

医疗健康知识图谱需求与技术构建研究

蒋君 肖宇锋 门佩璇 黄雅兰 张 玟

(中国医学科学院/北京协和医学院医学信息研究所/图书馆 北京 100005)

[摘要] **目的/意义** 调查不同群体对医疗健康知识图谱的需求, 探讨知识图谱研究现状和不足, 为医疗健康知识图谱的构建提出建议。**方法/过程** 基于文献调查、问卷调查、专家访谈等方式, 从数据来源、关键技术、应用场景、待改进之处等方面分析医疗健康知识图谱。**结果/结论** 提出以国家政策为导向、以权威知识为依托, 支持用户全程参与, 注重个人隐私和知识产权保护, 构建多场景、可视化、多语种、多模态的医疗健康知识图谱和易于知识传播的标准体系及底层数据规范制度。

[关键词] 知识图谱; 医疗健康; 需求调研; 人工智能

[中图分类号] R-058 **[文献标识码]** A **[DOI]** 10.3969/j.issn.1673-6036.2024.03.007

Study on the Demands and Technical Construction of Medical Health Knowledge Graph

JIANG Jun, XIAO Yufeng, MEN Peixuan, HUANG Yalan, ZHANG Bin

Library & Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, China

[Abstract] **Purpose/Significance** The paper investigates the needs of different groups for the medical health knowledge graph, discusses the current situation and shortcomings of the research on the knowledge graph, and puts forward suggestions for the construction of the medical health knowledge graph. **Method/Process** Based on literature survey, questionnaire survey and expert interview, the medical health knowledge graph is analyzed from the aspects of data sources, key technologies, application scenarios and things to be improved. **Result/Conclusion** It is proposed to take national policies as guidance, rely on authoritative knowledge, support users' full participation, pay attention to personal privacy and intellectual property protection, and build a multi-scene, visual, multilingual, multimodal medical health knowledge graph, a standard system and underlying data specification system that are easy to disseminate knowledge.

[Keywords] knowledge graph; medical health; demand research; artificial intelligence

1 引言

知识图谱是基于图形的知识表述和组织方法, 用“实体-关系-属性”三元组形式描述物理世界中的概念及其相互关系, 构成网状知识结

构^[1-3]。知识图谱的信息来自于专家、知识库、网站、电子病历、教科书等, 通过实体识别^[4]、信息抽取^[5]、语义分析^[6]、关系抽取^[7]和融合^[8]等自然语言处理技术, 采用自顶向下或自底向上的方式构建, 实体与实体之间被赋予双向或者单向的连接关系。知识图谱从初期的工业化应用领域扩展至社会网络分析^[9]、推荐系统^[10]、问答系统^[11]等领域, 其中医疗健康领域已成为知识图谱重要的应用方向, 为医生提供影像判别和辅助诊断, 为普通公众提供医疗健康领域的权威知识,

[修回日期] 2023-11-20

[作者简介] 蒋君, 馆员, 发表论文 13 篇; 通信作者: 张玟, 副研究馆员。

为医保基金和诊疗行为提供监管支持，为下游医疗应用提供有力支撑等^[1]。

既往医疗健康知识图谱研究集中在内容遴选、关键技术实现等，但缺乏对用户需求等系统性的调研分析，对在卫生健康领域的应用阐述不够深入。本研究通过文献调研、问卷调研和专家访谈等方法，面向医疗管理机构、医院和医学院校、保险公司、药企、高科技企业和出版社，以及普通公众，从医疗健康知识图谱的研究现状、数据来源、关键技术、应用场景和存在的不足等方面进行调研分析，提出相应建议。

2 对象与方法

2.1 文献调研

为了探讨国内外科研人员有关医疗知识图谱领域的研究现状、进展和热点等，从文献角度总结和分析医疗知识图谱领域相关研究成果，利用 Web of Science 核心合集的 SCIE 数据库，检索近 10 年与医疗健康知识图谱相关的文献，检索日期为 2022 年 10 月 20 日，共检索到 3 075 篇文献，见图 1。

