DOI: 10.3969/j.issn.1671-3982.2018.06.005

·情报方法与技术·

医学百科知识图谱构建

刘 燕[1],傅智杰[2],李 姣[1],侯 丽[1]

[摘要] 利用实体识别、关系抽取、可视化分析等技术构建医学知识图谱,以期为知识服务系统提供知识的高效检索、组织和管理,为知识间关联关系的发现奠定基础。该图谱可提供力导向布局图及和弦图两种可视化展示百科知识的直观方式,应用于医药卫生知识服务系统平台取得很好的效果,系统“百科数据”访问量突破性增加,超过 20% 的用户关注并浏览知识图谱应用。

[关键词] 知识图谱; 实体识别; 关系抽取; 可视化

[中图分类号] TP391.1; R-05 [文献标志码] A [文章编号] 1671-3982(2018)06-0028-07

Generation of medical encyclopedia knowledge graph

LIU Yan[1], FU Zhi-jie[2], LI Jiao[1], HOU Li[1]

(1.Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China; 2.Chinese Academy of Enginering, Beijing 100088, China)

[Abstract] A medical encyclopedia knowledge graph was generated using the entity identification, relationship extraction and visualization techniques in order to provide effective knowledge retrieval, organization and management for knowledge service system and to lay a foundation for the discovery of correlation between different kinds of knowledge. The medical encyclopedia knowledge graph provides force-oriented layout graph and chord diagram which can intuitively show the visualized encyclopedia knowledge, and can thus be used in medical and health knowledge service platform.

[Key words] Knowledge graph; Entity identification; Relationship extraction; Visualization

随着互联网、大数据等技术的发展,各领域的数据和知识都呈爆炸式的增长,对知识进行高效组织和管理的需求不断增加。知识图谱是在大数据背景下产生的一种高效的知识表示和管理方式[1],能够支持综合性的知识检索、问答、决策支持、可视化分析等智能应用[2-4]。目前,知识图谱已经被应用到各行各业中,如医学、金融、农业等领域的信息检索、知识问答、知识推理等。但现有的知名知识图谱大多适用于通用领域,如谷歌知识图谱、Facebook 兴趣图谱、搜狗“知立方”等,专业学术领域相关的研究和应用还相对较少,无法满足科研人员的需求。因此,面向特定领域的知识图谱研究与实践变得尤为重要。

近年来,研究者围绕医学知识图谱的构建与应用开展了大量研究,如 Maya 等人提出了一种从电子医学病历中自动提取疾病和症状概念并自动构建知识图谱的方法[5],Meng Wang 等人通过构建层次化知识图谱来获取电子医学病历中患者、疾病和药物之间的关系[6],Longxiang Shi 等人探索了一种可以实现知识图谱中异构医学健康知识和服务自动检索的新模型[7],以及面向知识图谱的可视化分析[8-10]和应用研究[11-12]等。现有的医学知识图谱

[基金项目] 中国工程科技知识中心建设项目“医药卫生专业知识服务系统”(CKCEST-2018-1-16)

[作者单位] 1.中国医学科学院医学信息研究所,北京 100020; 2.中国工程院,北京 100088

[作者简介] 刘 燕(1990-),女,陕西榆林人,硕士,研究实习员,研究方向为医学数据挖掘和知识图谱构建。

研究多集中于临床数据和文献资源，而面向医学百科数据的知识图谱研究还较为匮乏。因此，本文将借鉴谷歌知识图谱构建的技术和经验，选取较为规范的医学百科数据作为知识图谱的应用案例，构建面向重大疾病的医学百科知识图谱，以期为相关人员提供知识的高效搜索，为知识间关联关系的发现奠定基础，并最终应用于医药卫生专业知识服务系统平台，辅助开展知识的语义关联和搜索，以及知识问答、智能诊断等更为深入的应用。

本文利用医学百科数据进行医学百科数据的知识图谱构建，从数据获取、实体识别、关系抽取、可视化展示等方面阐述医学知识图谱的构建流程，最后应用于“中国工程科技知识中心医药卫生专业知识服务系统平台”，实现医学知识图谱的应用。

