计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20148228
53(3):582-600,2016

知识图谱构建技术综述

刘 峤 李 杨 段 宏 刘 瑶 秦志光
(电子科技大学信息与软件工程学院 成都 610054)
(qliu@uestc.edu.cn)

Knowledge Graph Construction Techniques

Liu Qiao, Li Yang, Duan Hong, Liu Yao, and Qin Zhiguang
(*School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu* 610054)

Abstract Google's knowledge graph technology has drawn a lot of research attentions in recent years. However, due to the limited public disclosure of technical details, people find it difficult to understand the connotation and value of this technology. In this paper, we introduce the key techniques involved in the construction of knowledge graph in a bottom-up way, starting from a clearly defined concept and a technical architecture of the knowledge graph. Firstly, we describe in detail the definition and connotation of the knowledge graph, and then we propose the technical framework for knowledge graph construction, in which the construction process is divided into three levels according to the abstract level of the input knowledge materials, including the information extraction layer, the knowledge integration layer, and the knowledge processing layer, respectively. Secondly, the research status of the key technologies for each level are surveyed comprehensively and also investigated critically for the purposes of gradually revealing the mysteries of the knowledge graph technology, the state-of-the-art progress, and its relationship with related disciplines. Finally, five major research challenges in this area are summarized, and the corresponding key research issues are highlighted.

Key words knowledge graph; semantic Web; information retrieval; semantic search engine; natural language processing

摘 要 谷歌知识图谱技术近年来引起了广泛关注,由于公开披露的技术资料较少,使人一时难以看清该技术的内涵和价值.从知识图谱的定义和技术架构出发,对构建知识图谱涉及的关键技术进行了自底向上的全面解析.1)对知识图谱的定义和内涵进行了说明,并给出了构建知识图谱的技术框架,按照输入的知识素材的抽象程度将其划分为3个层次:信息抽取层、知识融合层和知识加工层;2)分别对每个层次涉及的关键技术的研究现状进行分类说明,逐步揭示知识图谱技术的奥秘,及其与相关学科领域的关系;3)对知识图谱构建技术当前面临的重大挑战和关键问题进行了总结.

关键词 知识图谱;语义网;信息检索;语义搜索引擎;自然语言处理

中图法分类号 TP18

收稿日期:2014-11-06;**修回日期**:2015-04-08

基金项目:国家"八六三"高技术研究发展计划基金项目(2011AA010706);国家自然科学基金项目(61133016,61272527);教育部-中国移动科研基金项目(MCM20121041)

This work was supported by the National High Technology Research and Development Program of China (863 Program) (2011AA010706), the National Natural Science Foundation of China (61133016, 61272527), and Ministry of Education-ChinaMoblie Communications Corporation Research Funds (MCM20121041).

信息技术的发展不断推动着互联网技术的变革,Web 技术作为互联网时代的标志性技术,正处于这场技术变革的核心. 从网页的链接(Web 1.0)到数据的链接(linked data),Web 技术正在逐步朝向 Web 之父 Berners-Lee[1] 设想中的语义网络(semantic Web)演变.

根据 W3C 的解释,语义网络是一张数据构成的网络(Web of data),语义网络技术向用户提供的是一个查询环境,其核心要义是以图形的方式向用户返回经过加工和推理的知识①. 而知识图谱(knowledge graph)技术则是实现智能化语义检索的基础和桥梁. 传统搜索引擎技术能够根据用户查询快速排序网页,提高信息检索的效率. 然而,这种网页检索效率并不意味着用户能够快速准确地获取信息和知识,对于搜索引擎反馈的大量结果,还需要进行人工排查和筛选. 随着互联网信息总量的爆炸性增长,这种信息检索方式已经很难满足人们全面掌控信息资源的需求,知识图谱技术的出现为解决信息检索问题提供了新的思路.

知识图谱的概念是由谷歌公司提出的. 2012 年 5 月 17 日,谷歌发布知识图谱项目,并宣布以此为基础构建下一代智能化搜索引擎. 该项目始于 2010 年谷歌收购 Metaweb 公司,并籍此获得了该公司的语义搜索核心技术,其中的关键技术包括从互联网的网页中抽取出实体及其属性信息,以及实体间的关系. 这些技术特别适用于解决与实体相关的智能问答问题,由此创造出一种全新的信息检索模式.

虽然知识图谱的概念较新,但它并非是一个全新的研究领域. 早在 2006 年,Berners-Lee 就提出了数据链接(linked data)的思想,呼吁推广和完善相关的技术标准如 URI(uniform resource identifier),RDF(resource description framework),OWL(Web ontology language),为迎接语义网络时代的到来做好准备②. 随后掀起了一场语义网络研究热潮,知识图谱技术正是建立在相关的研究成果之上的,是对现有语义网络技术的一次扬弃和升华.

我国对于中文知识图谱的研究已经起步,并取得了许多有价值的研究成果. 早期的中文知识库主要采用人工编辑的方式进行构建,例如中国科学院计算机语言信息中心董振东领导的知网(HowNet)项目,其知识库特点是规模相对较小、知识质量高、但领域限定性较强. 由于中文知识图谱的构建对中文信息处理和检索具有重要的研究和应用价值,近年来吸引了大量的研究. 例如在业界,出现了百度知心、搜狗知立方等商业应用. 在学术界,清华大学建成了第 1 个大规模中英文跨语言知识图谱 XLore③、中国科学院计算技术研究所基于开放知识网络(OpenKN)建立了"人立方、事立方、知立方"原型系统、中国科学院数学与系统科学研究院陆汝钤院士提出知件(Knowware)的概念、上海交通大学构建并发布了中文知识图谱研究平台 zhishi.me④、复旦大学 GDM 实验室⑤推出的中文知识图谱项目等[2],这些项目的特点是知识库规模较大,涵盖的知识领域较广泛,并且能为用户提供一定的智能搜索及问答服务.

随着近年来谷歌知识图谱相关产品的不断上线,这一技术也引起了业界和学术界的广泛关注. 它究竟是概念的炒作还是如谷歌所宣称的那样是下一代搜索引擎的基石,代表着互联网技术发展的未来方向? 为了回答这一问题,首先需要对知识图谱技术有完整深刻的理解. 本文的目的就是从知识图谱的构建角度出发,深度剖析知识图谱概念的内涵和发展历程,帮助感兴趣的读者全面了解和认识该技术,从而客观地做出判断.

1 知识图谱的定义与架构

维基百科对知识图谱给出的词条解释仍然沿用了谷歌的定义,即:知识图谱是谷歌用于增强其搜索引擎功能的辅助知识库. 然而从业界的发展动态来看,这个定义显得过于简单. 微软在 2013 年 7 月发布了自己的 Satori 知识库之后,必应(Bing)搜索引擎产品的高级主管 Weitz 公开表示,发布 Satori 只是表明微软已有类似的技术,然而目前这一技术本身还存在许多问题,微软希望取得领导地位,而不是追随谷歌⑥. 这一表态,折射出该领域背后的技术竞争

① http://www.w3.org/standards/semanticweb/data

② http://www.w3.org/DesignIssues/LinkedData.html

③ http://xlore.org/index.action

④ http://zhishi.apexlab.org

⑤ http://gdm.fudan.edu.cn

⑥ http://en.wikipedia.org/wiki/Knowledge_Graph

十分激烈,从当前披露出来的商业产品,也能看出业界对此的普遍重视.表1给出了当前主流的知识库产品和相关应用,其中,包含实体数最多的是WolframAlpha知识库,实体总数已超过10万亿条.谷歌的知识图谱拥有5亿个实体和350亿条实体间的关系,而且规模在不断地增加.微软的Probase包含的概念总量达到千万级,是当前包含概念数量最多的知识库.Apple Siri,Google Now等当前流行的智能助理应用正是分别建立在WolframAlpha知识库和谷歌的知识图谱基础之上.值得注意的是:国内也涌现出一些知识图谱产品和应用,如搜狗的知立方,侧重于图的逻辑推理计算,能够利用基于语义网三元组推理补充实体数据,对用户查询进行语义理解以及句法分析等[3].

Table 1 Knowledge Graph and Similar Products

表1 知识图谱及相关类似产品

Knowledge Base	Products	Data Source
Knowledge Vault	Google Seach Engine Google Now	Wikipedia, Freebase, Web Open Data
Wolfram Alpha	Apple Siri	Mathematica
Satori/Probase	Bing Seach Engine Microsoft Cortana	Wikipedia, Web Open Data
Watson KB	IBM Watson System	Web Dictionaries The World Book Encyclopedia
DBpedia KB	DBpedia	Wikipedia
YAGO KB	YAGO	Wikipedia
NELL KB	NELL	Web Open Data
Facebook KB	Shopycat	Social Network Data
Zhilifang KB	Sougou Seach Engine	Web Open Data
Zhixin KB	Baidu Zhixin Platform	User Generated Content
Cross-Lingual KB	XLORE	Chinese/English Encyclopedia, Wikipedia
Zhishi. me KB	Zhishi. me	Chinese Encyclopedia

从表1可以看出,除传统搜索服务提供商之外,包括Facebook,Apple,IBM等互联网领军企业也加入了竞争.由于相关技术和标准尚未成熟,其应用也处于探索阶段,因此知识图谱的概念目前仍处在发展变化的过程中,通过对现有的研究成果进行比较和提炼,本文提出知识图谱的定义.

1.1 知识图谱的定义

定义1. 知识图谱.是结构化的语义知识库,用于以符号形式描述物理世界中的概念及其相互关系.其基本组成单位是"实体-关系-实体"三元组,以及实体及其相关属性-值对,实体间通过关系相互联结,构成网状的知识结构.

通过知识图谱,可以实现Web从网页链接向概念链接转变,支持用户按主题而不是字符串检索,从而真正实现语义检索.基于知识图谱的搜索引擎,能够以图形方式向用户反馈结构化的知识,用户不必浏览大量网页,就可以准确定位和深度获取知识.

定义1包含3层含义:

1) 知识图谱本身是一个具有属性的实体通过关系链接而成的网状知识库.从图的角度来看,知识图谱在本质上是一种概念网络,其中的节点表示物理世界的实体(或概念),而实体间的各种语义关系则构成网络中的边.由此,知识图谱是对物理世界的一种符号表达.

2) 知识图谱的研究价值在于,它是构建在当前Web基础之上的一层覆盖网络(overlay network),借助知识图谱,能够在Web网页之上建立概念间的链接关系,从而以最小的代价将互联网中积累的信息组织起来,成为可以被利用的知识.