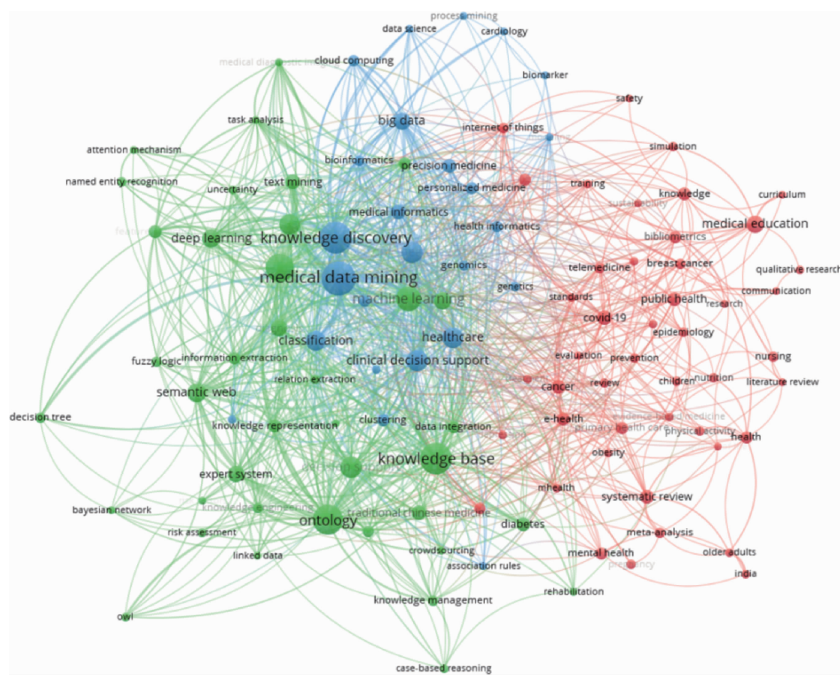


图 1 医疗健康知识图谱近 10 年研究热点

通过聚类发现，国内外研究重点主要集中在利用现有知识图谱软件对疾病进行文献计量和可视化分析（红色部分），知识图谱的算法和相关技术（绿色部分），知识图谱用于基因组学、遗传学和临床决策支持等前沿领域（蓝色部分）。

2.2 问卷调研

为了解普通公众对医疗健康知识图谱的认知程度和需求建议，采用在线问卷的形式进行不记名抽样调查。共回收有效问卷 110 份，调研对象的职业、学历、地域、年龄等覆盖广泛。被调研者遍及

15 个省、自治区、直辖市，主要分布在北京市、内蒙古自治区和四川省，以及东部地区，共约占全国被调研者的 50%；年龄主要集中在 35 岁以下，约占 75%；学历以本科、硕士和博士为主，占 80% 以上；以女性用户居多，大约 67%。被调研者主要工作在高等院校、医院和科研院所，约占 50% 以上，还有医药企业、保险公司、政府机关以及企业、银行和退休人员等。

2.3 专家访谈

在文献调研和问卷调研基础上，访谈医疗相关

领域专家, 根据不同行业特点制定不同的访谈提纲。共访谈来自医疗管理机构、医院和医学院校、保险公司、药企、高科技企业和出版社等不同类型机构的 17 名专家, 其结合自身工作, 发表对医疗健康知识图谱的意见和看法。

