知识图谱的发展与构建

李 涛[1,2],王次臣[1,2],李华康[1,2]

(南京邮电大学 1. 计算机学院; 2.江苏省大数据安全与智能处理实验室,江苏 南京 210003)

摘 要: 知识图谱作为一种智能、高效的知识组织方式,能够帮助用户迅速、准确地查询到自己需要的信息。本文通过回顾学者及科研机构或公司对知识图谱的研究内容,对知识图谱的发展和构建方法作了全面的介绍,包括知识图谱概念的起源、发展以及最终形成;构建知识图谱的数据来源;构建过程中涉及的方法,包括本体和实体的抽取,图谱的构建、更新、维护,以及面向知识图谱的内部结构挖掘和外部扩展应用。最后,对知识图谱的未来发展方向和面临的挑战作了展望。虽然现在已经有很多知识图谱被应用到各类系统中,但是其基础理论和应用技术,仍需展开进一步的研究。

关键词: 知识图谱; 构建方法; 实体; 知识挖掘; 扩展应用

中图分类号: TP39 **文章编号:** 1005-9830(2017)01-0022-13

DOI: 10.14177/j.cnki.32-1397n.2017.41.01.004

Development and construction of knowledge graph

Li Tao[1,2], Wang Cichen[1,2], Li Huakang[1,2]

(1.School of Computer Science; 2.Jiangsu Province Key Lab of Big Data Security and Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Knowledge graph, as an intelligent and efficient way for knowledge organization, enables users to quickly and accurately query the information they need. In this paper, we provide a comprehensive survey on the development and construction of knowledge graph by reviewing and summarizing recent advances in the research and practice of knowledge graph systems in the relevant literature. In particular, our introduction includes the concept origin, development, and eventual formation of the knowledge graph, various data sources for the knowledge graph, the ontology construction and the entity extraction, and the process of knowledge mining, updating, and maintenance. Finally, we discuss the technical challenges, development trends, and future research directions of knowledge

收稿日期: 2016-07-25 **修回日期:** 2016-12-18

基金项目: 国家自然科学基金(61502247,11501302,61502243,91646116);中国博士后科学基金(2016M600434);江苏省科技支撑计划(社会发展)项目(BE2016776);江苏省“六大人才高峰”项目(XYDXXJS-CXTD-006);江苏省博士后科研基金(1601128B)资助

作者简介: 李涛(1975-),男,博士,教授,主要研究方向:数据挖掘,E-mail: towerlee@njupt.edu.cn。

引文格式: 李涛,王次臣,李华康.知识图谱的发展与构建[J].南京理工大学学报,2017,41(1):22-34.

投稿网址: http://zrxuebao.njust.edu.cn

graph.In summary,the theory and the associated techniques of knowledge graph is of great research significance.However,there are still many technical challenges,which need further investigation,in building and using the knowledge graph.

Key words: knowledge graph; construction methods; entity; knowledge mining; extended application

20 世纪中叶,普赖斯等人[1]提出使用引文网络来研究当代科学发展脉络的方法,首次提出了知识图谱的概念。1977 年,知识工程的概念在第五届国际人工智能大会上被提出,以专家系统为代表的知识库系统开始被广泛研究和应用[2]。到 20 世纪 90 年代,机构知识库[3]的概念被提出,自此关于知识表示、知识组织的研究工作开始深入开展起来。机构知识库系统被广泛应用于各科研机构和单位内部的资料整合及对外宣传工作[4]。

进入 21 世纪,随着互联网的蓬勃发展以及知识的爆炸式增长,搜索引擎被广泛使用。但面对互联网上不断增加的海量信息,仅包含网页和网页之间链接的传统文档万维网已经不能满足人们迅速获取所需信息的需求。人们期望以更加智能的方式组织互联网上的资源,期望可以更加快速、准确、智能地获取到自己需要的信息。为了满足这种需求,知识图谱应运而生。它们力求通过将知识进行更加有序、有机的组织,对用户提供更加智能的访问接口,使用户可以更加快速、准确地访问自己需要的知识信息,并进行一定的知识挖掘和智能决策。从机构知识库到互联网搜索引擎,近年来不少学者和机构纷纷在知识图谱上深入研究,希望以这种更加清晰、动态的方式展现各种概念之间的联系,实现知识的智能获取和管理[5,6]。

2012 年 11 月 Google 公司率先提出知识图谱(Knowledge graph,KG)的概念,表示将在其搜索结果中加入知识图谱的功能。据 2015 年 1 月统计的数据,Google 构建的 KG 已拥有 5 亿个实体,约 35 亿条实体关系信息,已被广泛用于提高搜索引擎的搜索质量。另一个代表性的知识图谱系统是微软公司构建的 Probase[7]。根据微软公司官网上的数据显示,截至 2016 年 4 月,Probase 已拥有总量超过千万级的概念,其中核心概念大概有 270 万个,Probase 已成为知识库系统中拥有概念数最多的系统。上海交通大学的 zhishi.me 是国内构建的最早的知识库,zhishi.me 知识库通过整合维基百科(中文)、百度百科、互动百科中的数据以提供关联开放数据(Linking open data,LOD)的服务给知识库用户。中国科学院机构知识库(Chinese academy of sciences institutional repository,CAS-IR)对 DSpace 软件进行的二次开发[8]。截止到 2013 年 9 月,CAS-IR 累计采集和保存超过 44 万个的科研成果,其中,超过 70%的科研成果可获取全文,CAS-IR 是目前国内机构知识库网络中规模最大的一个;此外,国内知名搜索引擎公司也纷纷投入对知识图谱的构建,并在其搜索引擎中添加了知识图谱的功能,比如百度的“知心”和搜狗的“知立方”。

本文通过整理国内外学者对知识图谱相关概念的研究,以及搜集各科研机构或公司部署知识图谱系统的相关技术资料,对知识图谱的发展历史、构建过程、知识挖掘以及更新维护作了系统全面的介绍,希望对国内中文知识图谱系统的发展起到一定的推动作用。

1　知识图谱的发展历程

1.1　起源:知识图谱(Mapping Knowledge Domain)

1955 年,加菲尔德发表了一篇题为《Citation Indexes for Science: A New Dimension in Docu-mentation through Association of Ideas》的论文[9],提出了将引文索引应用于检索文献的思想。1965 年,普赖斯在《Networks of Scientific Papers》一文[1]中指出,引证网络——科学文献之间的引证关系,类似于当代科学发展的“地形图”,从此分析引文网络开始成为一种研究当代科学发展脉络的常用方法,进而形成了知识图谱(Mapping Knowledge Domain)的概念[10]。杨思洛等人[11]使用知识图谱的思想方法对中国知网中的关于知识图谱的论文从论文的发表日期、发表的期刊、作者、所属机构、高影响因子论文的分布等角度分析了知识图谱研究在中国的发展现状。通过可视化的分析,最终得出了未来研究应着重在理论研究方面、方法和工具以及应用研究方面的预测。

1.2　发展:知识库(Knowledge Base)