3) 知识图谱的应用价值在于,它能够改变现有的信息检索方式,一方面通过推理实现概念检索(相对于现有的字符串模糊匹配方式而言);另一方面以图形化方式向用户展示经过分类整理的结构化知识,从而使人们从人工过滤网页寻找答案的模式中解脱出来.

1.2 知识图谱的架构

知识图谱的架构,包括知识图谱自身的逻辑结构以及构建知识图谱所采用的技术(体系)架构,后者是本文讨论的重点.

首先介绍知识图谱的逻辑结构,从逻辑上将知识图谱划分为2个层次:数据层和模式层.在知识图谱的数据层,知识以事实(fact)为单位存储在图数据库.例如谷歌的Graphd和微软的Trinity都是典型的图数据库.如果以"实体-关系-实体"或者"实体-属性-性值"三元组作为事实的基本表达方式,则存储在图数据库中的所有数据将构成庞大的实体关系网络,形成知识的"图谱".

模式层在数据层之上,是知识图谱的核心.在模式层存储的是经过提炼的知识,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之间的联系.本体库在知识图谱中的地位相当于知识库的模具,拥有本体库的知识库冗余知识较少.

接下来从知识图谱构建的角度,介绍知识图谱的一般技术架构.图 1 给出了知识图谱技术的整体架构,其中虚线框内的部分为知识图谱的构建过程,同时也是知识图谱更新的过程.如图 1 所示,知识图谱的构建过程是从原始数据出发,采用一系列自动或半自动的技术手段,从原始数据中提取出知识要素(即事实),并将其存入知识库的数据层和模式层的过程.这是一个迭代更新的过程,根据知识获取的逻辑,每一轮迭代包含 3 个阶段:信息抽取、知识融合以及知识加工.

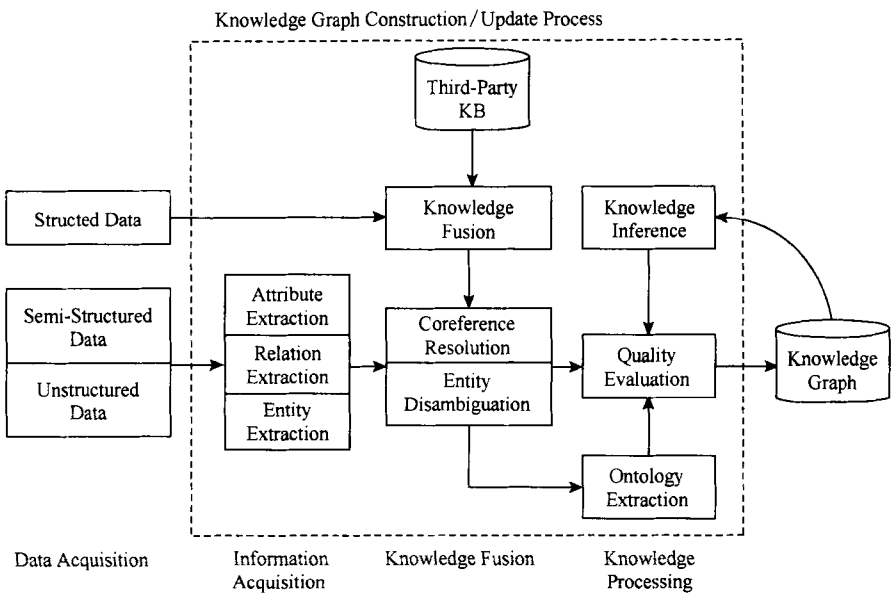

Fig. 1　Technical architecture of knowledge graph.

图 1　知识图谱的技术架构

知识图谱有自顶向下和自底向上 2 种构建方式.所谓自顶向下构建是指借助百科类网站等结构化数据源,从高质量数据中提取本体和模式信息,加入到知识库中;所谓自底向上构建,则是借助一定的技术手段,从公开采集的数据中提取出资源模式,选择其中置信度较高的新模式,经人工审核之后,加入到知识库中.

在知识图谱技术发展初期,多数参与企业和科研机构都是采用自顶向下的方式构建基础知识库,例如,Freebase 项目就是采用维基百科作为主要数据来源.随着自动知识抽取与加工技术的不断成熟,目前的知识图谱大多采用自底向上的方式构建,其中最具影响力的例子包括谷歌的 Knowledge Vault 和微软的 Satori 知识库,都是以公开采集的海量网页数据为数据源,通过自动抽取资源的方式来构建、丰富和完善现有的知识库.

因此,本文主要介绍自底向上的知识图谱构建技术,按照知识获取的过程分为 3 个层次:信息抽取、知识融合以及知识加工.

2　知识图谱的构建技术

如 1.2 节所述,采用自底向上的方式构建知识图谱的过程是一个迭代更新的过程,每一轮更新包括3 个步骤:1)信息抽取,即从各种类型的数据源中提取出实体(概念)、属性以及实体间的相互关系,在此基础上形成本体化的知识表达;2)知识融合,在获得新知识之后,需要对其进行整合,以消除矛盾和歧义,比如某些实体可能有多种表达,某个特定称谓也许对应于多个不同的实体等;3)知识加工,对于经过融合的新知识,需要经过质量评估之后(部分需要人工参与甄别),才能将合格的部分加入到知识库中,以确保知识库的质量.新增数据之后,可以进行知识推理、拓展现有知识、得到新知识.

2.1　信息抽取

信息抽取(information extraction)是知识图谱构建的第 1 步,其中的关键问题是如何从异构数据源中自动抽取信息得到候选知识单元.信息抽取是

一种自动化地从半结构化和无结构数据中抽取实体、关系以及实体属性等结构化信息的技术[4]. 涉及的关键技术包括:实体抽取、关系抽取和属性抽取.

2.1.1 实体抽取

实体抽取,也称为命名实体识别(named entity recognition, NER),是指从文本数据集中自动识别出命名实体. 实体抽取的质量(准确率和召回率)对后续的知识获取效率和质量影响极大,因此是信息抽取中最为基础和关键的部分.

早期对实体抽取方法的研究主要面向单一领域(如特定行业或特定业务),关注如何识别出文本中的人名、地名等专有名词和有意义的时间等实体信息[5]. 1991 年,Rau[6]采用启发式算法与人工编写规则相结合的方法,首次实现了从文本中自动抽取公司名称的实体抽取原型系统. 然而,基于规则的方法具有明显的局限性,不仅需要耗费大量人力,而且可扩展性较差,难以适应数据的变化. 随后,人们开始尝试采用统计机器学习方法辅助解决命名实体抽取问题,例如,Liu 等人[7]利用 K-最近邻(K-Nearest Neighbors)算法和条件随机场模型,实现了对 Twitter 文本数据中包含实体的识别. 然而迄今为止,单纯基于有监督学习的实体抽取方法,在准确率和召回率上的表现都不够理想,且算法的性能依赖于训练样本的规模,对此类方法的发展形成了制约. 最近有学者采用有监督学习与规则(先验知识)相结合的方法,取得了一些积极的研究成果,例如 Lin 等人[8]采用字典辅助下的最大熵算法,在基于 Medline 论文摘要的 GENIA 数据集上取得了实体抽取准确率和召回率均超过 70%的实验结果.

随着命名实体识别技术不断取得进展,学术界开始关注开放域(open domain)的信息抽取问题,即不再限定于特定的知识领域,而是面向开放的互联网,研究和解决全网信息抽取问题. 为此,需要首先建立一个科学完整的命名实体分类体系,一方面用于指导算法研究;另一方面便于对抽取得到的实体数据进行管理. 早在 2002 年,Sekine 等人[9]就提出了一个层次结构的命名实体分类体系,将网络中所有的命名实体划分为 150 个分类. 该项成果引起了学术界对建立命名实体分类体系的重视,并对后续的命名实体识别研究产生了深远的影响. 2012 年,Ling 等人[10]借鉴 Freebase 的实体分类方法,归纳出 112 种实体类别,并基于条件随机场模型进行实体边界识别,最后采用自适应感知机算法实现了对实体的自动分类,其实验结果显著优于 Stanford NER 等当前主流的命名实体识别系统.

然而,互联网中的内容是动态变化的,Web 2.0 技术更进一步推动了互联网的概念创新,采用人工预定义实体分类体系的方式已经很难适应时代的需求. 面向开放域的实体抽取和分类技术能够较好地解决这一问题,该方法的基本思想是对于任意给定的实体,采用统计机器学习的方法,从目标数据集(通常是网页等文本数据)中抽取出与之具有相似上下文特征的实体,从而实现实体的分类和聚类[11].

在面向开放域的实体识别和分类研究中,不需要(也不可能)为每个领域或每个实体类别建立单独的语料库作为训练集. 因此,该领域面临的主要挑战是如何从给定的少量实体实例中自动发现具有区分力的模式. 针对该问题,Whitelaw 等人[12]提出了一种迭代扩展实体语料库的解决方案,基本思路是根据已知的实体实例进行特征建模,利用该模型对处理海量数据集得到新的命名实体列表,然后针对新实体建模,迭代地生成实体标注语料库.

另一种思路是通过搜索引擎的服务器日志获取新出现的命名实体. 例如 Jain 等人[13]提出了一种面向开放域的无监督学习算法,即事先并不给出实体分类,而是基于实体的语义特征从搜索日志中识别出命名实体,然后采用聚类算法对识别出的实体对象进行聚类,该方法已经在搜索引擎技术中得到应用,用于根据用户输入的关键字自动补全信息.

2.1.2 关系抽取

文本语料经过实体抽取,得到的是一系列离散的命名实体,为了得到语义信息,还需要从相关语料中提取出实体之间的关联关系,通过关系将实体(概念)联系起来,才能够形成网状的知识结构. 研究关系抽取技术的目的,就是解决如何从文本语料中抽取实体间的关系这一基本问题.

早期的关系抽取研究方法主要是通过人工构造语法和语义规则,据此采用模式匹配的方法来识别实体间的关系. 这种方法有 2 点明显的不足:1)要求制定规则的人具有良好的语言学造诣,并且对特定领域有深入的理解和认知;2)规则制定工作量大,难以适应丰富的语言表达风格,且难以拓展到其他领域. 为此学术界开始尝试采用统计机器学习方法,通过对实体间关系的模式进行建模,替代预定义的语法和语义规则. 例如 Kambhatla 等人[14]利用自然语言中的词法、句法以及语义特征进行实体关系建模,

通过最大熵方法成功地实现了不借助规则硬编码的实体关系抽取.

