医学知识图谱构建技术与研究进展

作者	袁凯琦，邓扬，陈道源，张冰，雷凯，沈颖
机构	北京大学深圳研究生院 深圳市云计算关键技术与应用重点实验室
基金项目	国家自然科学基金青年基金资助项目（61602013），深圳市科创委基础研究项目（JCYJ20160330095313861；JCYJ20151030154330711；JCYJ20151014093505032）
预排期卷	《计算机应用研究》 2018 年第 35 卷第 7 期
摘要	医学知识图谱是实现智慧医疗的基石，有望带来更高效精准的医疗服务。然而，现有知识图谱构建技术在医学领域中普遍存在效率低，限制多，拓展性差等问题。针对医疗数据跨语种，专业性强，结构复杂等特点，对构建医学知识图谱的关键技术进行了自底向上的全面解析，涵盖了医学知识表示、抽取、融合和推理以及质量评估五部分内容。此外，还介绍了医学知识图谱在信息检索、知识问答、智能诊断等医疗服务中的应用现状。最后，结合当前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	知识图谱；知识获取；知识融合；知识推理；自然语言处理
作者简介	袁凯琦（1993-），女，广东湛江人，硕士研究生，主要研究方向为知识图谱，知识推理；邓扬（1993-），男，硕士研究生，主要研究方向为自然语言处理、知识图谱、自动问答、信息抽取；陈道源（1994-），男，硕士研究生，主要研究方向为知识图谱、知识表示学习；张冰（1993-），女，硕士研究生，主要研究方向为知识图谱，知识融合；雷凯（1976-），男，主任，副研究员，硕导，博士，主要研究方向为计算机网络、命名数据网络、社交网络、文本处理等；沈颖（1984-），女（通信作者），博士，主要研究方向为自然语言处理，知识图谱，医学信息学，大数据管理 .
中图分类号	TP391
访问地址	http://www.arocmag.com/article/02-2018-07-068.html
发布日期	2017 年 8 月 17 日
引用格式	袁凯琦，邓扬，陈道源，张冰，雷凯，沈颖. 医学知识图谱构建技术与研究进展[J/OL]. 2018, 35(7). [2017-08-17]. http://www.arocmag.com/article/02-2018-07-068.html.

医学知识图谱构建技术与研究进展 *

袁凯琦，邓 扬，陈道源，张 冰，雷 凯，沈 颖†

(北京大学深圳研究生院 深圳市云计算关键技术与应用重点实验室，广东 深圳 518055)

摘 要：医学知识图谱是实现智慧医疗的基石，有望带来更高效精准的医疗服务。然而，现有知识图谱构建技术在医学领域中普遍存在效率低，限制多，拓展性差等问题。针对医疗数据跨语种，专业性强，结构复杂等特点，对构建医学知识图谱的关键技术进行了自底向上的全面解析，涵盖了医学知识表示、抽取、融合和推理以及质量评估五部分内容。此外，还介绍了医学知识图谱在信息检索、知识问答、智能诊断等医疗服务中的应用现状。最后，结合当前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：知识图谱；知识获取；知识融合；知识推理；自然语言处理

中图分类号：TP391

Construction techniques and research development of medical knowledge graph

Yuan Kaiqi, Deng Yang, Chen Daoyuan, Zhang Bing, Lei Kai, Shen Ying†

(*Institute of Big Data Technologies Shenzhen Key Lab for Cloud Computing Technology & Applications. School of Electronic and Computer Engineering* (*SECE*)*Peking University*, *Shenzhen Guangdong* 518055, *China*)

Abstract: Medical knowledge graph is the cornerstone of Artificial Intelligence and Smart Healthcare, and leading to more efficient and accurate medical service. However, current constructing techniques of knowledge graph have some common defects in efficiency, scalability and applicability. Considering the specific features of medical data, this paper analyzes and classifies the key techniques and methods involved in the construction of medical knowledge graph in a bottom-up way, including representation, extraction, fusion and reasoning of medical knowledge and quality assessment of medical knowledge graph. Furthermore, the paper also introduces the research and application of search engine, question-answering system and decision support system based on medical knowledge graph. Finally, the paper summarizes challenges and major problems of medical knowledge graph, and prospects for future development.

Key Words: knowledge graph; knowledge extraction; knowledge fusion; knowledge reasoning; natural language processing

0 引言

自 1998 年万维网之父 Tim Berners-Lee 提出语义网，人们不断在网络等电子载体上表达和修正对客观世界的理解，形成了一个概念标准化的过程，同时随着链接开放数据 (Linked Open Data)的规模激增，互联网上散落了越来越多的知识元数据。

知识图谱就是在这样的大数据背景下产生的一种知识表示和管理的方式，强调语义检索能力。近年来，在人工智能的蓬勃发展下，知识图谱涉及到的知识抽取、表示、融合、推理、问答等关键问题得到一定程度的解决和突破，知识图谱成为知识服务领域的一个新热点，受到国内外学者和工业界广泛关注。

知识图谱的前身是语义网，它吸收了语义网、本体在知识组织和表达方面的理念，使得知识更易于在计算机之间和计算机与人之间交换、流通和加工。具体来说，一个知识图谱由模式图、数据图及两者之间的关系组成：模式图对人类知识领域的概念层面进行描述，强调概念及概念关系的形式化表达，模式图中节点是概念实体，边是概念间的语义关系，如 part-of；数据图对物理世界层面进行描述，强调一系列客观事实。数据图中的节点有两类，一是模式图中的概念实体，二是描述性字符串，数据图中的边是具体事实的语义描述；模式图和数据图之间的关系指数据图的实例与模式图的概念之间的对应，或者

基金项目：国家自然科学基金青年基金资助项目（61602013），深圳市科创委基础研究项目（JCYJ20160330095313861；JCYJ20151030154330711；JCYJ20151014093505032）

作者简介：袁凯琦（1993-），女，广东湛江人，硕士研究生，主要研究方向为知识图谱，知识推理；邓扬（1993-），男，硕士研究生，主要研究方向为自然语言处理、知识图谱、自动问答、信息抽取；陈道源（1994-），男，硕士研究生，主要研究方向为知识图谱、知识表示学习；张冰（1993-），女，硕士研究生，主要研究方向为知识图谱，知识融合；雷凯（1976-），男，主任，副研究员，硕导，博士，主要研究方向为计算机网络、命名数据网络、社交网络、文本处理等；沈颖（1984-），女（通信作者），博士，主要研究方向为自然语言处理，知识图谱，医学信息学，大数据管理.

说模式图是数据图的模具。

著名的通用知识图谱中有，谷歌“Knowledge Graph”[1]、搜狗“知立方”[1]、YAGO[2]、DBpedia[3]等，它们具有规模大、领域宽，包含大量常识等特点。目前，医学是知识图谱应用最广的垂直领域之一，如上海曙光医院构建的中医药知识图谱[4]、本体医疗知识库 SNOMED-CT[2]，IBM Watson Health[3]等应用近两年也开始进入人们视线。

知识图谱是智能大数据的前沿研究问题，它以独有的技术优势顺应了信息化时代的发展，比如渐增式的数据模式设计；良好的数据集成；现有 RDF、OWL 等标准支持；语义搜索和知识推理能力等。在医学领域，随着区域卫生信息化及医疗信息系统的发展，积累了海量的医学数据。如何从这些数据中提炼信息，并加以管理、共享及应用，是推进医学智能化的关键问题，是医学知识检索、临床诊断、医疗质量管理、电子病历及健康档案智能化处理的基础。

本文首先介绍了医学知识图谱的构建技术，主要涉及到医学知识表示；医学知识抽取，包括实体、关系、属性的抽取；医学知识融合；医学知识推理和质量评估五个部分。然后介绍了基于医学知识图谱的包括检索、问答、决策等的应用现状。最后对医学知识图谱的研究、应用重点，面临的挑战以及未来的发展趋势进行了展望。

