计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200700010

知识图谱构建技术:分类、调查和未来方向

杭婷婷[1,2]　冯　钧[1]　陆佳民[1]
1 河海大学计算机与信息学院　南京 211100
2 无人机开发及数据应用安徽高校联合重点实验室　安徽 马鞍山 243031
(httsf@hhu.edu.cn)

摘　要　知识图谱的概念由谷歌于2012年提出,随后逐渐成为人工智能领域的一个研究热点,已在信息搜索、自动问答、决策分析等应用中发挥作用。虽然知识图谱在各领域展现出了巨大的潜力,但不难发现目前缺乏成熟的知识图谱构建平台,需要对知识图谱的构建体系进行研究,以满足不同的行业应用需求。文中以知识图谱构建为主线,首先介绍目前主流的通用知识图谱和领域知识图谱,描述两者在构建过程中的区别;然后,分类讨论图谱构建过程中存在的问题和挑战,并针对这些问题和挑战,分类描述目前图谱构建过程中的知识抽取、知识表示、知识融合、知识推理、知识存储5个层面的解决方法和策略;最后,展望未来可能的研究方向。

关键词:知识图谱;知识抽取;知识表示;知识融合;知识推理;知识存储
中图法分类号　TP391.1

Knowledge Graph Construction Techniques:Taxonomy,Survey and Future Directions

HANG Ting-ting[1,2],FENG Jun[1] and LU Jia-min[1]
1 School of Computer and Information College,Hohai University,Nanjing 211100,China
2 Key Laboratory of Unmanned Aerial Vehicle Development and Data Application of Anhui Higher Education Institutes,Maanshan,Anhui 243031,China

Abstract　With the concept of knowledge graph proposed by Google in 2012,it has gradually become a research hotspot in the field of artificial intelligence and played a role in applications such as information retrieval,question answering,and decision analysis. While the knowledge graph shows its potential in various fields,it is easy to find that there is no mature knowledge graph construction platform currently. Therefore,it is essential to research the knowledge graph construction system to meet the application needs of different industries. This paper focuses on the construction of the knowledge graph. Firstly,it introduces the current mainstream general knowledge graphs and domain knowledge graphs and describes the differences between the two in the construction process. Then,it discusses the problems and challenges in the construction of the knowledge graph according to various types. To address the above-mentioned issues and challenges,it describes the five-level solution methods and strategies of knowledge extraction,knowledge representation,knowledge fusion,knowledge reasoning,and knowledge storage in the current graph construction process. Finally,it discusses the possible directions for future research on the knowledge graph and its application.

Keywords　Knowledge graph,Knowledge extraction,Knowledge representation,Knowledge fusion,Knowledge reasoning,Knowledge storage

1　引言

知识图谱的概念是由谷歌公司于2012年5月17日首次提出的,最早被应用于信息搜索领域[1]。知识图谱可以形式化定义为:$G=\{E,R,F\}$,其中E,R和F分别是实体集合、关系集合和事实集合。事实可以表示为$(h,r,t)\in F$[2]。知识图谱通过对数据的整合与规范,向人们提供有价值的结构化信息,已被广泛应用于信息搜索、自动问答、决策分析等领域,是推动数据价值挖掘和支撑智能信息服务的重要基础技术。随着社会和企业对知识图谱构建的需求逐渐加大,近几年研究

到稿日期:2020-07-02　返修日期:2020-10-29　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划(2018YFC0407901);安徽省高等学校自然科学研究重点项目(KJ2019A1277);江苏省研究生科研创新计划(2019B64214)
This work was supported by the National Key R&D Program of China(2018YFC0407901),University Natural Science Research of Anhui(KJ2019A1277) and Graduate Research Innovation Support Program of Jiangsu(2019B64214).
通信作者:冯钧(fengjun@hhu.edu.cn)

人员将人工智能、深度神经网络、自然语言处理与数据库技术相结合，在对知识图谱构建的研究方面取得了一些成果。然而，在高效构建知识图谱的过程中，不难发现目前缺乏成熟的知识图谱构建平台，需要对知识图谱的构建体系进行研究和分析，以满足不同行业的应用需求。那么，如何高效地构建知识图谱，成为了人工智能领域的一个重要的研究课题。

