

知识图谱研究综述^①

黄恒琪¹, 于娟¹, 廖晓², 席运江³

¹(福州大学 经济与管理学院, 福州 350108)

²(广东金融学院 互联网金融与信息工程学院, 广州 510521)

³(华南理工大学 工商管理学院, 广州 510641)

通讯作者: 于娟, E-mail: yujuan@fzu.edu.cn



摘要: 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, 是语义搜索、智能问答、决策支持等智能服务的基础技术之一。目前, 知识图谱的内涵还不够清晰; 且因建档不全, 已有知识图谱的使用率和重用率不高。为此, 本文给出知识图谱的定义, 辨析其与本体等相关概念的关系。本体是知识图谱的模式层和逻辑基础, 知识图谱是本体的实例化; 本体研究成果可以作为知识图谱研究的基础, 促进知识图谱的更快发展和更广应用。本文罗列分析了国内外已有的主要通用知识图谱和行业知识图谱及其构建、存储及检索方法, 以提高其使用率和重用率。最后指出知识图谱未来的研究方向。

关键词: 知识图谱; 本体; 通用知识图谱; 行业知识图谱; 知识图谱构建

引用格式: 黄恒琪, 于娟, 廖晓, 席运江. 知识图谱研究综述. 计算机系统应用, 2019, 28(6): 1-12. <http://www.c-s-a.org.cn/1003-3254/6915.html>

Review on Knowledge Graphs

HUANG Heng-Qi¹, YU Juan¹, LIAO Xiao², XI Yun-Jiang³

¹(School of Economics and Management, Fuzhou University, Fuzhou 350108, China)

²(School of Internet Finance and Information Engineering, Guangdong University of Finance, Guangzhou 510521, China)

³(School of Business Administration, South China University of Technology, Guangzhou 510641, China)

Abstract: A knowledge graph is a knowledge base that represents objective concepts/entities and their relationships in the form of graph, which is one of the fundamental technologies for intelligent services such as semantic retrieval, intelligent answering, decision support, etc. Currently, the connotation of knowledge graph is not clear enough and the usage/reuse rate of existing knowledge graphs is relatively low due to lack of documentation. This paper clarifies the concept of knowledge graph through differentiating it from related concepts such as ontology in that the ontology is the schema layer and the logical basis of a knowledge graph while the knowledge graph is the instantiation of an ontology. Research results of ontologies can be used as the foundation of knowledge graph research to promote its developments and applications. Existing generic/domain knowledge graphs are briefly documented and analyzed in terms of building, storage, and retrieval methods. Moreover, future research directions are pointed out.

Key words: knowledge graph; ontology; generic knowledge graph; domain knowledge graph; knowledge graph building

1 引言

知识图谱 (knowledge graph) 是以图的形式表现客观世界中的实体 (概念、人、事物) 及其之间关系的知

识库。2012 年, 谷歌提出了知识图谱的概念。自此, 知识图谱得到了广泛关注和应用研究, 现已发展成为语义搜索、智能问答、决策支持等智能服务的基础技术之

^① 基金项目: 国家自然科学基金 (71771054); 福建省社会科学规划项目 (FJ2016C044)

Foundation item: National Natural Science Foundation of China (71771054); Social Science Planning Project of Fujian Province (FJ2016C044)

收稿时间: 2018-12-25; 修改时间: 2019-01-03; 采用时间: 2019-01-17; csa 在线出版时间: 2019-05-25

一.但是,目前知识图谱的概念界定还不清晰,常与本体(ontology)技术一起提及,二者之间的关系尚未得到明确.另一方面,本体的研究起始时间更早,已积累了较多的研究成果和应用经验.如能确定知识图谱与本体的相关关系,就可将已有的本体研究成果作为知识图谱研究的基础,促进知识图谱更快发展.

为此,在广泛查阅知识图谱和本体的国内外相关研究成果的基础上,本文定义了知识图谱与本体之间的关系,罗列已开发的知识图谱/本体,分析其构建、存储及检索方法与技术.本文主要内容:第2节给出知识图谱的定义,辨析其与本体、知识地图、科学知识图谱等相关概念的区别,研究知识图谱的构成,并罗列分析国内外的通用知识图谱与行业知识图谱.第3节说明知识图谱在语义搜索、智能问答和个性化推荐方面的应用;第4节解释知识图谱的两种构建方法:自顶向下法和自底向上法;第5节分析知识图谱的存储与检索方法:关系型数据库、图数据库与混合方法;最后指出未来研究方向.

2 知识图谱定义

知识图谱是以图的形式表现客观世界中的实体(概念)及其之间关系的知识库.知识图谱的研究起源于语义Web.在2000年的XML大会上, Tim Berners Lee 提出了语义Web的理念,目标是为Web网页添加语义,支持机器自动处理,以提供诸如信息代理、搜索代理、信息过滤等语义服务.此后,互联网逐步从仅包含网页与网页之间超链接的文档万维网转变为包含大量描述各种实体和实体之间丰富关系的数据万维网.基于关键词的传统搜索引擎技术也逐渐开始添加语义搜索功能.2005年,美国Metaweb公司成立,致力于开发用于Web语义服务的开放共享的世界知识库. Metaweb 基于诸如维基百科、美国证券交易委员会等的公开数据集,提取现实世界中的实体(人或事物)及其之间的关系,然后以图结构存储在计算机中.2010年谷歌收购了Metaweb,获得其语义搜索技术,并于2012年提出知识图谱的概念.

2.1 知识图谱相关概念

本体(ontology)是共享概念模型的显式说明^[1],描述概念与概念间的关系;是语义Web的关键技术,用于为Web网页添加语义.语义Web理念中的本体与知识图谱,二者密切相关.本体描述概念及概念间的关系,

是大多数知识图谱的模式层,是知识图谱的概念模型和逻辑基础.知识图谱与本体的相同之处在于:二者都通过定义元数据以支持语义服务.不同之处在于:知识图谱更灵活,支持通过添加自定义的标签划分事物的类别.本体侧重概念模型的说明,能对知识表示进行概括性、抽象性的描述,强调的是概念以及概念之间的关系.大部分本体不包含过多的实例,本体实例的填充通常是在本体构建完成以后进行的.知识图谱更侧重描述实体关系,在实体层面对本体进行大量的丰富与扩充.可以认为,本体是知识图谱的抽象表达,描述知识图谱的上层模式;知识图谱是本体的实例化,是基于本体的知识库.