1　基于医学百科数据的知识图谱构建

1.1　医学百科知识图谱的构建方法与流程

知识图谱的构建方法可归纳为自顶向下和自底向上两种[13]。自顶向下的方法是先构建知识图谱的本体，自底向上的方法则是从实体层开始构建[14]，然而实际构建过程中一般是 2 种方法结合着使用。知识图谱的构建涉及实体抽取和实体之间关系的建立[15]，首先需要从数据中提取出实体、关系和属性，然后利用图谱绘制软件或工具生成相应的图谱，可视化展示实体及实体间的关系。

医学百科知识图谱的构建流程与之相似，分为知识获取、知识处理和知识应用 3 部分。针对半结构化的医学百科数据，需通过命名实体识别、实体关系抽取等技术进行结构化处理，形成对应的知识三元组，然后利用相关软件和工具将其转换为另一种可视化、直观的表示形式，即知识图谱。具体流程如图 1 所示。

1.2　医学百科数据获取

知识获取即调研、收集拟处理的数据对象，以数据的可靠性、完整性、权威性为目标，以确保数据在后续处理、分析、评估及共享过程中的合理性和价值。随着健康问题的关注度持续上升，积极开展健康知识的研究也是实现“健康中国 2030”的重要保障。医学百科是公众获取健康知识的一种重要途径，好的展示方式将有助于用户更有效地了解相关知识，从而促进重大疾病的预防和筛查。考虑到医学百科数据量较大，本文拟选取目前疾病负担较重的肿瘤、心脑血管疾病、呼吸系统疾病等探索医学百科知识图谱的构建方法，并以发病率较高的哮喘为例进行详细说明。

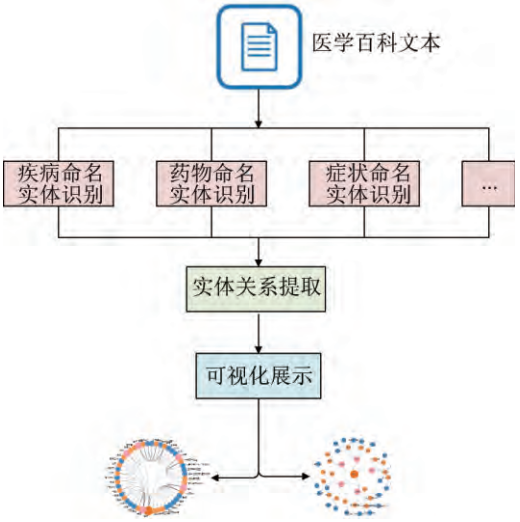

图 1　医学百科知识图谱构建流程

“百科名医网”（http://www.baikemy.com/）是国家卫健委（原“卫生计生委”）权威医学科普项目唯一的指定网站，涵盖了大量医学和养生知识，拥有严格的质量控制和审核机制，内容严谨、科学。鉴于其数据内容全面、质量可靠、权威性高，本文使用 Java 语言通过网络爬虫方法抓取“百科名医网”中与肿瘤、心脑血管疾病、呼吸系统疾病等主题相关的词条信息，为后续知识处理产生原始数据基础。

本文基于该爬虫程序构建了医学百科数据集，并采用人工剔除的方式辅助筛选出了 82 条词条信息，包括疾病名称、临床表现症状、原因、诊断、治疗、预防等内容。同时对采集的词条信息进行数据清洗、编辑、分组、排序、重复值删除、规约等一系列预处理操作，以保证数据的完整和准确。

1.3　医学百科数据处理

知识处理是指通过命名实体识别、实体关系提取等技术和方法对所收集的数据进行的规范化处理。其中，命名实体识别技术是信息抽取、机器翻译、问答系统等多种自然语言处理技术必不可少的组成部分，也是构建知识图谱的重要手段之一[16-18]；实体关系抽取的目的则是确定文本中实体对之间的关系，具体而言就是利用关系抽取技术，从