随着知识图谱的普及和应用，其构建过程受到了极大的关注，目前已取得了一些研究成果。文献[3]在全面阐述知识图谱定义和架构的基础上，对知识图谱中的知识抽取、知识表示、知识融合和知识推理四大核心技术的研究进展进行了介绍。文献[4]对信息抽取层、知识融合层和知识加工层所涉及的关键技术的研究现状进行了分类说明。文献[5]定义了知识图谱与本体的关系，并简述了已开发的国内外知识图谱。然而，上述研究都没有对知识图谱的构建体系进行系统性的总结和未来研究方向的展望。因此，本文依次从知识抽取、知识表示、知识融合、知识推理和知识存储5个层面对现有的知识图谱构建技术的研究成果进行了归纳和分析，各部分内容之间的总体路线图如图1所示。本文第2节以知识图谱构建过程为主线，对通用知识图谱和领域知识图谱进行了比较分析；第3节阐述了知识图谱构建过程中面临的问题与挑战；第4节描述了目前知识图谱构建过程中5个层面的解决方法和策略；最后对知识图谱构建的未来研究方向进行了展望。

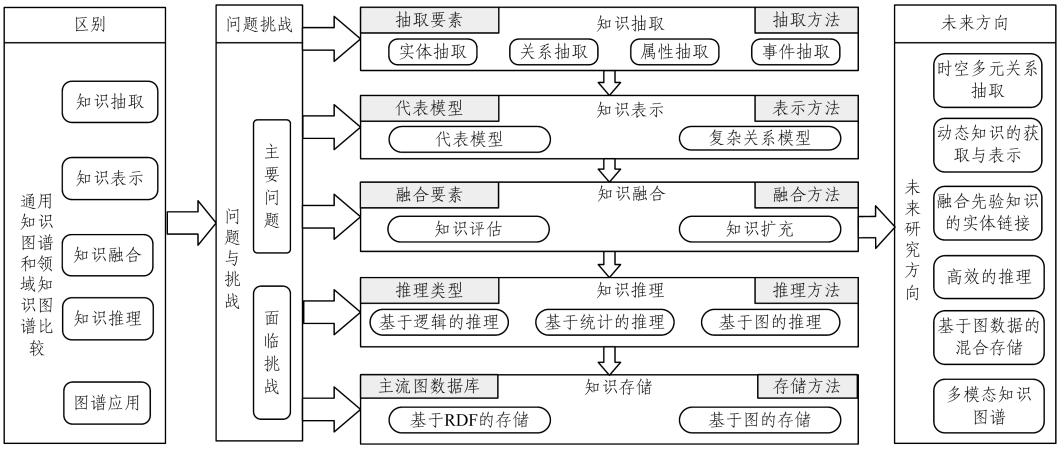

图1 本文各部分内容的总体路线图

Fig. 1 Roadmap of contents of this survey

2 通用知识图谱和领域知识图谱的区别

目前，国内外多个研究机构建立了一些大型通用知识图谱。在国内，代表性的通用知识图谱包括搜狗知立方[6]、百度知心[7]、Zhishi. me[8]、OpenKN[9]、CN-DBpedia[10]等。在国外，代表性的通用知识图谱包括WordNet[11]、DBpedia[12]、Freebase[13]、YAGO[14-15]、Probase[16]、Knowledge vault[17]等。上述通用知识图谱的比较如表1所列。通过对现有研究成果的整理和比较可以发现，目前知识图谱的概念规模仍处于发展变化阶段。

表1 通用知识图谱的比较

Table 1 Comparison of general knowledge graph

通用知识图谱	数据源	规模	语言	构建特点
WordNet	普林斯顿大学认知实验室	WordNet3.0已经包含15万个词和20万个语义关系	英文	主要用于词义消歧，构建名词、动词、形容词和副词之间的关系
DBpedia	Wikipedia	2800万个实体，30亿RDF三元组	127种语言	从多种语言的维基百科中抽取信息，并将其以关联数据的形式发布到互联网。它构建一个包含人、地点、音乐、电影、组织机构、物种、疾病等定义的本体
Freebase	Wikipedia, NNDB, MusicBrainz	19亿条三元组	英文	主要采用社区成员协作方式来构建，构建过程自顶向下
YAGO	Wikipedia, WordNet, GeoNames	1000万个实体，1.2亿条三元组知识	10种语言	构建一个更加丰富的实体分类体系，为很多条目增加时间和空间维度的属性描述
Probase	Wikipedia, Web Open Data	1000万个概念，1600万种isA关系	中文、英文	微软公司发布的基于概率化构建的知识库，是包含概念最多的知识库
Knowledge vault	Wikipedia, Freebase, Web Open Data	4500万个实体，4469种关系，2.7亿条三元组	多语言	通过算法自动搜集网上的信息，通过机器学习把数据变成可用知识
知立方	Web Open Data	整合海量的互联网碎片化信息	中文	国内首家知识库搜索产品，引入语义理解技术，理解用户搜索意图
知心	User Generated Content	已有教育、医疗、游戏等多个知识集群	中文	百度公司构建的中文知识图谱，侧重于深度搜索和实体推荐
Zhishi. me	Chinese Encyclopedia	1400多万个实体来自于百度百科，550多万个实体来自于互动百科，90多万个实体来自于中文维基	中文	从开放的百科数据中抽取结构化数据，提供一个SPARQL终端供用户查询
OpenKN	网页、百科、核心词汇	3000万个实体，10亿条边	中文	通过网络不断获取新知识，不仅更新自身包含的知识，而且存储其他的开放知识库中有用的知识
CN-DBpedia	Chinese Encyclopedia	1600多万个实体，2亿多个关系数，产生10多亿次API调用	中文	从纯文本页面中提取信息，经过滤、融合、推断等操作后，最终形成高质量的结构化数据

