计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(6):1−12 [doi: 10.15888/j.cnki.csa.007431]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

领域知识图谱研究综述①

刘烨宸, 李华昱

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通讯作者: 李华昱, E-mail: lycssbcom@163.com

摘　要: 知识图谱由 Google 公司提出, 作为增强其搜索功能的知识库, 在近几年得到了迅速发展. 随着知识图谱价值不断地被发掘, 各类领域知识图谱也迅速建设起来. 本文通过领域知识图谱和通用知识图谱的对比来清晰化领域知识图谱的定义, 介绍了领域知识图谱的架构, 并以医学知识图谱为例讲解了领域知识图谱的构建技术. 最后, 本文介绍了当前热门的领域知识图谱的发展状况和应用, 对当前领域知识图谱状况进行了较为全面的总结.

关键词: 领域知识图谱; 架构; 构建技术

引用格式: 刘烨宸,李华昱.领域知识图谱研究综述.计算机系统应用,2020,29(6):1–12. http://www.c-s-a.org.cn/1003-3254/7431.html

Survey on Domain Knowledge Graph Research

LIU Ye-Chen, LI Hua-Yu

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: The knowledge graph has been proposed by Google as a knowledge base to enhance its search function, and has been rapidly developed in recent years. As the value of knowledge graphs is constantly being explored, many domain knowledge graphs are rapidly being built. This study clarifies the definition of domain knowledge graphs by comparing domain knowledge graphs with general knowledge graphs, introduces the architecture of domain knowledge graphs, and uses medical knowledge graphs as an example to explain the construction techniques of knowledge graphs. Finally, this study introduces the development of the current popular domain knowledge graphs, and provides a comprehensive summary of the current domain knowledge graph status.

Key words: domain knowledge graph; architecture; construction techniques

1　引言

知识图谱的概念要追溯到上世纪六十年代提出的一种知识表示形式——语义网络 (semantic network), 它由相互连接的节点和边组成, 节点表示概念或对象, 边表示节点与节点之间的关系. 在表现形式上, 语义网络和知识图谱相似, 但语义网络侧重于描述概念与概念之间的关系, 知识图谱侧重于描述实体与实体之间的关系[1]. 除了语义网络之外, 语义网 (semantic web) 和链接数据 (linked data) 也为知识图谱的诞生提供了支撑.

知识图谱分为通用知识图谱与领域知识图谱两类. 这两种知识图谱主要存在覆盖范围和使用方式上的差异. 通用知识图谱面向通用领域, 主要包含了大量的现实世界中的常识性知识, 覆盖面广. 领域知识图谱又称为行业知识图谱或垂直知识图谱, 是面向某一特定领域的, 是由该领域的专业数据构成的行业知识库, 因其基于行业数据构建, 有着严格而丰富的数据模式, 所以对该领域知识的深度、知识准确性有着更高的要求.

① 基金项目: 国家自然科学基金面上项目 (61572522); 国家科技重大专项 (2017ZX05013001)
Foundation item: General Program of National Natural Science Foundation of China (61572522); National Science and Technology Major Program (2017ZX05013001)
收稿时间: 2019-10-24; 修改时间: 2019-11-20; 采用时间: 2019-12-05; csa 在线出版时间: 2020-06-10

本文通过介绍领域知识图谱的定义与架构, 首先对领域知识图谱有个基本了解. 然后以医学知识图谱的构建为例介绍信息抽取、知识融合和知识加工 3 个核心技术. 最后列举了几大热门领域知识图谱的现状并对知识图谱的应用做出说明.

2 领域知识图谱的定义与架构

2.1 领域知识图谱的定义

要说明什么是领域知识图谱, 首先应该阐述什么是知识图谱. 其实, 工业界和学术界都没有对于知识图谱给出一个严格的定义. 本文在这里借用“Exploiting Linked Data and Knowledge Graphs in Large Organisations”[2]这本书对知识图谱的定义: “A knowledge graph consists of a set of interconnected typed entities and their attributes.”, 即知识图谱是由一些相互连接的实体以及它们的属性构成的. 知识图谱是由一条条知识组成, 而知识需要有其表达形式, 目前主流的知识表达形式有两种: W3W 制定的资源描述框架 (Resource Description Framework, RDF) 和网络本体语言 (Web Ontology Language, OWL). 本质上, 知识图谱是一种揭露实体之间关系的语义网络. 但是又不同于上世纪五六十年代产生的语义网络, 它之所以成为了新兴技术, 其中的关键就是知识规模. 知识图谱是大数据时代催生的, 其规模之大决定了其效用之大. 当前已经建成了多个大规模知识图谱: DBpedia, YAGO, XLORE, Freebase, Google KG 等. 表 1 统计了部分知识图谱的数据规模.

表 1 部分知识图谱规模统计

知识图谱	概念数量 (K)	实例数量 (M)	属性数量	三元组
DBpedia	250	4	6000	500 M
YAGO	350	10	100	120 M
XLORE	663	16	70 000	1 B
Freebase	15	40	4000	1 B
Goole KG	15	600		20 B

领域知识图谱 (domain-specific knowledge graph) 作为知识图谱的一个分支, 它把知识的覆盖范围和使用方式都聚焦于某一特定领域, 因此其对该领域知识的深度和精度都有很高的要求. 通用知识图谱则更注重广度, 强调融合更多的实体, 其精确度不够高, 且受概念范围的影响, 很难借助本体库对公理、规则以及约束条件的支持能力规范其实体、属性、实体间的关系等[3]. 领域知识图谱具有许多不同的数据模式以适应不同的业务场景和使用人员

表 2 总结了领域知识图谱和通用知识图谱在知识表示、知识获取和知识应用 3 个方面的区别.

表 2 通用知识图谱和领域知识图谱比较[4]

比较层面	比较维度	领域知识图谱	通用知识图谱
知识表示	广度	窄	宽
	深度	深	浅
	粒度	细	粗
知识获取	质量要求	苛刻	高
	专家参与	重度	轻度
	自动化程度	低	高
知识应用	推理链条	长	短
	应用复杂性	复杂	简单

知识表示的 3 个维度中比较重要的一个维度是知识粒度, 知识粒度反映了基本知识单元的大小. 不同领域中粒度大小往往是不相同的, 也难以形成一个统一标准. 在传统的知识搜索领域中, 知识粒度往往是文档级别, 这也就表现为搜索结果是一堆文档的罗列. 而在引入知识图谱后的搜索结果可以直接给出答案的名词以及答案的相近关系, 这也就是知识表示粒度细化到单个实体乃至是实体的某个属性的表现 (如图 1). 一般来说, 知识表示的细腻程度与表达能力成正比, 与获取难度成反比. 领域知识图谱往往要求更细的知识粒度, 这也就造成了知识获取的困难. 所以领域知识图谱的构建更加花费资源[4].