1 医学知识图谱构建

本文将医学知识图谱构建技术归纳为五部分，即医学知识的表示、抽取、融合、推理以及质量评估。通过从大量的结构化或非结构化的医学数据中提取出实体、关系、属性等知识图谱的组成元素，选择合理高效的方式存入知识库。医学知识融合对医学知识库内容进行消歧和链接，增强知识库内部的逻辑性和表达能力，并通过人工或自动的方式为医学知识图谱更新旧知识或补充新知识。借助知识推理，推出缺失事实，自动完成疾病诊断与治疗。质量评估则是保障数据的重要手段，提高医学知识图谱的可信度和准确度。

1.1 医学知识表示

知识表示是为描述世界所做的一组约定，是知识符号化、形式化、模式化的过程[5]，主要研究计算机存储知识的方法，其表示方式影响系统的知识获取、存储及运用的效率。然而医学数据种类繁杂，存储方式不一，电子病历格式和标准不同，经常涉及交叉领域等特点，导致医学领域与其他领域在知识表示方面有所差异，同时也给医学领域的知识表示带来极大的挑战。

早期医疗知识库运用的知识表示方法有：谓词逻辑表示法，产生式表示法，框架表示法，语义网表示法等等。比如，SNOMED-CT，早期的 MYCIN 系统[6]，大肠杆菌数据库 EcoCyc[7]等。随着知识图谱中知识增长、关系复杂化，这些方法由于表示能力有限且缺乏灵活性，不再作为主要的知识表示方法，更多是作为医学知识表示的辅助或补充。

本体表示法以网络的形式表示知识，即以（实体 1，关系，实体 2）三元组来表示相关联的两个节点（实体），在知识图谱[8]提出之后逐渐得到认可。它借鉴了语义网表示法但又有所区别：本体关注的是实体固有特征，比后者更聚焦，更深入，因而也具有更大的发展潜力。而本体的描述语言也多种多样：主要有 RDF 和 RDF-S、DAML、OWL 等。使用本体表示医学术语可以提升数据整合能力：建立强大、可互操作的医疗信息系统；满足重用共享传输医疗数据的需求；提供基于不同语义标准的统计聚合。 医学领域本体的构建，需要深入分析医学术语的结构和概念，才能将晦涩甚至是跨语言的医学知识有效地表达出来。目前的医学知识本体库有：医学概念知识库 LinkBase[9]，TAMBIS 本体库 (TaO)[10]等等。

知识图谱的节点个数影响着网络的结构复杂度及推理的效率和难度。知识表示学习借助机器学习，将研究对象的语义信息表示为稠密低维向量，有效解决数据稀疏问题，从而提升知识融合和推理性能[11]。低维向量表示是一种分布式表示（distributed representation）[12]，它模仿人脑中使用多个神经元存储对象的工作机制，使用多维度向量表示对象的语义信息。

知识表示学习中的代表模型有：结构化表示法（Structure Embedding，SE）[13]，单层神经网络模型（single layer model，SLM）[14]，隐变量模型（latent factor model，LFM）[15]，基于 TransE[16]的翻译模型等等。这些模型考虑实体间的协同性和计算开销，用向量表示实体，再对表示实体的向量或关系进行相应的矩阵变换，提出评价函数来衡量实体间的相关性，并为之后的知识补全和推理提供重要参考。Kleyko 等人[17]证明了分布式表示方法表示医学图像进行分类精度能够与最佳经典方法相同；Henriksson 等人[18]对比使用多种知识表示方法表示 EHR 中 4 类记录：诊断记录，药物使用记录，治疗方法和病程记录。显然，知识表示学习无疑为医学知识图谱的知识表示开辟了新思路。

1.2 医学知识抽取

医学知识图谱的构建主要是从非结构化数据中人工或自动地提取实体、关系和属性。人工提取是通过专家依据一定规则收集并整理相关信息，提取知识。目前通过人工构建的医学知识库包括临床医学知识库[19]、SNOMED-CT、ICD-10 等。自动提取则是利用机器学习、人工智能、数据挖掘等信息抽取技术，从数据源中自动提取出知识图谱的基本组成元素。自动构建医学知识库的典型例子有一体化医学语言系统 UMLS[20]。人工提取的代价太大，知识的自动提取是目前重点的研究方向，也是将来构建知识图谱的趋势。本节主要介绍如何自动从数据源中抽取知识和信息，包括实体、关系和属性抽取。

1.2.1 实体抽取

识别文本中的生物医学实体，其目的在于通过识别关键概念进一步提取关系和其他信息，并将识别的概念以标准化的形

[1] https://www.sogou.com/
[2] http://www.snomed.org/
[3] http://www-935.ibm.com/industries/hea lthcare/index.html

式表示出来。医学领域的实体抽取是从医学数据源中提取出特定类型的命名实体。本节将医学实体的抽取方法归纳为三类：基于医学词典及规则的方法、基于医学数据源的统计学和机器学习方法以及深度学习方法。

1）基于医学词典及规则的方法

该方法通过人工定义规则和模式匹配生成词典或使用现有医学词典从语料中抽取医学实体，该方法是具有挑战性的。首先，目前没有完整的字典囊括所有类型的生物命名实体，所以简单的文本匹配算法是不足以应对实体识别的。其次，相同的单词或短语其意义可根据上下文的改变而指代不同的物体（如，铁蛋白可以是生物物质或实验室测试方法）。再次，许多生物或药物实体同时拥有多个名称（如 PTEN 和 MMAC1 指代相同的基因）。因此，基于医学词典及规则只在最早期被广泛使用。Friedman 等人[21]通过自定义语义模式和语法来识别电子病历中的医学信息。Wu[22]等人使用了 CHV[23]和 SNOMED-CT 两个医学词典得到了不错的实验结果。虽然该方法能达到很高的准确度，但无法彻底解决上述问题，也过分依赖专家编写的词典和规则，无法适应医学领域词汇不断涌现的现实情况。

2）基于医学数据源和数学模型的机器学习方法

该方法通过使用统计学和机器学习方法，结合医学数据源的特点训练模型，进行实体识别。在英文医学实体抽取方面，最具代表性的标注语料是 I2B2 2010[24]发布的英文电子病历标注语料。另外，还有 SemEval[4]、NTCIR[5]等评测，以及 NCBI[25]语料库等，都提供了英文医学实体标注数据。

目前常用方法有隐马尔可夫模型（HMM），条件随机场模型（CRF），支持向量机模型（SVM）等。Kazama 等人[26]使用 SVM 模型进行生物医学命名实体识别，引入了 POS，词缀存，无监督训练得到的 HMM 状态等特征。该方法在 GENIA 语料库中准确率高于最大熵标记方法，并能较高效地应用于大规模语料集。Zhou 等人[27]通过一系列特征训练 HMM 模型，包括词的构成特征，形态特征，POS，语义触发，文献内名称别名等。其识别准确率达 66.5%，在 GENIA 语料库中的召回率达 66.6%。综合以上方法，Chen 和 Friedman[28]利用 MEDLEE 系统来识别与生物医学文本中与表型信息相对应的短语。该系统使用自然语言技术来识别期刊文章摘要中存在的表型短语。生物医学的实体识别常常可使用较小的表型相关术语的知识库。Chen 和 Friedman[29]自动导入与语义类别相关的数千个 UMLS 术语，如细胞体功能和细胞功能障碍，以及哺乳动物本体中的几百个术语；并手动添加了几百个术语。实验结果表明，其实体识别准确率达 64.0%，召回率达 77.1%。虽然结果不高，但为之后的研究人员提供了一条可行的思路。

