http://www.biomedrxiv.org.cn/article/doi/bmr.202110.00001
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

面向医学教育的可循证医学图书图谱化表示探讨

孙月萍[1]，董良广[2]，徐晓巍[1]，李露琪[1]，康宏宇[1]，李姣[1]，侯丽[1]*

1. 中国医学科学院医学信息研究所/北京协和医学院图书馆，北京 100020;

2. 人民卫生出版社有限公司，北京

【摘要】以人民卫生出版社的结核病相关电子资源中的数据为研究对象，设计一套标准化的医学图书图谱化表示模型，从半结构化XML数据中抽取实体、关系和属性，构建分层次的医学图书知识图谱，探讨可循证的医学教育知识图谱构建方法，以期为国内医学类电子图书的图谱化表示工作提供参考，为医学教育模式创新提供支撑。

【关键词】医学图书图谱; 可循证; 医学教育模式创新

Analysis of evidence-based knowledge graph construction of medical books for medical education

Sun Yueping[1], Dong Liangguang[2], Xu Xiaowei[1], Li Luqi[1], Kang Hongyu[1], Li Jiao[1], Hou Li[1]*

1. 1Institute of Medical Information, Chinese Academy of Medical Sciences/Peking Union Medical College, Beijing 100020, China

2. People's Medical Publishing House Co., Ltd, Beijing 10021, China

【Abstract】Take tuberculosis related medical books published by People's Medical Publishing House as resource, this study designed a standard schema for medical book knowledge graph, extracted entities, relations and properties from semi-structured XML data, and constructed a hierarchy medical book knowledge graph. The evidence-based graph construction method was investigated to provide hints for domestic medical book graph construction work and to provide supports for medical education innovation.

【Key words】Medical book knowledge graph; Evidence-based; Medical education mode innovation

1 医学知识图谱助力医学教育模式创新

1.1 医学图书的知识呈现方式有待改进

互联网+背景下，医学教育要在教学形式、知识获取途径、更新、实时获得、虚拟训练和智能化、网络化管理等方面进行改革创新。在医学教育实践中，一方面，医学课程大多概念多而抽象,知识点之间联系复杂,由于内容快速更新导致其理论架构越来越复杂，而另一方面国内院校多数仍单纯采取传统授课方式,即根据教学大纲要求并依赖教科书进行常规授课[1]。随着医学数字出版的发展，以Lippincott Williams & Wilkins（LWW）[2]、爱思唯尔（Elsevier）[3]、人民卫生出版社[4]为代表的专业医学出版社均开展了医学图书的电子化工作，并联合信息检索平台OVID平台等开展基于图书的信息检索服务，但其研究重点侧重于电子图书的结构化处理以及信息处理，医学电子图书的表示和应用大多在信息层，缺乏对医学电子图书的知识表示、加工及应用研究，医学电子图书的结构化呈现对医学教育模式创新的促进作用有限。因此，应用知识组织新技术进一步改变医学教科书的知识组织形式，从数据中提炼医学知识，并用一种用户友好的方式呈现知识点之间的复杂联系并支持可循证的推理工作，辅助教师、学生合理高效地对其进行管理和共享，对当今的医学教育模式创新有着重要的应用价值。

1.2 医学知识图谱构建及应用进展

知识图谱是以图的形式表现客观世界中的概念和实体及其之间关系的知识库[5]，作为语义搜索、智能问答、决策支持等智能服务的基础技术，近几年医学知识图谱技术发展迅速，在通用领域、各垂直领域得到了广泛应用。例如，医疗信息搜索引擎（Google，搜狗名医）、开放式医疗问答系统（IBM的Watson系统[6]、微软的小冰[7]）、医疗决策支持系统（iOSC3[8]，iPharma用药助手）等均应用了医学知识图谱技术。另有研究基于知识图谱和数据分析技术的科学出版物管理模型来提高医学科技论文检索的效率 [9]。已有的基于知识图谱的应用，受

bmr.202110.00001V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202110.00001
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