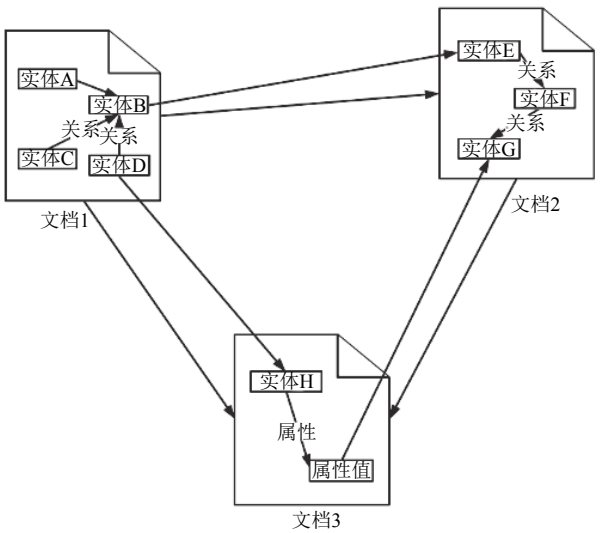

图 1 知识图谱以文档中的实体作为知识单元

从知识获取层面看, 领域知识图谱对知识质量要求更加苛刻, 这是因为领域内的应用容错率更低. 比如

教育领域, 某一知识点的错误还可能导致与其关联知识产生偏差. 对质量要求苛刻自然也就需要更多的专家参与, 这也是领域知识图谱准确度的保障. 但重度专家参与并不意味着完全由专家建设, 充分发挥专家在该领域的专业性, 自动化建设与人力补充才是构建领域知识图谱的正确思路.

由于领域知识图谱知识覆盖范围较小, 知识深度更深, 所以知识点更加密集, 这就导致领域知识图谱的推理链条更长. 领域知识图谱往往是为了某一专业领域而构建的, 其应用复杂度自然更复杂一些.

2.2 领域知识图谱的架构

领域知识图谱的架构分两种: 一种是领域知识图谱自身的逻辑结构; 另一种是领域知识图谱的构建技术 (体系) 架构, 如图 2 所示.

从逻辑上看, 知识图谱分为数据层和模式层. 在数据层中, 知识以事实为单位进行存储. 事实通常以三元组的形式进行存储在图数据库中. 像 Neo4J、ArangoDB、OrientDB 都是当前主流的图数据库. 模式层制定了数据层应该遵守的约束规范. 通常采用本体库来管理知识图谱的模式层, 借助本体库对公理、规则和约束条件的支持能力来规范实体、关系以及实体的类型和属性等对象间的联系[5]. 知识图谱的技术 (体系) 架构是指其构建模式结构, 通常有自底向上构建和自顶向下两种构建方式. 自底向上的构建方式是直接进行数据抽取, 将所得实体、关系、属性等经审核后整合到知识库中. 自顶向下的构建方式先定义顶层关系本体, 再将实体整合到顶层本体中.通用知识图谱为了融合更多的实体, 大多采用自底向上的方式构建[6]. 领域知识图谱面向特定领域, 对知识的质量和准确度要求苛刻, 因此要求领域知识图谱具有完备的本体层模式, 通常采用自顶向下和自底向上相结合的构建方式[6].

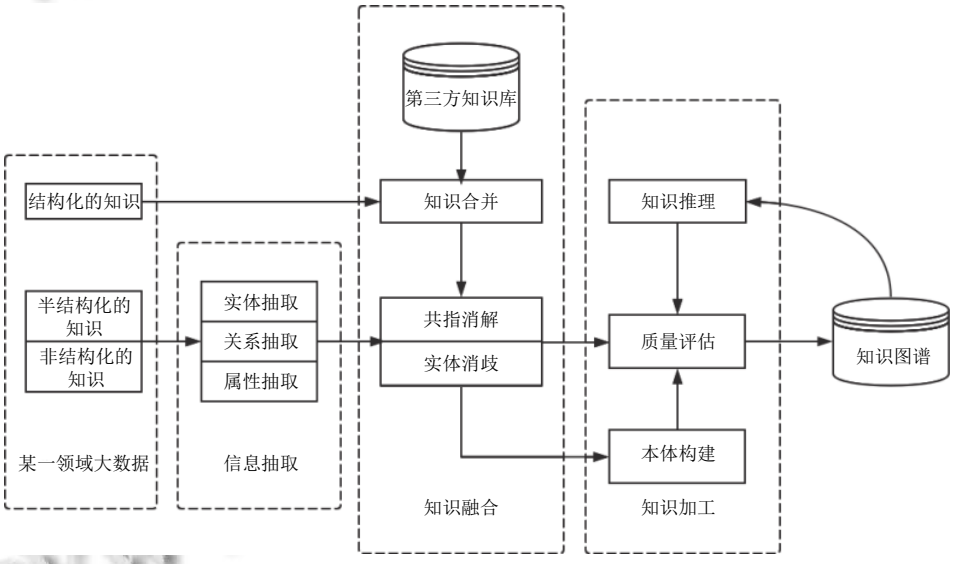

图 2 领域知识图谱体系架构

3 领域知识图谱构建

随着研究热度越来越高, 各类领域知识图谱迅速建设起来, 不但涉及领域范围广, 而且领域细分更加复杂. 比如医学知识图谱就可以细化为生物医学领域知识图谱、中医学科领域知识图谱、中文疾病知识图谱乃至乳腺肿瘤知识图谱、基于甲状腺知识图谱等. 虽然说领域知识图谱的构建总体遵循上文阐述的体系架构, 但是具体到各行业总会产生差异. 所以无法空泛地讲领域知识图谱的构建. 接下来, 本文将选取当前比较热门的医学领域为例, 阐述领域知识图谱的构建技术.

医学领域知识图谱是由多种信息处理技术共同构建而成. 通过医学信息抽取技术, 可以从包含医学数据的数据源中提取出实体、实体间的关系和实体的属性等要素. 通过医学知识融合技术, 可以使信息抽取中提取的事实歧义性更小、冗余度更低、错误更低. 但事实本身不等于知识, 要想最终获取结构化、网络化的

知识体系, 还要进行知识加工.

3.1 医学知识表示

知识表示是医学知识图谱构建之前确定下来的一组约定, 以便将知识以符号的形式存储. 知识表示的选择影响着医学知识图谱这个系统在信息抽取、存储以及应用的效率[7]. 知识表示方法有 3 类.

(1) 基于符号逻辑的知识表示. 该方法是早期医疗知识库使用的知识表示方法, 常见的逻辑方法有时间、概率、答案集编程、时间抽象等. 在文献[8]中, van der Heijden 等提出基于时态逻辑的知识表示方法来规范化具有生理背景知识的临床指南. 在文献[9]中, Merhej 等提出了一种基于答案集编程 (ASP) 方法, 该方法在处理复杂搜索问题时取得了较好的成效, 不仅可以用于检测不同治疗方法的冲突, 还可以检测治疗方法间的相互作用. 但是基于符号逻辑的知识表示很难使用机器生成推理规则, 仅仅在数据规模较小的时期使用较广, 现在面对规模庞大的医学知识库建设、面对具有挑战性的临床患者数据和基因组数据时, 仅作为辅助形式存在[7].