随后,出现了大量基于特征向量或核函数的有监督学习方法,关系抽取的准确性也不断提高.例如,刘克彬等人[15]借助知网(HowNet)提供的本体知识库构造语义核函数,在开放数据集上对ACE定义的6类实体关系进行抽取,准确率达到了88%.然而,有监督学习方法也存在明显不足,为了确保算法的有效性,需要人工标注大量的语料作为训练集.因此,近年来的研究重点逐渐转向半监督和无监督的学习方式.例如,Carlson等人[16]提出了一种基于Bootstrap算法的半监督学习方法,能够自动进行实体关系建模.陈立玮等人[17]针对弱监督学习中标注数据不完全可靠的问题,基于Bootstrapping算法设计思想,提出了一种协同训练方法,通过向传统模型中引入N-Gram特征进行协同训练,实现了对弱监督关系抽取模型的强化,在中文和英文数据集上关系抽取性能均得到了提升.Zhang等人[18]采用基于实例的无监督学习方法,在公开语料库上获得了较好的实验结果,能够对实体间的雇佣关系、位置关系以及生产关系等多元关系进行精准识别.

以上研究成果的共同特点是需要预先定义实体关系类型,如雇佣关系、整体部分关系以及位置关系等.然而在实际应用中,要想定义出一个完美的实体关系分类系统是十分困难的.为了解决这一制约关系抽取技术走向实际应用的关键问题,2007年,华盛顿大学图灵中心的Banko等人[19]提出了面向开放域的信息抽取方法框架(open information extraction,OIE),并发布了基于自监督(self-supervised)学习方式的开放信息抽取原型系统(TextRunner).该系统采用少量人工标记数据作为训练集,据此得到一个实体关系分类模型,再依据该模型对开放数据进行分类,依据分类结果训练朴素贝叶斯模型来识别"实体-关系-实体"三元组,经过大规模真实数据测试,取得了显著优于同时期其他方法的结果.

面向开放域的关系抽取技术直接利用语料中的关系词汇对实体关系进行建模,因此不需要预先指定关系的分类,这是一个很大的进步,例如,Wu等人[20]在OIE的基础上,发布了面向开放域信息抽取的WOE系统,该系统能够利用维基百科网页信息框(infobox)提供的属性信息,自动构造实体关系训练集,性能优于早期的TextRunner系统,这项工作也为批量构造高质量的训练语料提供了新的思路.Fader等人[21]通过对TextRunner系统和WOE系统的实体关系抽取结果进行分析,发现其中错误的部分主要是一些无意义或不合逻辑的实体关系三元组,据此引入语法限制条件和字典约束,采用先识别关系指示词,然后再对实体进行识别的策略,有效提高了关系识别准确率.Mausam等人[22]针对上述系统均无法识别非动词性关系的局限,通过引入上下文分析技术,提出了一个支持非动词性关系抽取的OILLIE系统,有效提高了自动关系抽取的准确率和召回率.

由于当前的面向开放域的关系抽取方法在准确率和召回率等综合性能指标方面与面向封闭领域的传统方法相比仍有一定的差距,因此有部分学者开始尝试将两者的优势结合起来.例如Banko等人[23]提出了一种基于条件随机场的关系抽取模型(H-CRF),当目标数据集中拥有的关系数量不大,而且有预先定义好的实体关系分类模型可用的情况下,采用传统的机器学习算法进行关系抽取,而对于没有预先定义好的实体关系模型或者关系数量过多的情况,则采用开放域关系抽取方法.微软公司人立方项目所采用的StatSnowball模型也是基于这种策略实现其关系抽取功能[24].

当前流行的OIE系统在关系抽取方面存在2个主要问题.1)当前研究的重点是如何提高二元实体间关系(三元组模式)的抽取准确率和召回率,很少考虑到在现实生活中普遍存在的高阶多元实体关系;2)所采用的研究方法大多只关注发掘词汇或词组之间的关系模式,而无法实现对隐含语义关系的抽取.对此,学术界有着清醒的认识,例如Alan等人[25]采用N元关系模型对OIE系统进行改进,提出了KRAKEN模型,能够有效提高OIE系统对多元实体关系的识别能力.在隐含关系识别方面,McCallum[26]提出采用后期关系推理的方法,提高OIE系统对隐含实体关系的发现能力.这些工作都是该领域值得重视的研究动向,然而在OIE关系抽取研究领域,要实现算法性能由量变到质变的飞跃,还需要一段时间的积累.

2.1.3　属性抽取

属性抽取的目标是从不同信息源中采集特定实体的属性信息.例如针对某个公众人物,可以从网络公开信息中得到其昵称、生日、国籍、教育背景等信息.属性抽取技术能够从多种数据来源中汇集这些信息,实现对实体属性的完整勾画.

由于可以将实体的属性视为实体与属性值之间的一种名词性关系,因此也可以将属性抽取问题视为关系抽取问题.例如郭剑毅等人[27]将人物属性抽取问题转化为实体关系抽取问题,采用支持向量机算法实现了人物属性抽取与关系预测模型.

百科类网站提供的半结构化数据是当前实体属性抽取研究的主要数据来源.例如 Suchanek 等人[28]设计了基于规则和启发式算法的属性抽取算法,能够从 Wikipedia 和 WordNet 网页信息框中自动提取属性名和属性值信息,据此得到了扩展性良好的本体知识库(YAGO),其抽取准确率高达 95%.受 YAGO 和 Freebase 项目的启发,DBpedia 项目以维基百科作为研究对象,从维基百科网页信息框中抽取了超过 458 万个实体和超过 30 亿条实体关系信息.作为 Linked Data 项目的重要组成部分,DBpedia 构建了一个维基百科之上的知识网络,反过来促进了维基百科的应用创新,如关系查询、多维度搜索等,DBpedia 也因此成为了目前世界上最庞大的多领域本体知识库之一[29].

尽管可以从百科类网站获取大量实体属性数据,然而这只是人类知识的冰山一角,还有大量的实体属性数据隐藏在非结构化的公开数据中.如何从海量非结构化数据中抽取实体属性是值得关注的理论研究问题.一种解决方案是基于百科类网站的半结构化数据,通过自动抽取生成训练语料,用于训练实体属性标注模型,然后将其应用于对非结构化数据的实体属性抽取[30];另一种方案是采用数据挖掘的方法直接从文本中挖掘实体属性与属性值之间的关系模式,据此实现对属性名和属性值在文本中的定位.这种方法的基本假设是属性名和属性值之间有位置上的关联关系,事实上在真实语言环境中,许多实体属性值附近都存在一些用于限制和界定该属性值含义的关键词(属性名),在自然语言处理技术中将这类属性称为有名属性,因此可以利用这些关键字来定位有名属性的属性值[31].

2.2 知识融合

通过信息抽取,实现了从非结构化和半结构化数据中获取实体、关系以及实体属性信息的目标,然而,这些结果中可能包含大量的冗余和错误信息,数据之间的关系也是扁平化的,缺乏层次性和逻辑性,因此有必要对其进行清理和整合.知识融合包括 2 部分内容:实体链接和知识合并.通过知识融合,可以消除概念的歧义,剔除冗余和错误概念,从而确保知识的质量.

2.2.1 实体链接

实体链接(entity linking)是指对于从文本中抽取得到的实体对象,将其链接到知识库中对应的正确实体对象的操作[32].

实体链接的基本思想是首先根据给定的实体指称项,从知识库中选出一组候选实体对象,然后通过相似度计算将指称项链接到正确的实体对象.早期的实体链接研究仅关注如何将从文本中抽取到的实体链接到知识库中,忽视了位于同一文档的实体间存在的语义联系,近年来学术界开始关注利用实体的共现关系,同时将多个实体链接到知识库中,称为集成实体链接(collective entity linking).例如 Han 等人[33]提出的基于图的集成实体链接方法,能够有效提高实体链接的准确性.

实体链接的一般流程是:1)从文本中通过实体抽取得到实体指称项;2)进行实体消歧和共指消解,判断知识库中的同名实体与之是否代表不同的含义以及知识库中是否存在其他命名实体与之表示相同的含义;3)在确认知识库中对应的正确实体对象之后,将该实体指称项链接到知识库中对应实体.

1) 实体消歧

实体消歧(entity disambiguation)是专门用于解决同名实体产生歧义问题的技术.在实际语言环境中,经常会遇到某个实体指称项对应于多个命名实体对象的问题,例如“李娜”这个名词(指称项)可以对应于作为歌手的李娜这个实体,也可以对应于作为网球运动员的李娜这个实体,通过实体消歧,就可以根据当前的语境,准确建立实体链接.实体消歧主要采用聚类法.

聚类法是指以实体对象为聚类中心,将所有指向同一目标实体对象的指称项聚集到以该对象为中心的类别下.聚类法消歧的关键问题是如何定义实体对象与指称项之间的相似度,常用方法有 4 种.

① 空间向量模型(词袋模型).典型的方法是取当前语料中实体指称项周边的词构成特征向量,然后利用向量的余弦相似度进行比较,将该指称项聚类到与之最相近的实体指称项集合中.例如 Bagga 等人[34]采用该方法,在 MUC6(Message Understanding Conference)数据集上取得了很高的消歧精度(F 值高达 84.6%).然而该方法的缺点在于没有考虑上下文语义信息,这种信息损失会导致在某些情况下算法性能恶化,如短文本分析.

② 语义模型.该模型与空间向量模型类似,区

别在于特征向量的构造方法不同,语义模型的特征向量不仅包含词袋向量,而且包含一部分语义特征.例如 Pedersen 等人[35]采用奇异值分解技术对文本向量空间进行分解,得到给定维度的浅层语义特征,以此与词袋模型相结合,能够得到更精确的相似度计算结果.

③ 社会网络模型.该模型的基本假设是物以类聚、人以群分,在社会化语境中,实体指称项的意义在很大程度上是由与其相关联的实体所决定的.建模时,首先利用实体间的关系将与之相关的指称项链接起来构成网络,然后利用社会网络分析技术计算该网络中节点之间的拓扑距离(网络中的节点即实体的指称项),以此来判定指称项之间的相似度.例如 Malin 等人[36]利用随机漫步模型对演员合作网络数据进行实体消歧,得到了比基于文本相似度模型更好的消歧效果.

④ 百科知识模型.百科类网站通常会为每个实体(指称项)分配一个单独页面,其中包括指向其他实体页面的超链接,百科知识模型正是利用这种链接关系来计算实体指称项之间的相似度.例如 Han 等人[37]利用维基百科条目之间的关联关系计算实体指称项之间的相似度,实验结果表明这种方式能够有效消除同名实体间的歧义.Bunescu 等人[38]以维基百科作为知识库,基于实体所在页面的上下文信息和指称项所在语料的上下文信息,利用词袋模型构造特征向量作为实体链接时进行相似度比较的依据,实现了实体消歧.在此基础上,Sen[39]进一步采用主题模型作为相似度计算依据,在维基百科人物数据集上获得了高达 86% 的消歧准确率.Shen 等人[40]提出的 Linden 模型则同时考虑到了文本相似性和主题一致性,基于维基百科和 Wordnet 知识库,取得了当前最好的实体消歧实验结果.然而,由于百科类知识库中的实体数非常有限,此类方法的推广性较差.