在医学领域，命名实体识别的痛点在于数据质量的良莠不齐以及人工标注的专业性要求高。目前有专门研究如何降低对于数据标注依赖的研究，其原理主要是利用海量未标注数据持续提升模型性能，从小样本中进行学习，自我探索逐步学习新知识，形成一个交互学习过程。

3）深度学习方法

深度学习近年来开始被广泛应用于命名实体识别，最具代表性的模型是 2011 年 Collobert[30]提出的一个深层神经网络模型，其效果和性能超过了传统算法。Sahu 等人[31]所提出的 CNN 与 RNN 级联的方法生成词嵌入特征，其结果优于目前最好的算法且不需要过多的特征工程。

在医学领域，We 等人[32]基于 CRF 和双向 RNN 生成特征，再使用 SVM 进行疾病命名实体识别。目前医学信息命名实体识别任务中最主流的深度学习模型是 BiLSTM-CRF 模型，Jagannatha 等人[33]对比了 CRF，BiLSTM，BiLSTM-CRF 三种模型，以及一些它们的改进模型，在英文电子病历命名实体识别的效果，实验结果表明所有基于 LSTM 的模型都比 CRF 效果更好，并且 BiLSTM 结合 CRF 模型能够进一步提高评测结果 2%-5%的准确率。

1.2.2 实体抽取

本文将医学实体关系抽取归结为两类：a）同类型医学实体层级关系抽取，如疾病的“肠胃病-慢性胃炎”等；b）不同类型关系抽取，如“疾病-症状”等。

1）同类型医学实体层级关系抽取

同类型医学实体层级关系相对较为单一，主要是 is-a 和 part-of 关系。由于医学有其严谨的学科体系和行业规范，因此此类关系往往在医学词典、百科、信息标准中进行。

ICD-10[34]、SNOMED 等医疗词典或医疗数据库重点关注医学专业术语、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，是抽取层级间实体关系的首选来源。针对具体的医疗词典、知识库提供的数据格式和开放 API 接口，可通过爬虫、正则表达式、D2R 映射等技术从中抽取分层结构，抽取三元组来匹配、添加上下位关系。

2）不同类型医学实体关系抽取

不同类型医学实体间的语义关系识别大致基于两大不同数据源而实现。一是百科或其他结构化数据源，如 Medline，UMLS 等；二是半结构化的电子病历。

医学实体类型相对有限（主要是疾病、症状、治疗、药品等），目前通常在两个实体间预定义好要抽取的关系类型，再将抽取任务转换为分类问题来处理。如何预定义实体关系目前尚未有统一的标准，这取决于医学知识图谱构建过程中模式图的设置、实体识别情况、语料来源、构建目的及应用场景等，如在 I2B2 2010 评测中，将电子病历中的实体关系分成了医疗问题与医疗问题、医疗问题与治疗、医疗问题与检查三类。

近几年，Uzuner 团队[35]在句子层面抽取了六类医疗实体关系，使用实体顺序和距离、链接语法和词汇特征来训练 6 个 SVM 分类器，通过对比实验，指出词汇特征在实体关系识别中的重要作用。在此基础上，基于 Medline 摘要，Frunza 等[36]

[4] http://www.senseval.org/
[5] http://research.nii.ac.jp/ntcir

抽取了疾病、治疗间的三种关系，并引入 UMLS 生物和医疗实体特征，取得了不错的实验结果。而 Abacha 等[37]在同样的任务中使用人工模板和 SVM 的混合模型，取得了 94.07%的平均 F 值。该研究指出，在样本数较少时，模板匹配方法起主要作用，而面向海量样本时则 SVM 起主要作用。

此外，在关系识别的分类方法对比研究中，Bruijn 等人[38]在 I2B2 2010 评测中对比研究了有监督分类和基于 Self-training 的半监督分类的表现，表明了 UMLS、依存句法分析结果和未标记数据对关系识别有着显著影响。除了预定义关系然后转换为分类任务来处理的方法，还有少量研究采用了模板匹配、统计共现等方法来抽取关系。如在 Medline 摘要中通过统计基因名的共现来提取关系，并根据共现矩阵生成了关系图[39]，或在 Medline 摘要中通过语法依赖树进行图的模式匹配，进而抽取因果关系[40]。

1.2.3 属性抽取

属性抽取是指对属性和属性值对（attribute-value pair，AVP）的抽取，其中属性的抽取是指为医学实体构造属性列表，如药品的属性包括适应症、禁忌症等。属性值的抽取是指为各实体附加具体的属性值，如阿莫西林是青霉素过敏者禁用。常见的抽取方法包括从开放链接数据提取[41]、从结构化数据库提取、从百科类站点提取、从垂直网站进行包装器归纳、以及利用模式匹配从查询日志中提取等。对于医学知识图谱来说，主要通过上文提及的医学词典和主流医学站点来进行。值得一提的是，前者关于属性和属性值对的描述相对比较稀疏（特别对于中文领域），因此需从主流医学站点进一步抽取整合。

对于 AVP 结构化程度比较高的网站，如维基百科、A+医学百科[6]等，有规整的信息框（InfoBox），可以方便地爬取识别 InfoBox 部分，直接提取该实体对应的属性名和属性值，置信度高，但规模较小。而对于更多形式各异、半结构化的医药站点和垂直文本来说，通常是构建面向站点的包装器，从待抽取站点采样并标注几个典型的详细页面(Detailed Pages)，利用这些页面通过模式学习自动构建出一个或多个类 Xpath 表示的模式，然后将其应用在该站点的其他详细页面中从而实现自动化的 AVP 抽取。

1.3 医学知识融合

知识融合是高层次的知识组织[42]，使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、更新等步骤[7]，目的是解决知识复用的问题，增强知识库内部的逻辑性和表达能力。针对知识图谱中不同粒度的知识对象，知识融合可细分为实体对齐、知识库的融合等。

1.3.1 实体对齐

医疗知识图谱中知识来源的多样性导致了知识重复、知识质量良莠不齐、知识间关联不够明确等问题。 医学实体在不同的数据源中存在严重的多元指代问题，例如阿奇霉素在百度百科中被称为希舒美，在 A+医学百科中别名有阿齐霉素、阿奇红霉素、叠氮红霉素等，商品名有泰力特、希舒美、舒美特等。因此实体对齐是医学知识融合中非常重要的一步。实体对齐[43]是判断多源异构数据中的实体是否指向真实世界同一对象的过程。

现有的对齐算法有成对实体对齐与集体实体对齐两类。成对实体对齐方法只考虑实例及其属性相似度，包括基于传统概率模型的实体对齐方法和基于机器学习的实体对齐方法。前者以 Fellegi 等人[44]将基于属性相似性评分的实体对齐问题转换为分类问题的工作为基础，至今仍应用于很多实体对齐工作中；后者常用的对齐方法有分类回归树算法[45]、ID3 决策树算法[46]、SVM 分类方法[47]、集成学习框架[48]等基于有监督学习的方法和无监督学习下的层次图模型[49]等基于无监督学习的方法。

集体实体对齐在成对实体对齐的基础上在计算实体相似度时加入了实体间的相互关系，分为局部集体实体对齐与全局集体实体对齐。前者典型算法是使用向量空间模型和余弦相似度计算实体相似性[50]，准确率不高，但召回率和运行速度比较可观。后者通过不同匹配决策之间的相互影响来调整实体间的相似度，又分为基于相似性传播和基于概率模型的集体实体对齐方法。基于相似性传播的方法通过初始匹配以“bootstrapping”方式迭代地产生新的匹配[51]。Lacoste-Julien 等人[52]在此基础上提出的 SiGMa 算法更适合大规模知识库，但需要一定的人工干预。基于概率模型的方法通过为实体匹配关系和决策建立复杂的概率模型，包括关系贝叶斯网络模型[53]、LDA 分配模型[54]、CRF 模型[55, 56]和 Markov 逻辑网模型[57, 58]等，可以提高匹配效果，但效率还有待提高。