3 调研结果

3.1 国内外典型的医疗健康知识图谱

按照概念定义和应用程度将医疗健康知识图谱分为两种类型。一是构建医疗行业的底层概念。如美国国家医学图书馆建立的统一医学语言系统, 是生物医学领域最大的本体知识库, 整合了 218 个来源词表, 含有 54 种语义关系、133 种语义类型、440 万个概念、1 600 万个术语, 涉及 25 种语言^[13]; 医学系统命名法 - 临床术语 (systematized nomenclature of medicine, clinical terms, SNOMED CT) 含有临床相关的 34 万个概念, 并且与《国际疾病分类》第 10 版 (International Classification of Diseases, Version - 10, ICD - 10)、观测指标标识符逻辑命名与编码系统 (logical observation identifiers names and codes, LOINC) 等建立了概念之间的映射关系^[14-15]; 2018 年世界卫生组织发布《国际疾病分类》第 11 版 (ICD - 11), 含编码条目 55 000 余条, 各条目定义清晰, 作为疾病诊断、死因统计等参照编码^[16]; 中国医学科学院构建了中文一体化医学语言系统 (Chinese unified medical language system, CUMLS), 整合了 10 余部医学词表, 收录医学主题词 3 万余条、医学术语 10 万余条、医学词汇素材 30 余万条^[17]; 开放医疗与健康联盟构建了“七巧板”医学术语集, 已发布 102 万个概念、133 万个术语、300 万个关系、207 万个映射扩展集^[18]。二是医疗健康知识图谱的应用程序, 其研发机构主要分为两类。一类是互联网医疗健康企业, 其在知识图谱应用程序开发领域扮演重要角色, 通过大数据、人工智能等技术手段, 构建庞大的医学知识库, 为医生和患者提供便捷的信息查询和决策支持。例如, 腾讯医典^[19]、京东健康^[20]、阿里健康^[21]、百度灵医智惠等^[22-26], 通过整合各类医疗

资源, 为用户提供在线问诊、药品查询、健康管理等一站式服务。另一类是医院和医学院校, 其根据自身的特点和需求自主或合作研发知识图谱应用程序, 侧重于满足医院内部的管理和运营需求, 如电子病历系统、医院信息系统等。通过构建知识图谱, 医院可以更好地整合和管理患者信息、医疗资源、医疗流程等数据, 提高医疗服务的协同性和效率。此外, 知识图谱还有助于医院开展临床研究和健康素养培训等工作, 推动医学科学的进步和发展。

3.2 医疗健康知识图谱的数据来源

医疗健康知识图谱的底层数据来源于多种渠道, 既包括诊疗指南、临床路径、医学专业教材、医学百科全书、药品说明书等专业化资源, 又包括电子病历、可穿戴设备获取的健康信息和网络医学信息等个性化资源。专业化资源经过同行评议或审核, 内容严谨、完整且权威, 但是信息相对滞后, 语言过于专业, 不易被普通公众理解, 且大多是非结构化的信息。个性化资源如电子病历, 包含病案首页、病程记录、检查结果、医嘱等重要信息, 准确且全面, 但是数据获取相对较难; 可穿戴设备具有心率、睡眠、血压、血氧、血糖等监测功能, 是持续性的监测过程, 但只是一些基础信息, 不能提供完整的诊断结果; 网络医学信息更新速度快、方便获取扩充知识, 但是信息真假难辨、良莠不齐, 缺乏权威性和可信度, 没有严格的审查机制。因此, 在构建医疗健康知识图谱时要考虑不同来源的数据, 利用其优势, 整合出新的知识图谱供用户参考。

3.3 知识图谱构建相关技术

知识图谱构建基本流程包括知识表示、知识抽取、知识映射、知识整合和知识推理, 涉及机器学习、自然语言处理、人机交互、图像识别、分布式存储等关键技术。知识图谱的发展演变过程主要分为以下几个阶段: 1960 年语义网络作为知识表示的一种方法被提出, 用于自然语言理解领域; 1980 年“本体”被引入人工智能领域, 用来刻画知识;

1989 年万维网 Web1.0 实现了文本间的链接；1998 年语义网被提出，实现从超文本链接到语义链接；2006 年 Tim Berners - Lee 定义了在互联网上链接数据的 4 条原则，建立开放数据之间的链接；2012 年谷歌发布了基于知识图谱的搜索引擎，用于搜索、问答、决策、智能推理等方面^[27]。未来，将区块链、ChatGPT 等新技术融入知识图谱，去中心化实现数据的分布式记录、存储和更新，可为用户提供更好的体验^[28]。