(2) 使用语义网的知识表示. 这种方法当前认可度比较高, 使用也很广泛. 文献[10]就使用语义网络技术从计算机可解释的准则中评估护理行为并检测个性化过程中的潜在矛盾, 而文献[11]则使用语义网络技术通过医疗行为和治疗数据的层次结构进行推理以检测主要的替代干预措施. 在文献[12]中, 作者使用 UMLS 尤其是其语义网络来检测临床指南中的模式. 使用语义网的知识表示主要包括用于可扩展标记语言 XML、描述 Web 资源的资源描述框架 RDF 和本体语义描述语言 WOL. RDF 假定任何复杂的语义都可以通过若干个三元组的组合来表达. RDF作为一个统一且无歧义的语义定义方式, 能够促进语义网不同知识的相互链接, 克服了 XML 必须需要足够详细的 XML 解释文档才能解释语义的困难. 当前在工业界大规模应用的是基于 RDF 三元组的表示方法.

(3) 表示学习. RDF 方法虽然得到了大规模应用, 但是由于知识图谱中节点个数影响着推理的效率和难度, 所以 RDF 方法在应用于医学领域时会出现计算效率低等问题. 表示学习可以将医学研究对象的语义信息表示为稠密低维的实数值向量. 通过在低维空间中计算和推理, 能有效解决数据稀疏的问题, 适应了大数据环境下知识计算效率问题, 更容易解决不同源的异质信息融合问题. 医学知识图谱按照计算方式不同可以分为距离平移模型 (translational distance model) 和语义匹配模型 (semantic matching model)[6]. 其中距离平移模型通过设计距离评估函数判断知识的合理性, 平移模型的代表是 Bordes 提出的 TransE 模型. 语义匹配模型包括单层神经网络模型 (Single Layer Model, SLM)、隐因子模型 (Latent Factor Model, LFM)、神经张量模型 (Neural Tensor Model, NTM)、矩阵分解模型 (Matrix Factorization, MF) 等[6]. 这方面的研究有: Henriksson 等[13]证明基于电子病历中的临床事件的深度学习表示法可以对更高性能的预测模型进行后续训练. 可见表示学习在知识表示方面效果不错.

3.2 医学信息抽取

医学信息抽取主要是通过人工或者自动方式从非结构化或者半结构化的数据中提取医学知识单元[7]. 人工抽取可以通过基于访谈或焦点小组的工具辅助方法或定性方法来获取知识. 目前临床医学知识库、ICD-10 和上文提到的 SNOMED-CT 知识库都是采用这种方法抽取构建的. 自动抽取借助可以使用机器学习 (ML) 或基于案例的推理 (CBR) 技术从医学信息源中自动提取出医学知识单元以构建知识库. 采用这种方式构建的医学知识库有一体化医学语言系统 UMLS. 随着机器学习和深度学习技术的发展, 医学知识自动抽取的效率越来越高, 但不利于自动化抽取的数据, 人工抽取也是必不可少的. 接下来本文将从实体抽取、关系抽取和属性抽取 3 个方面介绍自动抽取技术.

3.2.1 实体抽取

实体抽取又称为命名实体识别 (named entity recognition), 旨在从医学信息源中识别出特定的医学实体. 实体抽取是医学信息抽取中至关重要的一环. 医学实体抽取主要有 3 种方法.

(1) 基于医学规则和医学词典的方法

早期医学实体抽取研究的主要方向是从医学信息文本中识别出疾病、症状、治疗、专家这些关键的实体信息, 为后续实体关系抽取奠定基础. Friedman 等[14]开发了一种通用的自然语言处理器来识别叙事报告中的临床信息并将其映射为包含临床术语的结构化表示形式. 基于医学规则和医学词典的实体抽取方法需要大量的人医学专家编写提取规则. 但是这些规则往往依赖于具体语言和文本风格, 这就造成了系统的可移植性不好, 限制了其使用, 现在这种方法逐渐被另外两

种方法取代[3]. 但在文献[15]中, 提出了一种将令牌级词典功能整合到神经模型中以进行命名实体识别的方法, 使基于词典的实体抽取方法得到发展.

(2) 基于机器学习与统计学算法结合的方法

机器学习诞生后, 研究者尝试通过使用机器学习中的监督算法结合一些医学规则从医学数据源中提取实体. 这种方法取得了不错的效果, 其中最具代表性的是 2010 年美国国家集成生物与临床信息学研究中心 (I2B2) 给出的电子病历命名实体语料标注. 除此之外, 文献[16]中提到 Azalia 使用朴素贝叶斯分类器的命名实体识别, 对圣训的印度尼西亚语翻译中的名称索引. 使用机器学习从带有命名实体的手动注释的语料库中学习. 但是, 手动注释语料库非常昂贵且费力. 文献[17]中提出了一种无需任何人工注释即可用于训练临床 NER 系统的新颖方法. 它仅需要原始文本语料库和诸如 UMLS 之类的资源, 即可提供命名实体及其语义类型的列表. 使用这两个资源, 将自动获取注释以训练机器学习方法. 该方法在 i2b2 2010 和 SemEval 2014 的 NER 共享任务数据集上进行了评估. 其精度可以与过去使用人工注释进行训练的许多监督系统相媲美.

(3) 基于深度学习的方法

深度学习方法是当前使用很广泛的实体抽取方法, 该方法的思路是从目标数据集中将有相似上下文特征的实体进行聚类操作. 这个方法的缺陷是需要使用大量的标准语料进行模型训练, 当给定的实体实例较少时将面临困难. 在智能医疗领域, 在这个问题上取得比较好的突破的是哥伦比亚大学的 Zhang CW 和腾讯的 Li YL[18]. 他们在 2018 年引入了一种生成式的视角来研究关系医学实体对发现问题, 旨在在最小化数据需求的同时, 扩大高质量而又新颖的结构化新医学知识的规模. 基于此提出了 (CRVAE) 模型, 通过利用已标注的实体三元组在自然语言表述上的特点, 将医学实体和关系输入编码器, 通过训练模型, 对每一种医疗关系的不同实体对进行编码, 再通过解码器进行共同训练, 重建实体对, 最后得到未被标注的实体三元组. 这种方法即使在仅有少量外部资源的情况下也能有不错的判别效果. Zhang 等的实验表明: 该方法能够在降低外部资源的条件下, 以 92.91% 的支持度生成属于某个特定医疗关系的实体三元组, 其结果产生了 61.93% 的新样本, 准确率也达到了 77.17%. 要正确地识别实体, 形态分析 (MA) 是必不可少的步骤. 文献[19]提出了同时执行 MA 和 NER 的集成神经网络模型, 重新设计了 MA 和 NER 的执行顺序, 该模型优于独立的 MA 模型和独立的 NER 模型, 可以有效缓解流水线架构中经常发生的错误传播问题.