无结构的海量文本中提取出格式统一的数据,然后借助计算机快速处理文本,抽取实体之间的语义关系,从而构建出众多实体之间的关联信息[19-20]。尽管目前面向命名实体识别、实体关系抽取任务的工具有很多,如针对疾病的工具 DNorm-0.06、针对药物的工具 tmChemM1-0.02 等,但大多只适用于某些特定的应用场景。因此,根据实际需要选取相应的工具提取实体及实体间的关系信息至关重要。

1.3.1 命名实体识别

本文通过中文命名实体识别工具 Stanford NLP 识别出有效的疾病、症状等实体,为后续实体关系的抽取奠定基础。此外,为了确保数据质量,聘请专业人员对识别结果进行审核、校对,修改未能正确识别的命名实体。

通过对上述百科数据集进行症状、诊断、病因等命名实体的识别与校对,共得到 1 876 个实体。本文以哮喘的部分文本为例进行分析说明。图 2 为哮喘百科文本中识别出的实体情况,包括病因、症状、诊断等相关概念实体,每个实体又包含了实体的名称、实体的类型等。

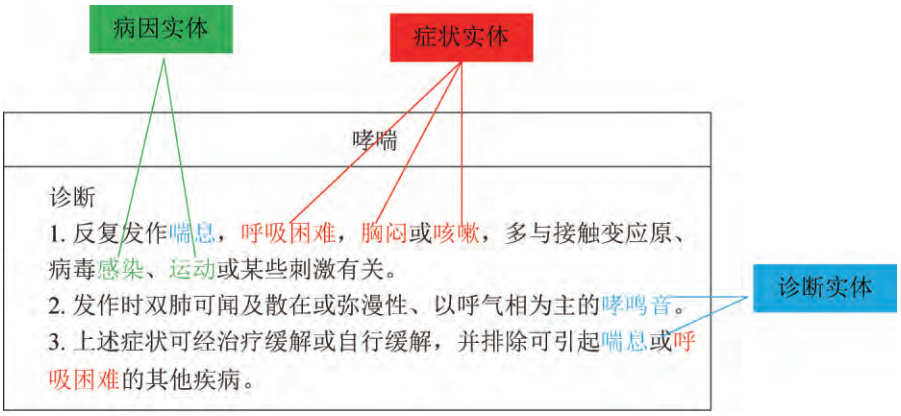

图 2 “哮喘”百科文本中相关命名实体

通过对上述“哮喘”百科文本数据进行命名实体的识别,共识别出 41 个相关实体。其中病因实体 15 个、症状实体 9 个、治疗实体 9 个、诊断实体 2 个、检查实体 6 个。

1.3.2 实体关系抽取

实体关系抽取是构建知识图谱的重要环节之一,主要是根据实体的属性、类别、消歧信息、关键词等特征确定实体的所属关系类别。

医学领域数据具有内容丰富、信息量大、潜在价值高等特点,因此对该领域的数据进行关系抽取具有非常重要的意义。如医学百科的关系抽取可以帮助公众快速了解疾病的病因和症状,电子病历的关系抽取可以用于临床决策支持等。

医学领域实体关系抽取的任务主要是抽取疾病和药物、疾病和症状、疾病和基因、疾病和疾病、药物和症状、药物和药物等实体间的关联关系,从而为患者和领域专家提供支持。

首先,抽取每个实体所对应的特征和关键词等信息。百科中的实体都对应着一些结构化和半结构化的特征,

本文将采用这些特征来表示实体的类型(图 3),如“哮喘”百科文本中描述的语义特征包括类别特征、上下文特征、关系特征、别名等。另外,文本中的关键词也能对实体之间的关系起到一定的提示作用,如两个实体之间的关系为“治疗”,那么句子中就可能包含消除、减缓、恢复、控制等关键词[21]。

然后根据抽取出的实体及其特征和关键词信息进行实体关系的标注,并用 RDF 三元组表示,如“哮喘”的症状表现为“胸闷”、检查方式有“肺活量”等;同时能够展示层次化的关系,如“哮喘”的病因有“敏感原”,‘敏感原”又包括“花粉”等(图 4)。