3.4 医疗健康知识图谱的应用场景和功能

医疗健康知识图谱可以为不同用户提供个性化功能。对于普通公众，可以提供医疗健康类科普知识、用药咨询、化验报告或体检报告解读、饮食建议等，通过可穿戴设备监测分析健康数据；对于医疗监管部门，可以对比不同医院医生之间的诊疗行为差异，为医保基金管理和诊疗行为监管提供支持，进行医疗舆情监控和公共卫生事件预警，帮助发现潜在的公共威胁，从源头上预防和降低舆情风险；对于医院，可以用于患者教育，帮助医生进行辅助诊疗、分诊导诊、影像识别、临床干预提醒、病历处方质控、智能随访、科研决策支持等；对于保险类企业，可以用于商保理赔审核中的商保控费、投保时的保险核保；对于医药企业，可以辅助药物研发，通过知识聚类提出新的假说，加速药物研发的过程，降低研发成本；对于互联网医疗企业，可以提供智能问诊、用药咨询等服务，给用户更好的体验，增加用户粘性；对于医学信息工作者，可以快速了解领域全貌，通过可视化分析定制个性化医疗健康知识图谱。

3.5 现有医疗健康知识图谱待改进之处

通过调研发现人们对现有医疗健康知识图谱的满足率不高，31% 的被调研者认为满足率为 20% ~ 50%，28% 的被调研者认为满足率可以达到 51% ~ 70%。不能满足的原因主要是无法深入了解相关信息（58%）、信息不权威（55%）、信息不准确（38%）、信息更新不及时（25%）、大量投放广告（20%）、检索功能不好用（20%）、网站访问较慢

或出错（8%）等，影响用户体验。

此外，个人隐私和知识产权保护方面也有待提高，欧洲颁布的《通用数据保护条例》（General Data Protection Regulation, GDPR）、美国的《健康保险携带和责任法案》（Health Insurance Portability and Accountability Act, HIPAA），中国的《个人信息保护法》等规定了个人数据的使用目的、用户同意、数据安全等方面的要求，应严格遵守。知识图谱技术也存在不足，文本提取较为困难，算法精度低、约束多、可扩展性差、实体对应不准确等。总体来看，医疗健康知识图谱底层数据质量和可信度有待改进，数据的规范性、知识的权威性、更新的及时性、界面的易操作性、人工智能决策过程的透明性、用户个人隐私和知识产权保护等方面，都是当前存在的主要问题。

4 讨论与建议

4.1 医疗健康知识图谱的信息来源应真实权威并及时更新

权威的医疗健康知识是知识图谱构建的关键。网络信息泛滥，甚至有些信息相互矛盾，在海量信息中找到正确权威的知识是用户目前最迫切的需求。尤其是在涉及生命健康的医学领域，这种需求更加强烈，所以知识来源的权威性和专业性成为关注重点，基于医学领域权威书籍、由专业人员构建的知识库将会获得广泛认可。任何事物都在发展变化，医学领域的发展更为迅速，为了使用户了解准确信息，知识库要及时更新数据，保证准确性和有效性。

4.2 医疗健康知识图谱构建过程需要用户全程参与

知识图谱从设计、研发到使用都与用户有关，因此整个过程都需要用户参与。例如，设计时开展调研了解用户需求，以解决用户痛点；研发中邀请用户参与测试，体验图谱的设计逻辑和功能模块，提出改进意见。文化、语言、社会环境、工作背景和教育体系等因素会影响个人健康素养，也会较大程度上决定知识获取和使用，所以要考虑不同类型

的人群特征。对于有医学背景的人群，倾向于提供专业术语和更深入的医学知识，发挥辅助决策作用；对于没有医学背景的人群，更倾向于提供通俗易懂的知识，发挥健康科普作用。

4.3 信息获取避免侵犯个人隐私和知识产权

知识图谱在获取信息时，可能存在侵犯用户个人隐私、泄露商业私密、侵犯他人知识产权等违法行为，因此在构建过程中要考虑数据来源的合法性，防止侵权等行为发生^[29]。收集个人信息时要合理合法，对收集到的信息严格保密，未经用户同意不得擅自泄露和篡改，并采取技术手段和其他必要措施防止信息泄露、毁损和丢失。对于网络上的公共资源，要遵守相关的法律法规和版权规定，避免侵犯他人知识产权。获取数据时必须征得著作权人许可，并签订著作权许可使用合同，如不能确定著作权人，尽量不使用，如必须使用应注明出处。

