗性、机器学习陷阱等难点^[3-4]。近年来小数据思维^[5]逐渐引起国内图书情报领域广泛关注,出现科研用户小数据^[6]、读者小数据^[7]、智库用户小数据^[8]等方面研究,反映出科学数据的多面性。小数据与大数据既有相通也有差异,如前者以个体为对象,重点在于深度^[9],后者则侧重于某领域大范围、大规模的数据,重点在于广度。根据研

〔修回日期〕 2021-10-11

〔作者简介〕 赖昕,硕士,助理馆员,发表论文16篇;通信作者:范美玉,工程师。

〔基金项目〕 国家重点研发计划中医药现代化研究重点专项子课题“10个中医药优势病种临床与基础数据宝库构建”(项目编号:2019YFC1709802)。

究对象的差异有针对性地采用两种理论都可以获取数据规律和价值。因此针对学科用户特点,本文拟从小数据视角讨论医院优势学科信息资源建设,阐述小数据概念、分类、获取与集成管理以及基于小数据的学科信息资源建设策略。

2 小数据概念与研究现状

2.1 概念

小数据 (Small Data) 一词最早由 Deborah Estrin 于 2014 年提出^[5]。Deborah Estrin 及团队认为用户小数据是其日常行为活动的全部表征,通过对个体数据的全方位收集、监测、跟踪,能得到用户在不同时间段的个人健康状况信息,从而为个性化决策提供依据^[10],在精准医学、预测建模和多靶点整合医学干预等方面^[11]展现出潜在应用优势。

2.2 研究现状

国内小数据研究不多,主要集中在图书情报领域:小数据用于图书馆个性化、精准信息服务^[12-13];小数据用于档案用户挖掘、档案资源利用^[14-15];科研小数据融合研究^[16]等。针对小数据内涵,李立睿和邓仲华^[6]提出科研用户小数据是个体研究工作者在项目推进过程中全部行为和状态的数字记录集合,主要涵盖个体基本信息、行为记录、习惯偏好、情绪变化、性格特点、研究问题等;刁羽^[8]认为智库用户小数据是基于用户某个历史时间段利用个性化智库型信息服务过程中产生的思维活动、需求表达及利用行为等数据的集合;曹霞^[17]提出小数据是一种基于个人或单个团队的新兴数据,是有选择性、可靠、可控、增值的数字化信息,包括与分析对象有关的基本特征数据、行为模式数据、情景感知数据、社会关系数据等。

3 小数据在医院优势学科信息资源建设中的意义

3.1 新环境的因势利导

泛在信息环境下,信息的来源、载体、可获取

的渠道呈现多元化趋势。随着云计算、大数据、物联网等技术在各领域的应用,信息行为向现代化、多样化、数据化、智慧化方向发展。多维、全面、动态的小数据思维正与当前信息环境相适应。通过跟踪、记录、研究各角度、各时间段、各类型学科人员小数据,构建具有个性化特征的数据资源库和信息管理系统,可以解析、预测个体行为特征、关系网络、决策需求等并根据动态变化进行实时调整。

3.2 与大数据优势互补

学科小数据来源于学科本身和学科人员。数据量小、易采集,降低计算机设备与技术要求、减少投入成本,弥补了大数据投入成本高的缺点。同时学科小数据能够有效弥补大数据个性化、针对性弱的短板,使得决策更加精准。此外一般不同属性的小数据采集、处理和利用是在相对封闭环境中进行,能够有效避免大数据无限度地采集和监测个人信息,有效实现学科用户个体数据隐私保护。

4 优势学科小数据分类、获取、管理、分析

4.1 小数据分类与获取

4.1.1 学科人员线上线下信息行为小数据 借阅馆藏纸质书与期刊、访问馆藏电子数据库资源、馆际互借和文献传递、参加院内讲座和论坛等行为是学科人员个性化小数据的主要来源。获取这些数据的方法如下:通过图书馆信息管理系统导出数据;通过文献传递服务群,借助文本数据分析工具对聊天记录文件进行活跃人群、活跃时间段、用户科室分布、文献主题领域、文献传递完成情况等方面的分析;借助网络爬虫获取学科人员访问图书馆官方网站的痕迹数据,借助开源 Web 日志分析工具进行流量分析、离站链接数量分析、页面浏览次数分析、访问时间分析等;通过图书馆工作记录 and 数据库公司提供的资源使用情况,统计馆际互借和文献传递数据、参加院内讲座和论坛的人员数据等。