知识图谱采用三元组描述事实,所使用的描述语言大多是已研发的本体语言,如RDFS^[2]、OWL^[3]等.知识图谱也可以通过RDFS或OWL定义规则用于知识推理.知识图谱的关键技术也与本体很相似,涉及:(1)知识图谱构建阶段的实体抽取、关系抽取、语义解析等机器学习和自然语言处理方法和算法,(2)用于知识图谱存储的知识表示、图数据库和知识融合等方法和技术,(3)知识图谱应用阶段的数据集成、知识推理等.

除了本体之外,与知识图谱相关的概念还有知识地图和科学知识图谱.知识地图(knowledge map)将特定组织内的知识索引通过“地图”的形式串联在一起,揭示相关知识资源的类型、特征以及相互关系^[4,5].知识地图的主要功能在于实现知识的快速检索、共享和再重用,充分有效地利用知识资源^[6].知识地图是关于知识的来源的知识^[7].知识并非存储在知识地图中,而是存储在知识地图所指向的知识源中.知识地图指向的知识源包含数据库、文件以及拥有丰富隐性知识的专家或员工.有的企业应用知识地图来揭示知识的结构,实现对知识及其相关知识的检索.另外,知识地图在文献学中也有应用,即科学知识图谱.

科学知识图谱(mapping knowledge domain)是用来显示知识演化进程和知识结构的图形化与序列化的知识谱系^[8].1955年,加菲尔德提出可以将引文索引应用于检索文献的思想^[9].1965年,普赖斯等人^[10]指出,引证网络类似于当代科学发展的“地形图”,并提出使用引文网络来研究当代科学发展脉络的方法.从此,分析引文网络成为一种研究科学发展脉络的常用方法.2005年,刘则渊等^[11]引入了“科学知识图谱”的概念.科学知识图谱结合应用计量学引文分析和共现分析、图

形学、可视化技术、信息科学等学科的理论与方法^[12], 图形化的展示各领域的学科结构、各学科的研究内容、学科间的关系、识别和分析学科的发展新趋势以

及预测前沿等。

表1比较分析了知识图谱、本体、知识地图以及科学知识图谱的异同。

表1 知识图谱相关概念表

比较内容	知识图谱	本体	知识地图	科学知识图谱
学科范围	计算机科学、人工智能	人工智能、知识管理	图情学、知识管理	图情学、科学计量学
定义	描述客观世界的实体与其间关系	共享概念模型的明确的形式化规范说明	知识指南与知识管理工具	科学知识结构及演化的可视化
形式化	图、OWL	OWL	图	图
应用	语义搜索、智能问答等	知识共享、语义服务、自动推理	组织内知识索引、文献知识分析	科学社会网络分析、学科知识趋势分析

2.2 知识图谱的构成

知识图谱由数据层 (data layer) 和模式层 (schema layer) 两部分构成^[13]。

模式层是知识图谱的概念模型和逻辑基础, 对数据层进行规范约束。多采用本体作为知识图谱的模式层, 借助本体定义的规则和公理约束知识图谱的数据层。也可将知识图谱视为实例化了的本体, 知识图谱的数据层是本体的实例。如果不需支持推理, 则知识图谱 (大多是自底向上构建的) 可以只有数据层而没有模式层。在知识图谱的模式层, 节点表示本体概念, 边表示概念间的关系。

在数据层, 事实以“实体-关系-实体”或“实体-属性-属性值”的三元组存储, 形成一个图状知识库。其中, 实体是知识图谱的基本元素, 指具体的人名、组织机构名、地名、日期、时间等。关系是两个实体之间的语义关系, 是模式层所定义关系的实例。属性是对实体的说明, 是实体与属性值之间的映射关系。属性可视为实体与属性值之间的 hasValue 关系, 从而也转化为以“实体-关系-实体”的三元组存储。在知识图谱的数据层, 节点表示实体, 边表示实体间关系或实体的属性。

2.3 知识图谱的分类

从构建过程是否依赖自动抽取技术来看, 知识图谱大致可分为两类。一类是早期的本体, 如 WordNet^[14]、CYC^[15]、HowNet^[16]等。这类知识图谱大多由专业人士手工构建, 规模较小; 但其知识质量高, 能够确保准确性与完整性。另一类是从开放的互联网信息中自动抽取实体与关系构建的, 如 YAGO^[17]、DBPedia^[18]等。这类知识图谱规模大; 但因其数据源的复杂多样及自动

抽取算法的不完全准确, 可能会有大量不完整信息、噪声等。近年来, 随着知识图谱成为学界及商界的热点, 国内也重视知识图谱的研究发展, 中文的知识图谱纷纷涌现, 如 CN-DBpedia^[19]、zhishi.me^[20]等。

从覆盖范围来看, 本体可分为通用本体 (generic ontology) 与领域本体 (domain ontology); 同样的, 知识图谱也可分为通用知识图谱与行业知识图谱两类^[21]。通用知识图谱 (generic knowledge graph) 描述全面的常识性的知识, 主要应用于语义搜索, 对知识的准确度要求不高, 如百科类的 DBpedia、zhishi.me 和语言学类的 WordNet、大词林等。通用知识图谱强调知识的广度, 大多采用自底向上的方式构建, 侧重实体层的扩充, 因此也导致其大部分较难构建规范的本体层^[21]。行业知识图谱 (domain knowledge graph) 面向特定领域, 能够进行知识推理, 实现辅助分析及决策支持等功能, 如 GeoNames^[22]、中医医案知识图谱等。行业知识图谱对专业性与准确度的要求高, 这也要求其必须有严格的本体层模式, 通常采用自底向上与自顶向下结合的方式进行构建。通用知识图谱可作为行业知识图谱的构建基础, 行业知识图谱也可在构建完成后补充融合至通用知识图谱中。一般来说, 通用知识图谱的使用率更高, 是现有知识图谱的基础; 而行业知识图谱则推进了知识图谱技术融入生活, 服务于民。表2总结了部分典型中文通用知识图谱与中文通用本体; 表3总结了部分典型外文通用知识图谱与通用本体; 表4总结了部分典型中文行业知识图谱与中文领域本体; 表5总结了部分典型外文行业知识图谱与外文领域本体。