为了充分利用海量公开数据中包含的实体区分性证据,Li 等人[32]基于生成模型提出了一种增量证据挖掘算法,在 Twitter 数据集上实现了实体消歧准确率的大幅提升.该方法降低了消歧算法对于知识库的依赖,提供了一种很有希望的算法新思路.

实体消歧技术能够帮助搜索引擎更好地理解用户的搜索意图,从而给出更好的上下文推荐结果,提高搜索服务质量.其中还有一个很重要的问题是如何对存在歧义的实体进行重要性评估,以确定推荐内容的优先级.当前的主要研究思路是为实体赋予权重,用于表示该实体出现的频率或先验概率.例如 Ratinov 等人[41]通过统计维基百科中的实体出现的频率以此作为实体推荐时排序的依据.Ochs 等人[42]则借助搜索引擎的关键词日志和 DBpedia 知识库,构建了一个知名人物本体库,据此实现了一个本体搜索引擎原型系统,为解决人物实体的重要性评估提供了一种新的思路.

2) 共指消解

共指消解(entity resolution)技术主要用于解决多个指称项对应于同一实体对象的问题.例如在一篇新闻稿中,"Barack Obama","president Obama","the president"等指称项可能指向的是同一实体对象,其中的许多代词如"he","him"等,也可能指向该实体对象.利用共指消解技术,可以将这些指称项关联(合并)到正确的实体对象.由于该问题在信息检索和自然语言处理等领域具有特殊的重要性,吸引了大量的研究努力,因此学术界对该问题有多种不同的表述,典型的包括:对象对齐(object alignment)、实体匹配(entity matching)以及实体同义(entity synonyms).

共指消解问题的早期研究成果主要来自自然语言处理领域,近年来统计机器学习领域的学者越来越多地参与到这项工作中.基于自然语言处理的共指消解是以句法分析为基础的,代表性方法是 Hobbs 算法和向心理论(centering theory).Hobbs 算法是最早的代词消解算法之一,主要思路是基于句法分析树进行搜索,因此适用于实体与代词出现在同一句子中的场景,有一定的局限性.早期的 Hobbs 算法完全基于句法分析(朴素 Hobbs 算法),后来则加入了语义分析并沿用至今[43].向心理论的基本思想是:将表达模式(utterance)视为语篇(discourse)的基本组成单元,通过识别表达模式中的实体,可以获得当前和后续语篇中的关注中心(实体),根据语义的局部连贯性和显著性,就可以在语篇中跟踪受关注的实体[44].向心理论的提出最初并不是为了解决代词消解问题,而是为了对语篇中关注中心的局部连贯性进行建模,因此它虽然一段时间内成为主要的代词消解手段,但却不是最佳的理论模型.近年来,学术界开始尝试在向心理论的基础上,利用词性标注和语法分析技术,提高实体消解方法的适用范围和准确性.例如 Lappin 等人[45]基于句法分析和词法分析技术提出了消解算法,能够识别语篇中的第 3 人称代词和反身代词等回指性代词在语篇中回

指的对象,其性能优于 Hobbs 算法和基于向心理论的实体消解方法.

随着统计机器学习方法被引入该领域,共指消解技术进入了快速发展阶段. McCarthy 等人[46]首次将 C4.5 决策树算法应用于解决共指消解问题,结果在 MUC-5 公开数据集的多数任务中均取得了优胜. Bean 等人[47]通过实验发现,语义背景知识对于构造共指消解算法非常有帮助,他们利用 Utah 大学发布的 AutoSlog 系统从原始语料中抽取实体上下文模式信息,应用 Dempster-Shafer 概率模型对实体模式进行建模,在 2 个公开数据集上(MUC-4 的恐怖主义数据集和路透社自然灾害新闻数据集)分别取得了 76%和 87%的共指消解准确率.

除了将共指消解问题视为分类问题之外,还可以将其作为聚类问题来求解. 聚类法的基本思想是以实体指称项为中心,通过实体聚类实现指称项与实体对象的匹配. 其关键问题是如何定义实体间的相似性测度. Turney[48]基于点互信息(pointwise mutual information, PMI)来求解实体所在文档的相似度,并用于求解 TOEFL 和 ESL 考试中的同义词测试问题,取得了 74%的正确率. Cheng 等人[49]通过对搜索引擎的查询和点击记录进行研究,发现可以根据用户查询之后的点击行为对实体进行区分. 据此,通过查询和点击记录建立实体指称项与相关网页 URL 之间的关联,进而计算出实体指称项之间的点击相似度(click similarity),结果表明该方法能够有效实现共指消解,从而提高搜索覆盖率.

基于统计机器学习的共指消解方法通常受限于 2 个问题:训练数据的(特征)稀疏性和难以在不同的概念上下文中建立实体关联. 为解决该问题,Pantel 等人[50]基于 Harris 提出的分布相似性模型,提出了一个新的实体相似性测度模型,称为术语相似度(term similarity),借助该模型可以从全局语料中得到所有术语间的统计意义上的相似性,据此可以完成实体合并,达到共指消解的目的. Chakrabarti 等人[51]则将网页点击相似性和文档相似性这 2 种测度相结合,提出了一种新的查询上下文相似性测度(query context similarity),通过在 Bing 系统上进行测试,该测度能够有效识别同义词,并显著提高查全率. 值得注意的是,上述 2 种方法均支持并行计算,二者均采用了 MapReduce 框架,其中,前者在 200 个 4 核处理器上,用时 50h 得到了 5 亿条术语的相似度矩阵,而后者则已经在 Bing 搜索引擎的商品和视频搜索中取得应用.

2.2.2 知识合并

在构建知识图谱时,可以从第三方知识库产品或已有结构化数据获取知识输入. 例如,关联开放数据项目(linked open data)会定期发布其经过积累和整理的语义知识数据,其中既包括前文介绍过的通用知识库 DBpedia 和 YAGO,也包括面向特定领域的知识库产品,如 MusicBrainz 和 DrugBank 等.

1) 合并外部知识库

将外部知识库融合到本地知识库需要处理 2 个层面的问题. ①数据层的融合,包括实体的指称、属性、关系以及所属类别等,主要的问题是如何避免实例以及关系的冲突问题,造成不必要的冗余;②通过模式层的融合,将新得到的本体融入已有的本体库中[52].

为促进知识库融合的标准化,Mendes 等人[53]提出了开放数据集成框架(linked data integration framework, LDIF),用于对 LOD 知识库产品进行融合. 其中包括 4 个步骤:①获取知识;②概念匹配,由于不同本体库中的概念表达使用的词汇可能不同,因此需要对概念表达方式进行统一化处理;③实体匹配,由于知识库中有些实体含义相同但是具有不同的标识符,因此需要对这些实体进行合并处理;④知识评估,知识融合的最后一步是对新增知识进行验证和评估,以确保知识图谱的内容一致性和准确性,通常采用的方法是在评估过程中为新加入的知识赋予可信度值,据此进行知识的过滤和融合.

2) 合并关系数据库

在知识图谱构建过程中,一个重要的高质量知识来源是企业或者机构自己的关系数据库. 为了将这些结构化的历史数据融入到知识图谱中,可以采用资源描述框架(RDF)作为数据模型. 业界和学术界将这一数据转换过程形象地称为 RDB2RDF,其实质就是将关系数据库的数据换成 RDF 的三元组数据. 根据 W3C 的调查报告显示,当前已经出现了大量 RDB2RDF 的开源工具(如 Triplify, D2R Server, OpenLink Virtuoso, SparqlMap 等),然而由于缺少标准规范,使得这些工具的推广应用受到极大制约[54]. 为此,W3C 于 2012 年推出了 2 种映射语言标准:Direct Mapping (A direct mapping of relational data to RDF)和 R2RML (RDB to RDF mapping language). 其中,Direct Mapping 采用直接映射的方式,将关系数据库表结构和数据直接输出为 RDF 图,在 RDF 图中所用到的用于表示类和谓词的术语与关系数据库中的表名和字段名保持一致.

而 R2RML 则具有较高的灵活性和可定制性,允许为给定的数据库结构定制词汇表,可以将关系数据库通过 R2RML 映射为 RDF 数据集,其中所用的术语如类的名称,谓词均来自定义词汇表.

除了关系型数据库之外,还有许多以半结构化方式存储(如 XML,CSV,JSON 等格式)的历史数据也是高质量的知识来源,同样可以采用 RDF 数据模型将其合并到知识图谱当中.当前已经有许多这样的工具软件,例如 XSPARQL 支持从 XML 格式转化为 RDF,Datalift 支持从 XML 和 CSV 格式转化为 RDF,经过 RDF 转化的知识元素,经实体链接之后,就可以加入到知识库中,实现知识合并[55].

2.3 知识加工

通过信息抽取,可以从原始语料中提取出实体、关系与属性等知识要素.再经过知识融合,可以消除实体指称项与实体对象之间的歧义,得到一系列基本的事实表达.然而,事实本身并不等于知识,要想最终获得结构化、网络化的知识体系,还需要经历知识加工的过程.知识加工主要包括 3 方面内容:本体构建、知识推理和质量评估.

2.3.1 本体构建

本体(ontology)是对概念进行建模的规范,是描述客观世界的抽象模型,以形式化方式对概念及其之间的联系给出明确定义.本体的最大特点在于它是共享的,本体中反映的知识是一种明确定义的共识.虽然在不同时代和领域,学者们对本体曾经给出过不同的定义,但这些定义的内涵是一致的,即:本体是同一领域内的不同主体之间进行交流的语义基础[56].本体是树状结构,相邻层次的节点(概念)之间具有严格的"IsA"关系,这种单纯的关系有助于知识推理,但却不利于表达概念的多样性.在知识图谱中,本体位于模式层,用于描述概念层次体系是知识库中知识的概念模板[57].