1977 年,在第五届国际人工智能会议上,美

国计算机科学家 B.A.Feigenbaum 首次提出知识工程的概念。知识工程是通过存储现存的知识来实现对用户的提问进行求解的系统,其中最典型和成功的知识工程的应用是专家系统。作为知识工程的重要组成部分,知识库是经过分类和序化,根据一定格式将相互关联的各种知识存储在计算机中。和一般的数据库系统相比,知识库添加了对知识结构的分析功能,对于知识的组织更加强调针对性和目的性,是更高效、更智能,对用户更加友好的知识服务系统。

对于知识库的研究首先需要讨论的是知识的表达和组织的基本问题。鄢珞青等人[12]重点研究了知识库中知识表达的问题,并提出了知识点的概念,此外还讨论了各种知识表达的类型等。王知津等人[13]在全面、系统、深入地对知识组织的理论、方法及应用进行了分析之后,提出了科学性原则、多维性原则等十大原则。王军等人[14]着重对互联网环境下知识的组织结构进行了系统化讨论,针对网络知识组织系统(Network knowledge organization system,NKOS)的类型和表示、互操作性、标准和规范、生成和维护以及其应用作了详细的介绍。知识的表示和组织需要服从于知识库系统整体的需求定位以及框架。目前,较通用的知识表示框架通常采用面向对象的思想,将知识拆解成实体、实体属性以及实体之间的关系。

近年来,深度学习的理论方法取得了重大的成功,知识的表示学习也逐渐成为目前研究的热点。知识表示学习旨在对于知识库中的实体和关系进行表示学习,将知识中蕴含的语义信息表示为稠密低维实值向量,从而在低维空间中实现高效计算实体和关系的语义联系,不但有效解决数据稀疏的问题,而且使知识获取、融合和推理的效果得到显著的提升。刘知远等人[15]系统地介绍了知识表示学习的进展和主要的表示学习算法,并对知识表示学习的未来发展作了展望。国外关于知识库的研究更加侧重于实践方面,并且主要是针对 NKOS 进行了相关的研发工作,例如对于在线图书馆的研究等[16]。

对于特定的机构,其内部的特定领域的知识相对较少,容易通过知识库的理论和方法有效地组织和管理知识。作为机构知识基础设施,知识库对于机构内部知识的保存、管理、访问、宣传、答疑等工作都能起到重大的作用,同时可以用于预测和决策支持。据不完全统计,截至 2013 年 11 月,全球提供开放服务的机构知识库已经超过 3 500个,从开放服务的机构知识库中可以获得的科研文献已经超过5 200万篇[4]。根据全球机构库统计网站开放获取知识库名录(The directory of open access repositories,OpenDOAR)的数据,截至 2014 年 4 月,大约有2 616个知识库已在该网站注册,这其中包含机构知识库2 212个,占 84.56%。在国内,中科院知识库始建于 2007 年,建设完毕以来,全民可免费阅读、下载和利用中科院 100 多个研究所在知识库中分享的科研成果。另外,许多高校也已经构建或开始构建自己的知识库系统。

1.3 形成:知识图谱(Knowledge Graph)

2012 年,Google 率先提出知识图谱(Knowledge Graph)的概念。知识图谱由知识以及知识之间的关系组成。知识或者说实体的内部特性使用属性—值对(Attribute-value pair,AVP)来表示。知识之间的关系通过两个实体之间相连接的边来表示。

这里的知识图谱,即 Knowledge Graph,与最开始的用于可视化科学文献引用网络的知识图谱,即 Mapping Knowledge Domain,在概念上已经有了较大的变迁。在下文的讨论中,除非另外说明,所说的知识图谱均指 Knowledge Graph。

知识图谱与知识库在理论和方法上都存在相似的地方,即都是通过更加有效和智能地保存、管理已有的知识,同时对外提供一个便捷访问所需知识的接口,满足人们对于所需知识高效地、准确地获取需求。然而,知识图谱和知识库的区别也是明显的。知识库更多的建立在机构内部,为机构内部人员和需要访问该机构的人们提供服务,知识库中所包含的知识都是该机构领域内的知识。从这个角度讲,作为互联网搜索引擎高度发展之后衍生出来的一个概念,知识图谱的含义要更加的宽泛很多。可以说,知识图谱是一个更大的、包含世界上所有机构知识库的知识集合。知识图谱里面的知识应该包含人们生活中的万事万物,包含人类文明所发现和创造的所有知识。当然要建立这样一个庞大的知识图谱不是一蹴而就的事情,但这正应该是知识图谱的愿景。

与传统的基于关键字匹配的搜索引擎工作原理不同的是,知识图谱利用概念、实体的匹配度返回给用户与搜索相关的更全面的知识体系,其工作原理如图 1 所示。

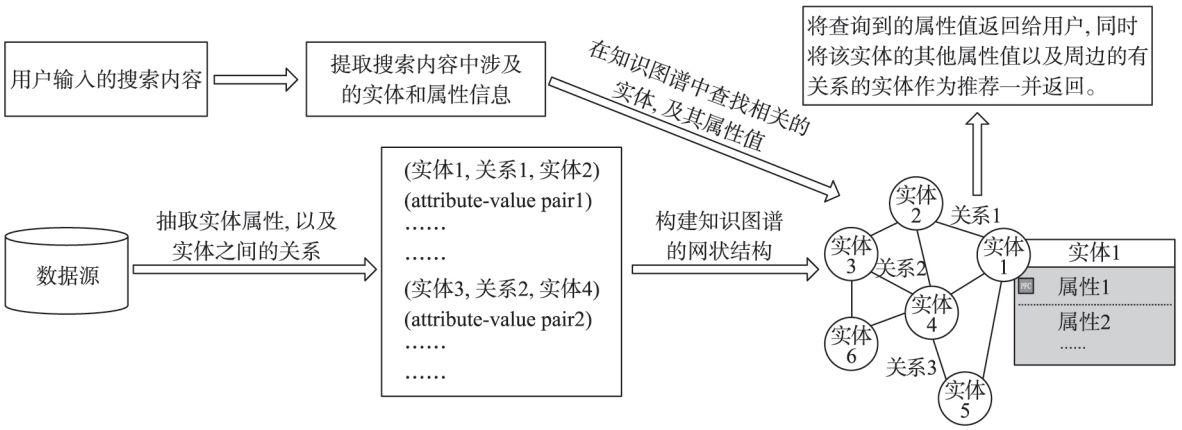

图 1　知识图谱的工作原理

Google 公司作为全球最大的搜索引擎公司，拥有最多数量的互联网用户，是最有需要也是最有资源去建立一个庞大的知识图谱的。Google 采用多种语言对知识图谱中的实体、属性和实体之间的关系进行描述，逐步将知识图谱的理论方法运用到搜索引擎中来提高搜索质量。紧随 Google 的步伐，国内的搜索引擎公司也陆续着手构建自己的知识图谱。百度的“知心”通过筛选搜索结果，将内容相近的信息组织在一起，以知识图谱的形式重新呈现出来，达到搜索直接得到答案的效果。搜狗的“知立方”可以处理海量的互联网碎片化信息，通过“语义理解”重新优化计算搜索结果，向用户呈现最核心的信息。但美中不足的是，国内的“知心”和“知立方”，并不支持多语言，知识图谱的规模也小于 Google 的 Knowledge Graph。