4.1.2 学科人员基本信息与需求小数据 基本信息数据主要包括年龄、性别、学历、专业、职称、

职务、研究方向、发表论文、出版专著、参与课题、申请专利、性格特点、学科建设主要分工、工作阶段性目标等。这些大数据适合采用封闭式（如职务）与开放式（如工作阶段性目标）问题结合的简单问卷调查收集。在数据获取时，根据问卷初稿抽取数名学科内各职称和岗位有代表性的人员进行预调查，如情况良好则进行下一步，如有问题将返回重新完善设计问卷内容。预调查完成后将在学科内进行正式问卷调查。学科人员网络或媒体信息行为搜寻偏好与习惯、学科建设工作中遇到的问题和心理状态、希望从学科服务人员处获得的参考咨询意见与服务、对学科建设工作中的困惑与意见等则是学科人员业务需求数据，适合通过半结构化访谈获取。

4.1.3 学科信息沟通与服务中产生的大数据 学科人员与其他工作人员进行信息互通时也会产生交互、情境大数据，如学科政策文件解读、学科建设阶段性任务布置、学科绩效指标、学科人员变动与分工调整、院内信息系统设置与平台使用方法、电子资源使用与投稿咨询、学科人员对信息系统与信息服务的感受与反馈，以及所有对学科人员提供的学科服务数据等。这部分数据可通过院内即时消息平台、企业微信、邮件等渠道获取，也可在学科建设工作中实时使用电子工具记录与保存。

4.1.4 学科人员自身产生的灰色数据 关注学科人员自身产生的未公开数据，包括未发表的论文、纸质手稿、科研推导数据、实验步骤数据、长尾数据、科研进展记录、项目阶段报表、病例分析、研讨笔记、会议记录、阅读笔记、研讨心得等。广州中医药大学第二附属医院已搭建机构知识库，向用户设置开放个人存储空间。用户具有上传数据、数据访问自控等权限。学科人员选择公开部分不涉及医院内部私密信息的灰色数据，可以被采集作为学科大数据。

4.2 大数据的管理

4.2.1 预处理 由于真实环境与活动较复杂、数据获取方式有限等原因，大数据存在数据噪声问题。因此必须先对数据进行预处理，减少数据噪

声，提升价值密度和可用性。主要方式有：（1）清洗。针对因填写不规范、隐私保护等产生的数据缺失、重复等情况，采取删除重复数据、补充缺失数据、去除异常数据等操作。（2）变换。由于原始数据来源不一、类型多样、采集方式不同，原始数据无法满足学科建设分析需求，需要采取变量派生、变量转换、数据标准化等方法加以转换。（3）规约。出于对一般数据处理标准、数据价值和现实资源的综合考量，在保证原有数据完整性与有效性的基础上，采用维规约、数量规约等方式以有效降低数据规模、精简数据量。（4）其他处理。其他必需的预处理操作。

4.2.2 利用与安全 数据利用可通过统计、分析、归类等方法实现：使用 Excel 或 EpiData 软件录入预处理的数据后导入 SPSS 软件进行统计与分析；按照数据类型和主要内容将整理后的统计结果以不同模块进行萃取、归类。此外由于采集的大数据中含有用户身份特征数据，涉及用户隐私，数据安全尤为重要。因此可以对数据库登录进行权限设置，对所存储数据进行安全保护，避免信息泄露。

5 大数据导向的优势学科信息资源建设策略

5.1 信息资源优化配置

信息资源建设理论认为信息资源系统功能的发挥取决于各种信息资源的质量与构成^[18]。因此医院优势学科建设的资源配置必须既关注质量也关注结构。通过大数据可从以下两方面优化资源配置：一是将学科人员行为与需求数据作为购买纸质文献的参考。采编馆员此前一直是从供应商提供的目录中直接选购纸质文献，文献整体利用率低，造成空间和资金浪费。可通过统计学科人员借阅行为，向学科人员征集荐购内容，整理学科人员平时关注、正在研究的主题及学科发展趋势获得大数据，并在此基础上采购最新、最前沿、最适应读者需求的文献，既践行了精准采购^[19]和读者决策采购^[20]，又节省了经费。二是网络时代数字资源不断普及，大众阅读习惯随之改变，应增加馆藏电子资源。根据学科人员对信息资源的需求比例和利用率