当来自不同知识源的数据出现数据冲突时，需要考虑知识源的可靠性以及不同信息在各知识源中出现的频度等因素。阮彤等人[59]在构建中医药知识图谱时对数据源的可信度进行评分，结合数据在不同来源中出现的次数，对数据项进行排序，并补充到相应的属性值字段中。

随着知识库规模扩大和实体数量的增加，知识库中的实体对齐越来越受到重视，如何准确高效地实体对齐是未来知识融合的研究重点之一。

1.3.2 知识库融合

构建知识库时需求和设计理念不同会导致知识库中数据的多样性和异构性。对于庞杂的医疗知识来说，当前多数知识库都是针对某个科室或者某类疾病或药物来构建的，比如脾胃病知识库[60]、中医药知识图谱[4]等，若要得到更完善的医疗知识图谱，需要对不同的医疗知识库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是一个不断迭代更新的过程。

知识库融合的研究工作始于“本体匹配”，初期针对本体类别的语义相似性进行匹配[61]。随着知识库规模扩大和结构复杂化，类别、属性以及实体和它们之间的相互关系等也成为考虑的因素。Suchanek 等人提出的基于概率的知识融合算法

[6] http://www.a-hospital.com/
[7] 张坤.面向知识图谱的搜索技术(搜狗) http://www.cipsc.org.cn/kg1/

PAIRS[62]以两个知识库作为输入，能够高效地跨本体同时对齐类别、实例、属性和关系。但 PARIS 需要一定的人工参与。由于人类的精力和认知有限，自动地从 Web 中获取知识并进行融合十分必要。Dong 等人[63]提出了将以消除歧义的三元组的形式从整个网络中提取事实与使用 PRA 和神经网络模型两种方法从 Freebase 图得到的先验知识融合在一起的知识融合方法，可达到自动构建 Web 规模的概率知识库的水准，提升了效率。

在医疗领域，Dieng-Kuntz 等人[64]将医疗数据库转换为医疗本体，然后对其他文本语料使用半自动的语言工具进行语义提取，在人工控制下对本体进行扩展和补全，并用启发式规则自动建立知识的概念层次。Baorto 等人[65]将数据源添加到临床信息系统时先确定数据的控制术语是否已经存在，然后将新术语添加到 MED(Medical Entities Dictionary)，同时建立审计流程以保证引入数据的一致性。

目前医疗领域知识图谱的融合技术虽有一些有意义的尝试，但仍需要大量人工干预，高效的知识融合算法有待进一步研究。医疗领域的知识图谱也可以考虑采用众包的方式进行知识融合。

1.4 医学知识推理

推理是从已有知识中挖掘出隐含信息，而知识推理更注重知识与方法的选择与运用，尽量减少人工参与，推出缺失事实，完成问题求解。在医学知识图谱中，知识推理帮助医生完成病患数据搜集、疾病诊断与治疗，控制医疗差错率。然而，即使对于相同的疾病，医生也会根据病人状况作出不同的诊断，即医学知识图谱必须处理大量重复矛盾的信息，这就增加了构建医学推理模型的复杂性。

传统的知识推理方法有基于描述逻辑（Description Logic，DL）推理[66]，基于规则推理（Rule-based Reasoning，CBR）[67]与基于案例推理（Case-based Reasoning，CBR）[68]等等。Bousquet C 等人使用基于 DAML+OIL 描述逻辑执行术语推理来改进药物警戒系统中信号检测[69]；Chen R 等人采用 RBR 方法开发出糖尿病诊断系统以提供用药建议[70]；CARE-PARNER 系统则是基于 CBR 给出诊断结果与治疗方案[71]等等。

传统的知识推理方法虽在一定程度推动医疗诊断自动化进程，但是也存在学习能力不足，数据利用率不高，准确率待提升等明显缺陷，远未达到实际应用的要求。面对日益增长的医疗数据，诊断时不可避免地会出现信息遗漏，诊断时间延长等问题。而人工智能，尤其是人工神经网络（Artificial Neural Networks，ANNs）拥有从海量数据挖掘有用信息的天然优势。ART-KNN（ART-Kohonen neural network）与 CBR 相结合可以提高后者在推理故障情况的效率和准确度[72]。神经张量网络模型（neural tensor networks）在 FreeBase 等开放本体库上对未知关系推理的准确率可达 90.0%[73]。Karegowda A G 等人在 Pima 印第安人糖尿病知识库（PIDD）中使用了遗传算法（Genetic Algorithm，GA）和反向传播网络（Back Propatation Network，BPN）的混合模型，诊断准确率提高 7%左右[74]。

与深度学习将知识图谱作为数据源不同，基于图的推理则将知识图谱视为图，以实体为节点，以关系或属性为边，利用关系路径来找到节点间的多步路径，Path Ranking 算法[75]，PTransE[76]就是利用这样的原理来推断实体间的语义关系。图数据库[77]使知识图谱能以图的数据结构进行存储，与传统数据库相比，前者在高维度关联查询的效率明显提高。然而图数据库尚未成熟，暂无法完成太复杂的知识推理。较为流行的图数据库有 Neo4j[78]，Titan[79]，OrientDB[80]和 ArangoDB[81]等。王昊奋等人在医疗质量与患者安全辅助监控系统中，就将医疗本体的数据存放在 AllegroGraph 图数据库中[27]。

与通用知识图谱相类似，医学知识图谱也还有对跨知识库知识推理、基于模糊本体的知识推理等问题的研究。

1.5 质量评估

数据的质量直接影响数据的运用，质量评估是保障数据的重要手段，可以量化数据质量，筛选出置信度高的数据。医学诊断对数据和医学知识图谱的可信度和准确度提出了更高的要求。质量评估并不是构建医疗知识图谱的最后一步，而是贯穿在知识图谱的整个生命周期：早在 2013 年本体峰会（Ontology Summit2013）就对本体的生命周期的各个阶段所要进行的评估工作做了相关的说明[82]。

目前，知识图谱/本体的评估方法可分为四大类：基于黄金标准的方法，基于本体任务/应用的方法，数据驱动的方法，和基于指标的方法。表 1 对比了这几种本体评估方法。

具体到医学领域，本体评估方法因应用场景而异：Clarke 等人使用基于任务评估方法来分析基因本体的从 2004 至 2012 年的性能[83]；Bright 等使用本体设计原则和领域专家审查意见作为指标来评估本体在抗生素决策支持系统中的效果[84]；Gordon 等通过将电子病历，诊断案例和临床实践等来构造“黄金标准”来评估，改进传染疾病本体 BCIDO[85]等。为方便用户进行本体评估和加快本体评估自动化进程，本体评估工具封装了评估方法。不同的工具从不同的视角对本体的不同指标进行评估。针对评估的侧重点，选择合适的工具才能对本体作出符合应用要求的评估。

比较常用的本体评估工具有：ODEval[86]，OOPS！[87]，OntoManager[88]，Core[89]等。除了医学本体，医学知识图谱还涵盖其他复杂多样的信息，因此，质量评估还包括对数据质量，专家信息，知识库等方面进行评估。

相比通用领域的知识图谱，医学领域的知识图谱评估存在以下特殊性：a）鉴于医学的严谨性，评估往往综合多种方法进行多角度的评估，如 Bright 等利用本体设计原则和领域专家审查来进行等级评估[84]；b）往往需设置等级较高的警告（alerts），如与处方相关的警报包括抗生素-微生物不匹配警报、用药过敏警报、非推荐的经验性抗生素治疗警报、治疗方案-症状间不匹配警报等;c）除了从形式方面评价知识图谱以外，也注重于检验知识差距，因为知识的全面性和准确性将直接影响临床决策支持的置信度。此外，医学知识图谱是融合计算机科学，医学