本文将识别的疾病、症状、诊断、治疗等相关的实体和概念使用 xml 技术存储于数据库中(图 5),然后基于 dom4j、XPath 等技术对 xml 文件进行解

析,构造相应参数,为后续的可视化分析和展示奠定基础。

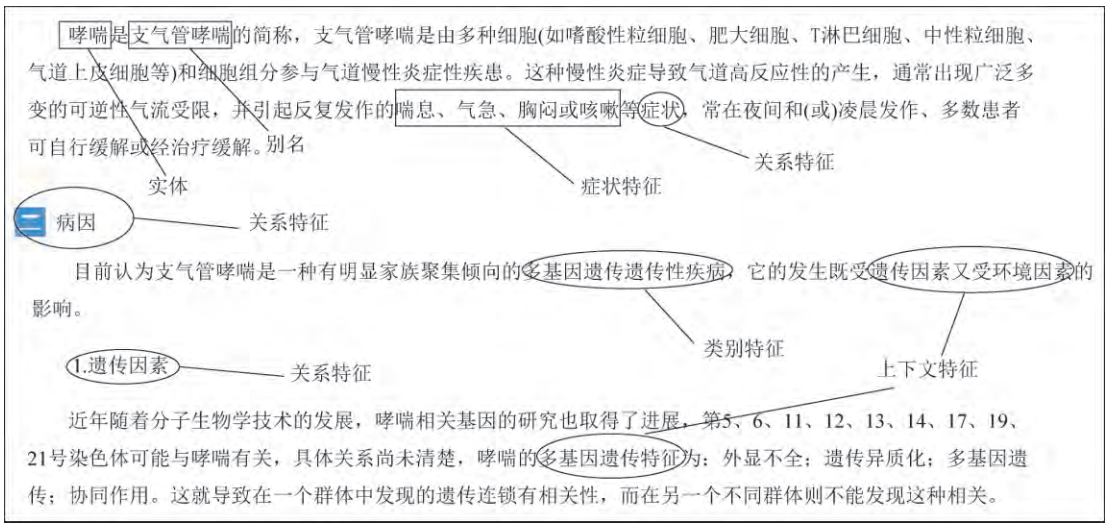

图 3　百科中“哮喘”的实体特征

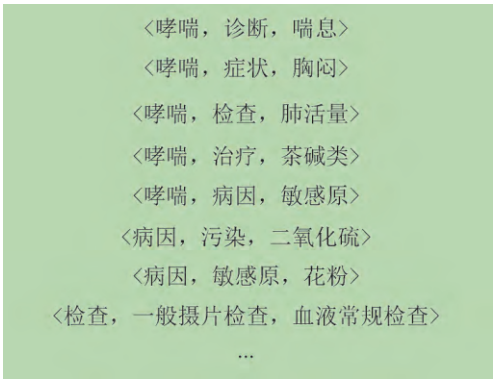

图 4　RDF 三元组表示

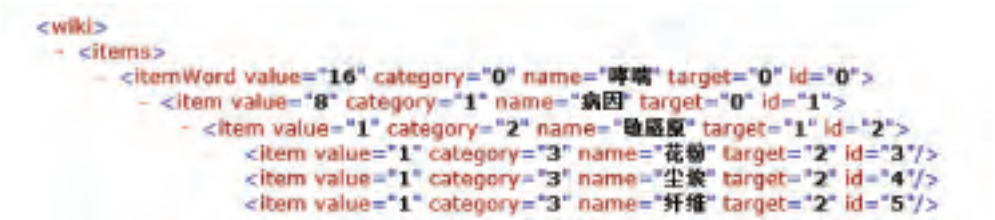

图 5　知识存储情况

2　知识图谱可视化应用

知识应用是基于上述知识提供的知识图谱、辅助语义搜索、可视化分析、智能问答、专家系统等功能和应用。其中医学知识图谱是一种新型、直观的实体关系展示方式,可基于实体的概念、属性、关系等生成多元的可视化知识图谱;可视化技术可以利用计算机技术将医学数据转换为图形或图像,提高交互能力。通过对疾病、药物、症状等医学数据的可