限于医学知识图谱的知识数量和质量，构建完备、可靠的医学知识图谱是关键[10]。在诸多医学电子资源中，医学电子图书由于其特有的专业性和权威性，被广泛认为是一种兼顾完备性和可靠性的优质电子资源。同时，受益于医学图书的电子化，已有整合包括医学电子书在内资源的医学知识图谱构建工作[11]， Elsevier 为医学决策应用构建了医疗健康知识图谱，数据来源包括医学书籍、临床试验数据、药品数据库等[12]，但不支持面向医学教育的医学知识表示及知识溯源的应用需求。同时，已有部分应用于教育的知识图谱构建及应用工作[13,14,15]，知识图谱的构建来源以课程标准、教科书、维基百科等为主，但尚未发现为医学教育应用服务的基于中文医学图书的医学图谱。因此，本研究将以人卫出版社半结构化电子书为数据基础，构建基于医学电子资源的知识图谱，以期为医学教育模式创新提供支撑。

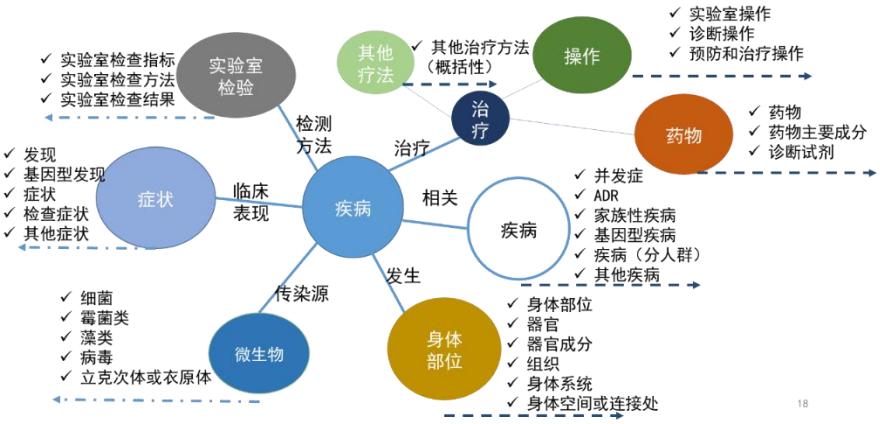
2 可循证的医学图书知识图谱建设

区别于传统的科学出版模型，本研究构建了一个可循证的医学图书知识图谱模型，以医学图书知识图谱为底层支撑，在医学图书知识图谱的基础上，进一步通过数据分析和挖掘搭建应用，为顶层用户（教师、学生、公众）提供服务，例如医学可视化分析、医学知识搜索、构建医学题库、医学知识问答、医学知识推荐、医学辅助决策等，为教育模式创新提供支持。

2.1 医学图书知识图谱数据模型设计

通过对一系列结核病相关主题电子书的内容分析，围绕结核病诊断、防治中的关键要素开展顶层框架模型设计。医学图书知识图谱Schema由实体、概念以及实体之间的关系构成。本研究设计了9类实体，并对每一类实体涵盖的范畴参考UMLS进行界定，具体包括：疾病、症状、身体部位、微生物、实验室检验、药物、操作、其他疗法（非药物治疗手段）。其中疾病实体包括疾病、疾病副作用、家族性疾病、基因型疾病、疾病（分人群）和其他疾病，症状实体的范畴界定为：发现、基因型发现、症状、检查症状以及其他症状；微生物包括细菌、霉菌类、藻类、病毒以及立克次体或衣原体；药物包括药物、药物主要成分以及诊断试剂。此外，本研究共并定义了8类实体之间的9种语义关系，具体包括：疾病-相关-疾病、疾病-发生-身体部位、疾病-传染源-微生物、疾病-临床表现-症状、疾病-检查方法-实验室检验、疾病-治疗方法-药物、疾病-治疗方法-操作、疾病-治疗方法-其他疗法。该知识图谱数据模型Schema如图1所示：

图1 医学图书知识图谱数据模型



2.2 医学图书知识图谱构建流程

bmr.202110.00001V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202110.00001
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

由于医学图书术语、数据相对比较清晰，本研究采用自顶向下的方式构建知识图谱（如图2所示），包括数据解析、数据模型构建、知识抽取、实体对齐、质量评估、知识图谱应用六个步骤。第一步，解析医学电子图书XML格式源数据，从医学图书元数据中抽取关键元数据信息，包括书名、别名、图书版本、国际标准书号、出版商、出版时间、作者等，构建元数据层schema。同时，将医学图书分解为章节，其中，每个章节保留章节序列号（供溯源使用）、章节标题以及章节内容，供构建医学图书知识图谱使用。第二步，根据已有医学知识，结合医学图书具体内容，构建数据模型，即医学图书知识图谱数据组织框架。第三步，运用语义处理技术，自动从半结构化XML数据中抽取实体、关系和属性，实体和实体关系如上节所示，其中实体抽取采用基于正向最大匹配加检索树的方法，实体关系抽取采用基于共现的方法，如果两个实体在同一章节中共现，则认为二者存在对应的语义关系。第四步，实体对齐，对实体进行实体消歧和共指消解，为不同实体类型定义重要性级别，着重处理同一个指称，但是不同语义类型的实体。第五步，质量评估的主要目的是量化知识的可信度，舍弃置信度低的知识以保证知识图谱中知识的质量。结合语料库中已标注的参考答案，对构建好的三元组进行自动质量评估，此外，针对语料库没有标注的语义类型（身体部位，微生物、操作、其他疗法），采用人工质量评估方法。第六步，知识图谱应用，为医学图书用户提供医学可视化分析等。