等众多学科的交叉学科，评价指标不能简单地照搬某个的学科 的指标，而是应该综合考虑众多因素。

表 1 知识图谱/本体评估方法

评估方法	介绍	评价层次	优点	缺点
基于黄金标准的本体评估	将所构建的本体与一个现有公认的比较成熟的“黄金标准”（本体）比较，罗列出其不足并进行改进。	词汇数据层，层级分类层 ，语法层，语义层	客观	“黄金指标”可能难以寻找或并不存在。
基于本体任务/应用的本体评估	一个特定应用环境中，测试一组本体，看哪个本体最适合该应用，这些应用包括语义网、信息抽取、信息检索等。	词汇数据层，层级分类层，语义层，内容或应用层	直接面向本体应用，富有针对性	依赖具体应用，可移植性和通用性较差
数据驱动的本体评估	通过衡量本体与领域语料的匹配度或本体的领域覆盖度来评估本体，或使用其他参考数据来辅助本体评估过程，这种方法常常和文本分析、 机器学习技术结合。	词汇数据层，层级分类层，语义层	相对于专家主观评估更为客观，方法新颖多样，易于验证	领域本体的选择本身也存在评估问题，评估范围主要限定于本体概念覆盖度
基于指标的本体评估	基于一套预先定义好的原则，准则，标准等进行评估的方法，其多是从构建本体的原则来评估本体。	词汇数据层，层级分类层，语义层，内容或应用层，语法层，系统架构设计层	指标体系具有全面性和通用性	会出现指标过多难以统一的问题

因此，除了借鉴通用知识图谱质量评估方法，Ammenwerth E 等人结合医学本身的特点，就评价医学知识图谱时面临的评价对象的复杂度，评价指标与利益相关者的关系，评价动机三大问题出发，分别提出解决方案[90]。

不难看出，医学知识图谱质量评估的研究主要集中在方法，工具以及的研究这三方面。然而，现有的质量评估缺乏系统化，鲜有从医学领域特性的角度对知识图谱进行综合评估，多是借鉴通用的质量评估方法从微观上去评估某一项指标，具有分散性，片面性。

2 医学知识图谱应用

随着人们对健康问题的愈发重视，医疗保健费用、需求的增长与优质医疗资源不足之间的矛盾亟待解决。对此，由于近几年人工智能的飞速发展以及精准医疗、智慧医疗的提出，医学知识图谱应用正受到国内外企业、学界的广泛关注，有望带来更廉价、高效、精准的医疗建议和诊断。本文可以从已有的医学知识图谱的应用中一窥其巨大潜力。

2.1 医疗信息搜索引擎

传统的医疗搜索引擎需要对百亿计的医疗相关网页进行检索、存储、处理，但难以理解用户的语义查询。而基于医学知识图谱的搜索，不仅提供用户网页间超链接的文档关系，还包括不同类型实体间丰富的语义关系。

知识图谱对于传统信息搜索的优化主要体现在查询扩展，从知识图谱中抽取与查询相关的若干实体及实体关系和属性进行扩展查询，以更好的理解用户的查询需求。Aronso 等人[91]较早将信息检索技术结合 UMLS 进行查询扩展，并应用于医学文献检索。Díazgaliano 等人[92]在生物医学信息搜索中加入了医学本体 MeSH[93]来进行查询扩展，包含同义、近似同义和密切相关概念的实体和关系，改进了信息检索的效果。Huang 等人[94]在医学本体的基础上，使用 LSA 自动挖掘实体间的语义关系，如药物诱导疾病关系、药物间相互作用关系等，对实体-关系、实体-实体的查询进行了扩展。中国中医科学院的贾李蓉等于 2002 年开始研制中医药学语言系统[95]，构建了包含 12 万多个概念、60 余万术语以及 127 余万语义关系的中医药知识图谱[96]。它通过在检索系统中嵌入“知识卡片”以及一个“知识地图”展示系统，将中医领域概念可视化，用户可以选择其中的概念开始构造查询或搜索。

目前，国外典型的医疗专用搜索引擎有 WebMd[8]、OmniMedicalSearch[9]、Healthline[10]等。其中 WebMd 和 OmniMedicalSearch 分别属于全文索引和目录索引类型的传统搜索引擎，Healthline 是一个基于知识库的医学信息搜索引擎，其知识库涵盖超 850，000 项医疗元数据和 50,000 条相互关联的概念。Google 率先提出将知识图谱应用于搜索，在医疗搜索应用上，当用户搜索疾病或症状时，Google 提供超过 400 种健康状况的数据，通过一张信息卡片，给出典型症状以及如是否严重、是否具有传染性、影响哪些年龄段的人等细节信息。而国内主流医疗搜索引擎有搜狗名医[11]、360 良医[12]，这两者都是结合了元搜索索引方式和知识库的搜索引擎，聚合权威的知识、医疗、学术网站，为用户提供包括维基百科、知乎问答、国际前沿学术论文等权威、真实内容。

基于知识图谱的搜索引擎已成为现今搜索引擎的主要形式，其技术框架也在不断改进和完善。目前的医疗搜索引擎主要受限于医学知识图谱的知识数量和质量，构建完备的医学知识图谱是其关键。

[8] http://www.webmd.com/
[9] http://www.omnimedicalsearch.com/
[10] http://www.healthline.com/
[11] http://mingyi.sogou.com/
[12] http://ly.so.com/

2.2 医疗问答系统

问答系统也是知识图谱的典型应用场景。目前在基于知识图谱的问答系统中采用的方法主要包括：基于信息提取的方法[97]，利用问句信息结合知识库资源获取候选答案；基于语义解析的方法，将自然语言问句解析成一种逻辑表达形式[98]，通过这种结构化表达从知识库中寻找答案；基于向量空间建模的方法[99]，使用向量空间描述自然语言问句以及知识图谱中的实体和关系，通过机器学习、深度学习等方法生成问答模型进行回答。

早期的医疗问答系统的研究主要集中于信息检索、提取和摘要技术[100-102]。知识图谱概念的提出，致使问答系统研究热点转移到基于知识图谱的问答系统研究。Terol[103]等人使用了 UMLS 以及 WordNet 两种知识库，设置了 10 类医学问题类型，利用自然语言处理技术的应用来生成和处理问题的逻辑形式，从知识库中提取答案。Abacha 等人[104]对比了基于医学本体的医疗问答系统，结合医学本体、领域知识，NLP 相关技术和语义关系，实现了一个医疗自动问答系统。华东理工大学阮彤、王昊奋等人[4]与上海曙光医院合作构建了包括疾病库、证库、症状库、中草药库和方剂库的中医药知识图谱，并基于该知识图谱进行中医药问答和辅助开药。该系统通过基于知识图谱的分词、模板匹配、模板的翻译执行来回答概念、实体、属性、属性值的模板组合问题，并将图谱中存储的数据自动转换成推理引擎适用的推理规则，再结合医生工作站传来的病人事实数据，辅助医生开方。

在医疗垂直领域的问答系统研究中，受限于现有医学知识图谱的推理能力，以及医学知识表示的复杂性，尚未出现重大的突破。而反观开放领域，对特定领域专业知识要求不高的情况下，IBM 的 Watson 系统、微软的小冰[13]都是融合传统问答系统和知识图谱的成功案例。

2.3 医疗决策支持系统

借助医疗知识图谱，医疗决策支持系统可以根据患者症状描述及化验数据，给出智能诊断、治疗方案推荐及转诊指南，还可以针对医生的诊疗方案进行分析、查漏补缺，减少甚至避免误诊。