3.2.2 关系抽取

RDF 知识表示方式中包含 (实体, 关系, 实体) 格式的三元组, 其中的关系就有关系抽取产生. 医学关系抽取就是从医学数据中抽取两实体关系以实现实体间语义联结. 早期的医学关系抽取方法类似于“实体抽取中基于医学规则和医学词典的方法”, 通过人工构造规则和模板进行关系抽取. 现阶段医学领域关系抽取方法有 3 种.

(1) 基于机器学习的方法

基于机器学习的方法是通过解决分类问题实现关系抽取, 常用的分类方法有基于特征和基于核两种.

基于特征的方法是从文本中生成句法和语义等特征向量, 分类器接受向量并判断实体对之间关系. 基于核的方法是根据某种结构 (比如序列、树、图、依存关系路径等) 来表示实体关系, 通过函数来计算对象相似度, 并称这种函数为核.

基于特征分类的方法抽取效果较好、速度很快, 但是选择合适的特征的会耗费许多时间和精力, 而选取特征的好坏关系着关系抽取的质量. 基于核的分类方法特征选取很灵活, 但关系抽取速度慢, 不适合大数据集的关系抽取.

(2) 基于深度学习的方法

基于深度学习的关系抽取方法是目前医学关系抽取主要的方法. 常见的深度学习模型有卷积神经

网络 (CNN) 和递归神经网络 (RNN). 卷积神经网络依靠卷积核获取局部特征, 适用于短句子实体关系抽取; 递归神经网络善于学习长期依赖特征, 适合处理长句子, 文献[20]中提出了一种结构块驱动的卷积神经学习的新型轻量级关系提取方法, 通过在两个数据集 SemEval2010 和 KBP37 上的实验, 证明了该方法的显着优势.

(3) 基于机器学习和深度学习相结合的方法

近年来, 为了充分发挥机器学习和深度学习的优势, 医学专家们将两种关系抽取方法结合起来, 以实现更高效的关系抽取. 李智恒等设计的从化学文献中抽取化学物质致病关系的系统——CDRExtractor, 就是

将基于特征的分类方法和基于核的分类方法结合起来进行 CID 关系抽取. 该系统在 BioCreative V CDR 测评任务 CID 子任务提供的测试集上达到了 67.72% 的 F 值[21]. Zhang Y 等[22]提出了一种混合模型, 采用 RNN 和 CNN 相结合的方式, 实现检测和提取生物医学关系, 实验结果表明, RNNs 和 CNNs 在生物医学关系提取中的优势是互补的. 针对处理长句子和句子中的多个实体时当前模型出现问题较多的情况, 文献[23]中使用具有分段注意力和实体描述的循环神经网络, 有效的克服了上述两个问题, 并将 F1 分数提高约 3%.

3.2.3 属性抽取

属性抽取的主要任务是获取 (实体, 属性, 属性值) 类型三元组中的属性和属性值. 对于医学实体, 药品的规格、剂量、用法用量等都可以看作药品实体的属性. 通过属性抽取建立完整的实体描述. 由于实体的属性可以看成是实体和属性值之间的一种名称性关系, 因此可以将实体属性的抽取问题转换为关系抽取问题. 比如张元博在文献[24]中探索到属性及其属性值存在共同特征, 采用基于特征的机器学习方法来实现医学实体的属性提取.

3.3 医学知识融合

医学知识融合的目的是将医学信息抽取中获得的不同来源、不同结构、不同表示方式的数据进行整合, 最终将这些异构医学数据实现在同一框架下的规范表示[7], 如图 3 所示. 知识融合分为共指消解和实体消歧.

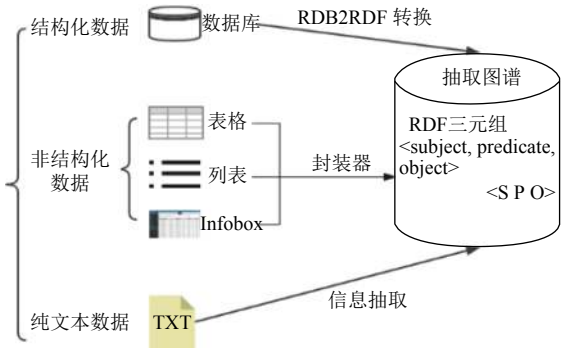

图 3 不同数据转化为三元组示意图

3.3.1 共指消解

共指消解的主要目的是当多个名称对应同一实体的时候, 将这些名称对应到正确的规范化的实体上, 也就是解决异名同物问题. 比如扑热息痛片又名泰诺林、必理通等, 它们都指的是学名为对乙酰氨基酚的药物. 在信息抽取完后产生了这些别名, 这时候就需要共指消解技术把它们关联到对乙酰氨基酚实体上. 共指消解问题可以通过把其看作聚类问题来求解. 该方法以规范化的实体为中心, 通过实体聚类实现规范实体与它的别名实体的匹配[25]. 这方面的研究有: 在文献[26]中, 提出了一种获取健康消费者术语并将其与标准医学术语保持一致的方法. 2015 年, 在文献[27]中提出了结合奇异值分解和多分类器针对共指消解问题的新方法, 该方法可以获得 72.1 的平均准确率.

3.3.2 实体消歧

实体消歧是专门用于解决异构数据的实体产生歧义问题的技术, 也就是针对同名异物问题. 比如止吐药 dogmatilum(舒必利, 止吐灵) 叫“舒宁”, 而抗焦虑药 oxazepam(N-去甲羟基安定) 也叫“舒宁”, 这种问题不加以解决会造成严重的后果. 实体消歧的主要思想是聚类, 基本过程如图 4 所示. 关键在于评估实体和指标的相似度, 度量实体对象与指称项之间相似度的常用的方法有 4 种: 空间向量模型 (实体的上下文), 语义模型 (实体的上下文语义), 社会网络模型 (利用关联实体的关系构建指标网络), 百科知识模型 (网站超链接)[25].