视化处理,能够实现更好的视觉效果,同时也更容易地被用户所关注,帮助用户快速发现感兴趣的内容、挖掘知识间的关联关系。常用的可视化工具有 CiteSpace[22]、Tableau[23]、R 语言[24]等,各类工具的功能和视觉效果都各有侧重,应具体情况具体分析。

图形化展示是知识图谱构建的重要部分。本文选取开源免费的 ECharts 可视化图表工具对上述百科实体及其关系进行可视化展示,直观、生动、可交互、可个性化定制,还赋予用户进行数据挖掘和整合的权限。

2.1 人机交互的可视化服务

为实现上述百科实体关系的可视化应用与服务,本文设计了可用于人机交互展示的系统功能架构,并对各百科知识图谱进行平台系统实现,在“医药卫生专业知识服务系统(http://med.ckcest.cn)”中新增“知识图谱”应用(图 6)。

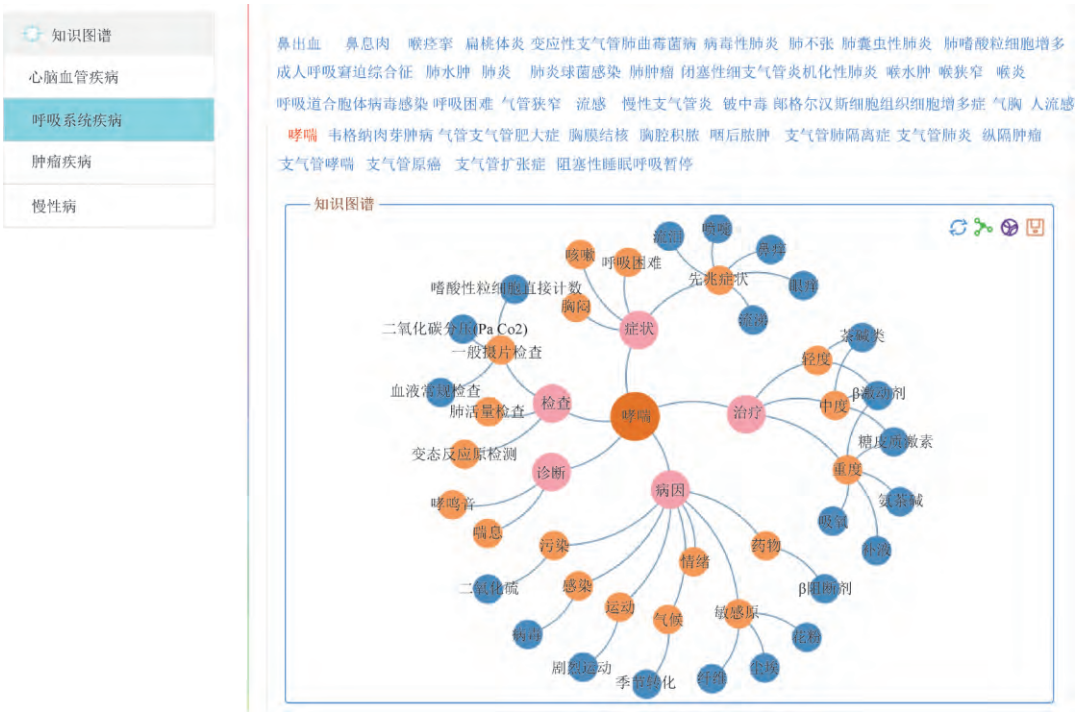

图 6 知识图谱应用系统页面

该知识图谱的应用实现了面向心脑血管疾病、肿瘤、呼吸系统疾病、慢性病等 4 个主题的医学领域重大疾病的百科知识图谱展示,以“哮喘”为例,提供“力导向布局图”和“和弦图”两种可视化功能,用户可根据需要随意切换,如图 7、图 8 所示。其中不同颜色、不同尺寸的圆圈表示不同级别的实体,“边”表示实体间的关联关系。如果想了解某两个实体间的关系,只需将鼠标悬浮于此即可查看。其中,“力导向布局图”还支持节点的拖拽功能,用户可按需调整各实体的位置以达到最好的展示效果。另外,两种图提供图片保存功能,方便后续开展对比分析和深入研究。