随着不断的探索研究，知识图谱作为一种知识管理的新思路，已经不再仅仅局限在搜索引擎应用，在各种智能系统（如 IBM Watson），以及数据存储（如 graph database，neo4j）领域也已崭露头角。

2　知识图谱的数据来源

知识图谱的数据来源是构建知识图谱面临的第一个问题。总体上可以将知识图谱的数据来源分为从网页上爬取数据和从数据库等结构化的数据集合中抽取得到。

2.1　网页数据

互联网中的无数的网页是知识图谱最大的数据来源。根据获取网页数据的难易程度，又可以细分为百科类网页、普通网页以及深网数据。

以维基百科为代表的百科类网站包含大量的知识，而且这些知识基本是以结构化的形式存储，加上不同垂直网站上特有领域的海量数据，可囊括很大一部分的常识性知识[5]。百科类网站的页面结构是由它们自己的数据模式生成的，对每一个百科网站，都可以用一个页面模版来提取其中的数据。百科类网站中的内容虽然是以形式化的网页存在，但其中包含了许多结构化的信息，如文章标题、分类标签、分类系统、信息模块、重定向、消歧、摘要等[6]。另一个可以作为百科类网站数据来源的是 Freebase。Freebase 的知识数量多于维基百科，并且拥有组织良好的知识结构，可以不使用任何复杂的实体识别规则即可得到高价值的知识。各组织机构包括百度和搜狗在构建知识图谱的过程中通常都会引入 Freebase 中的知识。

百科类网站中虽然包含了大量的常见的规范化知识，但并不能满足知识图谱的需求。结构化程度较低的普通网页是知识图谱构建的又一大信息来源。由于普通网页的样式千差万别，其中包含了大量的冗余信息、不可信的信息，所以从网页的非结构数据中提取知识的准确度很低。为了方便地抽取隐藏普通网页中的知识，一个通用的方法是构造面向网站的包装器[17]。其基本思想是：一个网站的风格（包括页面的布局和知识的组织）具有相似性，利用这种相似性，只需要对当前网站的几个有代表性的网页进行标注，然后就可以利用模式学习算法实现对网站中所有网页的自动化的知识抽取。显然，这种方式的缺陷是可能漏掉一些重要的知识或者产生错误的抽取结果。为了克服包装器带来的固有缺陷，一种比较通用的做法是，通过手动调整或添加适当的模式来挖掘知识。最后，利用手动对挖掘到的数据进行评

价,也可以手动增添一些标注过的网页,达到主动学习效果[6]。

卡耐基梅隆大学 Tom Mitchell 教授主持的“永不停止的语言学习”(Never-ending language learning,NELL)[18]项目从上亿个网页中根据用户提交的内容挖掘知识实体以及这些实体之间的关联,截至2015年8月,NELL项目已包含了超过5千万条知识。

深网数据是另一个有价值的数据来源。然而,深网数据使用通用的爬虫通常难以获得,有三方面原因。第一,数据的数量巨大,当超过站点对于网络带宽的限制之后,将导致网络站点拒绝访问;第二,有关站点的数据可能涉及知识产权的问题,从而站点采用相关的反爬技术,导致爬取错误或无法爬取;第三,网络爬虫算法需要根据不同的网站内容组织形式,采用不同的解析算法解析爬取的内容。因此,如果要使用深网中的数据,需要有专门的获取方式。现在已经出现了专门从事数据买卖的公司,同时,也有一些公司通过收购来获取更多的数据[19],像阿里收购高德地图,可以获得各种地理知识库数据。

2.2 数据库

在数据库技术广泛应用的今天,比较大型的网站通常通过部署数据库服务器,将几乎所有的数据都存储于数据库中。数据库包括关系数据库和面向对象数据库,包含结构整齐,顺序存储的数据,便于知识图谱对数据重新组织。常见的关系数据库采用的是经典的关系模型,二维形式的表格非常易于理解,通常供行业内部使用,主要用于构建行业知识图谱。在行业知识图谱构建时,通常以这些结构化的数据为起点,进而扩充其他数据。

资源描述框架(Resource description framework,RDF)是通用的用于描述实体信息的表示框架。在RDF框架中使用(实体1,关系,实体2)的三元组表示实体以及实体之间的关系。LOD[20]在实现语义网知识库的过程中往往通过设置RDF链接来完成。对于通常可以达到上亿规模的RDF三元组,一般采用数据库进行有效存储和查询。因此,LOD中结构完好的RDF三元组是又一个重要的数据来源。Google公司在利用搜索引擎积累的知识的基础上,通过整合LOD中的数据,来扩大实体的数量,扩充知识图谱的知识储备。曾锦麒[21]采用RDF对语义知识进行表示,并将其应用到网上求职招聘系统中,构建了网上招聘系统的本体模型。在采用RDF表示知识时,对于知识的时效性问题要单独进行考量,需要设计快速动态更新三元组的算法。

搜索日志是另一个重要的知识图谱的数据来源。搜索日志记录的是用户对知识图谱的各种查询,其中往往包含着各种最新的实体和其属性。日志记录除了保存在数据库中,也可以采用XML等纯文本方式保存,通常都具有良好的数据结构。从搜索日志中挖掘实体的一般办法是被称为基于Bootstrapping的多类别协同模式学习[19]。通过分析用户的检索词和点击浏览行为,可以推测出用户认可的或者偏好的相关知识对象。

3 知识图谱的构建

知识图谱的构建过程可以分为自顶向下和自底向上两种方式。自顶向下的构建过程如图2所示,首先从数据源中学习本体,得到术语、顶层的概念、同义和层次关系以及相关规则,然后进行实体学习的过程,将实体纳入前面的概念体系中。自底向上的构建过程与此相反,从归纳实体开始,进一步进行抽象,逐步形成分层的概念体系。在实际的构造过程中,可以先后混合使用两种方式,来提高实体抽取的准确度。

3.1 本体学习

本体的概念最先起源于哲学领域,表示的是客观存在的一个全面的说明。后来,在人工智能和信息技术的发展过程中引用了本体的概念,同时赋予本体新的含义。比如,利用本体的思想,在灾难信息管理中对短文本进行处理[22]以及日志挖掘[23]的工作;Jiang Yexi 等人[24]探讨了利用本体来进行云服务的智能推荐和配置。本体学习的过程主要包括术语、同义词、概念、分类关系以及公理和规则抽取。

3.1.1 术语抽取

术语抽取是本体构建的第一步。术语是知识图谱中的概念、实体或属性的语言学上的表示形式,术语抽取的目标是找到用于表示概念、实体或属性的标记集合。比如爬行动物、性别等都可以作为一个术语。术语抽取的实现方法有多种,主要包括下面几类。基于字典的方法通过定义一些包含术语的字典,从待处理文本查找字典中定义的术语;基于规则的方法则通过定义术语在语法

上的一些规则，从待处理文本中找到匹配规则的术语；基于统计的方法一般是通过统计术语出现的次数来对待处理文本中的潜在术语进行预测；基于机器学习的方法可以对术语的语法规则或者上下文的特征进行学习，从而实现对待处理文本的术语抽取。