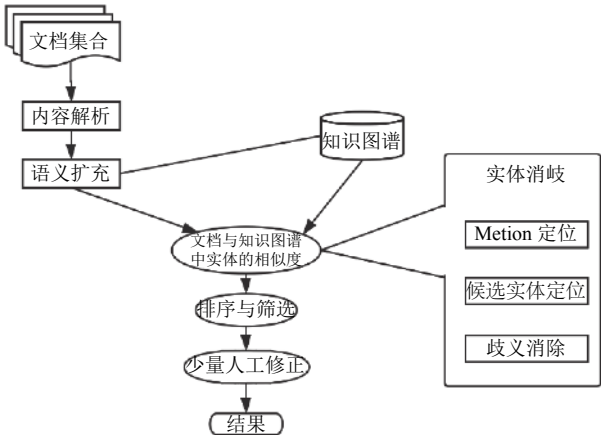

图 4 实体消歧的基本方法过程

近年来, 实体消歧技术也与深度学习相结合. 比如在文献[28]中, 将实体消歧定义为分类任务, 开发了一种新的基于 LSTM 的体系结构, 结果表明与其他方法 (例如文献[29]的 HAC) 相比, 基于 RNN 对句子含义进行编码更适合于实体消歧的任务.

3.3.3 知识合并

知识合并的主要任务是把结构化的知识或者第三方知识库的知识整合到知识图谱中. 结构化的知

识符合知识规范, 实用度高. 第三方知识库也能为知识图谱构建提供可靠的知识来源, 像 WebMD、“好医生”智能医学数据库、家庭医生在线等都可以看作是第三方医学知识库, 其中包含高质量、规范化的医学知识.

本文参考 Mendes 等对 LOD 进行知识合并的方法[30], 把合并第三方知识库的流程归类为: 获取知识; 概念匹配; 实体匹配; 知识评估. 其中概念匹配和实体匹配都是对第三方数据库中获得知识的概念和实体进行归一化处理, 知识评估是对新获得知识一致性和准确性的检测[25].

将原有的关系数据库转化为知识图谱的知识表示也是知识合并的重要任务. 在图数据库未使用之前, 使用比较普遍的都是关系型数据库. W3C 的 RDB2RDF 小组制定了 direct mapping 和 R2RML 两个标准, 用于将关系型数据库的数据转换为 RDF 格式的数据. Direct mapping 采用直接映射的方式, 实现表→类、列→属性、行→实例、单元格值→属性值的映射. Direct mapping 不能将数据库的数据映射到我们自己定义的本体上, R2RML 通过自主编辑和设置映射规则解决了这个问题. 从 RDB 到 RDF 的常用转化工具有 D2RQ、SquirrelRDF、OpenLink Virtuoso 等.

3.4 医学知识加工

医学知识加工的目的是把信息抽取和知识融合中获得的知识加工成高质量的知识. 知识加工包括本体构建、质量评估和知识推理 3 部分[25].

3.4.1 医学本体构建

医学本体是对于医学领域之中医学概念及其相互之间关系的形式化表达. 医学本体可以通过人工方法构建也可以通过数据驱动自动构建. 人工方法构建的本体很适应目前大数据的形式, 所以本文着重介绍下自动化的本体构建技术.

自动化构建本体的方法主要包括中心扩展法、由局部到全体、直接抽取文档构建本体等方法[31]. 本文将不同的本体构建方法汇总在表 3.

表 3 不同的本体构建方法比较[31]

构建方法	期望目标	构建方法	语言分析
TextOnt	自动构建本体 (德文)	从局部到全局	对数据源概念及其关系
Waste H 方法	自动构建初始核本体 (英文)	中间扩展法	对相似概念进行不断扩展
Jean 方法	自动构建本体 (英文)	由局部到全体	对概念间关系
Chang-Shing LI 方法	自动构建本体 (繁体中文)	直接抽取文档构建本体	对汉字语境、语形间关系
陆汝钤半自动构建本体方法	对已有本体进行增添删减 (中文)	利用仿生学, 动态更新领域本体	对语义间关系
杨争库方法	自动构建本体 (中文)	由局部到全体	对概念间关系的映射
王磊方法	自动构建文档本体 (中文)	统计分词, 特征提取, FCA 构建文档本体	对预料内容, 概念层次关系

就医学知识图谱的本体构建来看, 目前存在一些问题: ① 医学领域本体的构建需要医学专家的参与, 并没有实现真正的自动化, 还是以半自动化为主; ② 医学领域本体自动化构建具体实现较少, 大多数研究还是理论研究; ③ 语言分析软件较少, 不能满足现在大规模医学图谱构建的需求. 目前来看本体构建技术的发展和知识图谱的发展热度不匹配, 本体构建也应该尽快实现理论到实践的转换, 以适应构建大规模知识图谱的需求.

3.4.2 质量评估

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才能保证知识图谱中知识的质量[32]. 为了促进知识选择, 应该使用系统来自动 (或半自动化) 用于特定目的的最佳知识的选择. 这需要基于一组特定标准来评估本体质量的方法. 这些标准必须是可量化的, 以便系统而不是人来完成它. 文献[33]研究提出并开发了一种基于符号学的分层本体度量标准套件, 它可以为有效属性提供总体得分的度量, 可以结合使用手动计算和自动化来计算指标, 尽管只有某些指标可以完全自动化的方式计算. 该文章中提到, 此套件已正式确定并在由模块组成的排名系统中实现.

3.4.3 医学知识推理

知识推理是根据已有知识库, 采用相关算法, 实现对知识图谱的探索和挖掘. 在医学知识图谱中, 知识推理要有搜集数据、诊断疾病、提供治疗方法的功能. 而在医学方面, 病情往往因人而异, 对于具体疾病的诊断往往是依靠医生的从医经验, 所以医学知识推理的

构建难度还是很高的.

传统的知识推理方法包括基于描述逻辑的推理、基于规则的推理、基于分布式的知识推理等, 各方法的比较见表 4.

表 4 推理方法的比较[34]

推理方法	推理方式	方法优点	方法缺点
基于描述逻辑的推理	借助 TBox 和 ABox 工具将推理问题归结为 ABox 的一致性检验问题	达到了表达能力与推理复杂度的平衡	数据利用率不高
基于规则的推理	根据简单规则和统计特征的推理	可解释性强; 规则正确时, 准确率高	对规则的依赖度高, 规则制定困难且适用性较低
基于分布式的推理	对知识图谱的低维向量进行推理	计算方便快捷	由于忽视深层次的组合语义信息, 推理能力低
混合推理	结合多种方法的推理	结合了不同方法的优势	结合效率差, 即:一种推理方法为主, 另一种推理方法为辅, 缺乏更深层次的结合

这些方式很难满足医学大数据下的快速推理和对于增量知识和规则的快速加载, 所以现在应用更为广泛的是结合人工智能技术的知识推理模型, 常见的有人工神经网络模型 (artificial neural netword model)、遗传算法 (genetic algorithm) 和反向传播网络模型 (back propagation) 等. 文献[35]中就提出了一种表示本体, 以将文献抽象数据表征为 4 个知识元素 (背景, 目标, 解决方案和发现). 案例研究表明, 所提出的本体模型可以用来表示嵌入在文献摘要中的知识, 并且可以通过 NLP 模型自动提取本体元素. 所提出的框架可以增强文献计量分析, 以从文献中探索更多知识, 实现知识推理的功能.