表2 部分典型中文通用知识图谱/中文通用本体一览表

知识图谱	研究机构	研发时间	实体数量	关系数量	知识源	说明	应用
HowNet	中国科学院计算机语言信息中心	1999	1.1万		专家人工构建	人工编辑、小规模、常识知识库	语义倾向计算 ^[23]
THUOCL ^[24]	清华大学自然语言处理与社会人文计算实验室		15.7万		主流网站的社会标签、搜索热词、输入法词库等	开放的中文词库, 可供免费使用与下载	
大词林 ^[25]	哈尔滨工业大学社会计算与信息检索研究中心				百度百科、同义词词林、Web 网页	包含同义、同类、上下位关系的动态层级知识体系, 可自动扩充	
Zhishi.me ^[26]	上海交通大学	2011	1000+万	1.2亿	百度百科、互动百科、中文维基百科	中文通用知识图谱	
CN-Probase ^[27]			1700万	3300万	CN-DBpedia、海量中文网页语料等	大规模中文概念图谱	自然语言问答、实体概念标注、智能知识库验证
CN-DBpedia	复旦大学知识工场实验室		1686万	2228.6万	中文百科类网站(如百度百科、互动百科、中文维基百科等)的纯文本网页	大规模通用领域百科知识图谱	短文本依存分析
PKU-PIE	北京大学计算机科学技术研究所		70+万	900+万	维基百科, 德国 DBpedia, 百度百科	百科类知识库	
搜狗知立方	搜狗	2012			百科类知识、Web 网页		语义搜索
百度知心	百度	2013			百科类知识、Web 网页		语义搜索、自动文章生成
XLore ^[28]	清华大学计算机系知识工程研究室		1.49亿	51万	中、英文维基百科、百度百科、互动百科、Wikipedia	中英文双语的百科知识图谱	

表3 部分典型外文通用知识图谱/通用本体一览表

知识图谱	研发机构	研发时间	实体数量	关系数量	知识源	说明	应用
CYC ^[29]	美国 Cycorp 公司	1984	50万	1.5万	专家人工构建	常识性知识库	
WordNet ^[30]	普林斯顿大学	1985	15万	20万	专家人工构建	人工编辑, 英文词典, 按词义组织; 155 327 个单词, 同义词集 117 597 个, 同义词集之间由 22 种关系连接	词义消歧、语义搜索 ^[31]
ConceptNet ^[32]	麻省理工媒体实验室	1999		2800万	专家人工构建	由网络用户协作半自动生成的在线常识知识库	语音识别 ^[33] 、机器翻译 ^[34] 、词义消歧 ^[35]
Freebase	Metaweb 公司, 后被谷歌收购	2005	6800万	10亿	Wikipedia	大规模开放结构数据集	
DBpedia	德国莱比锡大学与曼海姆大学	2007	458万	30亿	Wikipedia	大规模跨语言的知识库, 支持多达 125 种语言	语义标注 ^[36] 、领域共享与服务 ^[37]
YAGO ^[38]	德国马普研究所	2007	1000万	1.8亿	Wikipedia、WordNet、GeoNames	大规模跨语言的语义知识库	
TextRunner	华盛顿大学国灵中心	2007			Web 网页	从非结构化数据中自动获取知识	
WolframAlpha ^[39]	沃尔夫勒姆研究公司	2009	10万亿			基于计算的知识搜索引擎	知识搜索
ProBase ^[40]	微软	2010	265+万	2075万	Web 网页	大规模通用知识库	语义消歧
NELL ^[41]	卡内基梅隆大学	2010	519万	5亿	Web 网页	基于开放文本抽取技术, 持续更新	
WikiData ^[42]	维基	2012	453.6万		Wikipedia、用户编辑	开放、多语言、协作、结构化, 支持 281 种语言	语义搜索 ^[43]

Google KnowledgeVault	谷歌	2012	5 亿	180 亿	Wikipedia、Freebase	大规模	语义搜索
BabelNet ^[44]	罗马大学计算机科学语言计算实验室		606+万	19 亿	Wikipedia、WordNet	多语言词典知识库,覆盖 50 多种语言, 1400 万个同义词集合和 7.46 亿个词义	多语词义消歧、计算语义相关性 ^[45]
Microsoft Concept Graph ^[46]	微软	2016	1255 万	8760 万	Web 网页	以概念层次体系为中心的 知识图谱	

表 4 部分典型中文行业知识图谱/中文领域本体一览表

知识图谱	研究机构	领域	说明
基于 CNSchema 的城市知识图谱	浙江大学	交通	涵盖上海市公交站点、地铁站点的静态数据、事件流动态数据 (如公交车进出站、刷卡进出地铁站)
空气质量语义描述		环境	将空气质量数据增加上下文语义描述, 外链至气象、相关疾病、健康等数据
SciKG	清华大学	科研	展现计算机科学领域的发展, 实现计算机领域中专家与论文的搜索和推荐
影视双语知识图谱 ^[47]	清华大学计算机系知识工程实验室	影视	融合 LinkedIMDB、百度百科、豆瓣等数据源
中国旅游景点知识图谱		旅游	包含中国主要旅游景点知识 ^[48]
哺乳动物脑结构公共本体	中国科学院自动化研究所	动物学	在哺乳动物脑结构公共本体基础上, 增添中文翻译、为术语分配了 URI 且链接至德国的 DBpedia
上海图书馆名人手稿档案关联开放数据集	上海图书馆	人文	包括上海图书馆馆藏的 24 万余种手稿及档案的元数据
华人家谱关联数据集 ^[49]			从家谱文献档案中串联起多部家谱, 梳理出家族迁徙图
有色行业产业链图谱	深圳市爱智慧科技有限公司	产业	展示有色行业 (包括有色金属、贵金属、稀土、金属非金属新材料的产品产业链信息) 内部结构的关系与产品上下游的关系
微观经济学知识库		经济学	关于微观经济学的知识和逻辑图谱, 目标是应用于经济金融垂直领域的问答和推理
中文症状库	华东理工大学		包含症状实体和症状相关三元组的数据集
乳腺癌知识图谱 ^[50]	Vrije University of Amsterdam		集成乳腺癌相关知识, 包括乳腺癌临床试验数据、乳腺癌医学指南、乳腺癌电子病历、乳腺癌临床试验数据语义标注、乳腺癌医学文献等
疾病术语集	开放医疗与健康联盟	医学	包含疾病实体及疾病相关同义词, 术语集的数据来源于国内权威的临床医学术语网站
中医药学语义网络 ^[51]			以中医药学语言系统为骨架, 将中医药领域现有的术语资源和数据库资源融合起来, 构成大规模知识图谱
中医药知识图谱 ^[52]	中国中医科学院中医药信息研究所		由中医医案知识图谱、中医特色诊疗技术知识图谱、中医美容知识图谱、中医学术传承知识图谱、中医养生知识图谱、中医经方知识图谱、中医特色疗法知识地图和中医药学语言系统组成。