2.2 “知识图谱”应用效果

通过对“医药卫生专业知识服务系统”后台日志的分析,发现“知识图谱”应用的嵌入使该系统中“百科数据”的访问量有了突破性增加,有超过 20% 的用户关注并浏览过知识图谱应用,且用户较为关注的内容主要包括哮喘、乳腺癌、肺癌、脑卒中等疾病。

由此可见,可视化的知识图谱应用符合领域人员的关注点,能为用户更好地发现知识间的关联关系提供帮助,并可大大提升平台的访问率和用户体验。结合各领域技术开展百科数据知识间的关联与挖掘,可为相关人员了解疾病的发病与关联因素,开展创新研究等提供新的视角,帮助公众直观获取相关疾病的知识,提高对疾病的认知,为减少或预防高

发疾病的发生提供依据，为国家工程科技智库在医药卫生、公众健康、科技创新等方面开展宏观发展策略研究提供多元、智能的知识和服务，为我国医药卫生事业发展做好服务支撑工作。

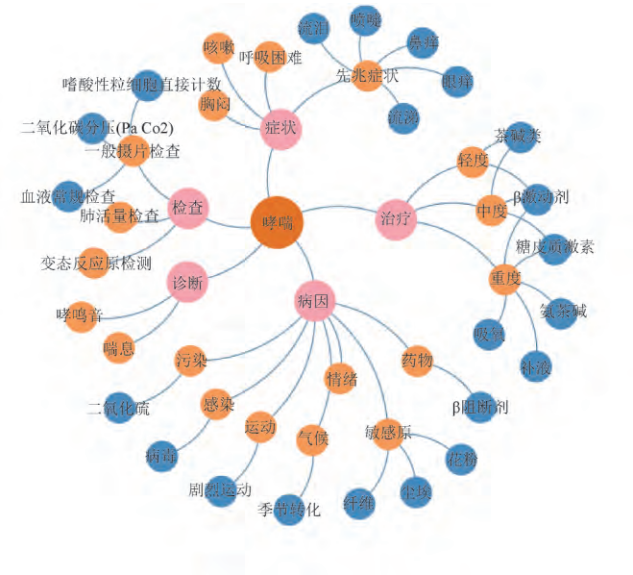

图 7　“力导向布局图”可视化展示

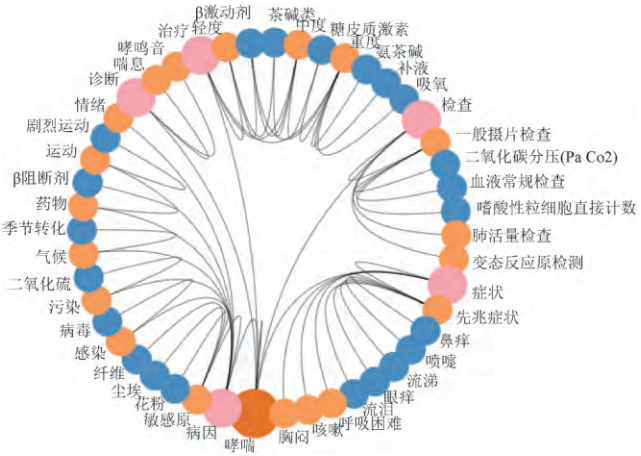

图 8　“和弦图”可视化展示

3　结论

本文通过构建面向重大疾病的医学百科知识图谱，实现了相关疾病、症状、药物、病因、诊断、治疗等知识的关联，提供了力导向布局图和弦图两种可视化展示方式，且支持人机交互以及图片下载保存的功能。在知识图谱可视化的基础上，用户可以直观获取上述知识间的关系，并能通过人机交互的方式、根据个性化需求生成满意的图谱，从而为深入开展相关科学研究和知识潜在关联关系的发现奠定基础。

目前，知识图谱技术仅应用于“医药卫生专业知识服务系统”中“百科数据”的展示，尚未实现真正意义上的语义搜索。下一步，我们将继续拓展医学专家、研究机构、专利、报告、文献等类型资源，抽取各类资源所包含的概念、实体、属性及其关系，围绕共同概念和实体整合不同类型、不同来源的知识，形成丰富、多元的知识库，从而构建全面、巨大的知识图谱，为“中国工程科技知识中心医药卫生专业

知识服务系统”提供全面的知识检索和实体链接方法，提高知识检索和获取的效率。

【参考文献】

[1] 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，2018，35(7)：1929-1936.