无论是传统的知识推理方法还是人工智能技术的推理方法都是以知识图谱作为数据源进行推理, 而图挖掘计算则是基于图论的相关算法, 把知识图谱看作图, 把医学实体看作节点, 实体间的关系看作边, 实现对图谱的探索和挖掘, 更有利于解决大规模的图数据分析问题[36]. 基于此, Jagvaral 于 2019 年提出具有注意机制的 CNN-BiLSTM 方法用于知识图谱基于路径的推理[37]. 论文中提到, 他们研发的路径编码器从大型图形的路径中提取特征更有效, 更是说明了应用多步推理在基于路径的推理中可能会有用. 此项研究只使用一种类型来表示实体, 而大多数知识图谱中的实体具有多种类型, 因此, 多种类型合并到路径编码中的路径推理推理还有待研究.

以上为比较具体的领域知识图谱构建流程, 虽然领域知识图谱应用比较广, 但目前还尚未实现自动构建, 而在 2018 年, 清华大学知识工程实验室发表一篇名为“一种准确而高效的领域知识图谱构建方法”的文章[38], 介绍了一种快速构建较高质量的领域知识图谱的方法, 为领域知识图谱构建提供另一种思路, 该方法称为“四步法”: ① 领域本体构建; ② 众包半自动语义标注; ③ 外源数据补全; ④ 信息抽取. 在领域知识图谱构建过程中, 权衡效率和准确率, 平衡自动化和人工构建, 以高效地构建图谱, 这是当前面临的一个很大问题.

3.5 知识图谱绘制工具

图 5 是以心律失常为关键词绘制的医学领域知识图谱, 它展现了知识图谱力导向布局图的视图形式.

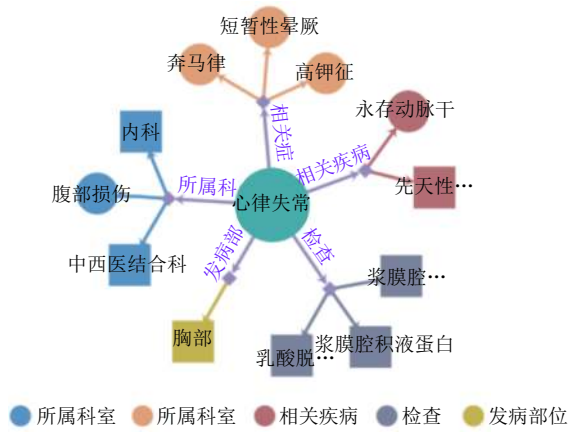

图 5 医疗领域知识图谱举例

知识图谱的绘制工具可分为两大类: 通用软件, 如 SPSS、Ucinet、PajekWordsmithTools 和 GIS 等. 另一类是专门用于知识图谱绘制的软件, 也有许多类型, 有些是针对某些特定领域, 有些是个人未公开的. 表 5 对知识图谱绘制工具做一个汇总.

表 5 知识图谱绘制工具[32]

绘制工具	说明	功能描述
SPSS[39]	美国 SPSS 公司 20 世纪 80 年代开发的大型统计分析软件, 商业用途.	具有完整的数据输入、编辑、统计分析、报表、图形绘制等功能. 常用于多元统计分析、数据挖掘和数据可视化.
Bibexcel[39]	瑞典科学计量学家 Persoon 开发的科学计量学软件, 用于科学研究免费软件.	具有文献计量分析、引文分析、共引分析、耦合分析、聚类分析和数据可视化等功能. 可用于分析 ISI 的 SCI、SSCI 和 A&HCI 文献数据库.
Histcite[40]	Eugene Garfield 等于 2001 年开发的科学文献引文链接分析和可视化系统, 免费软件.	可对 ISI 的 SCI、SSCI 和 SA&HCI 等文献数据库的引文数据进行计量分析, 生成文献、作者和期刊的引文矩阵和实时动态引文编年图. 直观的反映文献之间的引用关系、主题的宗谱关系、作者历史传承关系、科学知识发展演进等.
CiteSpace[41]	陈超美博士开发的专门用于科学知识图谱绘制的免费软件.	可用于追踪研究领域热点和发展趋势, 了解研究领域的研究前沿及演进关键路径, 重要的文献、作者及机构. 可用于对 ISI、CSSCI 和 CNKI 等多种文献数据库进行分析.
Pajek[42]	来自斯洛文尼亚的分析大型网络的社会网络分析免费软件.	Pajek 基于图论、网络分析和可视化技术, 主要用于大型网络分解, 网络关系展示, 科研作者合作网络图谱的绘制.

4 领域知识图谱的现状和应用

4.1 领域知识图谱的现状

随着近几年知识图谱技术的发展, 知识图谱研究与落地发生了一些转向. 其中一个重要变化就是领域知识图谱的建设成为主流. 知识图谱技术与各行业的深度融合已经成为一个重要趋势[4].

接下来, 本文对搜索、医疗、电商、社交、教育这几个热门领域规模比较大的知识图谱进行汇总, 见表 6.

表 6 热门领域知识图谱汇总

领域	知识图谱
搜索领域	百度知识图谱、搜狗百科知识图谱、Google 知识图谱
医疗领域	中医药知识图谱 (http://www.tcmkb.cn/kg/index.php)、中医药知识地图 (http://www.tcmkb.cn/knowledge_maps/index.php)、中文医学知识图谱 (http://zstp.pcl.ac.cn:8002/)
电商领域	阿里知识图谱、京东商品知识图谱
社交领域	Facebook 推出的 Graph Search
教育领域	中国基础教育知识图谱 (edukg.org)、课程图谱 (http://coursegraph.com/)、百度教育大脑、

医疗领域是当前建设很火热的领域, 仅是对中文医学知识图谱的相关检索就达 200 多条, 大到中文疾病知识图谱, 小到甲状腺知识图谱, 医疗领域知识图谱的理论实践化是有原因的: (1) 医疗信息化浪潮. 步入信息化社会以来, 医疗信息化的发展从未停歇过, 从最初的医院信息系统开始, 电子病历、临床智慧医疗等技术层出不穷. (2) 庞大的医学数据. 除医院提供的病例信息, 基因学研究, 蛋白组学也给医疗领域贡献了大量的数据. (3) 人工智能出现后, 为体量庞大的医学数据处理提供方向. 知识图谱正是作为大数据到人工智能的理想桥梁. 整合异构数据, 建立语义关系, 最重要的是知识推理, 医疗知识图谱在智慧医疗的建设中起到越来越重要的作用, 通过知识问答, 知识推理将更好的为社会服务. 所以医疗知识图谱发展迅速. 与之相似, 教育领域同样具有数据量大, 面临信息化建设等优点, 相信教育知识图谱也将会得到越来越多的关注.

4.2 领域知识图谱的应用

知识图谱作为近十年内新兴的概念, 其可以将各种信息和数据整合为知识, 为各研究领域提供可视化分析, 各类大规模知识图谱在智能搜索、智能问答、智能推荐、情报分析等方面发挥了重要作用.