表 5 部分典型外文行业知识图谱/外文领域本体一览表

知识图谱	研究机构	领域	说明
Linked Life Data	保加利亚 Ontotext 公司与 LarKC 项目		超 100 亿三元组, 包含基因知识库、蛋白质知识库、疾病的知识库
UMLS	美国国家医学图书馆	生物医学	一体化医学语言系统, 可一体化检索病案记录、书目数据库、事实数据库以及专家系统中的电子式生物医学情报
Diseasome	法国国家科学研究中心、Gephi 小组、Linkfluence 和 WebAtlas 公司		以地图导向方式探索疾病之间的关系
Linked Movie Dataset		电影	61.5 万三元组数据, 描述关于演员、电影等知识
Bio2RDF ^[53]	加拿大基因组、魁北克基因组	生命科学	构建生命科学领域数据库的关联数据网络
Gene Ontology	基因本体联盟		定义一套结构化的、定义精确的、通用的受控词表, 可用于描述任何有机生物体中的基因和基因产物
FOAF ^[54]	Libby Miller and Dan Brickley	社交	通过构建一个机器可读的本体, 描述了人与人之间的关系

3 知识图谱的应用

知识图谱提供了一种更好的组织、管理和理解互联网信息的能力,可用于语义搜索、智能问答、个性化推荐等,在社交和电子商务等领域中实现价值。基于知识图谱的应用是信息领域当前的研究热点,也是促进人工智能发展的基础技术之一。

3.1 语义搜索

知识图谱是语义搜索的大脑^[55]。传统搜索引擎基于用户输入的关键词检索后台数据库中的 Web 网页,将包含搜索关键词的网页的链接反馈给用户。语义搜索(也称为语义检索)则首先将用户输入的关键词映射至知识图谱中的一个或一组实体或概念,然后根据知识图谱中的概念层次结构进行解析和推理,向用户返回丰富的相关知识。谷歌提出语义搜索后,国内百度的“知心”与搜狗的“知立方”也致力于利用知识图谱技术提升用户的搜索体验。

基于知识图谱的语义搜索能够实现^[56]:(1)以知识卡片的形式提供结构化的搜索结果。例如,当用户搜索清华大学时,知识卡片呈现出的内容,包括有学校的地址、邮编、简介、创办年份等相关信息。(2)理解用户用自然语言描述的问题,并且给出相应的答案,即简单的智能问答。例如,当用户在搜索中以提问的方式输入“世界上最大的湖泊是?”,反馈的页面能够精确地给出里海相关的信息。(3)通过已有知识图谱中实体的关联,扩展用户搜索结果,发现更多内容,反馈丰富的关联结果。例如,当用户搜索达芬奇时,除了达芬奇的个人简介之外,语义搜索还能返回他的相关画作:《最后的晚餐》、《蒙娜丽莎》及其相关人物等信息。

3.2 智能问答

智能问答指用户以自然语言提问的形式提出信息查询需求,系统依据对问题的分析,从各种数据资源中自动找出准确的答案。问答系统是一种信息检索的高级模式,能提升效率、降低人工参与成本。问答系统将知识图谱看做一个大型知识库,首先对用户用自然语言提出的问题进行分析,进而将其转化成对知识图谱的查询,最后在知识图谱中查询答案。

百度知识图谱将智能问答具体划分为5类^[57]:(1)实体检索:问题是围绕着一个实体的属性方面展开的,如“魔兽世界世界的上映时间?”(2)知识推理:问题需要通过一定的推理分析,如针对“zippo 可否带上飞机?”这个问题,需要先推理得出 zippo 是一种打火机,再针对

打火机能否上飞机这个问题进行回答。(3)开放领域问答:问题没有固定的标准答案,如“形容不会说话的成语”,“与美人鱼相关的电影”等。(4)语义解析:首先需要理解问题的语义,即按照语法规则解析问题,最后转化成数据库对应的查询,如针对“魔兽啥时候出?什么时候能看魔兽?”这样的问题可通过语义解析将其转化为“魔兽世界世界的上映时间?”这一简单查询。(5)End-to-end 在线问答,通过某些已知条件来查询实体,如“知识即美德是谁提出来的?鞠躬尽瘁死而后已说的是谁?”。

除百度外,目前还有许多问答平台引入了知识图谱技术,如华盛顿大学的 Paralex 系统^[58]、亚马逊的自然语言助手 Evi^[59]、苹果智能语音助手 Siri^[60]和出门问问^[61]手机应用等。

3.3 个性化推荐

个性化推荐是指基于用户画像,不同的用户会看到不同的推荐结果,有着重要的商业价值。电子商务网站是运用个性化推荐最典型的应用,能通过行业知识图谱的丰富知识帮助实现精准营销与推荐。如:基于商品间的关联信息以及从网页抽取的相关信息,构建知识图谱,当用户输入关键词查看商品时,基于知识图谱向用户推荐可能需要的相关知识,包括商品结果、使用建议、搭配等,通过“你还可能感兴趣的有”、“猜您喜欢”或者是“其他人还在搜”进行相关的个性化推荐。

4 知识图谱的构建

4.1 构建方法

知识图谱的构建方法有三种:自底向上、自顶向下和二者混合的方法。

4.1.1 自底向上法

自底向上的构建方法,从开放链接的数据源中提取实体、属性和关系,加入到知识图谱的数据层;然后将这些知识要素进行归纳组织,逐步往上抽象为概念,最后形成模式层。自底向上法的流程如图1所示。

图1中的知识抽取,类似于本体学习,采用机器学习技术自动或半自动地从一些开放的多源数据中提取知识图谱的实体、关系、属性等要素。知识抽取包含实体抽取、关系抽取和属性抽取。实体抽取自动发现具体的人名、组织机构名、地名、日期、时间等实体,也称为命名实体学习或命名实体识别。实体抽取的准确率和召回率将直接影响到知识图谱的质量。关系抽取是指利用语言学、统计学、信息科学等学科的方法

技术,从文本中发现实体间的语义关系.属性抽取是针对实体而言的,实体属性的抽取问题可转化为关系抽取问题.

知识融合,类似于本体集成.由于知识图谱在进行知识抽取工作时所使用的数据源是多样化的,因此可能产生知识重复、知识间关系不明确等问题.知识融合可消除实体、关系、属性等指称项与事实对象之间的歧义,使不同来源的知识能够得到规范化整合.知识融合分为:(1)实体对齐:可用于判断相同或不同数据集中的多个实体是否指向客观世界同一实体,解决一个实体对应多个名称的问题.(2)属性值填充:针对同一属性出现不同值的情况,根据数据源的数量和可靠度进行决策,给出较为准确的属性值.