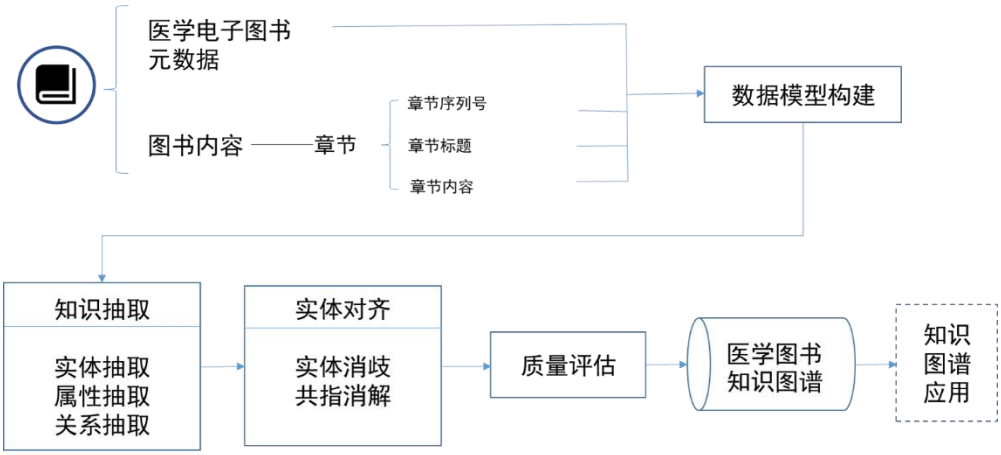


图2 构建医学图书知识图谱流程

bmr.202110.00001V1

2.3 主要软件工具

本研究主要使用基于 JAVA 的开源工具和基于 JAVA 的科学知识可视化工具 VOSviewer 实现。具体而言，JAVA 版本为 1.8.0_231，使用开源 JAVA XML API——Dom4J 读取解析 XML 图书，使用中文分词工具 Jieba（jieba-analysis-1.0.2）进行分词，医学词典使用 CMeSH，RDF 序列化的方式为 RDF/XML，图谱的可视化工具为 VOSviewer 1.6.15。

2.4 图谱构建结果展示

2.4.1 可循证的图谱数量统计 本研究共解析了41本由人民卫生出版社出版的结核病主题电子书，拆分出4746个章节。进一步通过基于正向最大匹配加检索树的方法抽取实体，保留了实体在原始电子书的原始位置、章节、书名，为后续深度开发与应用提供支撑。实体抽取结果统计如表1所示，共计抽取实体507087个，唯一概念实体10548个。其中，疾病、药物、症状和身体部位是本语料库中最常见的实体类型。抽取实体关系共计1091428，其中唯