4.4 构建多语种多模态的医疗健康知识图谱

随着全球化的发展，单一语种已不能满足需求，多语种的医疗知识图谱可以实现不同国家医学知识之间的相互沟通和交流，更有利于医学发展。现有医学知识图谱大多以文本和图片形式存储，但在医学临床中还存在大量的医疗影像和物联网数据，以及来自图书、期刊、网页、视频等的多模态信息，因此构建大规模、多语种、多模态的医学知识库是未来医疗知识图谱的研究热点之一。将真实世界数据与虚拟世界融合，通过触摸、移动、声音等方式，实现从宏观到微观、沉浸式、多维度、多感官的呈现，方便各年龄段患者接收医学信息，增加知识的可及性。

4.5 建立知识图谱领域标准体系

建立知识图谱技术相关标准规范，以国家为主导建立底层数据规范，便于不同知识图谱之间互通整合和知识传播。在知识图谱构建方面，从本体设计、知识抽取、知识映射、知识整合、知识存储、知识计算等维度进行审核，并建立相关审核标准；在知识图谱应用方面，考察其安全性、可靠性、响

应性、可移植性和易用性等，并建立相关检验标准；对于底层数据，紧随国家数据局推进数据基础制度建设，统筹数据资源整合共享和开发利用，使知识图谱领域更加规范和有序。

5 结语

知识图谱既是对语言的理解，又是对知识的展示；既是基于人脑神经网络的模拟，又是基于概念思维情感的理解。本文以“医疗健康知识的需求”为视角，通过文献调研、问卷调研和专家访谈，探讨医疗健康知识的获取渠道、应用场景和相应的技术发展状况，并分析现有图谱待改进之处，提出以国家政策为导向、以权威知识为依托，支持用户全程参与，注重个人隐私和知识产权保护，构建多场景、可视化、多语种、多模态的医疗知识图谱，建立便于知识传播的标准体系和底层数据规范制度。未来，知识图谱还将与区块链、ChatGPT 等新技术相结合，推动医疗数据的自动化处理和智能化使用，为医疗行业带来新的发展机遇，为医学知识获取和使用带来便利。

利益声明：所有作者均声明不存在利益冲突。

参考文献

- 1 LI L, WANG P, YAN J, et al. Real - world data medical knowledge graph: construction and applications [EB/OL]. [2023 - 11 - 20]. <https://www.sciencedirect.com/science/article/pii/S0933365719309546?via%3Dihub>.
- 2 NICHOLSON D N, GREENE C S. Constructing knowledge graphs and their biomedical applications [EB/OL]. [2023 - 11 - 20]. <https://www.sciencedirect.com/science/article/pii/S2001037020302804?via%3Dihub>.
- 3 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述 [J]. 计算机研究与发展, 2016, 53 (3): 582 - 600.
- 4 LUO L, YANG Z H, YANG P, et al. An attention - based BiLSTM - CRF approach to document - level chemical named entity recognition [J]. Bioinformatics, 2018, 34 (8): 1381 - 1388.
- 5 QU J. A review on the application of knowledge graph tech-