知识加工:是对已构建好的数据层进行概念抽象,即构建知识图谱的模式层.知识加工包括本体构建和质量评估.基于本体形成的知识库不仅层次结构较强,并且冗余程度较小.由于技术的限制,得到的知识元素可能存在错误,因此在将知识加入知识库以前,需要有一个评估过程.通过对已有知识的可信度进行量化,保留置信度高的知识来确保知识库的准确性.

4.1.2 自顶向下法

自顶向下的知识图谱构建方法如图2所示.方法的具体步骤为:

(1)本体构建,构建知识图谱的模式层.从最顶层的概念开始构建顶层本体,然后细化概念和关系,形成结构良好的概念层次树.需要利用一些数据源提取本体,即本体学习^[62].

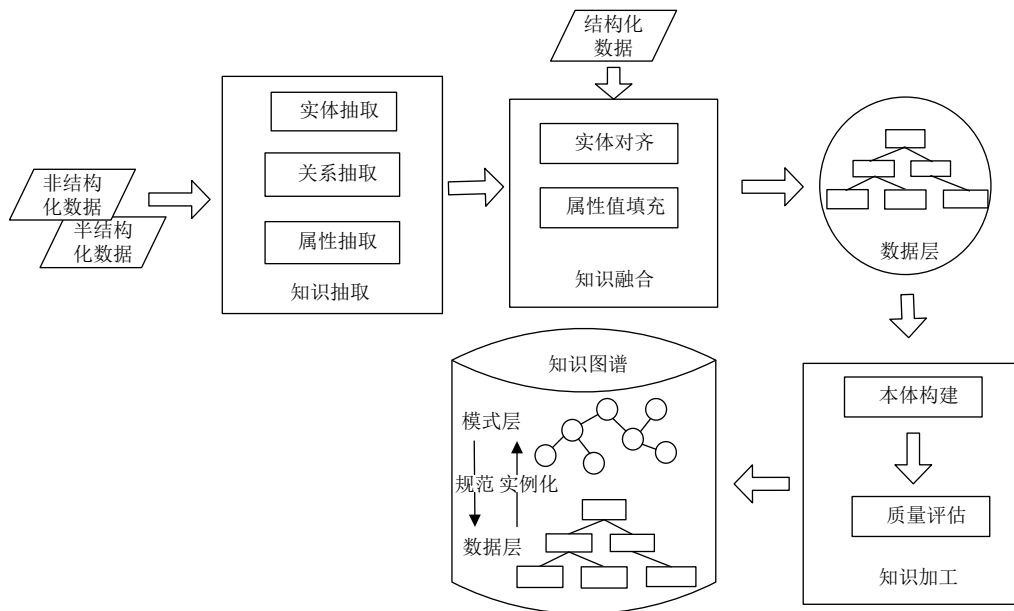


图1 自底向上的知识图谱构建流程

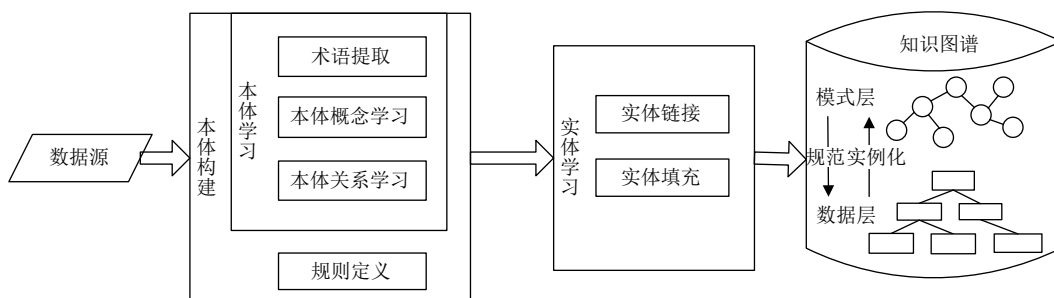


图2 自顶向下的知识图谱构建流程

(2) 实体学习, 将知识抽取得到的实体匹配填充到所构建的模式层本体中。

此外, 还可结合自顶向下和自底向上两种方法来构建知识图谱。首先, 在知识抽取的基础上归纳构建模式层, 之后可对新到的知识和数据进行归纳总结, 从而迭代更新模式层, 并基于更新后的模式层进行新一轮的实体填充。如百度百科图谱, 就是利用内外部以及用户数据采用混合方法构建所得^[58]。

上述三种知识图谱构建方法, 自顶向下法较好体现概念间层次, 但人工依赖性强、模式层更新受限, 仅适用于数据量小的知识图谱构建; 自底向上法更新快、支持大数据量的知识图谱构建, 但知识噪音大、准确性不高; 混合方法灵活性强, 但模式层构建难度大。

4.2 数据源

构建知识图谱的数据源主要有: 百科类半结构化的网页、结构化程度较低的普通网页、语料库、已构建好的知识库等。所用的数据源影响知识图谱的质量, 也影响构建过程。

4.2.1 百科网站

百科网站有固定的页面模版, 每一页面都针对某一概念或实体进行详细的介绍。这样的半结构化形式, 更易于实施知识抽取。并且, 大多数百科网站知识质量高, 权威性强, 出错率较低, 抽取所得知识的质量更高。广义来讲, 符合这些特征的网站都可视为百科网站, 如购物网站、电影、音乐网站等, 也可以通过相似的方法构建相关领域的知识图谱。

基于百科网站的知识图谱构建又可分为两类^[63]: (1) 单百科网站的深度知识抽取, 典型代表有 Dbpedia、Yago 和 CN-DBpedia。Dbpedia 和 Yago 以维基百科作为数据源; CN-DBpedia 以百度百科作为数据源。(2) 多百科网站的知识融合, 典型代表有 zhishi.me 和 X Lore。Zhishi.me 融合了百度百科、互动百科以及中文维基百科。X Lore 融合了百度百科、互动百科以及英文维基百科。这两类构建过程中的侧重点也不同, 前者更侧重于抽取, 后者不仅需要知识抽取还注重融合。

4.2.2 普通网页

尽管百科网站包含了大量的常见知识, 但其所覆盖的知识范围有限, 不能满足通用知识图谱和专业知识图谱的构建需求。因此, 结构化程度较低的普通网页是知识图谱的另一大数据来源。普通网页数据的格式

丰富多样, 没有较为一致的规范, 且包含的知识可能存在于大量的冗余和错误, 准确率较低。因此, 基于普通网页的知识抽取工作的复杂度较高, 知识融合的难度更大。此类知识图谱的构建工作重点在于知识抽取与知识融合。卡内基梅隆大学的——“永不停止的语言学习”项目 (Never-Ending Language Learning, NELL), 就是从上亿个网页中进行知识的抽取。