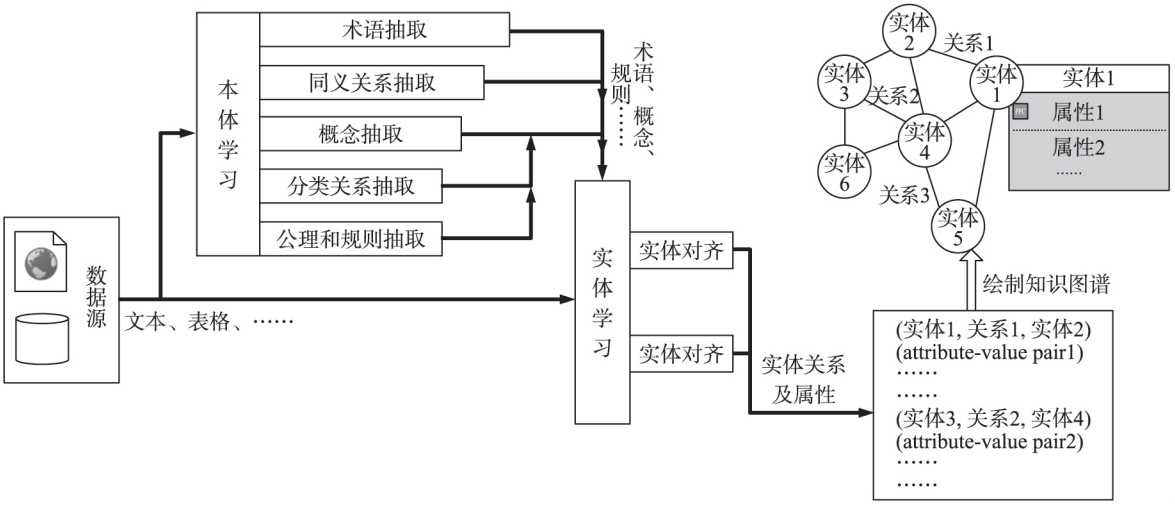

图 2　自顶向下构建知识图谱

梁铭等人[25]基于双语语料对齐的英汉平行语料库，其中中文经过了分词处理，设计了一种自动从待处理文本中实现抽取术语的算法，取得了较好的术语抽取效果。曹浩等人[26]提出了一种基于模式挖掘的机器学习方法，此方法的优点在于解决了固定模式带来的局限性，而且摆脱了开始时选择术语不当的影响，提升了术语抽取的正确率。这两种方法都依赖于使用训练语料库，在互联网环境下，对于新出现的文本特征往往不能及时有效地进行抽取。

3.1.2　同义关系抽取

同义关系是指在概念层面上相同或相似的实体，如“old man，daddy，dad，father，male parent”这一组词都是“父亲”的意思。同义关系抽取的目标是寻找那些字面不同但是指代同一概念、实体或属性的术语。

传统的基于模板的同义关系抽取方法灵活性不够，模板的覆盖率不高，导致该方法的正确率和召回率都比较低。孙霞等人[27]提出了一种自动抽取同义关系的机器学习算法。该方法具有领域自适应性，可以将训练好的分类器应用到不同的领域文本中，与基于模板的方法相比较，抽取结果的精度有了普遍提高。随着统计语言模型和深度算法的逐渐成熟，对于寻找自然语言中的同义关系的任务通常可以取得较满意的正确率。

3.1.3　概念抽取

概念的含义与实体相近，但在有的知识图谱系统中，有必要将概念和实体的含义相区别。概念的含义比实体的含义更加抽象，是比较普遍的想法、观念或充当命名实体、事件或关系的范畴或一类的实体。比如，城市是一个概念，而深圳应该作为一个实体。常见的概念抽取的方法包括使用语言学的方法、使用统计学的方法以及两者相结合的方法。

基于语言学的方法也称为基于规则的方法，当抽取的目标是具有语意知识的术语时，基于规则的方法准确率高，计算量小，但是该方法与具体的语言相关，因此，对语言有较高的依赖。统计的方法通常是基于假设的，其核心思想是认为词在领域内的相关程度可以用词在领域内出现的频率来代表，可以根据设定阈值，对于领域内的术语进行抽取。该方法能高效地识别术语，无需人工构建领域词典，没有语言限制。但是该方法计算量比较大，无法抽取低频术语，对于词组构成的术语抽取准确率不高。关键[28]给出了一种基于语言学和基于统计学的多策略的概念识别方法，该方法解决了单纯使用统计学无法抽取多词短语概念和低频概念的不足，提高了领域内部概念抽取的效果。概念的抽取类似于术语，但概念比术语更加抽象，通常需要从语言学和统计语言模型两个角度入手，才能取得较好的抽取效果。

3.1.4　分类学关系抽取

分类学关系可称为概念层次关系，主要有下面三种抽取方法。

基于词法模式的原理，根据语句构成成分之间的语义关系，预测语句整体的意义。比如通过对文本句法的分析，对于一个以动词为核心的短句，可以抽取出实体之间的潜在关系。例如发现“中兴”与“深圳”的关系一般是“中兴，总部位于，深圳”、“中兴，总部设置于，深圳”、以及“中兴，将其总部建于，深圳”。使用这种方式不需要事先定义实体之间关系的种类，但这也导致了实体的关系没有归一化的问题。例如，上例中发现的“总部位于”、“总部设置于”以及“将其总部建于”三个实体关系在语义上应该属于同一种关系。

另外一种共现分析的方法，是一种定量与定性相结合的分析方法。其具体步骤为，先将待处理文本转化为数字形式表达的信息，然后使用不同的数学方法对文本进行定量计算和分析，最后结合定性分析的结果对文本中的分类关系进行综合分析。

最后，还有基于开放链接数据和基于在线百科的方法，该方法通过百科类等网站规则的知识分类体系，定义或者学习知识分类的规则和特征，从而对新的文本中隐藏的分类关系进行准确地抽取。

3.1.5 公理和规则学习

公理和规则学习指的是对包含了一定的实体和属性的通用句式或者模板规则进行学习的过程。常用的公理和规则学习方法是通过自举的思想。比如，通过“X 是 Y 的首都”这种句式可以抽取出(中国，首都，北京)、(英国，首都，伦敦)等三元组实体关系，然后根据这些三元组实例对“中国-北京”和“英国-伦敦”发现更多的匹配句式，像“X 的首都是 Y”、“Y 是 X 的政治中心”等；从而利用新学习到的模板可以抽取得到新的三元组实例，如此循环下去可以抽取更多新的实例和句式。这种自举的思想简易高效，但也存在引入噪声实例和模板，从而引起语义漂移的问题。对于这些问题，比较常见的解决方法是，同时给出扩充多个不同类别的实例，比如同时扩充动物、时间和人物，规定一个实例只允许属于其中一个类别；另外，也有研究学者提出通过使用负实例来规避语义漂移的问题。