本体可以采用人工编辑的方式手动构建(借助本体编辑软件),也可以采用计算机辅助,以数据驱动的方式自动构建,然后采用算法评估和人工审核相结合的方式加以修正和确认.对于特定领域而言,可以采用领域专家和众包的方式人工构建本体.然而对于跨领域的全局本体库而言,采用人工方式不仅工作量巨大,而且很难找到符合要求的专家.因此,当前主流的全局本体库产品,都是从一些面向特定领域的现有本体库出发,采用自动构建技术逐步扩展得到的.例如微软发布的 Probase 本体库就是采用数据驱动的自动化构建方法,利用统计机器学习算法迭代地从网页文本数据中抽取出概念之间的"IsA"关系,然后合并形成概念层次.目前,Probase 中包含了超过 270 万条概念,准确率高达 92.8%,在规模和准确性方面居于领先地位[58].

数据驱动的自动化本体构建过程包含 3 个阶段:实体并列关系相似度计算、实体上下位关系抽取以及本体的生成[59].1)实体并列关系相似度是用于考察任意给定的 2 个实体在多大程度上属于同一概念分类的指标测度,相似度越高,表明这 2 个实体越有可能属于同一语义类别.所谓并列关系,是相对于纵向的概念隶属关系而言的.例如"中国"和"美国"作为国家名称的实体,具有较高的并列关系相似度;而"美国"和"手机"这 2 个实体,属于同一语义类别的可能性较低,因此具有较低的并列关系相似度.2)实体上下位关系抽取是用于确定概念之间的隶属(IsA)关系,这种关系也称为上下位关系,例如,词组(导弹,武器)构成上下位关系,其中的"导弹"为下位词,"武器"为上位词.3)本体生成阶段的主要任务是对各层次得到的概念进行聚类,并对其进行语义类的标定(为该类中的实体指定 1 个或多个公共上位词).

当前主流的实体并列关系相似度计算方法有 2 种:模式匹配法和分布相似度法.其中,模式匹配法采用预先定义实体对模式的方式,通过模式匹配取得给定关键字组合在同一语料单位中共同出现的频率,据此计算实体对之间的相似度.分布相似度(distributional similarity)方法的前提假设是:在相似的上下文环境中频繁出现的实体之间具有语义上的相似性[60].在具体计算时,首先将每个实体表示成 1 个 N 维向量,其中,向量的每个维度表示 1 个预先定义的上下文环境,向量元素值表示该实体出现在各上下文环境中的概率,然后就可以通过求解向量间的相似度,得到实体间的并列关系相似度.

实体上下位关系抽取是该领域的研究重点,主要的研究方法是基于语法模式(如 Hearst 模式)抽取 IsA 实体对[57].当前主流的信息抽取系统,如 KnowItAll,TextRunner,NELL 等,都可以在语法层面抽取实体上下位关系,而 Probase 则是采用基于语义的迭代抽取技术,以逐步求精的方式抽取实体上下位关系.基于语义的迭代抽取技术,一般是利用概率模型判定 IsA 关系和区分上下位词,通常会借助百科类网站提供的概念分类知识来帮助训练模型,

以提高算法精度[61]. 例如 Probase 在处理"domestic animals other than dogs such as cats"这样的句子时,可以通过抽取 IsA 实体对中的上下位词得到 2 个备选事实:(cat,IsA,dog)和(cat,IsA,domestic animal). 如果 Probase 中已经有关于这些实体的概念,就可以得到正确的结果[58].

除了数据驱动的方法,还可以用跨语言知识链接的方法来构建本体库. 例如 Wang 等人[62]利用跨语言知识链接方法得到的知识对,在分别生成中英文本体模型的过程中,使二者相互确认,同时提高了中文关系和英文关系预测的准确度.

当前对本体生成方法的研究工作主要集中于实体聚类方法,主要的挑战在于经过信息抽取得到的实体描述非常简短,缺乏必要的上下文信息,导致多数统计模型不可用. 例如 Wang 等人[63]利用基于主题进行层次聚类的方法得到本体结构,为了解决主题模型不适用于短文本的问题,提出了一个基于单词共现网络(term co-occurrence network)的主题聚类和上位词抽取模型(CATHY),实现了基于短文本的主题聚类. Liu 等人[64]则采用贝叶斯模型对实体关键词进行分层聚类,经过改进的算法具有近似线性的复杂度($O(n\log n)$),能够在 1 h 内从 100 万关键词中抽取出特定领域的本体.

2.3.2 知识推理

知识推理是指从知识库中已有的实体关系数据出发,经过计算机推理,建立实体间的新关联,从而拓展和丰富知识网络. 知识推理是知识图谱构建的重要手段和关键环节,通过知识推理,能够从现有知识中发现新的知识. 例如已知(乾隆,父亲,雍正)和(雍正,父亲,康熙),可以得到(乾隆,祖父,康熙)或(康熙,孙子,乾隆). 知识推理的对象并不局限于实体间的关系,也可以是实体的属性值、本体的概念层次关系等. 例如已知某实体的生日属性,可以通过推理得到该实体的年龄属性. 根据本体库中的概念继承关系,也可以进行概念推理,例如已知(老虎,科,猫科)和(猫科,目,食肉目),可以推出(老虎,目,食肉目).

知识的推理方法可以分为 2 大类:基于逻辑的推理和基于图的推理.

基于逻辑的推理主要包括一阶谓词逻辑、描述逻辑以及基于规则的推理. 一阶谓词逻辑建立在命题的基础上,在一阶谓词逻辑中,命题被分解为个体(individuals)和谓词(predication)2 部分. 个体是指可独立存在的客体,可以是一个具体的事物,例如奥巴马,也可以是一个抽象的概念,例如学生. 谓词是用来刻画个体的性质及事物关系的词,例如三元组(A,friend,B)中 friend 就是表达个体 A 和 B 关系的谓词. 举例来说,对于人际关系可以采用一阶谓词逻辑进行推理,方法是将关系视为谓词,将人物视为变元,采用逻辑运算符号表达人际关系,然后设定关系推理的逻辑和约束条件,就可以实现简单关系的逻辑推理.

对于复杂的实体关系,可以采用描述逻辑进行推理. 描述逻辑(description logic)是一种基于对象的知识表示的形式化工具,是一阶谓词逻辑的子集,它是本体语言推理的重要设计基础. 基于描述逻辑的知识库一般包含 TBox(terminology box)与 ABox(assertion box),其中,TBox 是用于描述概念之间和关系之间的关系的公理集合,ABox 是描述具体事实的公理集合. 借助这 2 个工具,可以将基于描述逻辑的推理最终归结为 ABox 的一致性检验问题,从而简化并最终实现关系推理[65].

当基于本体的概念层次进行推理时,对象主要是以 Web 本体语言(OWL)描述的概念,OWL 提供丰富的语句,具有很强的知识描述能力. 然而在描述属性合成和属性值转移方面,网络本体语言的表达能力就显得不足,为了实现推理,可以利用专门的规则语言(如 semantic Web rule language,SWRL)对本体模型添加自定义规则进行功能拓展. 例如 Lu 等人[66]借助 SWRL 规则向本体库添加实体隐含关系推理规则,据此实现了网络服务的匹配机制.

基于图的推理方法主要基于神经网络模型或 Path Ranking 算法. 例如 Socher 等人[67]将知识库中的实体表达为词向量的形式,进而采用神经张量网络模型(neural tensor networks)进行关系推理,在 WordNet 和 FreeBase 等开放本体库上对未知关系进行推理的准确率分别达到 86.2%和 90.0%.

Path Ranking 算法的基本思想是将知识图谱视为图(以实体为节点,以关系或属性为边),从源节点开始,在图上执行随机游走,如果能够通过一个路径到达目标节点,则推测源和目的节点间可能存在关系. 例如假设 2 个节点(X,Y)共有 1 个孩子 Z,即存在路径 $X \xrightarrow{\text{Parentof}} Z \xleftarrow{\text{Parentof}} Y$,据此推测 X 和 Y 之间可能存在 MarriedTo 关系[68].

开放域信息抽取技术极大地拓展了知识图谱的知识来源,知识库内容的极大丰富为知识推理技术

的发展提供了新的机遇和挑战,现有的知识推理技术已经明显滞后于需求.由于推理得到的知识准确性低、冗余度高,因此在将其加入到知识库之前,通常需要进行可证明性检查、矛盾性检查、冗余性检查以及独立性检查,以确保推理的知识加入知识库后不会产生矛盾和冗余[69].在实际应用中,知识库的构建者为保证知识库应用的时效性,通常仅保留部分与业务密切相关的知识,而放弃其他推理结果.

此外,跨知识库的知识推理也是大趋势,同时也带来新的挑战,已经有部分学者开始关注这一问题.例如卢道设等人[70]通过对描述逻辑的表现形式进行扩展,提出了一种基于组合描述逻辑的 Tableau 算法,基于概念的相似性对不同领域的概念进行关联.实验结果表明,基于组合描述逻辑的推理方法可以利用不同知识库中的已有知识进行推理,该成果为跨知识库的知识推理方法研究提供了新的思路.

2.3.3　质量评估

质量评估也是知识库构建技术的重要组成部分.1)受现有技术水平的限制,采用开放域信息抽取技术得到的知识元素有可能存在错误(如实体识别错误、关系抽取错误等),经过知识推理得到的知识的质量同样也是没有保障的,因此在将其加入知识库之前,需要有一个质量评估的过程;2)随着开放关联数据项目的推进,各子项目所产生的知识库产品间的质量差异也在增大,数据间的冲突日益增多,如何对其质量进行评估,对于全局知识图谱的构建起着重要的作用.引入质量评估的意义在于:可以对知识的可信度进行量化,通过舍弃置信度较低的知识,可以保障知识库的质量.

为解决知识库之间的冲突问题,Mendes 等人[53]在 LDIF 框架基础上提出了一种新的质量评估方法(Sieve 方法),支持用户根据自身业务需求灵活定义质量评估函数,也可以对多种评估方法的结果进行综合考评以确定知识的最终质量评分.

在对 REVERB 系统的信息抽取质量进行评估时,Fader 等人[21]采用人工标注方式对 1000 个句子中的实体关系三元组进行了标注,并以此作为训练集,得到了一个逻辑斯蒂回归模型,用于对 REVERB 系统的信息抽取结果计算置信度.

谷歌的 Knowledge Vault 项目从全网范围内抽取结构化的数据信息,并根据某一数据信息在整个抽取过程中抽取到的频率对该数据信息的可信度进行评分,然后利用从可信知识库 Freebase 中得到先验知识对先前的可信度信息进行修正,实验结果表明,这一方法可以有效降低对数据信息正误判断的不确定性,提高知识图谱中知识的质量[71].

对于用户贡献的结构化知识的评估,与通过信息抽取获得的知识评估方法稍有不同.谷歌提出了一种依据用户的贡献历史和领域,以及问题的难易程度进行自动评估用户贡献知识质量的方法.用户提交知识后,该方法可以立刻计算出知识的可信度.使用该方法对大规模的用户贡献知识的评估准确率达到了 91%,召回率达到了 80%[72].