4.3 知识更新

随着人们对客观世界的认知加深, 信息与知识量不断增加, 知识图谱的内容也需要与时俱进, 迭代更新, 增加新的知识, 删除过时的知识。

根据知识图谱的逻辑结构, 知识图谱的更新可分为模式层更新和数据层更新。模式层更新是指本体中元素的更新, 包括概念的增加、修改、删除, 概念属性的更新以及概念之间关系的更新等。其中, 概念属性的更新操作会直接影响到所有与其直接或间接相关的子概念和实体。因此, 模式层更新多数情况下是在人工干预的情况下完成的, 需要人工定义规则, 人工处理冲突等, 实施起来有一定的复杂度。数据层更新指的是实体元素的更新, 包括实体以及实体间关系和属性值的增加、修改、删除。由于数据层的更新对知识图谱的整体架构影响较小, 通过在可靠数据源 (如百科类网站) 自动抽取的方式即可完成。

根据更新的方式, 知识图谱的更新可分为增量更新和完全更新。增量更新是以知识图谱数据源 (维基百科等) 发布出的更新内容为基础对知识图谱进行部分更新。也可以基于用户在语义搜索平台上的行为, 如反馈信息过时或搜索了一个知识图谱中没有的新词而进行相应的更新。完全更新是指间隔一定的周期, 重新将知识图谱数据源的全部数据进行一次抽取解析。完全更新的优点在于: 能较大程度保证知识图谱更新过程中的逻辑一致性, 适用于模式层的更新。但该方法代价昂贵, 且耗时长, 不能保证时效性。

5 知识图谱的存储与检索

Web 本体语言 OWL 是当前存储本体的主要形式。OWL 也适用于知识图谱的存储, 但 OWL 文档不适用于大数据量的情况, 影响查询、修改和推理速度。作为知识库的知识图谱, 其基本元素是海量的各种关系联系在一起的实体, 需要创新性的存储方式。知识图谱可

采用关系型数据库或图数据库存储,也可采用混合的方式存储。

5.1 关系型数据库

将知识图谱存入关系型数据库,是一个将知识图谱转换为 RDF 三元组进行存储的问题,即,将知识图谱的关系分解为一个个<Subject, Predicate, Object>的三元组形式,并映射至关系表,存储于关系型数据库中^[64]。根据映射方法不同,基于关系型数据库的知识图谱存储可以采用四种方法^[65]:

(1) 三元组表:构建一张只有三列的表 (Subject, Predicate, Object),以元组为单位进行存储,将知识图谱所有的 RDF 三元组都存放在这个表中。这一方法简单直观,有很好的通用性,且语义表达较为明确。但在涉及关系的查询过程中可能会面临大量的自连接操作,导致查询性能低下。

(2) 水平存储^[66]:将知识图谱中的每个实体表示为数据库表中的一行,表中的列包括有 RDF 数据集中所有的属性。水平存储的优势在于设计简单,能够高效查询面向某单个实体的属性值。而其缺点在于:① 数据稀疏性,一个实体并不可能在所有的属性上都有属性值,从而导致表中的大量空值,增加存储负载也影响查询效率。② 水平存储中表的列数量是固定的,这样的固定模式使得实体在一个属性上只能存储一个值。但在实际应用中,属性往往都会存在多个值而与水平存储的模式不符。③ 固定的模式也会令数据的变化带来较大的更新成本。数据属性的更新可能涉及到整个表结构的变化,这也使水平存储无法很好应对这类问题。

(3) 属性表:属性表针对水平存储模式中表的数据稀疏问题作出优化。根据实体的属性对实体进行分类,将属性重合度高的实体分作一类,每一类仍采用水平存储方式,避免了表中列数过多等问题。属性表对于某些查询能够提高查询性能,但将导致大部分的查询都会涉及多个表的连接或合并操作。另外,实体与属性之间关联性不强的知识图谱不适合采用此种存储方法,易导致空值问题。

(4) 垂直分割:垂直分割指的是以关系进行划分,将三元组表重写为 N 张包含两列的表, N 指的是关系的个数。每一张表都以关系(属性)为表名,其第一列是所有在这个属性上有属性值的实体,第二列是该实体在这个属性上的值。垂直分割适应于多值数据,当一个实体在一个属性上有多个属性值时,只要将其存储为多行即可。结构化较差的知识图谱也同样适用此存储方法,如果一个实体未定义某个属性,那么这个记录就不会在这种存储方式中出现,避免了空值的产生。但垂直分割的缺点在于:① 增加了表连接的运算数。② 表的增多也增加了数据更新的难度,对一个实体的更新需要涉及多个表。

5.2 图数据库

目前,基于图数据库的知识图谱存储方法是学界研究的主流。图数据库增强了关系表达,能提供完善的图查询语言,支持各种图挖掘算法。采用图数据库存储知识图谱,能有效利用图数据库中以关联数据为中心的数据表达、存储和查询。适用于存储知识图谱的图数据库如表 6 所示。

表 6 已用于知识图谱存储的图数据库

图数据库	研发机构	说明
Neo4j ^[67]	美国 Neo technology	将数据存储在图模式中,处理数据间关系的存储与查询
Graph Engine(原 Trinity)	美国 微软	基于内存的分布式大规模图数据处理引擎
AllegroGraph ^[68]	美国 Franz.Inc	基于 W3C 标准的资源描述框架图数据库,可处理链接数据和 Web 语义,支持 SPARQL、RDFS 和 Prolog
GraphDB(原 OWLIM)	保加利亚 Ontotext	可扩展的语义图数据库,符合 W3C 标准,三元组存储,SPARQL 查询引擎,提供可配置的推理支持和性能
gstore ^[69]	中国 北京大学	基于图的 RDF 存储和 SPARQL 查询系统,用于管理大型图结构数据
Virtuoso ^[70]	-----	跨平台的关系数据库,增加对 RDF 存储与管理的功能,支持 SPARQL 查询

将知识图谱存入图数据库,是将知识图谱的实体或属性值存为图数据库中的节点,而实体间的关系或属性对应存为连接各个节点的边。图数据库的存储方式更适用于涉及多重关系的查询,还可通过经典的图

算法而实现更为复杂的图匹配查询与推理等。

除了图数据库,一些非关系型数据库也可用于知识图谱的存储。例如, MongoDB^[71]作为一个基于分布式文件存储的数据库,支持无模式的数据建模方式,即,

可以按需进行模式的添加或修改. 这样的特性也适合于自底向上地构建知识图谱.