3.2 实体学习

实体学习也可称为实体识别(Named entity recognition，NER)，指的是抽取文本数据中所涉及的对象信息。对于实体学习，一个关键的标准是能否准确把属于同一事物或概念的实体的不同表达方式进行归一化表示，以及区分同一表述方式在不同语境中指代的不同实体。其中，前者被称为实体对齐，后者可以通过实体填充来解决。寇月等人[29]给出了一种综合使用基于语义及统计分析的实体识别方法，通过文本粗略匹配过程、语义分析和分组统计三个过程，对于实体的识别结果不断求精，取得了较好的效果。在实体学习阶段，实体的对齐比较困难，同时实体对齐对于知识图谱的最终效果至关重要，通常需要借助概率语言模型在较大的语料库中进行学习达到。

3.2.1 实体对齐

实体对齐是知识图谱构建以及更新过程中的重要工作之一。通过实体对齐，同一个知识图谱内部的实体得到了精简，使得知识图谱的运转更加高效。同时，通过不同的知识图谱系统之间的实体对齐，可以实现知识图谱之间的链接与合并，从而实现构建一个更大规模，服务范围更广泛的知识图谱系统。

实体对齐实际上是涉及知识融合的一个过程，也就是对于物理世界中的同一个对象，要识别出它在不同语言，不同地域，不同数据源或者是同一个数据源下不同的表示形式，然后用一个全局唯一的编号来表征。实体对齐算法设计的主要思路是根据具体的知识图谱的特点和处理方法，利用不同的实体识别技术，具体有使用传统概率模型的方法、以及使用机器学习的方法，来完成实体对齐任务[30]。

3.2.2 实体填充

对于一个实体而言，如果仅有实体名称，实体的意义不大；为了使实体可以被人和机器所理解和区分，通常需要一定的方式来描述实体，主要包括实体的描述、图片、同义实体名和属性等。比如把“泰山”作为一个实体，当在搜索引擎上搜索“泰山”时，会出现泰山的简介、地理位置、海拔高度以及一个动画人物的图片、影片链接等信息。

4 知识图谱上的挖掘

在知识图谱建立完成之后，基于知识图谱的挖掘可以大大扩展知识图谱的知识覆盖率，基于知识图谱的挖掘主要包括知识推理和用户搜索意图的挖掘两个方面。

4.1　知识推理

知识推理可以分为对实体属性的推理和对实体关系的推理。对实体属性的推理主要包括对于会发生变化的实体属性值进行及时发现、推理、更新或者为实体创建新的属性; 对实体之间关系的推理则是对于实体之间潜在的关系进行推断和补充。

要完成知识推理则主要依赖于可扩展的规则引擎。对应知识推理的两个方面,推理规则包括针对实体属性的规则和针对实体关系的规则。其中,通过计算可获取针对属性的规则的属性值。例如,在知识图谱中,针对年龄属性,当前日期减去这个人的出生年月这一属性即可获得。实体间隐含关系的发现的另一种方法是通过链式的规则进行推断。例如,可以制定这样一条规则"父亲的父亲是爷爷"。然后,利用这条规则,当已知康熙对于雍正的关系是父亲,以及雍正对于乾隆的关系也是父亲时,就可以推理出康熙对于乾隆的关系是爷爷。

与知识推理的概念类似,也有学者将挖掘知识图谱隐含知识的过程称为知识图谱的补全。Liu Zhiyuan 等人[11]在知识表示模型 TransE 的基础上提出了 TransR 和 CTransR 表示模型,将实体空间和关系空间分离开,通过比较两个实体向量在一定的关系空间的投影距离来度量两个实体之间的某种关系,并进行了链路预测(Link prediction)、实体-关系对分类(Triple classification)、关系抽取(Relational fact extraction)等知识图谱补全的操作,取得了比原有的模型更准确的结果。传统的进行知识图谱补全的方法有利用因子分解的方式和基于机器学习的方法。Zhang Xiangling 等人[31]提出了一种通过计算实体相似度和公共语义模式的方法来评估一个实体-关系对的可能性的互补方法。该方法探查了实体的语义环境,在知识图谱补全方面取得了较大的提升。

知识推理是在知识图谱上进行数据挖掘,使知识图谱不断完善的重要手段,主要包括三个方面: 第一,线索挖掘; 第二,关系推理; 第三,关系预测[32]。

线索挖掘是指对于知识图谱中原来并没有关系的实体或概念,挖掘出它们之间的关系或关系模式,英文称为 Storytelling[33]。线索挖掘是对于在知识图谱构建过程中没有关联起来的实体进行相关性推理的过程,涉及到的处理方法主要有对于图的各种操作,比如查找子图、查找连通分支等。

Hossain 等人[34]提出了一种基于团(Clique)的关联方法。该方法通过构造一个两个实体之间的路径及路径上的相邻实体组成的团结构,为这两个实体添加了很多相邻的实体信息,从而为解释这两个实体之间的关联提供了更多的有用的线索。Fang 等人[35]则提出了一种通过对关系进行阐释的方法来实现关联知识图谱中的实体思想。该方法包含解释枚举和解释排序两个步骤。在解释枚举中,该方法通过定义一种称为覆盖路径模式集(Covering path pattern set)的结构来挑选出一些候选实体,这些候选实体将用于关联目标实体。在解释排序中,该方法对于上一步骤中选取的候选实体进行相关性的排序,排序标准包括分布和聚集性等多重度量指标。

随着知识图谱中实体规模的不断扩大,知识图谱中实体的关联,作为知识图谱补全的重要环节,将变得愈来愈重要。同时,由于对实体关联的高效性要求变得愈来愈高,以及知识图谱建设造成的不一致和噪声的干扰,实体关联的任务也会变得越来越复杂,需要研究出更加高效、更具抗噪声能力的实体关联线索挖掘方法。

关系推理是指根据知识图谱中已有的实体之间的关系推断出实体之间潜在的关系。例如,前文中提到的基于规则"父亲的父亲是爷爷"。然后根据已有的实体之间的关系,这里是康熙对于雍正的关系是父亲和雍正对于乾隆的关系是父亲,推断出康熙对于乾隆的关系是爷爷。

基于规则的方法,目前常用的方法是机器学习中的归纳逻辑编程技术,包括基于一阶 Horn 子句的方法或一阶归纳逻辑(FOIL)[36-38]。"永不停止的语言学习"(Never-ending language learning,NELL)项目中,Tom Mitchell 教授就是采用一阶 Horn 子句的方式来预测实体之间关联[39]。

德国马克斯·普朗克研究所的科研项目,YAGO[40]通过从维基百科网站和 WordNet 等数据集合中挖掘实体,截至 2010 年,已拥有千万级别的实体个数和上亿条实体的关系。YAGO2 在 YAGO 的基础上进行了进一步的扩展,为实体和事实构建了时空的属性。通过为实体和事实构建时间戳,它可以方便地计算出每一个实体或事实的存在的起止时间。同时,YAGO2 还通过

GeoNames数据源对实体和事实添加了空间上的属性。最后,YAGO2通过构建SPOTL模型,来表示时间和地点信息,进而完成知识图谱中实体和事实的时空信息进行快速地查询,同时也为基于已有的时空信息对实体之间或事实之间的隐含信息推理提供了方便有效的工具。

除了基于规则的方法之外,还有基于概率图的方法,包括随机游走[41]、Markov逻辑网络[42,43]等。例如,Schoenmackers等人[44]提出的Sherlock-Holmes方法利用Markov逻辑网络推理隐含关系以及关系的置信度,但仍然需要手动来事先整理一定的推理规则。

