基金项目论文

基于大数据技术的医学知识图谱构建方法

孙郑煜[1]，鄂海红[2]，宋美娜[2]，王 宁[2]

（1. 北方工业大学 信息学院，北京 100043；2. 北京邮电大学 计算机学院，北京 100876）

摘 要：为了解决医学知识图谱中知识重复、知识质量良莠不齐、知识间关联不够明确等问题，本文提出了一种大数据驱动下的医学知识图谱构建方法，同时针对医学知识图谱集成、演进、增强方面进行图谱知识融合和补全操作。然后，简单介绍医学知识图谱在医学领域的几个重要应用以及相关的人工智能技术的支持。最后，结合当前我国医学知识图谱构建技术面临的重大挑战和关键问题，对其发展前景进行了展望。

关键词：医学知识图谱；知识融合；知识补全；大数据驱动

中图分类号：TP391.1 **文献标识码：**A **DOI：**10.3969/j.issn.1003-6970.2020.01.003

本文著录格式：孙郑煜，鄂海红，宋美娜，等. 基于大数据技术的医学知识图谱构建方法[J]. 软件，2020，41（01）：13–17

The Method of Medical Knowledge Graphs Construction Based on Big Data Technology

SUN Zheng-yu[1], E Hai-hong[2], SONG Mei-na[2], WANG Ning[2]

(1. *College of Information, North China University of Technology, Beijing* 100043, *China*;
2. *College of Computer Science, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*)

【Abstract】: In order to solve the problems of knowledge duplication, uneven quality of knowledge, and unclear correlation between knowledge, this paper proposes a big-data-driven construction method of medical knowledge graphs,and carry out the knowledge graphs fusion and completion operation in terms of the integration, evolution and enhancement of medical knowledge graphs. Then, this paper briefly introduces several important applications of medical knowledge graphs in medical field and the support of related artificial intelligence technology. Finally, this paper summarized challenges and major problems of medical knowledge graph, and prospected for future development.

【Key words】: Medical knowledge graphs; Knowledge fusion; Knowledge completion; Big data driven

0 引言

医学知识图谱的构建主要是从非结构化的数据中人工或自动地提取实体、关系和属性。由于医学知识图谱的研究成果将有助于推进医学数据自动化和智能化处理，有着广阔的应用前景和社会价值，因此完善医学知识图谱的构建已经成为当前的一个研究热点。现有的基于深度学习的知识图谱融合与知识图谱补全方法已经取得了一定的成果，提升了融合和补全精度，降低了人工成本，加快了数据处理效率。但是在知识图谱融合和知识图谱补全领域中仍存在不少挑战。

1 医学知识图谱研究现状

本文主要针对面向大数据的医学知识图谱构建的持续演进，研究面向大数据与人工智能的知识图谱构建流程，同时设想研究基于图神经网络的知识融合、知识补全和动态知识更新表示问题。

目前基于图神经网络的知识图谱融合、补全和动态知识表示的相关研究还处于初级阶段，在医学知识图谱演进方向更未形成对应技术体系。下面就医学知识图谱的研究现状进行详细分析。

1.1 研究现状

医学知识数据集包括医学术语集（本体库）、医

基金项目：国家自然科学基金项目(批准号：61902034)

作者简介：孙郑煜(1998–)，女，本科生，主要研究方向：计算机软件及计算机应用，互联网技术；鄂海红(198–)，女，硕士生导师，主要研究方向：大数据与云计算，人工智能；宋美娜(1974–)，女，博士生导师，主要研究方向：大数据与云计算，人工智能；王宁(1995–)，女，硕士生，主要研究方向：大数据与云计算，人工智能。

学知识库和医学知识图谱。

其中目前的医学术语集（本体库）为医学知识库构建、医学知识图谱构建提供了医学专业术语、受限词汇的分类和概念标准化工作，权威且涵盖范围广，在数量和质量上都有所保障，被医疗行业广泛认可。

在医学知识库方面，目前国内外的医学知识库大多是基于某一专科领域的，但医学知识库是以结构化字段定义的方式存储医学知识，缺乏丰富的结构信息。而医学知识图谱是图状具有关联性的知识集合，实际上是基于语义网的知识库的形象化表示，重在抽取关系展示知识间的高关联性和高结构化的特征。由此医学知识图谱能够包含更加丰富的关系层次和关系链接，显著提升知识推理的精度及效果。