Ángel 等人[105]设计了一个本体驱动的，基于逻辑推理和概率统计优化的医学鉴别诊断系统 ODDIN，系统的知识库框架包含一个逻辑规则知识库和一个医学本体知识库，在逻辑规则知识库中，定义了四种逻辑规则，而医学本体知识库则是综合了多个医学本体资源，支持多种本体表示形式，包括 RDF、RDFS、OWL 和 SPARQL，诊断方法的核心是基于贝叶斯理论进行改进。Martnez-Romero 等人[106]设计的 iOSC3 系统是一个基于本体针对急性心脏病进行智能监控和诊断的系统，该系统分析患者的状况并提供最佳治疗方案的建议，其知识库由 OWL 本体和表示专家知识的一组 SWRL 规则所构成。王昊奋、张金康等人[27]通过搜集中文开放链接数据中的医疗信息（ICD9、ICD10 等）和主流医学站点中的医疗知识（39 健康网、寻医问药等）构建了医疗知识图谱，并将其应用于上海林康医疗信息技术有限公司的医疗质量与患者安全辅助监控系统和处方审核智能系统中，前者基于知识图谱来进行抗生素不合理使用的监控、危急值预测，后者快速判断处方为合理、疑似不合理和不规范处方，从而促进用药的合理性。

将知识图谱应用于医疗决策是目前的研究热点。但是，在实际应用中，主要存在着两方面的问题：一是缺少完备的全科医学知识图谱，二是医疗决策的可靠性。对于前者，目前基于知识图谱实际应用的医疗决策系统，主要还是针对于特定疾病类型的决策，无法广泛应用，如 IBM 的 Watson Health 主要面向肿瘤和癌症的决策支持，基于巨大的知识库和强大的认知计算能力，为临床医师提供快速的、个性化的循证肿瘤治疗方案。对于后者，医疗决策是直接关系到使用者的身体健康问题，依靠人工智能进行医疗决策对结果的准确性和可靠性有更高的要求。现阶段，基于知识图谱的医疗决策只是扮演着支持和辅助的角色。

3 挑战及研究展望

知识图谱是语义网与知识库的融合与升华。知识图谱的优势是具备强大的语义处理与开放互联能力。对医学领域而言，由于其专业性、规范性、术语有限性等特点，可以从医学词典、医学标准、电子病历等来源获取到高质量数据；另外在科教行业中，存在着知识管理、语义检索、商业分析、决策支持等需求，医学知识图谱的研究能推进海量数据的智能处理，催生上层智能医学的应用。医学知识图谱是大数据与医学的结合，将成为知识图谱和大数据智能的前沿问题。

1）人工智能

人工智能，尤其是深度学习已经在计算机视觉、语音识别、机器翻译等领域超越了传统算法，而其在医学领域中的知识融合与推理等方面的研究潜力还有待挖掘。拥有强大学习能力的深度学习与有着丰富数据储备的知识库相结合，将逐步扩展认知能力，为大规模知识图谱找到更宽广的应用场景。

2）复杂多样的数据源（跨语言，开放域）

互联网和电子病例带来了医疗数据的爆炸性增长，然而这些数据在语种、主题、存储等方面都存在较大差异。现有的研究成果多集中在特定数据集上，普遍存在算法准确率低、限制条件多、扩展性差等问题，同时知识复用、实体消歧等问题也有待进一步研究。跨语言医学知识图谱成为当前的研究热点，我国的研究者更应发挥自身在中文信息处理方面的优势。

3）众包技术

众包反馈机制不仅可以有效提高数据质量[43]，还能作为质量评估的重要补充。众包平台与知识库的有机结合，能够修正机器无法识别的错误，提高知识图谱的质量与性能。39 健康网，寻医问药网的问答区都有专业人士在线解答问题；A+医药

[13] http://www.msxiaoice.com/

百科中的药物信息由医学研究者共同编辑完成。而另一方面，如何评估和保障提交结果的质量，也开始受到国内外学者的关注。

4）可视化

知识图谱可视化的真正意义在于让人直观地了解推理的过程与结果。而医学知识图谱可视化站在医生或病人的立场，寻求最佳的知识展示方案：病人能够理解诊断结果，医生能够利用知识图谱的动态推理过程作出合理诊断。

4 结束语

随着医疗信息化的发展，医学电子数据有了一定的积累。构建医疗领域的知识图谱，可以从海量数据中提炼出医疗知识，并合理高效地对其进行管理、共享及应用，对当今的医疗行业有着重要意义，也是很多企业和研究机构的研究热点。本文从医疗知识图谱的构建与应用角度，综述了医疗知识图谱的相关背景、现有技术和应用，总结了目前医疗知识图谱面临的主要挑战，并对其未来的研究方向进行了展望。

医学知识图谱将知识图谱与医学知识进行结合，定会推进医学数据的自动化与智能化处理，为医疗行业带来新的发展契机。虽然目前对于医疗知识图谱的研究工作有了很多很有意义的尝试，但总的来说还不够完善和深入，需要更进一步的研究。希望本文能够为医疗知识图谱在国内的研究提供一些帮助与启发。

参考文献：

[1] Singhal A. Introducing the knowledge graph: things, not strings [EB/OL]. Official google blog, 2012. https://googleblog. blogspot. co. za/2012/05/introducing-knowledge-graph-things-not. html.

[2] Amarilli A, Galárraga L, Preda N, et al. Recent Topics of Research around the YAGO Knowledge Base [M]// Web Technologies and Applications. Springer International Publishing, 2014: 1-12.

[3] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A Nucleus for a Web of Open Data [M]// The Semantic Web. Springer Berlin Heidelberg, 2007.

[4] 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.

[5] 顾琳. 基于领域本体的亚健康中医辅助诊断系统的研究及应用 [D]. 昆明: 云南师范大学, 2008.

[6] Computer-based medical consultations: MYCIN [M]. Elsevier, 2012.

[7] Rédei G P. Encyclopedia of Genetics, Genomics, Proteomics and Informatics [M]. Springer Netherlands, 2008.

[8] Singhal A. Introducing the knowledge graph: things, not strings [EB/OL]. Official google blog, 2012. https://googleblog. blogspot. co. za/2012/05/introducing-knowledge-graph-things-not. html.

[9] Ceusters W, Martens P, Dhaen C, et al. LinkFactory: an advanced formal ontology management System [C]// Proc of Interactive Tools for Knowledge Capture Workshop. 2001: 175-204.

[10] Stevens R, Baker P, Bechhofer S, et al. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources [J]. Bioinformatics, 2000, 16 (2): 184.

[11] 刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展 [J]. 计算机研究与发展, 2016, 53 (2): 247-261.

[12] Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning [C]// Proc of the 48th Annual Meeting of the Association For Computational Linguistics. Association for Computational Linguistics, 2010: 384-394.

[13] Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases [C]// Proc of Conference on Artificial Intelligence. 2011.

[14] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion [C]// Advances in Neural Information Processing Systems. 2013: 926-934.

[15] Jenatton R, Roux N L, Bordes A, et al. A latent factor model for highly multi-relational data [C]// Advances in Neural Information Processing Systems. 2012: 3167-3175.

[16] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C]// Advances in neural information processing systems. 2013: 2787-2795.

[17] Kleyko D, Khan S, Osipov E, et al. Modality classification of medical images with distributed representations based on cellular automata reservoir computing [C]// Proc of IEEE International Symposium on Biomedical Imaging. 2017.

[18] Henriksson A, Zhao J, Dalianis H, et al. Ensembles of randomized trees using diverse distributed representations of clinical events [J]. BMC Medical Informatics and Decision Making, 2016, 16 (2): 69.

[19] 侯丽, 钱庆, 黄利辉, 等. 基于本体的临床医学知识库系统构建探讨 [J]. 医学信息学杂志, 2011, 32 (4): 42-47.