等数据分析,适当调整电子和纸质文献的比例,试用、购买部分学科小众电子资源。当然纸质图书在系统性、深阅读、文化传承等方面有不可比拟的重要性,尤其适应中医院保存大量古籍的需要。

5.2 信息资源个性推荐

根据学科人员小数据特征偏好开展个性化信息资源推荐,具体建议如下:一是根据学科人员的不同需求特点选择推荐内容。网络信息获取行为习惯:如为更倾向于使用手机浏览的用户推荐专业 APP 和公众号,为偏向于阅读纸质文献的用户推荐纸质新书或期刊。关注领域:如针对关注学科政策的用户推荐相关网站和评述文献,对专注临床研究的人员则推荐临床试验登记注册网站,对更关注学科前沿动态的人员推荐相关新闻动态和顶级期刊编译等。业务问题:如针对用户文献检索问题,推荐数据库检索讲座、教学文档、官方用户手册等;针对文章投稿问题可以推荐相关领域的期刊网站、征稿要求、同行投稿论坛和帖子等;针对数据分析问题可以推荐相关软件工具以及安装包、使用教程、参考范例等。二是系统地针对某个项目的进程或者个人需求将馆藏和网络资源进行筛选、整合、编辑,进行定期推送和动态更新。三是通过对学科人员个人小数据的挖掘与关联分析,建立个体兴趣预测与发现模型。

5.3 信息资源深度开发

深度开发信息资源是为医院优势学科建设提供精准知识服务的重要手段,也是小数据高价值密度和决策相关性的最大体现。基于学科人员小数据可以从以下几方面进行信息资源开发:一是提供嵌入式信息服务,全程跟踪具体项目或学科人员个人动态变化,对其信息需求进行实时检索、加工、传递。二是设置预测性信息服务,如挖掘学科前沿热点、绘制学科知识图谱、对相关学科文献进行编研、对未来学科政策进行趋势分析等。三是寻找学科建设可能的竞争对手或对标单位,对其各项情况进行检索、统计、分析进而形成竞争情报,辅助学科人员决策。四是积极开发院内机构成果库,完善模块设施、提

高界面友好度、增加应用功能等。在机构库建立优势学科门户模块,促进学科人员对本单位优势学科建设成果的了解。五是构建小数据资源库,部分已经深度开发的信息资源,经过加工整理成为具有可复用性的知识,由信息人员进行汇总、分类、存储后,学科人员随时查看和下载使用。

5.4 信息资源治理

学科用户小数据是用户在某个时间段内进行学科建设相关的思维活动、需求表达及行为方式等数据集合,呈现碎片化、个性化、多样化的特点^[8]。在小数据采集、存储、分析和利用的全生命周期中,如何保证数据的准确性、可获取性、安全性、适度共享和合规使用是高质量学科建设与服务的关键,因此有必要开展学科信息资源治理^[21]。这不仅能够帮助更有效地管理数据,而且能降低用户差异化服务成本,促进高质量数据的生成。标准化是数据有效管理和共享的前提和重要基础,贯穿数据全生命周期^[22],总体应借鉴国内外科学数据管理标准、建设标准及实践经验,尤其是卫生健康信息标准。在采集阶段根据应用场景的不同制定相应采集规则,包括采集的深度、广度、范围、清洗颗粒度、转换格式等,为数据挖掘奠定良好的基础。在存储阶段要降低数据噪声,如统一数据存储格式,保证数据质量,增强数据可用性。在分析阶段根据研究的差异化需求,确定挖掘的角度、层次、颗粒度等问题。学科小数据是以用户为核心的个人数据集合,个人隐私保护问题尤为重要^[23]。因此在学科小数据的全生命周期管理过程中,不仅要严格遵循《中华人民共和国个人信息保护法》《中华人民共和国数据安全法》等数据安全法律法规,而且要重视个人的知情同意。在具体实施中,应推动技术与管理并行,一方面要借助数据安全技术限定人员权限、防泄漏,另一方面提高用户的小数据知识产权保护意识,建立小数据管理与使用制度并适时更新,营造安全的数据全生命周期管理生态环境等。