总体来说，大规模、多领域、跨语言的专科医学知识图谱构建尚处于演进发展、不断增强阶段，若要得到更完善的医疗知识图谱，需要对不同的医疗本体库、知识库和图谱进行融合以及将尚未涵盖的知识和不断产生的新知识融合到已有的知识图谱中。医疗知识图谱的构建必须是一个不断迭代更新的过程。医学知识图谱演进所需的知识融合、知识补全、动态知识更新表示就变得迫切和亟需。

1.2　有效工具

目前知识图谱普遍采用了语义网框架中 RDF（Resource Description Framework，资源模式框架）模型来表示数据[1]。北京大学计算机所数据管理实验室研发了面向 RDF 知识图谱的开源数据库系统（通常称为 Triple Store）。不同于传统基于关系数据库的知识图谱数据管理方法，gStore 是直接开发面向 RDF 知识图谱数据的 Native 的知识图谱数据存储和查询系统（Native RDF 图数据库系统），考虑 RDF 知识图谱管理的特性，从数据库系统的底层进行优化[1]。它维持了原始 RDF 知识图谱的图结构，数据模型是有标签、有向的多边图，每个顶点对应着一个主体或客体。它将面向 RDF 的 SPARQL 查询转换为面向 RDF 图的子图匹配查询，利用其所提出的基于图结构的索引(VS-tree)来加速查询的性能[1]。

gStore 支持复杂的 SPARQL 查询及有效的增删改操作，支持 W3C 定义的 SPARQL 1.1 标准，包括含有 Union、OPTIONAL、FILTER 和聚集函数的查询；支持有效的增删改操作。同时，gStore 支持海量三元组规模的 RDF 知识图谱的数据管理任务，单机可以支持 5Billion（五十亿）三元组规模的 RDF 知识图谱的数据管理任务。分布式版本支持百亿边规模的分布式可扩展的部署模式[2]。由此非常有利于知识图谱的优化研究。

1.3　待解决问题

（1）由于不同医疗知识图谱的知识来源广泛，构建目的和方式也不同，使得单个知识图谱内存在知识质量低下、知识描述缺失等问题；不同知识图谱间又存在知识大量重复，异构性强等问题，给实体对齐算法的精度提升带来了困难。因此需要解决面向知识融合过程中信息缺失导致的实体对齐精度不高的问题。

（2）随着医学知识图谱不断地发展，越来越多大规模的医学知识图谱被构建出来。且知识图谱的规模不断增长，知识对齐算法的计算复杂度会呈现二次增长，因此，面向大规模医学知识图谱的高效处理问题有着重要的研究意义。

（3）医学知识图谱作为一种复杂的多关系图，含有丰富的图结构信息。而传统知识图谱补全的方法由于只考虑三元组的内部信息，而造成补全精度不高的问题。因此，如何更高效的利用图结构信息是进一步扩展图神经网络方法在知识图谱补全应用的关键点。

（4）当前动态知识图谱仅利用节点本身的结构信息，未能利用动态变化过程中的时序信息，造成表示精确度不高的问题；未考虑节点对相邻节点的传播影响，造成误差在时间序列中不断积累，从而影响最终表示，这是信息变化传播不充分的问题；同时，在更新知识图谱时，每次改变都需要对全局节点全部进行训练，造成更新代价大的问题。

2　大数据驱动下医学知识图谱构建

通过对大量的参考文献进行阅读、分析以及总结，将医学知识图谱构建的全流程总结为五个核心流程：医学数据采集、医学知识抽取、知识融合、构建图谱、知识更新。在医学领域知识图谱构建过程中存在着与之相对应的大数据处理流程，包括数据源与数据采集、数据处理和数据更新、以及支撑医学知识图谱构建全生命周期的数据存储。

下面简单归纳概括医学知识图谱构建过程中使用的大数据技术。医学知识图谱的数据主要来自网络，通过爬虫技术把信息抓取到 HDFS 或 MySQL 中，其他医学数据源（如部分标准医学数据库等）通过 Sqoop 导入 HDFS 或 MySQL 中，然后使用 MapReduce、Spark 等技术对数据进行处理，处理后的数据导入 Hive、Hbase 等，最后使用 Java、HiveQL、R 及 Spark 等进行数据分析与展示[3]。详细的图谱构建中大数据技术的使用将渗透在后文图谱构建生命