[20] Nadkarni P, Chen R, Brandt C. UMLS concept indexing for production databases [J]. Journal of the American Medical Informatics Association, 2001, 8 (1): 80-91.

[21] Friedman C, Alderson P O, Austin J H M, et al. A general natural-language text processor for clinical radiology [J]. Journal of the American Medical Informatics Association, 1994, 1 (2): 161-174.

[22] Wu S T, Liu H, Li D, et al. FOCUS on clinical research informatics: Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis [J]. Journal of the American Medical Informatics Association Jamia, 2012, 19 (e1): 149-56.

[23] Smith C A, Stavri P Z. Consumer Health Vocabulary [M]// Consumer Health Informatics. 2005: 122-128.

[24] Uzuner Ö, South B R, Shen S, et al. 2010 i2b2//VA challenge on concepts, assertions, and relations in clinical text. [J]. Journal of the American Medical Informatics Association, 2011, 18 (5): 552-6.

[25] Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. [J]. Journal of Biomedical

Informatics, 2014, 47 (2): 1-10.

[26] Kazama J, Makino T, Ohta Y, et al. Tuning support vector machines for biomedical named entity recognition [C]// Proc of Workshop on Natural language Processing in the Biomedical Domain-Volume 3. Association for Computational Linguistics, 2002: 1-8.

[27] Zhou, G. , Zhang, J. , Su, J. et al. Recognizing names in biomedical texts: A machine learning approach [J]. Bioinformatics, 2004, 20 (7) , pp. 1178 – 1190.

[28] Chen L, Friedman C. C: Extracting phenotypic information from the literature via natural language processing [J]. Studies in Health Technology & Informatics, 2004, 107 (2): 758-62.

[29] Liang L, Wang K, Meng D, et al. Active self-paced learning for cost-effective and progressive face identification [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2017, PP (99): 1-1.

[30] Collobert R, Weston J, Bottou L, et al. Natural Language Processing (Almost) from Scratch [J]. Journal of Machine Learning Research, 2011, 12 (1): 2493-2537.

[31] Sahu S K, Anand A. Recurrent neural network models for disease name recognition using domain invariant features [C]// Proc of Meeting of the Association for Computationa Linguistics. 2016.

[32] Wei Q, Chen T, Xu R, et al. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks [J]. Database the Journal of Biological Databases & Curation, 2016, 2016.

[33] Jagannatha A, Yu H. Structured prediction models for RNN based sequence labeling in clinical text [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2016: 856-865.

[34] Organization W H. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines [J]. Geneva World Health Organization, 1992, 10 (2): 86–92.

[35] Uzuner O, Mailoa J, Ryan R, et al. Semantic relations for problem-oriented medical records [J]. Artificial intelligence in medicine, 2010, 50 (2): 63-73.

[36] Frunza O, Inkpen D. Extraction of disease-treatment semantic relations from biomedical sentences [C]// Proc of Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics, 2010: 91-98.

[37] Abacha A B, Zweigenbaum P. A hybrid approach for the extraction of semantic relations from MEDLINE abstracts [C]// Proc of International Conference on Intelligent Text Processing and Computational Linguistics. Springer Berlin Heidelberg, 2011: 139-150.

[38] Bruijn B D, Cherry C, Kiritchenko S, et al. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010 [J]. Journal of the American Medical Informatics Association, 2011, 18 (5): 557-62.

[39] Stapley B J, Benoit G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts [C]// Pac Symp Biocomput. 2000, 5: 529-540.

[40] Khoo C S G, Chan S, Niu Y. Extracting causal knowledge from a medical database using graphical patterns [C]// Proc of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2000: 336-343.

[41] 王昊奋, 张金康, 程小军. 中文开放链接医疗数据的构建 [J]. 中国数字医学, 2013, 8 (4): 5-8.

[42] 徐绪堪, 房道伟, 蒋勋, 等. 知识组织中知识粒度化表示和规范化研究 [J]. 图书情报知识, 2014, 2014 (6): 101-106, 90.

[43] 庄严, 李国良, 冯建华. 知识库实体对齐技术综述 [J]. 计算机研究与发展, 2016, 53 (1): 165-192.

[44] Fellegi I P, Sunter A B. A Theory for Record Linkage [J]. Journal of the American Statistical Association, 1969, 64 (328): 1183-1210.

[45] Cochinwala M, Kurien V, Lalk G, et al. Efficient data reconciliation [J]. Information Sciences An International Journal, 2001, 137 (1–4): 1-15.

[46] Elfeky M G, Verykios V S, Elmagarmid A K. TAILOR: A Record Linkage Tool Box [C]// Proc of International Conference on Data Engineering. 2002: 17-28.

[47] Christen P. Automatic Training Example Selection for Scalable Unsupervised Record Linkage [M]// Advances in Knowledge Discovery and Data Mining. Berlin: Springer, 2008: 511-518.

[48] Chen Z, Kalashnikov D V, Mehrotra S. Exploiting context analysis for combining multiple entity resolution systems [C]// Proc of ACM SIGMOD International Conference on Management of Data. 2009: 207-218.

[49] Ravikumar P, Cohen W W. A Hierarchical Graphical Model for Record Linkage [J]. 2012, 24 (2) .

[50] Li J, Wang Z, Zhang X, et al. Large scale instance matching via multiple indexes and candidate selection [J]. Knowledge-Based Systems, 2013, 50 (30): 112-120.

[51] Bhattacharya I, Getoor L. Collective entity resolution in relational data [J]. ACM Transactions on Knowledge Discovery from Data (TKDD) , 2007, 1 (1): 5.

[52] Lacoste-Julien S, Palla K, Davies A, et al. Sigma: Simple greedy matching for aligning large knowledge bases [C]// Proc of the 19th ACM SIGKDD International Conference on Knowledge Discovery And Data Mining. ACM, 2013: 572-580.

[53] Tang J, Li J, Liang B, et al. Using Bayesian decision for ontology mapping [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2006, 4 (4): 243-262.

[54] Bhattacharya I, Getoor L. A latent dirichlet model for unsupervised entity resolution [C]// Proc of SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics. 2006: 47-58.

[55] Mccallum A, Wellner B. Conditional Models of Identity Uncertainty with Application to Noun Coreference [J]. Nips, 2004: 905-912.

[56] Domingos P. Multi-Relational Record Linkage [J]. Journal of Neuroscience, 2004, 25 (25): 1113-21.

[57] Singla P, Domingos P. Entity Resolution with Markov Logic [C]// International Conference on Data Mining. IEEE, 2006: 572-582.

[58] Rastogi V, Dalvi N, Garofalakis M. Large-scale collective entity matching [J]. Proceedings of the Vldb Endowment, 2011, 4 (4): 208-218.

[59] 阮彤，王梦婕，王昊奋，等. 垂直知识图谱的构建与应用研究 [J]. 知识管理论坛, 2016, 3: 007.

[60] 李敬华，易小烈，杨德利，等. 面向临床决策支持的中医脾胃病本体知识库构建研究 [J]. 中国医学创新, 2014, 2014 (27): 121-125.

[61] Shvaiko P, Euzenat J. Ontology Matching: State of the Art and Future Challenges [J]. IEEE Trans on Knowledge & Data Engineering, 2013, 25 (1): 158-176.

[62] Suchanek F M, Abiteboul S, Senellart P. PARIS: Probabilistic Alignment of Relations, Instances, and Schema [J]. Proceedings of the Vldb Endowment, 2011, 5 (3): 157-168.

[63] Dong X, Gabrilovich E, Heitz G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion [C]// Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014: 601-610.

[64] Dieng-Kuntz R, Minier D, Růzicka M, et al. Building and using a medical ontology for knowledge management and cooperative work in a health care network. [J]. Computers in Biology & Medicine, 2006, 36 (8): 871-892.