- nology in the medical field [EB/OL]. [2023 - 11 - 20]. <https://www.hindawi.com/journals/sp/2022/3212370/>.
- 6 ZHU G G, IGLESIAS C A. Exploiting semantic similarity for named entity disambiguation in knowledge graphs [EB/OL]. [2023 - 11 - 20]. <https://www.sciencedirect.com/science/article/pii/S0957417418300897>.
 - 7 ABU - SALIH B, AL - QURISHI M, ALWESHAN M, et al. Healthcare knowledge graph construction: a systematic review of the state - of - the - art, open issues, and opportunities [J]. *Journal of big data*, 2023, 10 (1): 81.
 - 8 ALOBAIDI M, MALIK K M, HUSSAIN M. Automated ontology generation framework powered by linked biomedical ontologies for disease - drug domain [EB/OL]. [2023 - 11 - 20] <https://www.sciencedirect.com/science/article/pii/S0169260717315791?via%3Dihub>.
 - 9 NICKEL M, MURPHY K, TRESP V, et al. A review of relational machine learning for knowledge graphs [J]. *Proceedings of the IEEE*, 2016, 104 (1): 11 - 33.
 - 10 ZHANG F Z, YUAN J, LIAN D F, et al. Collaborative knowledge base embedding for recommender systems [C]. San Francisco: 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2016.
 - 11 WANG X, HE X, CAO Y, et al. Knowledge graph attention network for recommendation [C]. New York: 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019.
 - 12 张元博. 医疗知识图谱构建与应用 [D]. 哈尔滨: 哈尔滨工业大学, 2018.
 - 13 JING X. The unified medical language system at 30 years and how it is used and published: systematic review and content analysis [J]. *JMIR medical informatics*, 2021, 9 (8): e20675.
 - 14 SNOMED International. SNOMED CT Starter Guide [EB/OL]. [2023 - 08 - 20]. <https://confluence.ihtsdotools.org/display/DOCSTART/1.+Introduction>.
 - 15 WANG K C. Standard lexicons, coding systems and ontologies for interoperability and semantic computation in imaging [J]. *Journal of digital imaging*, 2018, 31 (3): 353 - 360.
 - 16 世界卫生组织. ICD - 11 国际疾病分类第十一次修订本 [EB/OL]. [2023 - 08 - 20]. <https://icd.who.int/zh>.
 - 17 李丹亚, 胡铁军, 李军莲, 等. 中文一体化医学语言系统的构建与应用 [J]. *情报杂志*, 2011, 30 (2): 147 - 151.
 - 18 数字医学知识中心 - 术语组. “七巧板”医学术语集资源 [EB/OL]. [2023 - 08 - 20]. <http://wiki.omaha.org.cn/pages/viewpage.action?pageId=42336639>.
 - 19 环球网. 腾讯医疗副总裁黄磊: 从体系化知识图谱, 切入诊疗全流程 [EB/OL]. [2023 - 11 - 20]. <https://tech.huanqiu.com/article/3zVDJRTImkd>.
 - 20 环球网. 京东健康携手北医三院开发“PharmCoo”AI 药学知识图谱 [EB/OL]. [2023 - 11 - 20]. <https://baijiahao.baidu.com/s?id=1639674910368247005&wfr=spider&for=pc>.
 - 21 中国财经. 阿里健康启动医学知识图谱 承诺不带有商业化目的 [EB/OL]. [2023 - 11 - 20]. <https://finance.ifeng.com/c/7clVcDaq2nn>.
 - 22 灵医智慧. 医疗大数据解决方案 [EB/OL]. [2023 - 11 - 20]. <https://01.baidu.com/bigdata.html>.
 - 23 健康界. 知识图谱助力疫情防控 | 平安智慧医疗战“疫”背后的智慧“大脑” [EB/OL]. [2023 - 11 - 20]. <https://www.cn-healthcare.com/articlewm/20200414/content-1103602.html?appfrom=jkj>.
 - 24 央广网. 水滴上线 CONF 医疗知识图谱 构建健康保障数据中台 [EB/OL]. [2023 - 11 - 20]. <https://baijiahao.baidu.com/s?id=1681585105275348749&wfr=spider&for=pc>.
 - 25 知乎. 付子玉: 丁香园医疗领域图谱的构建与应用 [EB/OL]. [2023 - 11 - 20]. https://zhuanlan.zhihu.com/p/531133328?utm_id=0.
 - 26 左手医生. 智能自诊 [EB/OL]. [2023 - 11 - 20] https://open.zuoshouyisheng.com/self_diagnos?zy_channels=baidupc&bd_vid=11187117681277486036.
 - 27 王昊奋, 漆桂林, 陈华钧. 知识图谱方法、实践与应用 [M]. 北京: 电子工业出版社, 2019.
 - 28 朱光辉, 王喜文. ChatGPT 的运行模式、关键技术及未来图景 [J]. *新疆师范大学学报 (哲学社会科学版)*, 2023, 44 (4): 113 - 122.
 - 29 邓建鹏, 朱怿成. ChatGPT 模型的法律风险及应对之策 [J]. *新疆师范大学学报 (哲学社会科学版)*, 2023, 44 (5): 91 - 101, 2.