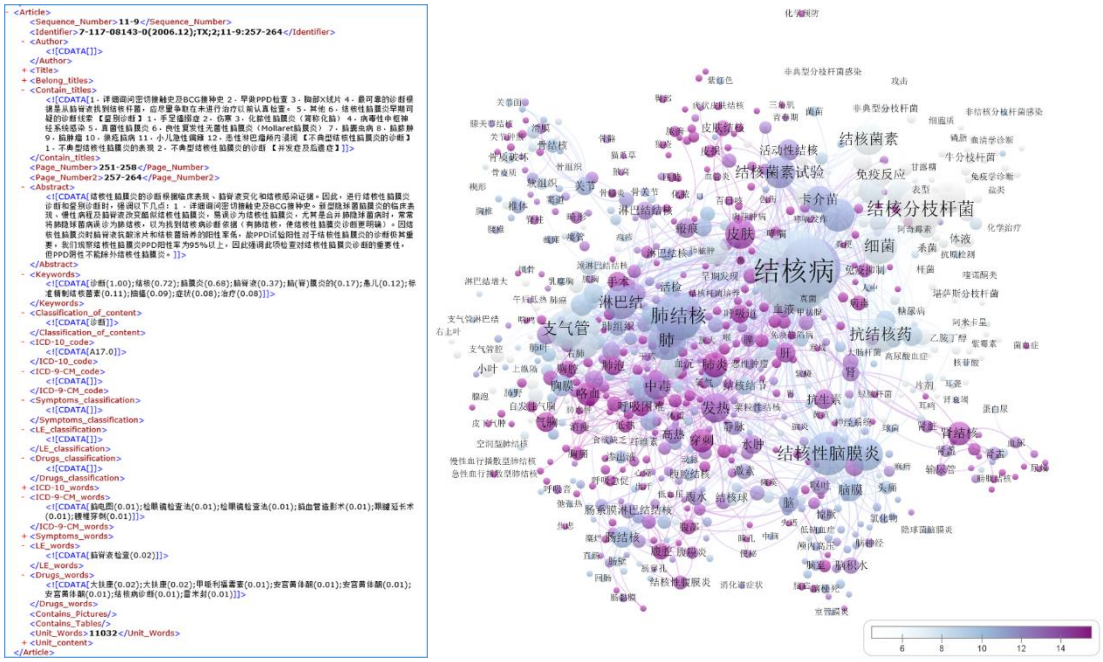
http://www.biomedrxiv.org.cn/article/doi/bmr.202110.00001
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

一实体关系424979，其中，疾病-相关-疾病，疾病-临床表现-症状，疾病-发生-身体部位关系数量最显著。

表1 实体抽取结果统计

实体类型	实体数	唯一概念数
疾病	145,623	2,905
药物	52,027	1,999
症状	30,795	1,170
检查	7,028	221
操作	29,193	726
微生物	28,762	294
身体部位	213,070	3,203
其他疗法	589	30
总计	507,087	10,548

2.4.2 图谱可视化展示 在经过实体对齐后，结合人卫卫生出版社对电子书的原有标注，对知识抽取结果进行质量评估以及改进。构建好的三元组底层数据以可视化展示的方式呈现，分别生成以书和章节为颗粒度的知识图谱。为了进一步查看医学图书知识图谱的可循证性，我们以2006年12月出版的《实用小儿结核病学》电子书构建的知识图谱为例，对构建的知识图谱进行了可视化展示。构建之前的数据样例如图3a所示，构建后的知识图谱如图3b所示，知识图谱以结核病、结核分支杆菌、肺结核、结核性脑膜炎等为中心，多实体类型、色彩区分度较高。元素的大小取决于实体出现的频次，右侧底部的颜色条代表实体所在的章节，颜色越深，实体所在章节越大，用一种直观的方式，展现了书的内容层次。此外，系统生成了每一个章节的知识图谱，概括展示章节内容。



（a） 图书的XML 数据　　（b）图谱可视化

图3 医学图书知识图谱示例—《实用小儿结核病学》（2006.12）

bmr.202110.00001V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202110.00001
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

2.5 应用场景

本研究构建的医学图书知识图谱，可辅助对现有医学图书出版物涵盖内容、知识点进行分析，帮助学生对知识点进行溯源并答题。以肺结核习题问题为例，关于原发性肺结核正确的是：A 多见于儿童，预后不良；B 病灶极少有吸收钙化；C多可形成干酪样坏死；D少数可发展为血行播散型肺结核；E 临床症状较重。其中，正确选项D中的血行血行播散型肺结核可在图3左下角找到，且与肺结核、原发性肺结核有一定关联，对应到三元组数据，包含三元组信息[‘急性血行播散型肺结核’, ‘相关’, ‘肺结核’]，[‘慢性血行播散型肺结核’, ‘相关’, ‘肺结核’]，[‘急性血行播散型肺结核’, ‘相关’, ‘慢性血行播散型肺结核’] ，[‘原发性肺结核’, ‘相关’，‘肺结核’]，说明选项D的证据支持较足，且每一条三元组都可以对应到XML电子书的具体章节。