4.2.1 智能搜索

基于知识图谱的智能搜索可以直接给出知识卡片而不是给出相关的链接序列. 在知识图谱的帮助下, 搜索引擎可以将搜索关键词映射到知识图谱中匹配度较高的一个或一组概念上, 最后以知识卡片的形式展现给用户. 知识卡片可以以 3 种形式展示知识[3]: ① 对于单一关键词的搜索, 返还用户查询的实体的结构化摘要. 比如搜索姚明, 将给出姚明的身份介绍以及主要关系介绍; ② 对于问题类的搜索, 知识卡片直接给出答案. 比如搜索“姚明的身高是多少?”, 搜索结构将是显示 226.0 cm 的知识卡片; ③ 对于模糊类的查询, 将给

出相关网页列表. 例如搜索“姚明最近的活动有哪些?”, 搜索结果是包含姚明活动的新闻网页.

4.2.2 智能问答

Gowild 狗尾草的 AI 虚拟生命“琥珀虚颜”和苹果的智能语音助手 Siri 都是知识图谱应用于智能问答方面的实例. 智能问答是信息检索系统的一种高级形式, 能够用自然语言为用户提供问题的解答或者实现人机交流. 目前, 语音助手研发十分火热, 比如百度自然语言部开发的小度机器人, 阿里巴巴人工智能实验室研发的天猫精灵, 亚马逊 Alexa 语音服务等都是为智能问答更加智能、准确做出地探究.

4.2.3 智能推荐

电商、教育、社交等行业都需要借助大数据行为分析进行用户画像, 以指导广告投放和提高用户体验. 相较于原先对关联性较差的数据进行用户行为分析, 知识图谱一个天然的优势就是更突出数据之间的关系, 这样就能根据知识关联关系获得更加精确的用户画像, 有助于精准营销、精细化运营. 除了用户画像, 智能推荐还要依靠商品之间的关联提供使用建议、搭配等.

4.2.4 情报分析

江苏大学刘桂峰利用 CiteSpace 软件信息可视化方法, 对 1990–2010 年间来自 Web of Science (SCIE) 数据库的太赫兹技术领域研究的文献数据进行统计和可视化分析, 揭示出该领域的领军人物、知识基础和研究前沿等信息[43]. 赵蓉英等[44]利用 CiteSpace II 的爆发词探测方法绘制知识图谱, 并绘制爆发词随时间演化的学科前沿发展趋势图, 进而发现学科前沿. 胡泽文等在文献[28]中借助通过 CiteSpace II 界定了改革开放来情报学的 3 个发展阶段. CiteSpace 是一款应用于科学文献中识别并显示科学发展新趋势和新动态的软件, 通过它绘制知识图谱, 能够发现经典文献、研究热点和研究前沿. 可见知识图谱用于情报分析方面有很大的发展潜力.

除此之外, 知识图谱应用于医学、教育等领域, 对于建设智能医疗、智慧教育起着支撑作用.

5 结语

知识图谱从最初作为辅助 Google 搜索的技术被提出, 到现在很多行业都在建设自己的知识图谱, 它的价值正在被慢慢挖掘出来. 知识图谱不是知识的终点, 但是它确实能解决很多学科领域的瓶颈问题, 成为智能化建设的基石.

结合医学知识图谱的构建和发展, 本文认为信息抽取技术仍是当前的研究热点, 最理想的信息抽取方式是结合实体抽取、关系抽取和属性抽取三者的联合抽取, 但该技术还没有典型代表. 而知识推理作为知识图谱最大的亮点和功能, 将其技术发展成熟还需要付出很大地努力. 在人工智能还有很大发展潜力的今天, 借助人工智能技术实现知识推理有很大的发展前景. 知识推理不仅是智能问答、智能推荐等应用的关键技术, 更是智能化建设的基石.

对于领域知识图谱的发展方向, 本文倾向于领域划分更精细, 领域交互更频繁的发展方向. 类比于医学领域中各种疾病的知识图谱, 也许教育领域会出现各种学科知识图谱, 因为越精细, 专业性越强, 知识越准确. 这也是越来越多的人主张建立企业知识图谱的原因. 此外, 各领域的知识图谱不该是独立存在的, 领域知识图谱之间有交互, 才能真正地构成知识网.

知识图谱仍在发展初期, 笔者仅希望通过本文的写作, 能抛砖引玉, 吸引更多人了解这门技术并投入到相关的研究中来.

参考文献

1 Huang HQ, Yu J, Liao X, *et al*. Review on knowledge graphs. Computer Systems & Applications, 2019, 28(6): 1–12.

2 Pan JZ, Vetere G, Gomez-Perez JM, *et al*. Exploiting Linked Data and Knowledge Graphs in Large Organisations. Cham: Springer, 2017.

3 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述. 电子科技大学学报, 2016, 45(4): 589–606. [doi: 10.3969/j.issn.1001-0548.2016.04.012]

4 肖仰华. 领域知识图谱落地实践中的问题与对策. 复旦大学知识工厂实验室. https://www.sohu.com/a/280006592_100099320. (2018-12-06)[2019-10-05].

5 李涓子, 侯磊. 知识图谱研究综述. 山西大学学报 (自然科学版), 2017, 40(3): 454–459.

6 侯梦薇, 卫荣, 陆亮, 等. 知识图谱研究综述及其在医疗领域的应用. 计算机研究与发展, 2018, 55(12): 2585–2599.

7 袁凯琦, 邓扬, 陈道源, 等. 医学知识图谱构建技术与研究进展. 计算机应用研究, 2018, 35(7): 1929–1936. [doi: 10.3969/j.issn.1001-3695.2018.07.002]

8 van der Heijden M, Lucas PJF. Extracting qualitative knowledge from medical guidelines for clinical decision-