近年来,在一些领域已经出现面向领域的知识图谱,包括电影领域的 IMDB[18]、生物医学领域的 DrugBank[19]、新闻领域的 ECKG[20]、学术领域的 Acemap[21]等。上述领域知识图谱的比较如表 2 所列。

表 2 领域知识图谱的比较

Table 2 Comparison of domain knowledge graph

领域知识图谱	种类	规模	内容	特点
IMDB	互联网电影资料库	共收集 210 多万部作品资料和 450 多万部人物资料	包含电影、电影演员、电视节目、明星和电子游戏等信息	资料库中的资料按照电影类型组织,每一个具体的条目包含详细的元信息
DrugBank	生物信息学和化学信息学知识库	截至 2020 年 4 月 22 日涵盖 13580 种药物条目,其中包含小分子药品、生物制剂、营养药品和实验药品	每个药品条目中包含详细的药品信息和药品所治疗的疾病	已被广泛用于药物目标发现、药物设计、药物筛选、药物作用预测等方面
ECKG	以事件为中心的知识图谱	捕获数十万个实体的发展与历史,并与传统知识图谱中的百科全书信息结合	维基新闻、FIFA 世界杯、全球汽车工业和空客 A380 飞机	通过以事件为中心的浏览器和可视化工具促进新闻故事情节的重建,对新闻隐藏事实进行调查
Acemap	学术知识图谱	描述超过 1 亿个学术实体、22 亿条三元组信息,提供数据集近 100GB	包含 6000 多万篇论文、5000 多万位学者、5 万多个研究领域、2 万多个学术研究机构	提供学术异构图谱,包含多样的学术实体和属性

通用知识图谱和领域知识图谱的区别主要体现在知识建模与覆盖范围上。

(1)通用知识图谱面向通用领域,以常识性知识为主,其构建过程高度自动化,通常采用自底向上的方式来构建。其关联的知识大多数是静态的、客观的、明确的三元组事实性知识。一般以互联网开放数据为基础,再逐步扩大数据规模。

(2)领域知识图谱面向某一特定领域,以行业数据为主,其构建过程是半自动化的,通常采用自顶向下和自底向上两种方式相结合的方式来构建。其关联的知识包含静态知识和动态知识。

此外,从图谱构建的具体子过程来看,通用知识图谱和领域知识图谱还存在以下区别。

(1)从知识抽取角度来看,通用知识图谱注重知识的广度,覆盖粗粒度的知识。其在实体抽取层面,关注更多的实体,准确度不高;在关系抽取层面,多采用面向开放域的关系抽取。领域知识图谱注重知识的深度,覆盖细粒度的知识。其在实体抽取层面,关注具有特定行业意义的领域数据,准确度高[3];在关系抽取层面,多采用预定义关系抽取。

(2)从知识表示角度来看,通用知识图谱将知识表示成多个互相关联的三元组。例如,(实体 1,关系,实体 2)或(实体,属性,属性值),各部分之间有明确的层次结构。领域知识图谱除了将知识表示为多个互相关联的三元组之外,还需要对专家经验知识、行业文本的语义信息进行表示。

(3)从知识融合角度来看,通用知识图谱对知识抽取的质量有一定容忍度,需要通过知识融合来提升数据质量。领域知识图谱从领域内部的结构化数据、半结构化数据、非结构化数据中抽取知识,并且有一定的人工审核校验机制来保证质量,需要通过知识融合来扩大数据层的规模。