3 对国内医学图书图谱化表示工作的启示

本文基于41本人卫电子出版社出版的结核病主题电子图书，构建了一个包含了8种实体类型、9种常用语义关系、可循证的中文医学图书知识图谱。该知识图谱的数据来源是半结构化的医学图书，在数据完备性和可靠性方面具有先天优势，是一种优质的知识图谱构建来源。此外，本文构建的医学图书知识图谱具有下列几个技术优势：1）医学图书知识图谱采用了国际规范的数据表达格式RDF，统一通过三元组来存储实体与实体之间的语义关系，方便后期的医学知识扩展应用，包括融合其他知识，系统功能扩充等；2）区别于常规知识图谱，医学图书知识图谱的数据可溯源到具体的章节，方便使用者在使用知识图谱的过程中溯源到具体的章节，为后期在医学教育方面的拓展应用提供了技术基础，可代替现有的大量的人工干预的枯燥工作。

以上医学图书知识图谱建设工作，对之后的工作启示如下：

①区别于通用领域知识图谱建设，在正式开展医学图书知识图谱构建工作之前，要根据待建设的资源涵盖的范畴以及建设知识图谱的目标，例如习题自动问答等，设计标准的知识图谱数据模型（schema），明确要抽取的知识（实体，关系以及属性）。

②医学知识图谱可视化最重要的目标之一是让人直观地了解推理的过程与结果，面向医学教育的知识图谱，知识溯源对知识推理结果的可靠性判断尤为重要，因此，在医学图书图谱化建设时，应重点保留该部分信息，构建不同篇章层次结构的图谱，供不同应用需求使用。

③在应用医学图书知识图谱解决实际应用问题过程中发现，现有的医学图书知识图谱构建以结核病为主题的医学教育书籍，但是在解决实际结核病相关习题过程中，除了结核病主题知识，还需要应用其他医学知识以及常识知识辅助实现语义推理功能，融合其他开放知识图谱，构建多学科、多主题交叉医学知识图谱，将进一步推进医学数据的自动化与智能化处理，为医学教育等应用带来新的发展契机。

参考文献

[1] 周莎,苏川.布朗大学医学免疫学教学分析研究与启示[J].基础医学教育,2019,21(11):880-882.

[2] OVID. [EB/OL]. https://www.ovid.com/platforms/point-of-reference/books.html. 2020-07-02

[3]https://www.elsevier.com/solutions

[4]http://books.ipmph.com/

bmr.202110.00001V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202110.00001
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

[5] 黄恒琪,于娟,廖晓,席运江.知识图谱研究综述[J].计算机系统应用,2019,28(06):1-12.

[6] Doyle-Lindrud S.Watson will see you now: a supercomputer to help clinicians make informed treatment decisions[J].Clin J of Oncol Nurs,2015, 19(1): 31–32.

[7] Zhou L,Gao J,Li D,et al.The design and implementation of xiaoice, an empathetic social chatbot[J]. Comput Linguist, 2020, 46(1): 53-93.

[8] Martínez-Romero M, Vázquez-Naya JM, Pereira J, et al. The iOSC3 system: using ontologies and SWRL rules for intelligent supervision and care of patients with acute cardiac disorders [J]. Comput and Mathem Methods in Med. 2013 ;2013:650671.

[9]Yang C,Yue Q,Rui S,et al.Knowledge Graph in Smart Education: A Case Study of Entrepreneurship Scientific Publication Management[J]. Sustainability, 2018, 10(4):995.

[10]袁凯琦，邓扬，陈道源,等. 医学知识图谱构建技术与研究进展[J]. 计算机应用研究，2018, 035(007):1929-1936.

[11] Shi L, Li S, Yang Xi, (若有三位及三位以上作者，请列出前三位作者姓名)，et al. Semantic Health Knowledge Graph: Semantic Integration of Heterogeneous Medical Knowledge and Services [J], Bio Med Res Inter, 2017: 1-12.

[12] Elsevier’s Healthcare Knowledge Graph and the Case for Enterprise Level Linked Data Standards. [EB/OL]. http://ceur-ws.org/Vol-2180/paper-85.pdf, 2020-07-02.

[13] Tay N N W, Yang S C, Lee C S, et al. Ontology-based adaptive e-textbook platform for student and machine co-learning[C].2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2018: 1-7.

[14] Chen P, Lu Y, Zheng V, et al. KnowEdu: A System to Construct Knowledge Graph for Education [J]. IEEE Access, 2018, 06: 31553-31563.

[15] Zheng W, Wang Z, Sun M, et al. Building a Large-Scale Knowledge Graph for Elementary Education in China[C]. Joint International Semantic Technology Conference. Springer, Singapore, 2019: 1-12.

bmr.202110.00001V1