support systems. Proceedings of AIME 2009 Workshop KR4HC 2009 Knowledge Representation for Health Care. Verona, Italy. 2017. 100–112.
9 Merhej E, Schockaert S, Mckelvey TG, *et al*. Generating conflict-free treatments for patients with comorbidity using answer set programming. Proceedings of HEC 2016 International Joint Workshop on Knowledge Representation for Health Care. Munich, Germany. 2017. 111–119.
10 Bonacin R, Pruski C, Da Silveira M. Careflow personalization services: Concepts and tool for the evaluation of computer-interpretable guidelines. Proceedings of AIME 2011 Workshop KR4HC 2011 Knowledge Representation for Health-Care. Bled, Slovenia. 2012. 80–93.
11 López-Vallverdú JA, Riaño D, Collado A. Detecting dominant alternative interventions to reduce treatment costs. Proceedings of AIME 2011 Workshop KR4HC 2011 Knowledge Representation for Health-Care. Bled, Slovenia. 2012. 131–144.
12 Kaiser K, Seyfang A, Miksch S. Identifying treatment activities for modelling computer-interpretable clinical practice guidelines. Proceedings of ECAI 2010 Workshop KR4HC 2010 Knowledge Representation for Health-Care. Lisbon, Portugal. 2011. 114–125.
13 Henriksson A, Zhao J, Dalianis H, *et al*. Ensembles of randomized trees using diverse distributed representations of clinical events. BMC Medical Informatics and Decision Making, 2016, 16(Suppl 2): 69.
14 Friedman C, Alderson PO, Austin JHM, *et al*. A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association, 1994, 1(2): 161–174. [doi: 10.1136/jamia.1994.95236146]
15 Mu XF, Wang W, Xu AP. Incorporating token-level dictionary feature into neural model for named entity recognition. Neurocomputing, 2019, 375: 43–50.
16 Azalia FY, Bijaksana MA, Huda AF. Name indexing in Indonesian translation of hadith using named entity recognition with Naïve Bayes classifier. Procedia Computer Science, 2019, 157: 142–149. [doi: 10.1016/j.procs.2019.08.151]
17 Ghiasvand O, Kate RJ. Learning for clinical named entity recognition without manual annotations. Informatics in Medicine Unlocked, 2018, 13: 122–127. [doi: 10.1016/j.imu.2018.10.011]
18 Zhang CW, Li YL, Du N, *et al*. On the generative discovery of structured medical knowledge. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2018:). London, UK. 2018. 2720–2728. [doi: 10.1145/3219819.3220010]
19 Lee HG, Park G, Kim H. Effective integration of morphological analysis and named entity recognition based on a recurrent neural network. Pattern Recognition Letters, 2018, 112: 361–365. [doi: 10.1016/j.patrec.2018.08.015]
20 Wang DS, Tiwari P, Garg S, *et al*. Structural block driven enhanced convolutional neural representation for relation extraction. Applied Soft Computing, 2020, 86: 105913. [doi: 10.1016/j.asoc.2019.105913]
21 李智恒, 桂颖溢, 杨志豪,等. 基于生物医学文献的化学物质致病关系抽取. 计算机研究与发展, 2018, 55(1):198–206.
22 Zhang Y , Lin H , Yang Z , *et al*. A hybrid model based on neural networks for biomedical relation extraction. Journal of Biomedical Informatics, 2018: S1532046418300534.
23 Li Z, Yang JS, Gou X, *et al*. Recurrent neural networks with segment attention and entity description for relation extraction from clinical texts. Artificial Intelligence in Medicine, 2019, 97: 9–18. [doi: 10.1016/j.artmed.2019.04.003]
24 张元博. 医疗知识图谱构建与应用[硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2018.
25 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述. 计算机研究与发展, 2016, 53(3): 582–600. [doi: 10.7544/issn1000-1239.2016.20148228]
26 Cardillo E, Serafini L, Tamilin A. A hybrid methodology for consumer-oriented healthcare knowledge acquisition. Proceedings of AIME 2009 Workshop KR4HC 2009 Knowledge Representation for Health-Care. Data, Processes and Guidelines. Verona, Italy. 2010. 38–49.
27 Zelaia A, Arregi O, Sierra B. Combining singular value decomposition and a multi-classifier: A new approach to support coreference resolution. Engineering Applications of Artificial Intelligence, 2015, 46: 279–286. [doi: 10.1016/j.engappai.2015.09.007]
28 Zuheros C, Tabik S, Valdivia A, *et al*. Deep recurrent neural network for geographical entities disambiguation on social media data. Knowledge-Based Systems, 2019, 173: 117–127. [doi: 10.1016/j.knosys.2019.02.030]
29 Delgado AD, Martínez R, Montalvo S, *et al*. Person name disambiguation on the web in a multilingual context. Information Sciences, 2018, 465: 373–387. [doi: 10.1016/j.ins.2018.07.024]
30 Mendes PN, Hleisen H, Bizer C. Sieve: Linked data quality assessment and fusion. Proceedings of the 2012 Joint EDBT/ICDT Workshops. Berlin, Germany. 2012. 116–123.

31 解峥, 王盼卿, 彭成. 本体的自动构建方法. 电子设计工程, 2015, 23(15): 39–41. [doi: 10.3969/j.issn.1674-6236.2015.15.012]

32 胡泽文, 孙建军, 武夷山. 国内知识图谱应用研究综述. 图书情报工作, 2013, 57(3): 131–137, 84. [doi: 10.7536/j.jssn.0252-3116.2013.03.024]

33 McDaniel M, Storey VC, Sugumaran V. Assessing the quality of domain ontologies: Metrics and an automated ranking system. Data & Knowledge Engineering, 2018, 115: 32–47.

34 官赛萍, 靳小龙, 贾岩涛, 等. 面向知识图谱的知识推理研究进展. 软件学报, 2018, 29(10): 2966–2994. [doi: 10.13328/j.cnki.jos.005551]

35 Chen HN, Luo XW. An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing. Advanced Engineering Informatics, 2019, 42: 100959. [doi: 10.1016/j.aei.2019.100959]

36 Yan D, Tian YY, Cheng J. Systems for Big Graph Analytics. Cham: Springer, 2017.

37 Jagvaral B, Lee WK, Roh JS, *et al*. Path-based reasoning approach for knowledge graph completion using CNN-BiLSTM with attention mechanism. Expert Systems with Applications, 2020, 142: 112960. [doi: 10.1016/j.eswa.2019.112960]

38 杨玉基, 许斌, 胡家威, 等. 一种准确而高效的领域知识图谱构建方法. 软件学报, 2018, 29(10): 2931–2947. [doi: 10.13328/j.cnki.jos.005552]

39 岳晓旭, 袁军鹏, 高继平, 等. 常用科学知识图谱工具实例对比. 数字图书馆论坛, 2014, (5): 66–72. [doi: 10.3772/j.issn.1673-2286.2014.05.011]

40 董立平, 郭继军. 利用 Histcite 的人胚胎干细胞引文编年图主要路径分析. 医学信息学杂志, 2010, 31(11): 38–40, 49. [doi: 10.3969/j.issn.1673-6036.2010.11.011]

41 陈超美, 陈悦, 侯剑华,等. CiteSpace Ⅱ: 科学文献中新趋势与新动态的识别与可视化. 情报学报, 2009, 28(3): 401–421.

42 Batagelj V, Mrvar A. Analiza Sieci Spolecznych Pajek. http://mrvar.fdv.uni-lj.si/pajek/. 2016.

43 刘桂锋, 杨国立. 基于 CiteSpace Ⅱ的国际太赫兹技术知识图谱研究. 图书情报研究, 2012, 5(3): 47–53.

44 赵蓉英, 许丽敏. 文献计量学发展演进与研究前沿的知识图谱探析. 中国图书馆学报, 2010, 36(5): 60–68.