6 结语

随着国家高校一流学科建设的深入推进和现代

医院高质量发展,越来越多的医院管理者已经认识到优势学科建设长远战略的作用。学科信息资源贯穿于学科建设和发展的全过程,是学科建设体系不可或缺的基础配置。基于小数据,通过了解学科人员的个性化行为与需求,图书馆能够掌握优势学科信息资源建设方向,从资源配置、推荐、开发等方面开展具体工作。目前关于学科小数据的研究不多,在实践中还有诸多问题,如需要哪些信息技术支撑,如何采集到更有利用价值的小数据,如何存储、分析和处理多层次的小数据,小数据隐私保护问题等。未来还需要展开进一步研究与工作,力求为“十四五”期间医院优势学科建设提供完善的资源保障,积极配合高校一流学科建设步伐,推动实现医院高水平发展。

参考文献

- 1 广东省中医院. 南粤杏林第一家——广东省中医院 [EB/OL]. [2021-10-11]. <http://www.gdhtcm.com/sitecn/yyjs/index.html>.
- 2 许茹. 知识管理视阈下“大数据”与“小数据”的比较分析 [J]. 办公室业务, 2019 (15): 63-64.
- 3 张铭. 大数据时代图书馆服务面临的机遇及应对措施 [J]. 图书馆工作与研究, 2019 (S1): 86-89.
- 4 姜会珍, 马琰, 朱卫国. 医疗大数据的“欺骗性”及其对策 [J]. 协和医学杂志, 2020, 11 (5): 542-546.
- 5 Estrin D. Small Data, Where N = Me [J]. Communications of the Association For Computing Machinery, 2014, 57 (4): 32-34.
- 6 李立睿, 邓仲华. 面向科研用户小数据的微知识服务研究 (上) [J]. 图书与情报, 2016 (3): 49-54.
- 7 陈臣. 基于小数据的读者兴趣发现与动态更新 [J]. 图书馆, 2017 (3): 23-27, 80.

- 8 刁羽. 小数据视角下高校图书馆个性化智库型信息咨询研究 [J]. 图书馆理论与实践, 2020 (2): 90-93.
- 9 刘朝, 马超群. 大数据与小数据深度融合的价值与路径 [J]. 人民论坛, 2021 (Z1): 30-33.
- 10 Estrin D, Thompson C W. Internet of You: Data Big and Small [J]. IEEE Internet Computing, 2015, 19 (6): 8-10.
- 11 Hiroko H D, Estrin D. Making Sense of Aging with Data Big and Small [J]. The Bridge, 2019, 49 (1): 39-46.
- 12 陈臣. 基于小数据决策支持的图书馆个性化服务 [J]. 图书与情报, 2015 (1): 82-86.
- 13 孙丹霞, 王伟军, 姜毅. 基于用户小数据的嵌入式学科服务研究 [J]. 图书馆工作与研究, 2019 (4): 84-90.
- 14 徐晨晨. 基于小数据的档案馆潜在用户信息推荐服务研究 [J]. 档案天地, 2021 (3): 30-33, 9.
- 15 吕元智. 基于小数据的数字档案资源知识集成服务研究 [J]. 档案学通讯, 2016 (6): 47-51.
- 16 李立睿, 邓仲华. “互联网+”背景下科研用户的小数据融合研究 [J]. 图书情报工作, 2016, 60 (60): 58-63.
- 17 曹霞. 国内图书馆界小数据研究进展 [J]. 图书情报工作, 2020, 64 (14): 145-151.
- 18 肖希明. 信息资源建设 [M]. 武汉: 武汉大学出版社, 2008: 39.
- 19 宋旅黄. 图书馆与馆配商合作效益研究 [D]. 武汉: 武汉大学, 2017.
- 20 肖希明, 尹彦力. 服务于“双一流”建设的高校图书馆信息资源建设 [J]. 图书馆建设, 2018 (4): 79-84.
- 21 张宁, 袁勤俭. 数据治理研究述评 [J]. 情报杂志, 2017, 36 (5): 129-134, 163.
- 22 李赞梅, 钱庆, 李姣, 等. 健康医疗科学数据共享标准体系框架构建 [J]. 医学信息学杂志, 2018, 39 (11): 49-53.
- 23 王敏. 小数据思维驱动下的公共图书馆信息服务模式研究 [J]. 图书馆学刊, 2015, 37 (12): 77-79.

欢迎订阅

欢迎赐稿