关系预测是指伴随时间的发展,从性质和数量两个角度对实体之间的关系作出推断。关系预测在依托社交网络构建的实体关系图谱中已经广泛使用,用以对用户之间未来是否会发生连接进行预测或者进行朋友推荐等。Jia等人[45]在微博平台上对这种链路预测(Link prediction)进行了深入地探讨和研究。然而,在知识图谱方面,还鲜有学者对关系预测进行过讨论。但是,社交网络中的连接预测理论可以作为一个借鉴,为知识图谱中搜索实体的相关实体推荐提供指导。

4.2 用户搜索意图的挖掘

用户的提问是用户接受知识图谱系统服务的接口。由于用户文化背景和表述习惯的不同,为了使用户能够更加快速、准确地查询到自己需要的知识,需要根据用户的提问,在这个接口层做好用户搜索意图的挖掘工作,也就是将用户的提问准确地匹配到知识图谱中相关的实体和概念上,更进一步地,还可以向用户推荐与搜索实体相关联的其他实体信息。

传统的挖掘用户搜索意图的方法大多致力于对用户的问题提取更丰富的特征。

Li[46]等提出了一种基于查询词与搜索结果之间的点击二分图以及部分已标注查询词的半监督学习方法。首先,根据部分已标注的查询词和查询词在点击图的邻接关系推断未标注的查询词,接着,利用这些标注的查询词自动地训练分类器,然后点击图的学习和分类器的训练协同工作,实现推断查询词对应的搜索意图的目的。Guo[47]等人提出了一种使用意图感知的思想,对查询关键字建模,实现查询目的的准确理解。在对用户的搜索内容进行扩充时,通常可以基于整网的搜索热词排行进行推荐。

同时,有不少的研究者从搜索系统的查询日志入手分析用户的查询意图。Chilton等[48]通过分析大量的用户搜索日志及点击记录,探索对于不同的查询结果用户点击行为的变化,提出了根据用户的点击行为来评价查询结果的价值以及应该在查询结果中关联和推荐的信息的方法,并进而判断用户之后的关键词输入行为所对应的准确搜索意图。He等[49]基于持续的部分可观测马尔可夫模型(Partially observable Markov model with duration,POMD)分析用户搜索日志中用户的阅读、跳过等行为,以及用户搜索的空间与时间信息,提出了两阶段训练算法和相应的贪婪段解码算法,给出了一个通过挖掘搜索日志感知用户的搜索行为的可行方法。另外,有学者通过将用户的查询关键字建立起语义模型来分析用户的查询目的。Wen[50]等人基于知识库系统Probase分析用户问题的语义信息,从知识图谱中自动检索出主题级别的知识标示。对于用户的搜索行为的学习和预测,需要在与用户的长期交互中对算法模型进行不断地调优,同时可以区分注册用户和未识别用户,强化对已注册用户的搜索行为的特征学习;基于已有的算法模型,对未识别的用户的搜索行为进行预测。

5 知识图谱的更新和维护

5.1 数据模式层的更新

数据模式层的更新是指概念层的更新,包括概念的层次关系,同义关系和概念的属性定义等。这些更新主要来自两个方面的原因,一是由于开放网络大数据导致各种数据源中结构化和半结构化的知识更新,知识图谱也需要做出相应的更新;二是在构建的时候预先确定的自动学习算法可能由于后期数据量和数据模式的快速变化而导致其性能下降,为了保持或增进知识图谱的性能需要对知识图谱数据模式进行更新。

对于概念的描述、图片及同义关系的变化,由于它们的变更所影响的仅为当前概念本身,因此,通常不需要进行额外处理,通过前面知识图谱构建方法进行更改。涉及概念的前后语境的变化会更新整个分类层次结构,因此需要谨慎处理;如果并未造成冲突,则不需要特殊处理,如果造成了闭环式的冲突,此时系统检测到以后,由知识图谱构建人员进行处理。对于概念的属性更新,如果是

新的属性添加，依据之前所述的方法进行处理即可；如果是对现有属性的更新，包括属性类型、值类型、值域的更新，系统检测到后由人工进行处理；如果是对属性的删除，若属于当前概念的所有实体中该属性均已被移除，则可以把概念的属性直接移除，否则仍然需要人工确认。

5.2　数据层的更新

信息具有时效性，知识图谱中的知识也不例外。如何对知识图谱中的知识进行及时刷新是一个棘手的问题。知识图谱的更新表现在两个方面，即新的实体或者实体间关系的加入引起的更新需求，以及需要对知识图谱中已包含的知识修正引起的更新需求。目前，对开放知识库的更新的研究还比较少，通常涉及对保存实体相关信息的数据库的更新。但是，知识图谱如何自发地、迅速地捕捉到知识更新的需求，也是衡量一个知识图谱系统的功能完善性的重要标准。目前从数据更新方式来分类，主要有两类：一类是通过人工手动更新，一类是利用知识图谱中保留的时间戳或者地理位置的信息而实现的自动更新。

基于人工的手动更新具有较高的准确率，但对于一个稍具规模的知识图谱，依靠一个专门的团队负责更新都是非常吃力的。通常可行的方法是充分利用现如今开放的网络和广大网民群体的力量，采用“众包”的机制。利用群体的智慧对数据进行处理，对于特定领域的数据，可能群体中的人员比知识图谱构造者更加了解该领域的知识体系结构，他们会对知识图谱的修正带来很多帮助。例如 Wikipedia[51]、知乎等在线问答系统，就充分利用了群体中的每一个人的积极性，完成一个包罗万象的知识体系的构建。

自动更新需要在知识图谱构建时制定更多的学习规则和实体属性，通常可以及时地捕捉到知识更新的需求。但是可以依赖自动更新的知识依然数量有限。NELL[52]系统从互联网上挖掘出 44 万个实体，准确率约为 86%，系统中存在的偏差通过自我纠正系统完成。YAGO2[53]是 YAGO 的扩展，不同点在于对于每一个实体，它都为实体添加了时间戳。例如通过顾客购买电脑的时间点和到售后维修的时间计算电脑使用时间，进而判断是否超出保修期。而对于各种事件则记录时间发生或结束的时间点和时间持续范围，比如记录北京奥运会开幕式和闭幕式时间点。利用这些时间点和查询语句对系统内各个实体和事件进行更新和维护。

在实际的知识图谱系统中，使用哪一种更新知识的方式，取决于系统的定位以及知识组织框架等因素。但就目前看，混合使用两种知识更新方式通常可以提高知识更新的准确率和效率。

6　结束语

知识图谱的理论方法可以使人们更加便捷、准确地获取到自己所需要的信息，具有重大的价值和研究意义。在未来信息爆炸的世界中，知识图谱作为人们访问知识信息的接口，将扮演越来越重要的角色。显然，现有的知识图谱系统还远不能满足人们的应用需求，构建一个健壮的、完善的知识图谱系统仍然面临诸多的挑战。