2.4　知识更新

人类所拥有的信息和知识量都是时间的单调递增函数,因此知识图谱的内容也需要与时俱进,其构建过程是一个不断迭代更新的过程.

从逻辑上看,知识库的更新包括概念层的更新和数据层的更新.概念层的更新是指新增数据后获得了新的概念,需要自动将新的概念添加到知识库的概念层中.数据层的更新主要是新增或更新实体、关系和属性值,对数据层进行更新需要考虑数据源的可靠性、数据的一致性(是否存在矛盾或冗余等问题)等多方面因素.当前流行的方法是选择百科类网站等可靠数据源,并选择在各数据源中出现频率高的事实和属性加入知识库.知识的更新也可以采用众包的模式(如 Freebase),而对于概念层的更新,则需要借助专业团队进行人工审核.

知识图谱的内容更新有 2 种方式:数据驱动下的全面更新和增量更新.所谓全面更新是指以更新后的全部数据为输入,从零开始构建知识图谱.这种方式比较简单,但资源消耗大,而且需要耗费大量人力资源进行系统维护;而增量更新,则是以当前新增数据为输入,向现有知识图谱中添加新增知识.这种方式资源消耗小,但目前仍需要大量人工干预(定义规则等),因此实施起来十分困难[52].

3　跨语言知识图谱的构建

随着英文知识图谱技术的快速发展,各语种的知识库建设也处在快速发展变化当中,跨语言知识图谱的构建技术也因此成为该领域的研究热点.对我国学者而言,更应发挥我们在中文信息处理方面的天然优势,面对挑战和机遇,做出应有的贡献.

研究构建跨语言知识图谱的意义在于:1)由于各语种知识分布不均匀,对其进行融合可以有效地弥补单语种知识库的不足;2)可以充分利用多语种在知识表达方式上的互补性,增加知识的覆盖率和

共享度;3)构建跨语言知识图谱可以比较不同语言对同一知识的表述,进而达到过滤错误信息,更新过时信息的目的.因此需要在多个语种间实现知识的融合,构建多语种知识间的映射关系.

跨语言知识图谱可以应用于跨语言的信息检索、机器翻译以及跨语言知识问答等.由于其广泛的应用前景,跨语言知识图谱的构建正得到学术界及业界的广泛重视.构建跨语言的知识图谱需要处理好3个关键问题:1)跨语言本体的构建;2)跨语言知识抽取;3)跨语言知识链接.其中,跨语言本体的构建可以参照2.3.1节介绍的本体构建方法,分别建立各语种的本体库,此处不再赘述.

3.1 跨语言知识抽取

由于不同语种间的知识分布存在不均衡性,将多语种知识进行融合可以有效地弥补单语种知识的不足,因此跨语言的知识抽取研究日益受到国内外重视,例如欧盟的Xlike项目和我国的XLore项目等.Xlike项目是由欧盟发起的框架项目,目的是对散布在世界各国主流媒体上的知识进行整合,实现跨语言的信息发布、媒体监督和商业智能服务,重点研究英、德、西、中、印等世界主流语言的跨语言知识抽取技术.XLore项目是清华大学构建的基于中英文的跨语言知识图谱,其中文知识源包括百度百科、互动百科以及中文维基百科,英文知识源为英文维基百科,该项目实现了跨语言知识图谱的构建,并能够提供中英文知识问答服务[73].

跨语言知识抽取的主要思路是借助于丰富的源语种知识自动化抽取缺失的目标语种知识.例如Nguye等人[74]采用基于翻译的跨语言知识抽取模型,该模型首先通过跨语言知识链接和属性对齐的方式将目标语种的相关内容映射到源语种知识库中所对应的内容,然后将相关知识翻译为目标语种,从而实现跨语言的知识抽取.这种方法的主要问题在于:1)受到不同语种间等价对象的数量以及源语种知识库中结构化信息(信息框)数量的限制;2)知识抽取的质量直接受机器翻译的质量限制.

针对跨语言知识抽取中存在的主题迁移和翻译错误问题,Wang等人[75]提出了一种基于迁移学习的跨语言知识抽取框架(WikiCiKE),该框架利用源语种知识库中丰富的无结构文本信息以及结构化信息,提高了目标语种知识库中信息抽取的数量和质量.通过与单语种知识抽取模型和基于翻译的跨语言知识抽取模型进行实验比较,WikiCiKE模型在4种典型属性信息(职业、国籍、母亲、故乡)上的信息抽取准确率和召回率分别提升了12.65%和12.47%,明显优于前2种抽取模型.

3.2 跨语言知识链接

知识链接是构建跨语言知识图谱需要解决的关键问题之一,其主要思想是将不同语言表示的相同知识链接起来,包括模式层的链接和数据层的链接.

模式层链接的核心是本体映射(对齐),其内涵是如果2个本体间如果存在语义上的概念关联,则通过语义关联实现二者之间的映射,本体映射的目的是实现知识的共享和重用.例如在合并2个本体知识库时,由于各自建立的依据不同,以及本体所对应的实例对象的个体丰富性,本体间的冲突在所难免,因此需要首先建立本体间的映射关系,然后再对知识图谱的数据层进行合并.当前主要研究的是单语种本体之间的映射,跨语言本体映射(cross-lingual ontology mapping或alignment)的研究还处于起步阶段.

跨语言本体映射研究的目标是实现不同语言的本体库之间的本体映射,当前主流的做法是使用翻译工具将其中一种语言的本体库翻译成另外一种语言,从而将跨语种本体映射问题转化为单语种本体映射问题.例如Fu等人[76]提出SOCOM方法分为3个阶段:1)将其他语言的本体翻译成目标语言的本体(称为rendering);2)执行单语言的实体对齐操作(称为matching);3)对映射的结果进行评估(称为matching audit),接受置信度高的映射结果.

Wang等人[77]提出了基于链接因子图模型的跨语言知识链接方法,根据本体的出链相似度、入链相似度、开放分类相似度以及作者兴趣相似度进行本体映射,实验表明该方法在多语维基百科上的预测准确率达到85.8%,召回率达到88.1%.同时,使用该模型可以在英文维基百科和中文百度百科中找到202 141组跨语言知识对.

基于链接相似度方法的准确性主要依赖于链接的结构和数量,使得一次发现的链接数量有限,会导致跨语言知识链接不准确.针对该问题,Wang等人[78]进一步提出了基于语义标注的增量式跨语言知识链接方法.1)利用少量的跨语言知识链接对,以及一些知识库内部链接作为种子;2)使用语义标注的方法丰富知识库内部链接;3)使用回归模型计算不同特征的权重,预测新的跨语言的知识链接,语义标注和知识链接预测结果相互迭代,不断增强.该方法在中英文维基百科数据集上有效提高了跨语言知识链接对的识别数量和质量.

4　知识图谱的应用

通过知识图谱,不仅可以将互联网的信息表达成更接近人类认知世界的形式,而且提供了一种更好的组织、管理和利用海量信息的方式.目前知识图谱技术主要用于智能语义搜索、移动个人助理(如Google Now,Apple Siri等)以及深度问答系统(如IBM Watson,Wolfram Alpha等),支撑这些应用的核心技术正是知识图谱技术.

在智能语义搜索应用中,当用户发起查询时,搜索引擎会借助知识图谱的帮助对用户查询的关键字进行解析和推理,进而将其映射到知识图谱中的一个或一组概念之上,然后根据知识图谱中的概念层次结构,向用户返回图形化的知识结构(其中包含指向资源页面的超链接信息),这就是我们在谷歌和百度的搜索结果中看到的知识卡片.

在深度问答应用中,系统同样会首先在知识图谱的帮助下对用户使用自然语言提出的问题进行语义分析和语法分析,进而将其转化成结构化形式的查询语句,然后在知识图谱中查询答案.对知识图谱的查询通常采用基于图的查询语句(如SPARQL),在查询过程中,通常会基于知识图谱对查询语句进行多次等价变换.例如,如果用户提问:"如何判断是否感染了埃博拉病毒?",则该查询有可能被等价变换成"感染埃博拉病毒的症状有哪些?",然后再进行推理变换,最终形成等价的三元组查询语句,如(埃博拉,症状,?)和(埃博拉,征兆,?)等,据此进行知识图谱查询得到答案.深度问答应用经常会遇到知识库中没有现成答案的情况,对此可以采用知识推理技术给出答案(参见2.4节).如果由于知识库不完善而无法通过推理解答用户的问题,深度问答系统还可以利用搜索引擎向用户反馈搜索结果,同时根据搜索的结果更新知识库,从而为回答后续的提问提前做出准备.

基于知识图谱的问答系统大致可以分为2类:基于信息检索的问答系统和基于语义分析的问答系统.其中,前者的主要代表是Jacana-Freebase系统①和华盛顿大学的Paralex系统②;后者的主要代表是斯坦福大学的SEMPRE系统③,分别介绍如下:

1)基于信息检索的问答系统的基本思路是首先将问题转变为一个基于知识库的结构化查询,从知识库中抽取与问题中实体相关的信息来生成多个候选答案,然后再从候选答案中识别出正确答案.Yao等人[79]基于Freebase知识库,对于一个给定的问题首先识别其中的疑问词、问题焦点词(暗示答案的类型)、问题主题词(知识库中的节点,即实体);识别问题中表示关系的词,并将关系词映射成Freebase中的关系谓词;根据问题主题词在Freebase知识库找到对应的节点和其相关的其他节点,以相关节点作为候选答案,遍历所有相关节点的属性和关系类型;从相关节点中识别出与关系词对应的节点作为答案.Berant等人[80]基于Freebase知识库,将给定问题转化为多个逻辑形式(logic form);根据抽取出的逻辑形式依据某种模式产生相对应的问题;计算产生的问题与原来输入问题的相似度.

2)基于语义分析的问答系统的基本思路是首先通过语义分析正确理解问题的含义,然后将问题转变为知识库的精确查询,直接找到正确答案.Fader等人[81]基于Freebase和Probase知识库,首先将给定的问题分解成小的问题,然后逐一进行解答,最后将答案合并.Berant等人[82]基于Freebase知识库,对于给定的问题,首先利用对齐规则将问题中实体、关系词、疑问词映射成知识库中的实体与关系谓词,然后将相邻的实体,关系谓词进行桥接,由此产生新的谓词,最后将问题中的所有谓词取交集形成一个精确的查询语句,再直接利用该查询得到答案.