周期的各个环节之中。

2.1 医学数据采集

如今，医疗信息技术飞速发展，医学数据数量急速增加，同时还有新知识不断产生，需要利用大数据技术进行数据源与数据采集。医学数据源主要分为三类，包括非结构化的文本数据，半结构化的表格、网页以及部分医疗信息系统的结构化数据。由于构建医学数据图谱所需的数据大多来源于网络，所以需要借助爬虫来获取，本文拟采用的方法是基于 Scrapy[4]框架实现爬虫，获取网络上的医学数据信息。此外，除了完成原始数据的采集，在数据采集过程中还通过一种基于百科类网站爬虫的同义实体扩充方法，构建一个准确且丰富的医学同义词库，以辅助实现知识融合中的实体链接[5]。

医学数据信息种类繁多，存储方式不一，因此采集来的医学数据信息有可能存在知识错误或者知识描述缺失等问题，尤其是对于非结构化的数据，这些来源于网络的纯文本数据通常需要使用自然语言处理(Natural Language Processing，NLP)技术进行预先处理。为了解决医学数据中可能出现的问题，需要用到 ETL 技术（数据抽取、转换、加载），而 Hive 作为一个可靠的 ETL 工具，在高效性、扩展性、容错性等方面的表现特别突出，进行数据预处理将原始数据转换为适合对其进行分析的数据模式对于保证数据质量起到了非常关键的作用[6]，这个步骤是是基于 Hive 完成的，是从数据采集向信息抽取的过渡流程。

Hive 是基于 Hadoop 的数据仓库工具，通过 Hive 可以使用传统的 RDBMS 的 SQL 语法来实现就 HDFS 的数据的 ETL 和数据模型的构建。并且 Hive 也支持 Spark 的计算引擎接口和分析展示的 R 包接口（RHive）来获取 Hive 构建好的模型表及逻辑。

2.2 医学知识抽取

医学知识抽取通过人工或自动化技术从半结构化和非结构化数据中抽取出可用的知识单元，这对应着大数据技术中的数据处理，实际上就是基于 Spark 完成医学数据非(半)结构化向结构化的转化。

原始数据采集完成进入 HDFS 后可能存在诸多问题，需要对数据进行预处理[6]，可以基于 Spark 调用机器学习模型完成实体、关系、属性的抽取完成医学数据非（半）结构化向结构化的转化。Spark 是在进行大规模数据处理时的高效引擎，Spark 数据计算在内存中完成，有效地解决了实时性问题[7]。同时，Spark 可以很好地和不同的数据源进行整合，比如 HDFS、HBase、Cassandra、S3 等，充分利用 Spark 计算引擎的特性[7]。

非（半）结构化数据向结构化数据转化的本质其实就是通过自然语言处理技术，从网络数据中大量的纯文本内容完成实体抽取、关系抽取和属性抽取。

（1）实体抽取从文本数据集中识别提取出命名实体，如医学文本中的疾病名、药物名、症状名等。

（2）关系抽取提取出实体之间的关联关系，例如医学中的疾病临床表现、疾病多发人群等，通过这些关系将一系列离散的医学实体联系起来形成网状的知识结构，从而解决医学实体间语义链接的问题。

（3）属性抽取则是从多种数据源中采集医学实体的属性信息来构造医学实体的属性列表，实现对医学实体的完整勾画。例如药品的属性包括适应症、不良反应、禁忌和慎用等。

基于 Hive 完成数据处理和基于 Spark 完成医学数据非（半）结构化向结构化的转化就可以得到较为完备的医学数据。

2.3 知识融合

由于医学数据库中的知识来源复杂，存在知识质量良莠不齐、不同数据源知识重复、知识间关联关系模糊等问题。知识融合就是完成对不同来源的知识在同一框架规范下进行数据整合、消歧、加工、推理验证、更新等操作，对数据进行剔粗取精，增强知识库内部的逻辑性和表达能力。知识融合的三个关键部分是实体对齐、实体链接和关系推演。