在构建和维护知识图谱的整个流程里，实体对齐问题仍然是一个难点。有效的实体对齐方法，不仅对于构建自身准确、高效的知识图谱系统具有重大作用，而且，对于知识图谱或知识库的跨系统融合，甚至跨语言融合都具有巨大的推动作用。当今，已有不少的公司或科研机构分别建立了自己的知识图谱系统。这些知识图谱系统各有特色，各有长处。通过有效的跨系统融合或跨语言融合，可以为用户提供更加完备的知识图谱系统，提升用户的搜索体验。这不仅需要在实体的对齐方面达到一致的标准方法，还需要在系统与系统之间制定统一的通信协议。最终，各知识图谱系统的互联互通将形成一个覆盖更广的知识范围的，更强大的知识图谱系统，提供给用户一个无所不知的搜索体验。

对于初步构建完成的知识图谱，进行实体关系及属性的深入挖掘是另一个面临的挑战。虽然在这一领域已经有了不少的研究成果，但对于比较复杂的实体关系（比如一对多，多对多等）的推理仍然还没有很好地实现。近年来，知识表示学习成为新的研究热点，众多的知识表示模型被提出并应用，在实体、关系的表示以及关系的预测方面取得了不错的效果，有望在处理实体之间复杂的关系推理方面取得进一步的突破。

在知识图谱的更新方面，需要解决对于新出现的知识的快速学习问题。互联网上的知识大多是以半结构的网页文本的形式展现，对于新产生的知识尤其如此。然而，互联网网页种类繁多，形式不一，很难通过定义一种或几种知识抽取规则，来对

新的知识进行学习。所以,利用互联网用户群体智慧的众包机制提供一种对知识图谱中的知识进行更新的有效的方法。通过众多的用户对知识图谱提供信息的纠错和编辑,可以使知识图谱不断更新。虽然,这种方式可以借鉴维基百科的成功经验,但是这种方式的滞后性往往比较严重。为了时刻为用户提供最新的,准确的知识,需要在秒级内实现知识的定位、统计和推理,这将是知识图谱系统走向应用所面临的巨大挑战。

另外,还有学者对于构建知识图谱过程中涉及到的个人隐私安全问题作了研究[54]。基于知识图谱构建的搜索引擎系统大多是面向所有互联网用户使用的,在对知识图谱中的人物实体进行构建的过程中,应该考虑到隐私保护的问题。

知识图谱的理论方法为知识信息提供了一种新的获取、存储、组织、管理、更新和展示的手段。在未来信息大爆炸的时代,以知识图谱理论为指导的知识信息管理方法是必然的发展趋势,它提供了一个更友好、更便捷的知识获取方式。本文介绍了诸多的学者、科研机构或公司已经对知识图谱的构建做出的不同尝试以及取得的成果,但整体而言,关于知识图谱的研究工作还处于探索阶段,仍然有大量的难题需要攻克。关于知识图谱,国外比较前沿的研究还包括DBpedia,基于维基百科的KG,google scholar等,而中文的知识图谱研究还相对滞后,希望本文对于知识图谱的介绍能为中文知识图谱的研究发展提供帮助。

参考文献:

[1] Djds B P. Networks of scientific papers [J]. Science, 2010, 149(3683): 510-515.

[2] 袁国铭,李洪奇,樊波. 关于知识工程的发展综述[J]. 计算技术与自动化,2011,34(1):138-143.
Yuan Guoming, Li Hongqi, Fan Bo. Survey on development of knowledge engineering system [J]. Computing Technology and Automation, 2011, 34(1): 138-143.

[3] 陈和. 机构知识库发展趋势探析[J]. 图书情报工作,2012,21:62-66.
Chen He. Development trends of the institutional repository [J]. Library and Information Service, 2012, 21: 62-66.

[4] 张晓林. 机构知识库的发展趋势与挑战[J]. 现代图书情报技术,2014,30(2):1-7.
Zhang Xiaolin. Trends and challenges for institutional repositories [J]. New Technology of Library and Information Service, 2014, 30(2): 1-7.

[5] 曹倩,赵一鸣. 知识图谱的技术实现流程及相关应用[J]. 情报理论与实践,2015,38(12):13-18.
Cao qian, Zhao Yiming. Technology implementation process and the related application of knowledge graph [J]. Information Studies: Theory & Application, 2015, 38(12): 13-18.

[6] 胡芳槐. 基于多种数据源的中文知识图谱构建方法研究[D]. 上海:华东理工大学计算机学院,2014.

[7] Wu W, Li H, Wang H, et al. Probase: a probabilistic taxonomy for text understanding [C]//Proc of the 2012 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2012: 481-492.

[8] 祝忠明,马建霞,卢利农,等. 机构知识库开源软件DSpace的扩展开发与应用[J]. 现代图书情报技术,2009(7-8):11-17.
Zhu Zhongming, Ma Jianxia, Lu Linong, et al. Expansion development and application of DSpace: the institutional knowledge base open source software [J]. New Technology of Library and Information Service, 2009(7-8): 11-17.

[9] Garfield E. Citation indexes for science: a new dimension in documentation through association of ideas [J]. International Journal of Epidemiology, 2006, 122(5): 1123-1127.

[10] 秦长江,侯汉清. 知识图谱——信息管理与知识管理的新领域[J]. 大学图书馆学报,2009(1):30-37.
Qin Changjiang, Hou Hanqing. Knowledge Graph—the new field of information and knowledge management [J]. Journal of Academic Libraries, 2009(1): 30-37.

[11] 杨思洛,韩瑞珍. 知识图谱研究现状及趋势的可视化分析[J]. 情报资料工作,2012,33(4):22-28.
Yang Siluo, Han Ruizhen. A visual analysis of the status quo and trend of knowledge mapping research [J]. Information and Documentation Services, 2012, 33(4): 22-28.

[12] 鄢珞青. 知识库的知识表达方式探讨[J]. 情报杂志,2003(4):63-64.
Yan Luoqing. Methods of knowledge expression in knowledge base [J]. Journal of Information, 2003(4): 63-64.

[13] 王知津,王璇,马婧. 论知识组织的十大原则[J]. 国家图书馆学刊,2012,21(4):3-11.
Wang Zhijin, Wang Xuan, Ma Jing. The ten principles

of know ledge organization [J]. Journal of The National Library of China, 2012, 21(4): 3-11.

[14] 王军,张丽.网络知识组织系统的研究现状和发展趋势[J].中国图书馆学报,2008,34(1):65-69.
Wang Jun, Zhang Li. Research status and development trend of network knowledge organization system [J]. Journal of Library Science in China, 2008, 34(1): 65-69.

[15] 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.
Liu Zhiyuan, Sun Maoshong, Lin Yankai, et al. Knowledge representation learning: a Review [J]. Journal of Computer Research and Development, 2016, 53(2): 247-261.

[16] Hodge G. Next generation knowledge organization systems: Integration challenges and strategies [C]//ACM/IEEE-CS Joint Conference on Digital Libraries. New York: ACM, 2005.

[17] 袁旭萍.基于深度学习的商业领域知识图谱构建[D].上海:华东师范大学商学院,2015.

[18] Wikipedia. Never-Ending Language Learning [EB/OL]. http://en.wikipedia.org/wiki/Never-Ending_Language_Learning, 2015.