5　问题与挑战

知识图谱是一个新概念,从2012年提出到现在不过2年时间,然而通过对知识图谱构建技术体系进行深入观察和分析,可以看出它事实上是建立在多个学科领域研究成果基础之上的一门实用技术,堪称是信息检索(information retrieval)、自然语言处理(natural language processing)、万维网(WWW)和人工智能(artificial intelligence)等领域交汇处的理论研究热点和应用技术集大成者.

虽然谷歌的Knowledge Vault和微软的Satori等项目已经部分揭示出知识图谱技术的魅力和前景,

① https://code.google.com/p/jacana

② http://knowitall.cs.washington.edu/paralex

③ http://www-nlp.stanford.edu/software

但通过以上分析不难看出,在知识图谱构建的各关键环节都面临着一些巨大的困难和挑战.

1) 在信息抽取环节,面向开放域的信息抽取方法研究还处于起步阶段,部分研究成果虽然在特定(语种、领域、主题等)数据集上取得了较好的结果,但普遍存在算法准确性和召回率低、限制条件多、扩展性不好的问题.因此,要想建成面向全球的知识图谱,第1个挑战来自开放域信息抽取,主要的问题包括实体抽取、关系抽取以及属性抽取.其中,多语种、开放领域的纯文本信息抽取问题是当前面临的重要挑战.

2) 在知识融合环节,如何实现准确的实体链接是一个主要挑战.虽然关于实体消歧和共指消解技术的研究已经有很长的历史,然而迄今为止所取得的研究成果距离实际应用还有很大距离.主要的研究问题包括开放域条件下的实体消歧、共指消解、外部知识库融合和关系数据库知识融合等问题.当前受到学术界普遍关注的问题是如何在上下文信息受限(短文本、跨语境、跨领域等)条件下,准确地将从文本中抽取得到的实体正确链接到知识库中对应的实体.

3) 知识加工是最具特色的知识图谱技术,同时也是该领域最大的挑战之所在.主要的研究问题包括:本体的自动构建、知识推理技术、知识质量评估手段以及推理技术的应用.目前,本体构建问题的研究焦点是聚类问题,对知识质量评估问题的研究则主要关注建立完善的质量评估技术标准和指标体系.知识推理的方法和应用研究是当前该领域最为困难,同时也是最为吸引人的问题,需要突破现有技术和思维方式的限制,知识推理技术的创新也将对知识图谱的应用产生深远影响.

4) 在知识更新环节,增量更新技术是未来的发展方向,然而现有的知识更新技术严重依赖人工干预.可以预见随着知识图谱的不断积累,依靠人工制定更新规则和逐条检视的旧模式将会逐步降低比重,自动化程度将不断提高,如何确保自动化更新的有效性,是该领域面临的又一重大挑战.

5) 最具基础研究价值的挑战是如何解决知识的表达、存储与查询问题,这个问题将伴随知识图谱技术发展的始终,对该问题的解决将反过来影响前面提出的挑战和关键问题.当前的知识图谱主要采用图数据库进行存储,在受益于图数据库带来的查询效率的同时,也失去了关系型数据库的优点,如SQL语言支持和集合查询效率等.在查询方面,如何处理自然语言查询,对其进行分析推理,翻译成知识图谱可理解的查询表达式以及等价表达式等也都是知识图谱应用需解决的关键问题.

6 结束语

互联网正从包含网页和网页之间超链接的文档万维网(Web of document)转变成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(Web of data).知识图谱作为下一代智能搜索的核心关键技术,具有重要的理论研究价值和现实的实际应用价值.本文从知识图谱构建的视角,对知识图谱的内涵,以及知识图谱构建关键技术的研究发展现状进行了全面调研和深入分析,并对知识图谱构建工作面临的重要挑战和关键问题进行了总结.

知识图谱的重要性不仅在于它是一个全局知识库,是支撑智能搜索和深度问答等智能应用的基础,而且在于它是一把钥匙,能够打开人类的知识宝库,为许多相关学科领域开启新的发展机会.从这个意义上来看,知识图谱不仅是一项技术,更是一项战略资产.本文的主要目的是介绍和宣传这项技术,希望吸引更多人重视和投入这项研究工作.

参 考 文 献

[1] Christian B, Heath T, Berners-Lee T. Linked data-the story so far [J]. International Journal on Semantic Web and Information Systems, 2009, 5(3): 1-22

[2] Chen Xueqi, Jin Xiaolong, Wang Yuanzhuo, et al. Survey on big data system and analytic technology [J]. Journal of Software, 2014, 25(9): 1889-1908 (in Chinese)
(程学旗, 靳小龙, 王元卓, 等. 大数据系统和分析技术综述[J]. 软件学报, 2014, 25(9): 1889-1908)

[3] Wang Yuanzhuo, Jia Yantao, Liu Dawei, et al. Open Web knowledge aided information search and data mining [J]. Journal of Computer Research and Development, 2014, 52(2): 456-474 (in Chinese)
(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, 2014, 52(2): 456-474)

[4] Cowie J, Lehnert W. Information extraction [J]. Communications of the ACM, 1996, 39(1): 80-91

[5] Chinchor N, Marsh E. Muc-7 information extraction task definition [C] //Proc of the 7th Message Understanding Conf. Philadelphia: Linguistic Data Consortium, 1998: 359-367

[6] Rau L F. Extracting company names from text [C] //Proc of the 7th IEEE Conf on Artificial Intelligence Applications. Piscataway, NJ: IEEE, 1991: 29-32

[7] Liu Xiaohua, Zhang Shaodian, Wei Furu, et al. Recognizing named entities in tweets [C] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2011: 359-367

[8] Lin Yifeng, Tsai Tzonghan, Chou Wenchi, et al. A maximum entropy approach to biomedical named entity recognition [C] //Proc of the 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics. New York: ACM, 2004: 56-61

[9] Sekine S, Sudo K, Nobata C. Extended named entity hierarchy [C] //Proc of the 3rd Language Resources and Evaluation Conf. New York: European Language Resources Association, 2002: 1818-1824

[10] Ling Xiao, Weld D. S. Fine-grained entity recognition [C] // Proc of the 26th Conf on Association for the Advancement of Artificial Intelligence. Menlo Park, CA: AAAI, 2012: 94-100

[11] Zhao Jun, Liu kang, Zhou Guangyou, et al. Open information extraction [J]. Journal of Chinese Information Processing, 2011, 25(6): 98-110 (in Chinese)
(赵军, 刘康, 周光有, 等. 开放式文本信息抽取[J]. 中文信息学报, 2011, 25(6): 98-110)

[12] Whitelaw C, Kehlenbeck A, Petrovic N, et al. Web-scale named entity recognition [C] //Proc of the 17th ACM Conf on Information and Knowledge Management. New York: ACM, 2008: 123-132

[13] Jain A, Pennacchiotti M. Open entity extraction from Web search query logs [C] //Proc of the 23rd Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2010: 510-518

[14] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations [C] //Proc of the 42nd Association for Computational Linguistics. Stroudsburg, PA: ACL, 2004: 1-22

[15] Liu Kebin, Li Fang, Liu Lei, et al. Implementation of a kernel-based chinese relation extraction system [J]. Journal of Computer Research and Development, 2007, 44(8): 1406-1411 (in Chinese)
(刘克彬, 李芳, 刘磊, 等. 基于核函数中文关系自动抽取系统的实现[J]. 计算机研究与发展, 2007, 44(8): 1406-1411)

[16] Carlson A, Betteridge J, Wang R C, et al. Coupled semi-supervised learning for information extraction [C] //Proc of the 3rd ACM Int Conf on Web Search and Data Mining. New York: ACM, 2010: 101-110

[17] Chen Liwei, Feng Yansong, Zhao Dongyan. Extracting relations from the Web via weakly supervised learning [J]. Journal of Computer Research and Development, 2013, 50(9): 1825-1835 (in Chinese)
(陈立玮 冯岩松 赵东岩. 基于弱监督学习的海量网络数据关系抽取[J]. 计算机研究与发展, 2013, 50(9): 1825-1835)

[18] Zhang Yiming, Zhou J F. A trainable method for extracting Chinese entity names and their relations [C] //Proc of the 2nd Workshop on Chinese Language Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2000: 66-72

[19] Banko M, Cafarella M J, Soderland S, et al. Open information extraction for the Web [C] //Proc of the 20th Int Joint Conf on Artificial Intelligence. New York: ACM, 2007: 2670-2676

[20] Wu Fei, Weld D S. Open information extraction using Wikipedia [C] //Proc of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2010: 118-127

[21] Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction [C] //Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2011: 1535-1545

[22] Mausam, Schmitz M, Bart R, et al. Open language learning for information extraction [C] //Proc of the Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg, PA: ACL, 2012: 523-534

[23] Banko M, Etzioni O. The Tradeoffs between open and traditional relation extraction [C] //Proc of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2008: 28-36

[24] Zhu Jun, Nie Zaijiang, Liu Xiaojiang, et al. StatSnowball: A statistical approach to extracting entity relationships [C] // Proc of the 18th Int Conf on World Wide Web. New York: ACM, 2009: 101-110

[25] Alan A, Alexander L. KrakeN: N-ary facts in open information extraction [C] //Proc of the Joint Workshop on Automatic Knowledge Base Construction and Web - scale Knowledge Extraction. Stroudsburg, PA: ACL, 2012: 52-56

[26] McCallum A. Joint inference for natural language processing [C] //Proc of the 13th Conf on Computational Natural Language Learning. Stroudsburg, PA: ACL, 2009: 1

[27] Guo Jianyi, Li Zhen, Yu Zhengtao, et al. Extraction and relation prediction of domain ontology concept instance, attribute and attribute [J]. Journal of Nanjing University: Natural Sciences, 2012, 48(4): 383-389 (in Chinese)
(郭剑毅, 李真, 余正涛, 等. 领域本体概念实例、属性和属性值的抽取及关系预测[J]. 南京大学学报: 自然科学版, 2012, 48(4): 383-389)

[28] Suchanek F M, 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

[29] Auer S, Bizer C, Kobilarov G, et al. Dbpedia: A nucleus for a Web of open data [C] //Proc of the 6th Int Semantic Web Conf. Berlin: Springer, 2007: 722-735

[30] Wu Fei, Weld D S. Autonomously semantifying wikipedia [C] //Proc of the 16th ACM Conf on Information and Knowledge Management. New York: ACM, 2007: 41-50

[31] Wang Yu, Tan Songbo, Liao Xiangwen, et al. Extracted domin model based named attribute extraction [J]. Journal of Computer Research and Development, 2010, 47(9): 1567-1573 (in Chinese)
(王宇, 谭松波, 廖祥文, 等. 基于扩展领域模型的有名属性抽取[J]. 计算机研究与发展, 2010, 47(9): 1567-1573)