[2] 于 彤，刘 静，贾李蓉，等.大型中医药知识图谱构建研究[J].中国数字医学，2015，10(3)：80-82.

[3] 陈 悦，刘则渊.悄然兴起的科学知识图谱[J].科学学研究，2005，23(2)：149-154.

[4] 陈 悦，刘则渊，陈 劲，等.科学知识图谱的发展历程[J].科学学研究，2008，26(3)：449-460.

[5] Rotmensch M，Halpern Y，Tlimat A，*et al*.Learning a health knowledge graph from electronic medical records[J].Scientific Reports，2017(7)：5994.

[6] Wang M，Zhang J，Liu J，*et al*.PDD Graph：Bridging electronic medical records and biomedical knowledge graphs via entity linking[C].Berlin，German：International Semantic Web Conference，2017.

[7] Shi L，Li S，Yang X，*et al*.Semantic health knowledge graph：semantic integration of heterogeneous medical knowledge and service[J].BioMed Research International，2017(4)：1-12.

[8] 黄 鑫，胡榜利，邓 莉，等.基于知识图谱的生物医学信息可视化研究进展[J].中国临床新医学，2012，5(11)：1090-1093.

[9] 石习敏，陈 娟，杨均雪，等.基于知识图谱的国内外医学数据挖掘研究可视化探析[J].中国全科医学，2017，20(21)：2623-2628.

[10] 谢华鑫，何小菁.基于知识图谱的医学人文研究热点分析[J].南京医科大学学报：社会科学版，2017，17(1)：47-51.

[11] 李新龙，刘 岩，何丽云，等.知识图谱研究概况及其在中医药领域的应用[J].中国中医药信息杂志，2017，24(7)：129-132.

[12] 张德政，谢永红，李 曼，等.基于本体的中医知识图谱构建[J].情报工程，2017，3(1)：35-42.

[13] 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.

[14] 吴运兵，阴爱英，林开标，等.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(3)：329-335.

[15] 吴运兵，杨 帆，赖国华，等.知识图谱学习和推理研究进展[J].小型微型计算机系统，2016，37(9)：2007-2013.

[16] 郑 强，刘齐军，王正华，等.生物医学命名实体识别的研究与进展[J].计算机应用研究，2010，27(3)：811-815.

[17] 康宏宇，李 姣.生物医学文献的知识发现与数据整合[J].中华医学图书情报杂志，2015，24(2)：15-20.

[18] 漆桂林，高 桓，吴天星.知识图谱研究进展[J].情报工程，2017，3(1)：4-25.

[19] 孙紫阳，顾君忠，杨 静.基于深度学习的中文实体关系抽取方法[EB/OL].(2017-10-27)[2018-04-16].http://kns.cnki.net/kcms/detail/31.1289.TP.20171027.1109.006.html.

[20] 徐 健，张智雄，吴振新.实体关系抽取的技术方法综述[J].现代图书情报技术，2008(8)：18-23.

[21] Nguyen DPT，Matsuo Y，Ishizuka M.Subtree mining for relation extraction from Wikipedia[C].New York：Human Language Technology Conference of the North American，2007.

[22] 袁 润，李广平.基于 CiteSpace 的图书馆微博与微信研究论文的比较分析[J].图书情报研究，2018，11(2)：69-74.

[23] 王 露，杨晶晶，黄 铭.基于 R 语言和 Tableau 的气象数据可视化分析[J].计算机与网络，2017，43(24)：69-71.

[24] 张永安，马 昱.基于 R 语言的区域技术创新政策量化分析[J].情报杂志，2017，36(3)：113-118.

[收稿日期：2018-05-30]

[本文编辑：黄思敏]