知识图谱的存储方式应考虑到其后续的使用效率. 实际应用中, 往往采用多种方法结合的方式来存储复杂的知识图谱. 例如, 采用关系数据库存储相对固定的数据, 如概念和实体的基本信息、介绍等; 采用图数据库存储属性关系类的数据; 采用内存数据库存储查询频繁的数据, 如概念的名称等.

6 结论

知识图谱提供了一种新的方法来实现知识的表示、存储和管理, 逐渐受到关注并获得一定的研究进展. 本文在介绍了知识图谱的构建、存储和应用等方面的基础上, 阐述了知识图谱与本体间的关系. 通过分析了国内外已有的知识图谱, 可见知识图谱的研究已有一定的成果, 未来的研究方向在于:

(1) 知识图谱的分布式存储. 由于知识图谱的结构特殊性, 随着数据量的增大, 如何将其进行分布式存储是一个具有现实意义的问题. 该问题包括有如何将知识图谱进行合理分割存储而不影响其功能, 知识图谱的负载均衡, 知识图谱的存储模型等.

(2) 知识图谱的推理. 知识图谱的构建推理能支持知识图谱的构建, 通过检测原有的知识是否存在逻辑矛盾以及从已知知识中发现未知的关系, 保证知识图谱的一致性与完整性, 又能丰富和扩展知识图谱. 另一方面, 知识图谱的应用推理也值得关注. 通过添加领域内的知识规则, 知识图谱的应用推理可以实现领域知识的推论, 辅助自动决策、智能问答和预测等.

(3) 目前已有的知识图谱大部分使用率和重用率不高, 甚至在构建工作完成以后就被搁置; 而另一方面, 有实际需求的企业却缺少构建知识图谱的渠道. 针对这种情况, 未来可以考虑: ① 加强知识图谱以及本体构建工程的知识理论体系建设, 以及相关人才的培养. ② 加大通用知识图谱的构建力度, 而行业知识图谱则只在有实际需求时再根据情况构建. ③ 继续加强知识图谱以及本体自动构建方法的研发, 提高构建过程的自动化程度.

参考文献

1 Gruber TR. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 1993, 5(2): 199–220.

- 2 Pan JZ, Horrocks I. RDFS(FA): Connecting RDF(S) and OWL DL. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(2): 192–206. [doi: 10.1109/TKDE.2007.37]
- 3 Mcguinness DL, Harmelen F. OWL Web ontology language overview. W3C Recommendation, 2004, 63(45): 990–996.
- 4 Brookes BC. The foundations of information science. Part III: Quantitative aspects: Objective maps and subjective landscapes. *Journal of Information Science*, 1980, 2(6): 269–275. [doi: 10.1177/016555158000200602]
- 5 马费成. 论布鲁克斯情报学基本理论. *情报学报*, 1983, 2(4): 314–324.
- 6 陈强, 廖开际, 奚建清. 知识地图研究现状与展望. *情报杂志*, 2006, 25(5): 43–46. [doi: 10.3969/j.issn.1002-1965.2006.05.016]
- 7 Zins C. Knowledge map of information science. *Journal of the Association for Information Science and Technology*, 2007, 58(4): 526–535.
- 8 陈悦, 陈超美, 刘则渊, 等. CiteSpace 知识图谱的方法论功能. *科学学研究*, 2015, 33(2): 242–253. [doi: 10.3969/j.issn.1003-2053.2015.02.009]
- 9 Garfield E. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 1955, 122(3159): 108–111. [doi: 10.1126/science.122.3159.108]
- 10 de Solla Price DJ. Networks of scientific papers. *Science*, 1965, 149(3683): 510–515. [doi: 10.1126/science.149.3683.510]
- 11 陈悦, 刘则渊. 悄然兴起的科学知识图谱. *科学学研究*, 2005, 23(2): 149–154. [doi: 10.3969/j.issn.1003-2053.2005.02.002]
- 12 秦长江, 侯汉清. 知识图谱——信息管理与知识管理的新领域. *大学图书馆学报*, 2009, 27(1): 30–37. [doi: 10.3969/j.issn.1002-1027.2009.01.007]
- 13 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]
- 14 Fellbaum C. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press, 1998.
- 15 Guha RV, Lenat D. CYC: A large-scale investment in knowledge infrastructure. *Applied Artificial Intelligence*, 1990, 5(1): 45–86. [doi: 10.1080/08839519108927917]
- 16 You B, Liu XR, Li N, *et al.* Using information content to evaluate semantic similarity on HowNet. *Proceedings of 2012 Eighth International Conference on Computational Intelligence and Security*. Guangzhou, China, 2013.
- 17 YAGO (database). <http://www.yago.com/>, [2015-01-08].