[65] Baorto D, Li L, Cimino J J. Practical experience with the maintenance and auditing of a large medical ontology [J]. Journal of biomedical informatics, 2009, 42 (3): 494-503.

[66] Giacomo G D, Lenzerini M. TBox and ABox Reasoning in Expressive Description Logics [C]// Proc of International Workshop on Description Logics. 1996: 37-48.

[67] Buchanan B, Shortliffe E. Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project [J]. 1984.

[68] Aamodt A, Plaza E. Case-based reasoning: Foundational issues, methodological variations, and system approaches [J]. AI communications, 1994, 7 (1): 39-59.

[69] Bousquet C, Henegar C, Lillo-Le Louët A, et al. Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach [J]. International journal of medical informatics, 2005, 74 (7): 563-571.

[70] Chen R C, Huang Y H, Bau C T, et al. A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection [J]. Expert Systems with Applications, 2012, 39 (4): 3995-4006.

[71] Bichindaritz I, Kansu E, Sullivan K M. Case-based reasoning in care-partner: Gathering evidence for evidence-based medical practice [C]// Proc of European Workshop on Advances in Case-Based Reasoning. Berlin: Springer, 1998: 334-345.

[72] Yang B S, Han T, Kim Y S. Integration of ART-Kohonen neural network and case-based reasoning for intelligent fault diagnosis [J]. Expert Systems with Applications, 2004, 26 (3): 387-395.

[73] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion [C]// Proc of International Conference on Intelligent Control & Information Processing. 2013: 464-469.

[74] Karegowda A G, Manjunath A S, Jayaram M A. Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes [J]. International Journal on Soft Computing, 2011, 2 (2): 15-23.

[75] Lao N, Mitchell T, Cohen W W. Random walk inference and learning in a large scale knowledge base [C]// Proc of Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2011: 529-539.

[76] Lin Y, Liu Z, Luan H, et al. Modeling relation paths for representation learning of knowledge bases [J]. arXiv preprint arXiv: 1506. 00379, 2015.

[77] Angles R, Gutierrez C. Survey of graph database models [J]. ACM Computing Surveys (CSUR) , 2008, 40 (1): 1.

[78] Holzschuher F, Peinl R. Performance of graph query languages: comparison of cypher, gremlin and native access in Neo4j [C]// Proc of the Joint EDBT//ICDT Workshops. ACM, 2013: 195-204.

[79] Chang C, Moon B, Acharya A, et al. Titan: a high-performance remote-sensing database [C]// Proc of the 13th International Conference on Data Engineering. 1997: 375-384.

[80] Tesoriero C. Getting Started with OrientDB [M]. Packt Publishing Ltd, 2013.

[81] Dohmen L, Klamma P D R, Celler F. Algorithms for large networks in the nosql database arangodb [D]. Aachen: Bachelors thesis, RWTH Aachen, 2012.

[82] Neuhaus F, Vizedom A, Baclawski K, et al. Towards ontology evaluation across the life cycle [J]. Applied Ontology, 2013, 8 (3): 179-194.

[83] Clarke E L, Loguercio S, Good B M, et al. A task-based approach for Gene Ontology evaluation [J]. Journal of biomedical semantics, 2013, 4 (1): S4.

[84] Bright T J, Furuya E Y, Kuperman G J, et al. Development and evaluation of an ontology for guiding appropriate antibiotic prescribing [J]. Journal of biomedical informatics, 2012, 45 (1): 120-128.

[85] Gordon C L, Pouch S, Cowell L G, et al. Design and evaluation of a bacterial clinical infectious diseases ontology [J]. AMIA Symposium. AMIA Symposium, 2013, 2013 (2013): 502.

[86] Corcho Ó, Gómez-Pérez A, González-Cabero R, et al. ODEval: a tool for evaluating RDF (S) , DAML+OIL, and OWL concept taxonomies [M]// Artificial Intelligence Applications and Innovations. Springer US, 2004: 369-382.

[87] Poveda-Villalón M, Suárez-Figueroa M C, Gómez-Pérez A. Validating ontologies with oops! [C]// Proc of International Conference on Knowledge Engineering and Knowledge Management. Berlin: Springer, 2012: 267-281.

[88] Stojanovic L, Stojanovic N, Gonzalez J, et al. OntoManager–a system for

the usage-based ontology management [C]// OTM Confederated International Conferences"On the Move to Meaningful Internet Systems. Springer Berlin Heidelberg, 2003: 858-875.

[89] Fernández M, Cantador I, Castells P. CORE: A tool for collaborative ontology reuse and evaluation [C]// Proc of International Workshop on Evaluation of Ontologies for the Web. 2006.

[90] Ammenwerth E, Gräber S, Herrmann G, et al. Evaluation of health information systems-problems and challenges. [J]. International Journal of Medical Informatics, 2003, 71 (3): 125-135.

[91] Aronson A R, Rindflesch T C. Query expansion using the UMLS Metathesaurus. [C]// Proc AMIA Annu Fall Symp, 1996: 485-9.

[92] Díazgaliano M C, Martínvaldivia M T, Ureñalópez L A. Query expansion with a medical ontology to improve a multimodal information retrieval system. [J]. Computers in Biology & Medicine, 2009, 39 (4): 396-403.

[93] Nelson S J, Johnston W D, Humphreys B L. Relationships in medical subject headings (MeSH) [M]// Relationships in the Organization of Knowledge. Springer Netherlands, 2001: 171-184.

[94] Huang C C, Lu Z. Exploring Query Expansion for Entity Searches in PubMed [C]// Proc of International Workshop on Health Text Mining and Information Analysis. 2016: 106-112.

[95] 贾李蓉，于彤，崔蒙，等. 中医药学语言系统研究进展 [J]. 中国数字医学, 2014, 2014 (10): 57-59.

[96] 贾李蓉，刘静，于彤，等. 中医药知识图谱构建 [J]. 医学信息学杂志，2015, 2015 (8): 51-53.

[97] Yao X, Durme B V. Information Extraction over Structured Data: Question Answering with Freebase [C]// Proc of Meeting of the Association for Computational Linguistics. 2014: 956-966.

[98] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs [C]. Proc of EMNLP, 2013.

[99] Bordes A, Weston J, Usunier N. Open Question Answering with Weakly Supervised Embedding Models [M]// Machine Learning and Knowledge Discovery in Databases. Springer Berlin, 2014: 165-180.

[100] Lee M, Cimino J, Zhu H R, et al. Beyond information retrieval-medical question answering. [J]. AMIA Symposium. AMIA Symposium, 2006, 2006: 469-73.

[101] Cao Y G, Liu F, Simpson P, et al. AskHERMES: An online question answering system for complex clinical questions [J]. Journal of Biomedical Informatics, 2011, 44 (2): 277.

[102] Cairns B L, Nielsen R D, Masanz J J, et al. The MiPACQ clinical question answering system. [J]. AMIA Symposium. AMIA Symposium, 2011, 2011 (2011): 171-80.

[103] Terol R M, Martínez-Barco P, Palomar M. A knowledge based method for the medical question answering problem. [J]. Computers in Biology & Medicine, 2007, 37 (10): 1511-1521.

[104] Abacha A B, Zweigenbaum P. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies [J]. Information Processing & Management, 2015, 51 (5): 570-594.

[105] Ángel García-Cresp, Rodríguez A, Mencke M, et al. ODDIN: Ontology-driven differential diagnosis based on logical inference and probabilistic refinements [J]. Expert Systems with Applications, 2010, 37 (3): 2621-2628.

[106] Martínez-Romero M, Vázquez-Naya J M, Pereira J, et al. The iOSC3 System: Using Ontologies and SWRL Rules for Intelligent Supervision and Care of Patients with Acute Cardiac Disorders [J]. Computational & Mathematical Methods in Medicine, 2013, 2013 (5904): 178-180.