[32] Li Yang, Wang Chi, Han Fangqiu, et al. Mining evidences for named entity disambiguation [C] //Proc of the 19th Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 1070-1078

[33] Han Xianpei, Sun Le, Zhao Jun. Collective entity linking in Web text: A graph-based method [C] //Proc of the 34th Int ACM Conf on Research and Development in Information Retrieval. New York: ACM, 2011: 765-774

[34] Bagga A, Baldwin B. Entity-based cross-document coreferencing using the vector space model [C] //Proc of the 17th Int Conf on Computational linguistics. Stroudsburg, PA: ACL, 1998: 79-85

[35] Pedersen T, Purandare A, Kulkarni A. Name discrimination by clustering similar contexts [G] //Proc of the 6th Int Conf on Intelligent Text Processing and Computational Linguistics. Berlin: Springer, 2005: 220-231

[36] Malin B, Airoldi E, Carley K. A network analysis model for disambiguation of names in lists [J]. Computational & Mathematical Organization Theory, 2005, 11(2): 119-139

[37] Han Xianpei, Zhao Jun. Named entity disambiguation by leveraging wikipedia semantic knowledge [C] //Proc of the 18th ACM Conf on Information and Knowledge Management. New York: ACM, 2009: 215-224

[38] Bunescu R, Pasca M. Using encyclopedic knowledge for named entity disambiguation [C] //Proc of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2006: 9-16

[39] Sen P. Collective context-aware topic models for entity disambiguation [C] //Proc of the 21st Int Conf on World Wide Web. New York: ACM, 2012: 729-738

[40] Shen Wei, Wang Jianyong, Luo Ping, et al. Linden: Linking named entities with knowledge base via semantic knowledge [C] //Proc of the 21st Int Conf on World Wide Web. New York: ACM, 2012: 449-458

[41] Ratinov L, Roth D, Downey D, et al. Local and global algorithms for disambiguation to wikipedia [C] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2011: 1375-1384

[42] Ochs C, Tian T, Geller J, et al. Google knows who is famous today—building an ontology from search engine knowledge and DBpedia [C] //Proc of the 5th IEEE Int Conf on Semantic Computing. Piscataway, NJ: IEEE, 2011: 320-327

[43] Hobbs J R. Resolving pronoun references [J]. Lingua, 1978, 44(4): 311-338

[44] Grosz B J, Weinstein S, Joshi A K. Centering: A framework for modeling the local coherence of discourse [J]. Computational Linguistics, 1995, 21(2): 203-225

[45] Lappin S, Shalom H J. An algorithm for pronominal anaphora resolution [J]. Computational Linguistics, 1994, 20(4): 535-561

[46] McCarthy J F, Lehnert W G. Using decision trees for coreference resolution [C] //Proc of the 14th Int Joint Conf on Artificial Intelligence. San Francisco: Morgan Kaufmann, 1995: 1050-1055

[47] Bean D L, Riloff E. Unsupervised learning of contextual role knowledge for coreference resolution [C] //Proc of the Human Language Technologies North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2004: 297-304

[48] Turney P. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL [C] //Proc of the 12th European Conf on Machine Learning. Berlin: Springer, 2001: 491-502

[49] Cheng Tao, Lauw H W, Paparizos S. Entity synonyms for structured Web search [J]. IEEE Trans on Knowledge and Data Engineering, 2012, 24(10): 1862-1875

[50] Pantel P, Crestan E, Borkovsky A, et al. Web-scale distributional similarity and entity set expansion [C] //Proc of the 2009 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2009: 938-947

[51] Chakrabarti K, Chaudhuri S, Cheng Tao, et al. A framework for robust discovery of entity synonyms [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2012: 1384-1392

[52] Deshpande O, Lamba D S, Tourn M, et al. Building, maintaining, and using knowledge bases: A report from the trenches [C] //Proc of the 32nd ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2013: 1209-1220

[53] Mendes P N, Mühleisen H, Bizer C. Sieve: Linked data quality assessment and fusion [C] //Proc of the 2nd Int Workshop on Linked Web Data Management at Extending Database Technology. New York: ACM, 2012: 116-123

[54] Sahoo S S, Halb W, Hellmann S, et al. A survey of current approaches for mapping of relational databases to RDF [R]. Cambridge, MA: The W3C RDB2RDF Working Group, 2009

[55] Michel F, Montagnat J, Faron-Zucker C. A survey of RDB to RDF translation approaches and tools [R]. Nice, France: Informatics, Signals & Systems Lab (I3S), University of Nice-Sophia Antipolis, 2014

[56] Studer R, Benjamins V R, Fensel D. Knowledge engineering: Principles and methods [J]. Data & Knowledge Engineering, 1998, 25(1): 161-197

[57] Wong W, Liu Wei, Bennamoun M. Ontology learning from text: A look back and into the future [J]. ACM Computing Surveys, 2012, 44(4): 20123915468506

[58] Wu Wentao, Li Hongsong, Wang Haixun, et al. Probase: A probabilistic taxonomy for text understanding [C] //Proc of the 31st ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2012: 481-492

[59] Shi Shuming. Automatic and semi-automatic knowledge extraction [J]. Communications of the CCF, 2013, 9(8): 65-73 (in Chinese)
(史树明. 自动和半自动知识提取[J]. 中国计算机学会通讯, 2013, 9(8): 65-73)

[60] Harris Z S. Distributional structure [J]. Word, 1954, 10 (23): 146-162

[61] Zeng Yi, Wang Dongsheng, Zhang Tielin, et al. CASIA-KB: A multi-source chinese semantic knowledge base built from structured and unstructured Web data [G] //Semantic Technology. Berlin: Springer, 2014: 75-88

[62] Wang Zhigang, Li Juanzi, Li Shuangjie, et al. Cross-lingual knowledge validation based taxonomy derivation from heterogeneous online wikis [C] //Proc of the 28th Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2014: 180-186

[63] Wang Chi, Danilevsky M, Desai N, et al. A phrase mining framework for recursive construction of a topical hierarchy [C] //Proc of the 19th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mning. New York: ACM, 2013: 437-445

[64] Liu Xueqing, Song Yangqiu, Liu Shixia, et al. Automatic taxonomy construction from keywords [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2012: 1433-1441

[65] Lee T W, Lewicki M S, Girolami M, et al. Blind source separation of more sources than mixtures using overcomplete representations [J]. Signal Processing Letters, 1999, 6(4): 87-90

[66] Lu Shaoyuan, Hsu K H, Kuo Lijing. A semantic service match approach based on wordnet and SWRL rules [C] // Proc of the 10th IEEE Int Conf on E-Business Engineering. Piscataway, NJ: IEEE, 2013: 419-422

[67] Socher R, Chen Dandi, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion [C] // Proc of Neural Information Processing Systems. Nevada, USA: NIPS, 2013: 926-934

[68] Lao Ni, Mitchell T, 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. Stroudsburg, PA: ACL, 2011: 529-539

[69] Yang Li, Hu Shouren. Knowledge base inference and maintain system [J]. Joural of National University of Defense Technology, 1991, 13(2): 127-133 (in Chinese)
(杨莉, 胡守仁. 知识库推理和维护系统 (KBIMS)[J]. 国防科技大学学报, 1991, 13(2): 127-133)

[70] Lu Daoshe, Yang Shihan, Wu Jinzhao, et al. Interdisciplinary reasoning on description logic [J]. Journal of Application Research of Computers, 2013, 29(12): 4503-450 (in Chinese)
(卢道设, 杨世瀚, 吴尽昭, 等. 基于描述逻辑的组合知识库推理[J]. 计算机应用研究, 2013, 29(12): 4503-4506)

[71] Dong Xin, Gabrilovich E, Heitz G, et al. Knowledge vault: A Web-scale approach to probabilistic knowledge fusion [C] //Proc of the 20th Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 601-610

[72] Tan C H, Agichtein E, Ipeirotis P, et al. Trust, but verify: Predicting contribution quality for knowledge base construction and curation [C] //Proc of the 7th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2014: 553-562

[73] Wang Zhigang, Li Juanzi, Wang Zhichun, et al. XLore: A large-scale english-chinese bilingual knowledge graph [C] // Proc of the 12th Int Semantic Web Conf. New York: ACM, 2013: 121-124

[74] Nguyen T, Moreira V, Nguyen H, et al. Multilingual schema matching for wikipedia infoboxes [J]. The Proceedings of the VLDB Endowment, 2011, 5(2): 133-144

[75] Wang Zhigang, Li Zhixing, Li Juanzi, et al. Transfer learning based cross-lingual knowledge extraction for wikipedia [C] //Proc of the 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2013: 641-650

[76] Fu B, Brennan R, Declan O S. Cross-lingual ontology mapping and its use on the multilingual semantic Web [C] // Proc of the 1st Workshop on the Multilingual Semantic Web, at the 19th Int World Wide Web Conf (WWW 2010). Tilburg, Netherlands: CEUR-WS, 2010: 13-20

[77] Wang Zhichun, Li Juanzi, Wang Zhigang, et al. Cross-lingual knowledge linking across wiki knowledge bases [C] // Proc of the 21st Int Conf on World Wide Web. New York: ACM, 2012: 459-468

[78] Wang Zhichun, Li Juanzi, Tang Jie. Boosting cross-lingual knowledge linking via concept annotation [C] //Proc of the 23rd Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2013: 2733-2739

[79] Yao Xuchen, Benjamin V D. Information extraction over structured data: question answering with freebase [C] //Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2014: 956-966

[80] Berant J, Liang P. Semantic parsing via paraphrasing [C] // Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2014: 1415-1425

[81] Fader A, Zettlemoyer L, Etzioni O. Open question answering over curated and extracted knowledge bases [C] // Proc of the 20th ACM Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 1156-1165

[82] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs [C] //Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2013: 1533-1544

Liu Qiao, born in 1974. PhD, associate professor. Member of China Computer Federation. His main research interests include machine learning, data mining, natural language processing, and social network analysis.

Li Yang, born in 1990. Master, student member of China Computer Federation. His main research interests include knowledge graph, machine learning and natural language processing (kedashqs@163.com).

Duan Hong, born in 1974. Master, lecturer. His main research interests include machine learning and data mining, natural language processing, and social network analysis(dhpro@sina.com).

Liu Yao, born in 1978. PhD, lecturer. Member of China Computer Federation. Her main research interests include social network analysis, data mining, and network measurement(liuyao@uestc.edu.cn).

Qin Zhiguang, born in 1956. PhD, professor. Senior member of China Computer Federation. His main research interests include information security and mobile computing(qinzg@uestc.edu.cn).