- 18 Auer S, Bizer C, Kobilarov G, *et al.* DBpedia: A nucleus for a web of open data. Proceedings of the Semantic Web, International Semantic Web Conference, Asian Semantic Web Conference, ISWC 2007+Aswc 2007. Busan, Korea. 2007. 722–735.
- 19 Xu B, Xu Y, Liang JQ, *et al.* CN-DBpedia: A never-ending Chinese knowledge extraction system. Proceedings of the 30th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Arras, France. 2017. 428–438.
- 20 Niu X, Sun XR, Wang HF, *et al.* Zhishi.me-weaving Chinese linking open data. Proceedings of the 10th International Semantic Web Conference. Bonn, Germany. 2011. 205–220.
- 21 胡芳槐. 基于多种数据源的中文知识图谱构建方法研究 [博士学位论文]. 上海: 华东理工大学, 2015.
- 22 GeoNames. <http://www.geonames.org/>, [2018-02-27].
- 23 朱嫣岚, 闵锦, 周雅倩, 等. 基于 HowNet 的词汇语义倾向计算. 中文信息学报, 2006, 20(1): 16–22.
- 24 THUOCL: 清华大学开放中文词库. <http://thuocl.thunlp.org/>, [2018-09-07].
- 25 大词林. <http://www.bigcilin.com/>, [2017-03-27].
- 26 Zhishi.me. <http://zhishi.me/api>. [2018-12–23].
- 27 CN-Probbase. <http://kw.fudan.edu.cn/>, [2016-05-18].
- 28 XLORE. <http://xlore.org/>, [2019-03-23].
- 29 Lenat DB. CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 1995, 38(11): 33–38. [doi: 10.1145/219717.219745]
- 30 Miller GA. WordNet: A lexical database for English. Communications of the ACM, 1995, 38(11): 39–41. [doi: 10.1145/219717.219748]
- 31 高雪霞, 炎士涛. 基于 WordNet 词义消歧的语义检索研究. 湘潭大学自然科学学报, 2017, 39(2): 118–121.
- 32 Speer R, Havasi C. ConceptNet 5: A large semantic network for relational knowledge. In: Gurevych I, Kim J, eds. The People’s Web Meets NLP. Berlin: Springer, 2013. 161–176.
- 33 Lieberman H, Faaborg A, Daher W, *et al.* How to wreck a nice beach you sing calm incense. Proceedings of the 10th International Conference on Intelligent User Interfaces. San Diego, CA, USA. 2005. 278–280.
- 34 de Medeiros Caseli H, Sugiyama BA, Anacleto JC. Using common sense to generate culturally contextualized machine translation. Proceedings of the NAACL Hlt 2010 Young Investigators Workshop on Computational Approaches To Languages of the Americas. Los Angeles, CA, USA. 2010. 24–31.
- 35 Speer R, Lowry-Duda J. ConceptNet at SemEval-2017 Task 2: Extending word embeddings with multilingual relational knowledge. arXiv: 1704.03560, 2017.
- 36 Nguyen TN, Takeda H, Nguyen K, *et al.* A novel method to predict type for DBpedia entity. In: Sieminski A, Kozierekiewicz A, Nunez M, *et al.*, eds. Modern Approaches for Intelligent Information and Database Systems. Cham: Springer, 2018.
- 37 朝乐门, 张勇, 邢春晓. DBpedia 及其典型应用. 现代图书情报技术, 2011, 27(3): 80–87.
- 38 Suchanek FM, Kasneci G, Weikum G. YAGO: A large ontology from wikipedia and WordNet. Journal of Web Semantics, 2008, 6(3): 203–217. [doi: 10.1016/j.websem.2008.06.001]
- 39 Thrasher EP, Perry AD. High-leverage apps for the mathematics classroom: WolframAlpha. The Mathematics Teacher, 2015, 109(1): 66–70. [doi: 10.5951/mathteacher.109.1.0066]
- 40 Wu WT, Li HS, Wang HX, *et al.* Probbase: A probabilistic taxonomy for text understanding. Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. Scottsdale, AZ, USA. 2012. 481–492.
- 41 NELL. <http://rtw.ml.cmu.edu>, [2015-12-27].
- 42 Erxleben F, Günther M, Krötzsch M, *et al.* Introducing wikidata to the linked data web. Proceedings of the 13th International Semantic Web Conference. Riva del Garda, Italy. 2014. 50–65.
- 43 贾君枝, 薛秋红. Wikidata 的特点、数据获取与应用. 图书情报工作, 2016, 60(17): 136–141, 148.
- 44 Navigli R, Ponzetto SP. BabelNet: Building a very large multilingual semantic network. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden. 2010. 216–225.
- 45 Navigli R, Ponzetto SP. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 2012, 193: 217–250. [doi: 10.1016/j.artint.2012.07.001]
- 46 Microsoft Concept Graph. <http://concept.research.microsoft.com/>.
- 47 王巍巍, 王志刚, 潘亮铭, 等. 双语影视知识图谱的构建研究. 北京大学学报(自然科学版), 2016, 52(1): 25–34.
- 48 Zeng Y, Zhang T, Hao H. Active recommendation of tourist attractions based on visitors interests and semantic relatedness. In: Ślęzak D, Schaefer G, Vuong ST, *et al.*, eds. Active Media Technology. Cham: Springer, 2014. 263–273.
- 49 华人家谱关联数据集. <http://openkg.cn/dataset/jp>.
- 50 乳腺癌知识图谱. <http://wasp.cs.vu.nl/BreastCancerKG/>.

- 51 中医药学语义网络. <http://www.tcminformatics.org/wiki/index.php/graph>.
- 52 中医药知识图谱. <http://www.tcmkb.cn/kg/index.php>.
- 53 Belleau F, Nolin MA, Tourigny N, *et al.* Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 2008, 41(5): 706–716. [doi: 10.1016/j.jbi.2008.03.004]
- 54 Sohn JS, Chung IJ. Dynamic FOAF management method for social networks in the social web environment. *The Journal of Supercomputing*, 2013, 66(2): 633–648. [doi: 10.1007/s11227-012-0847-x]
- 55 Amerland D. 谷歌语义搜索. 程龚, 译. 北京: 人民邮电出版社, 2015. 12–31.
- 56 曹倩, 赵一鸣. 知识图谱的技术实现流程及相关应用. *情报理论与实践*, 2015, 38(12): 127–132.
- 57 赵世奇. 百度百科图谱: 人工智能的知识心脏. 福州: 中国计算机大会, 2017.
- 58 Fader. Paralex. <http://knowitall.cs.washington.edu/paralex>, [2016-05-08].
- 59 Wikipedia. Evi. [http://en.wikipedia.org/wiki/Evi_\(software\)](http://en.wikipedia.org/wiki/Evi_(software)), [2016-03-18].
- 60 Siri. <https://www.apple.com/ios/siri/>, [2016-05-02].
- 61 Wikipedia. Mobvoi. <https://en.wikipedia.org/wiki/Mobvoi>, [2018-03-20].
- 62 Wong W, Liu W, Bennamoun M. Ontology learning from text: A look back and into the future. *ACM Computing Surveys*, 2012, 44(4): Article No.20.
- 63 徐波. 百科知识图谱构建. http://cips-upload.bj.bcebos.com/ccks2017/upload/百科知识图谱构建_徐波 CCKS2017.pdf.
- 64 RDF Access to Relational Databases. <http://www.w3.org/2003/01/21-RDF-RDB-access/>, [2003-01-21].
- 65 邹磊. 浅谈知识图谱数据管理. <http://blog.openkg.cn/邹磊-浅谈知识图谱数据管理/>, [2017-04-27].
- 66 Pan ZX, Heflin J. DLDB: Extending relational databases to support semantic web queries. *Proceedings of the First International Workshop on Practical and Scalable Semantic Systems*. Sanibel Island, FL, USA. 2003. 109–113.
- 67 Neo4j. <https://neo4j.com/>, [2018-03-06].
- 68 AllegroGraph. <https://allegrograph.com/>, [2018-02-08].
- 69 Zou L, Özsu MT, Chen L, *et al.* gStore: A graph-based SPARQL query engine. *The VLDB Journal*, 2014, 23(4): 565–590. [doi: 10.1007/s00778-013-0337-7]
- 70 Erling O, Mikhailov I. Virtuoso: RDF Support in a native RDBMS. de Virgilio R, Giunchiglia F, Tanca L. *Semantic Web Information Management*. Berlin, Heidelberg: Springer, 2010. 501–519.
- 71 mongoDB. <https://www.mongodb.com/>, [2018-03-08].