（1）实体对齐用于消除异构数据中的实体冲突、指向不明等不一致问题，医学实体在不同的数据源中存在严重的多元指代问题[8]，例如西药头孢哌酮钠，通用名称为头孢哌酮钠，商品名称可以为先抗、先锋必素、头孢氧哌唑、先锋必、先锋哌酮、氧哌羟苯唑、头孢菌素钠、先锋哌唑酮、先锋松等，因此实体对齐是医学知识融合中非常重要的一步。

（2）实体链接的主要作用是利用医学知识库中的实体对从医疗大数据的文本中获取的实体指代进行消歧，然后将对应的医学实体链接到医学知识库中的对应实体。

（3）关系推演的主要目标是将从医学大数据文本中获取的实体关系动态扩展到知识库中，有助于提高医学知识库的时新性、覆盖能力等，实现关系的扩充。

根据是否使用标记数据，知识融合方法可以分为有监督方法和无监督方法[9]。有监督方法是从标记数据中学习模型以进行实体对齐，主要分为基于属性比较的方法、基于聚类的方法和主动学习方法。无监督方法仅依靠少量种子集或不依靠种子集就可

完成模型的学习，主要分为传统的无监督方法和知识嵌入方法。

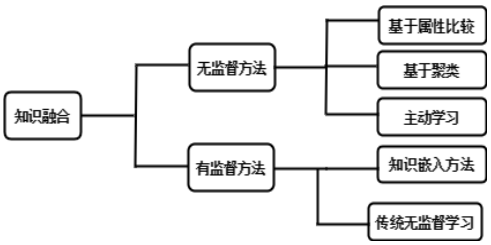

图 1　知识融合方法图
Fig.1　Knowledge fusion method diagram

针对上文提出的医学知识图谱构建过程中待解决的有关知识融合的科学问题，本文提出了相应的解决方向设想。拟采用基于高效图神经网络的知识融合模型来解决知识融合过程中信息缺失导致的实体对齐精度不高的问题和大规模医学知识图谱的高效处理问题。

2.4　图谱构建

图谱构建就是基于之前得到的关系型数据库模式转换成图数据库模式。下图是一个简单的关于糖尿病所构建的医学知识图谱，图中简单表示出了实体以及实体与实体之间的关系，如糖尿病临床表现为多饮、多食、多尿、疲乏无力等，糖尿病多发于老年人群体。

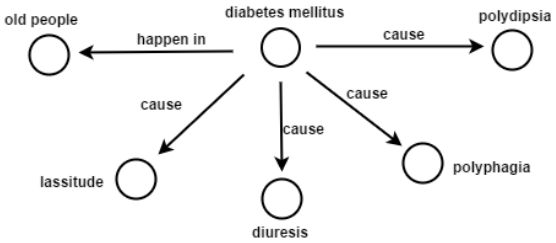

图 2　知识图谱示例
Fig.2　Knowledge graph example

2.5　知识存储更新与补全

数据存储与更新支撑着医学知识图谱构建的整个生命周期，在数据的更新过程中可以将数据划分为四个等级。其中，采集来的原始医学数据是一级数据，经过 Hive 处理过的三元组数据是二级数据，Hive 作为一个数据仓库的客户端工具，本身是不保存数据的，它所操作的表数据都存放在 HDFS 中[10]。构建出的医学图谱的实体、关系、静态属性以及动态属性是三级数据，图谱更新后的更新类型及三元组数据是四级数据[5]，其变迁流程以及存储位置如下图所示。

知识图谱补全是通过预测出三元组中缺失的部分，从而使知识图谱变得更加完整。知识图谱补全可以分为实体预测以及关系预测任务。静态知识图谱补全是补全已知实体之间的隐含关系或补全存在于知识图谱中的实体属性。动态知识图谱补全是能够建立知识图谱与外界的关联，从而扩大知识图谱的实体集、关系集以及三元组集。利用静态知识图谱补全可以对知识图谱中的实体属性和关系进行补全；现有的动态知识图谱补全能对新增实体的知识图谱中的数据进行更新。