[19] 王昊奋.知识图谱技术原理介绍[EB/OL].http://wenku. baidu. coni/view/b3858227c5d a50e2534d7fd8.html, 2015.

[20] 项灵辉.基于图数据库的海量 RDF 数据分布式存储[D].武汉:武汉科技大学计算机学院,2013.

[21] 曾锦麒.语义 WEB 的知识表示语言及其应用研究[D].长沙:中南大学管理学院,2004.

[22] Li Lei, Li Tao. An empirical study of ontology-based multi-document summarization in disaster management [J]. IEEE Transactions SMC: Systems, 2014, 44(2): 162-171.

[23] 李涛.数据挖掘的应用与实践[M].厦门:厦门大学出版社,2013.

[24] Jiang Yexi, Chang-Shing Perng, Anca Sailer, et al. CSM: a cloud service marketplace for complex service acquisition [J]. ACM Transactions on Intelligent Systems and Technology, 2016, 8(1): 1-25.

[25] 梁铭.基于英汉平行语料库术语词典的自动抽取[J].电脑知识与技术,2009,5(19):5081-5083.
Liang Ming. English-Chinese parallel corpora based on the automatic extraction of terms dictionary [J]. Computer Knowledge and Technology, 2009, 5(19): 5081-5083.

[26] 曹浩.基于机器学习的双语词汇抽取问题研究[D].天津:南开大学研究生学院,2011.

[27] 孙霞,董乐红.基于监督学习的同义关系自动抽取方法[J].西北大学学报,2008,38(1):35-39.
Sun xia, Dong Yuehong. Automatic extraction of synonymy relation using supervised learning [J]. Journal of Northwest University, 2008, 38(1): 35-39.

[28] 关键.面向中文文本本体学习概念抽取的研究[D].吉林:吉林大学,2010.

[29] 寇月,申德荣,李冬,等.一种基于语义及统计分析的 Deep Web 实体识别机制[J].软件学报,2008,19(2):194-208.
Kou Yue, Shen Derong, Li Dong, et al. A deep web entity identification mechanism based on semantics and statistical analysis [J]. Journal of Software, 2008, 19(2): 194-208.

[30] 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.
Zhuang Yan, Li Guoliang, Feng JianHua. A survey on entity a lignment of knowledge base [J]. Journal of Computer Research and Development. 2016, 53(1): 165-192.

[31] Zhang Xiangling, Du Cuilan, Li Peishan, et al. Knowledge graph completion via local semantic contexts [J]. Database Systems for Advanced Applications, 2011, 9642: 432-446.

[32] 王元卓,贾岩涛,刘大伟,等.基于开放网络知识的信息检索与数据挖掘[J].计算机研究与发展,2015,52(2):456-474.
Wang Yuanzhuo, Jia Yantao, Liu Dawei, et al. Open web knowledge aided information search and data mining [J]. Journal of Computer Research and Development, 2015, 52(2): 456-474.

[33] Kumar D, Ramakrishnan N, Helm R, et al. Algorithms for story telling [J]. IEEE Trans on Knowledge and Data Engineering, 2008, 20(6): 736-751.

[34] Hossain M, Butler P, Boedihardjo A, et al. Story telling in entity networks to support intelligence analysts [C]//Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2012: 1375-1383.

[35] Fang L, Sarma A D, Yu C, et al. REX: Explaining relationships between entity pairs [J]. VLDB Endowment, 2011, 5(3): 241-252.

[36] Quinlan J R, Cameron-Jones R M. FOIL: A midterm report [C]//Proc of the 5th European Conf on Machine Learning. Berlin: Springer, 1993: 3-20.

[37] Mitchell T M, Betteridge J, Carlson A, et al. Populating these mantic Web by macro-reading internet text [C]//Proc of the 8th Int Semantic Web Conf. Berlin:

Springer,2009: 998-1002.

[38] Cohen W W, Page D. Polynomial learnability and inductive logic programming: Methods and results [J]. New Generation Computing,1995,13(34): 369-409.

[39] Mitchell T M,Betteridge J,Carlson A,et al. Populating these mantic Web by macro-reading internet text [C] //Proc of the 8th Int Semantic Web Conf. Berlin: Springer,2009: 998-1002.

[40] Suchanek F,Kasneci G,Weikum G. YAGO—A core of semantic knowledge [C] //Proc of the 16th Int Conf on World Wide Web. New York: ACM,2007: 697-706.

[41] Lao N,Mitchell T M,Cohen W W. Random walk inference and learning in a large scale knowledge base [C] //Proc of the Conf on Empirical Methods in Natural Language Processing,EMNLP'11. Stroudsburg, PA: Association for Computational Linguistics,2011: 529-539.

[42] Schoenmackers S, Etzioni O, Weld D, et al. Learning first-order Horn clauses from Web text [C] //Proc of the 2010 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics,2010: 1088-1098.

[43] Schoenmackers S, Etzioni O, Weld D. Scaling textual inference to the Web [C] //Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2008: 79-88.

[44] Schoenmackers S, Etzioni O, Weld D, et al. Learning first-order horn clauses from Web text [C] //Proc of the 2010 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics,2010: 1088-1098.

[45] Jia Yantao, Wang Yuanzhuo, Li Jingyuan, et al. Structural-interaction link prediction in microblogs [C] // Proc of the 22nd Int Conf on World Wide Web Companion. New York: ACM,2013: 193-194.

[46] Li X, Wang Y Y, Shen D, et al. Learning with click graph for query intent classification [J]. ACM Transactions on Information Systems,2010,28(3): 1-20.

[47] Guo Jiafeng. Intent-aware query similarity [C] //Proc of the17th ACM Conf on Information and Knowledge Management (CIKM'11). New York: ACM, 2011: 259-268.

[48] Chilton L B, Teevan J. Addressing people's information needs directly in a web search result page [C] //Proceedings of the 20th International Conference on World Wide Web. New York: ACM,2011: 27-36.

[49] He Y, Wang K. Inferring search behaviors using partially observable Markov model with duration [C] // Proceedings of the Fourth ACM International Conference on Web Search and Data Mining-WSDM'11. New York: ACM,2011: 415-424.

[50] Wen Hua, Song Yangqiu, Wang Haixun, et al. Identifying users' topical tasks in Web search [C] // Proc of the 4th ACM Int Conf on Web Search and Data Mining, WSDM'13. NewYork: ACM,2013: 93-102.

[51] Tran T, Cao T H. Automatic detection of outdated information in Wikipedia infoboxes [J]. Research in Computing Science,2013,70: 183-194.

[52] Jia Y, Wang Y, Cheng X, et al. OpenKN: An open knowledge computational engine for network big data [C] //Advances in Social Networks Analysis and Mining, 2014 IEEE/ACM International Conference on. Washington DC: IEEE,2014: 657-664.

[53] Hoffart J, Suchanek F M, Berberich K, et al. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia [J]. Artificial Intelligence, 2013, 194: 28-61.

[54] 方滨兴,贾焰,李爱平,等. 网络空间大搜索研究范畴与发展趋势 [J]. 通信学报,2015,36(12): 1-8.
Fang Binxing, Jia Yan, Li Aiping, et al. Research progress and trend of cyberspace big search [J]. Journal on Communications,2015,36(12): 1-8.