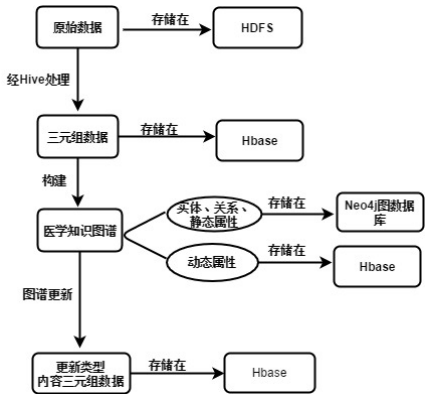

图 3　数据变迁及存储位置图
Fig.3　Data changes and storage locations

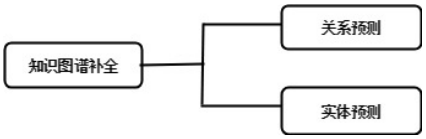

图 4　知识图谱补全方法图
Fig.4　Knowledge graph completion method diagram

针对上文提出的医学知识图谱构建过程中待解决的有关医学知识图谱补全的科学问题，本文提出了相应的解决方向设想。采用结合对抗学习和注意力机制的多关系图神经网络知识补全模型来解决知识图谱补全精度不高和结构信息利用不合理问题，采用基于共享权重机制多关系图神经网络知识图谱补全模型来解决动态知识图谱表示不精确、信息变化传播不充分和更新代价大的问题。

3　医学知识图谱的应用

知识图谱在医学领域的应用有助于提高医疗智能化的水平，目前医学知识图谱主要应用于临床决策支持系统、医疗智能语义搜索引擎、医疗问答系统[11]、医学知识科普等方面。

3.1　医学知识推理

人工智能技术的发展和应用，提高了医学知识图谱的构建效率和知识推理的准确率。医学知识图谱必须处理大量重复矛盾的医学信息，例如即使对于相同的疾病，医生也要根据患者病情状况作出不

同的诊断[12]，给出不同的解决方案，人工智能拥有从海量数据中挖掘有用信息的天然优势，知识推理注重知识与方法的选择和运用，能够推断出缺失事实完成对问题的求解[8]。

3.2 应用分析

利用医学知识图谱可以辅助医疗行业进行大数据分析与决策，根据患者症状以及检查结果等数据自动生成诊断和治疗方案，供医学专业人员参考，同时还可以对医生的诊疗方案进行智能化分析，有效减少误诊情况的发生[11]。

同时，从医学知识图谱中检索并查询相关的实体对、实体关系及属性进行扩展查询[11]，从而改善医疗信息搜索和查询结果的准确性，可以实现以自然语言形式为用户提供准确的问题的解答，辅助患者在就诊前得到相关的医学知识科普，帮助患者找到合适的医生，同时还可以一定程度上避免由于医学知识专业性强、医患信息沟通困难而导致的医患关系紧张，可有效改善患者就医体验，提高后续医疗服务的精准度和效率与患者就诊满意度[13]。

4 展望

知识图谱具有强大的语义处理和开放获取能力，是对语义网和知识库的改造和升华。医学知识图谱将医学知识与知识图谱结合起来，推动医学数据的智能化和自动化处理[14]，定会为医疗行业的发展带来新的契机。

知识图谱在医学领域的应用为医疗行业带来了新的机遇，同时也带带来了一系列挑战。目前，医学知识图谱构建的关键关节还面临着一些巨大的困难和挑战。例如，目前的应用于医学文本抽取的算法普遍存在着准确性低、限制条件多、扩展性差等问题，医学知识来源的多样性导致医学实体在不同的数据源中存在严重的多源指代问题，动态医学知识图谱表示不准确、信息变化传播不充分和更新代价大的问题，如何利用医学知识图谱可视化为医生寻求最佳的诊疗展示方案使病人理解展示结果也是一个挑战。

医学知识图谱是大数据、人工智能与医学的结合，在未来必将成为医疗行业与大数据智能研究的热点和前沿问题。

参考文献

[1] gStore团队. gStore系统使用手册[OL]. (2019-07-29)[2019-10-11]. http://gstore-pku. com/pcsite/index.html.

[2] 北京大学王选计算机研究所. gStore是什么？[OL]. (2019-06-06) [2019-10-11]. http://gstore-pku.com/pcsite/index.html.

[3] 张魁, 张粤磊, 刘未昕, 吴茂贵. 自己动手做大数据系统[M]. 北京: 电子工业出版社, 2016: 22.

[4] Guogang Zhang. Python Network Source Automatic Evaluation System[A]. Proceedings of 2016 4th International Conference on Electrical & Electronics Engineering and Computer Science (ICEEECS 2016)[C]. (Computer Science and Electronic Technology International Society), 2016: 5.

[5] 王宁. 基于Web的领域知识图谱构建平台的研究与实现[D]. 北京: 北京邮电大学, 2019: 36-42.

[6] 张魁, 张粤磊, 刘未昕, 吴茂贵. 自己动手做大数据系统[M]. 北京: 电子工业出版社, 2016: 152-154.

[7] 张魁, 张粤磊, 刘未昕, 吴茂贵. 自己动手做大数据系统[M]. 北京: 电子工业出版社, 2016: 281-282.

[8] 贾辛洪. 医学知识图谱构建技术与研究进度[OL]. (2019-08-21)[2019-10-11]. http://blog.csdn.net/jiaxinhong/article/details/81865768.

[9] Guan S, Jin X, Jia Y, et al. Self-learning and embedding based entity alignment[C]//2017 IEEE International Conference on Big Knowledge (ICBK). IEEE, 2017: 33-40.

[10] 张魁, 张粤磊, 刘未昕, 吴茂贵. 自己动手做大数据系统[M]. 北京: 电子工业出版社, 2016: 217.

[11] 侯梦薇, 卫荣, 陆亮, 兰欣, 蔡宏伟. 知识图谱研究综述及其在医疗领域的应用[J]. 计算机研究与发展, 2018, 55(12): 2587-2599.

[12] 袁凯琦, 邓扬, 陈道源, 张冰, 雷凯. 医学知识图谱构建技术与研究进展[J]. 计算机应用研究, 2018, 35(07): 1929-1936.

[13] 修晓蕾, 吴思竹, 崔佳伟, 邬金鸣, 钱庆. 医学知识图谱构建研究进展[J]. 中华医学图书情报杂志, 2018, 27(10): 33-39.

[14] 刘雷. 临床诊断决策需要知识图谱的“供养”[J]. 张江科技评论, 2019(04): 34-36.

[15] 吴运兵, 阴爱英, 林开标, 余小燕, 赖国华. 基于多数据源的知识图谱构建方法研究[J]. 福州大学学报(自然科学版), 2017, 45(03): 329-335.

[16] 杨林朋, 董一超, 赵祖桉, 崔雪宁, 刘新奎. 我国医学信息学领域的研究现状及其可视化分析[J]. 中国卫生产业, 2019, 16(22): 160-164.

[17] 李红艳, 皇甫慧慧. 基于知识图谱的全科医生研究可视化分析与展望[J]. 中国全科医学, 2019, 22(27): 3387-3394.

[18] 郁小玲, 张铁山, 吴彤, 等. 基于两位一体的中文电子病历命名实体识别[J]. 中国卫生信息管理杂志, 2017, 14(4): 552-556.

[19] 康准, 王德军. 基于知识图谱的生物学科知识问答系统[J]. 软件, 2018, 39(02): 7-11.

[20] 王雪鹏, 刘康, 何世柱, et al. 基于网络语义标签的多源知识库实体对齐算法[J]. 计算机学报, 2017(03): 169-179.

[21] 林海伦, 王元卓, 贾岩涛, 张鹏, 王伟平. 面向网络大数据的知识融合方法综述[J]. 计算机学报, 2017, 40(01): 1-27.

[22] 扩展知识图谱上的实体关系检索[J]. 王秋月, 覃雄派, 曹巍, 覃飙. 计算机应用. 2016(04)

[23] 朱国丞. 基于大数据平台的知识图谱存储访问系统的设计与实现[D]. 东南大学, 2018.

[24] 张龙斌. 面向成果转化的知识图谱研究及应用[D]. 杭州电子科技大学, 2018.

[25] 李涓子, 侯磊. 知识图谱研究综述[J]. 山西大学学报(自然科学版), 2017, 40(03): 454-459.

[26] 方阳, 赵翔, 谭真, 杨世宇, 肖卫东. 一种改进的基于翻译的知识图谱表示方法[J]. 计算机研究与发展, 2018, 55(01): 139-150.

[27] 聂莉莉, 李传富, 许晓倩, 朱川川, 徐志鹏, 武红利. 人工智能在医学诊断知识图谱构建中的应用研究[J]. 医学信息学杂志, 2018, 5(6): 7-12.

[28] 阮彤, 孙程琳, 王昊奋, 方之家, 殷亦超. 中医药知识图谱构建与应用[J]. 医学信息学杂志, 2016, 37(4): 8-13.