(4)从知识推理角度来看,由于通用知识图谱的知识覆盖范围较宽,深度较浅,从而导致图谱上的推理路径相对较短。而领域知识图谱的知识相对密集,这就导致图谱上的推理路径相对较长。当然,也存在一些特殊情况,例如 DBpedia 具有丰富的推理规则,推理路径比某些只有少量推理规则的领域知识图谱长。另外,推理路径上的区别体现在上层本体和垂直本体的比较上。

(5)从图谱应用角度来看,通用知识图谱主要应用在信息搜索和自动问答方面。领域知识图谱的主要应用除了上述方面,还包括决策分析、业务管理等。

通过上述通用知识图谱和领域知识图谱的比较,可以发现两者在构建过程中存在很多区别。但是,在实际的工程实践中,两者之间也存在着较强的联系。例如,构建领域知识图谱需要借鉴通用知识图谱的方法,需要引入通用知识库进行知识的融合。但是,全盘接收通用知识图谱中的数据,会引入大量领域不相关的信息,影响领域知识利用的效果。因此,可以利用通用知识图谱的广度结合领域知识图谱的深度,形成更加完善的知识图谱。

3 问题与挑战

通过上述两种知识图谱在构建过程中各个环节的对比,结合当前的研究热点与趋势,下文主要对知识抽取、知识表示、知识融合、知识推理和知识存储这 5 个层面的问题和挑战进行分析。

(1)知识抽取主要研究如何从多源异构数据中,自动或半自动地抽取出实体、属性、关系和事件等知识要素。近几年,主流的技术多是基于神经网络的抽取模型,这些模型在样本丰富的情况下抽取效果较好。但在小样本、零样本以及面向开放领域的环境下,如何保证抽取知识的准确率和覆盖率,尚需研究者去探索。

(2)知识表示主要研究如何对抽取出来的知识进行合理表示。在实际应用场景中,如何对多元组信息、稀疏知识以及动态知识进行表示,是知识表示需要解决的问题。

(3)知识融合主要研究如何对从不同来源抽取到的知识进行冲突检测和一致性检查,将验证正确的知识通过对齐、合并计算组织成知识库[22]。知识融合仍有大量问题需要研究:如何评估具有动态演化性的动态时序知识;当知识库的规模进一步增加时,其数据结构和数据特征也很复杂,那么如何对海量知识进行高效融合;如何对新增知识进行实时融合以及如何进行多语言融合。上述问题仍有广阔的研究空间。

(4)知识推理主要研究如何挖掘隐含的知识或识别出图谱中的错误知识,扩充、纠正已有知识图谱。近年来,国内外涌现出了很多基于神经网络的推理,并取得了一定进展。但是,神经网络具备黑盒特性,该技术的可解释性问题仍未得到

有效解决。另外,随着数据增长速度的加快,当数据规模增大到目前基于内存的服务器无法处理时,如何提升推理的效率和扩展性,如何保证一定时空及讨论域内知识图谱数据的完整性和正确性,仍是值得研究的问题。

(5)知识存储主要研究如何管理知识图谱中的数据,使其满足用户的查询、推理及各种应用需求。目前,没有一个统一的可以实现所有类型知识存储的方式,需要根据自身知识的特点,选择合适的存储方案来满足快速推理与图计算等应用的需要,这是知识存储过程中面临的一个挑战。

针对以上5个层面的问题与挑战,下文分别对现有研究成果中提出的解决方案进行了归纳和分析。

4 知识图谱的构建过程

4.1 知识抽取

知识抽取任务根据抽取要素的不同,可以分为实体抽取、属性抽取、关系抽取和事件抽取。

4.1.1 实体抽取

实体抽取指从文本语料中自动识别出实体。实体抽取方法可以分为:基于规则与词典的方法、基于统计机器学习的方法和面向开放域的抽取方法。

(1)基于规则与词典的方法

基于规则与词典的方法利用人工构建的规则与词典,面向单一领域从文本中识别出实体信息[23]。Raul[24]采用基于启发式算法和规则模板相结合的方法,实现了一套能从财经新闻中抽取公司名称的实体抽取系统。Chinchor等[25]使用已定义的规则,抽取出文本中的人名、地名、组织机构名、时间等实体。上述方法依靠领域专家编写规则模板,抽取出来的实体准确率较高,但是覆盖的范围有限,很难适应数据变化的新需求。

(2)基于统计机器学习的方法

基于统计机器学习的方法利用机器学习中的监督学习方法训练模型以识别实体。Lin等[26]提出了一种混合方法,将最大熵(Maximum Entropy,ME)和基于规则与词典的方法相结合,实现对Medline论文摘要的GENIA数据集上的实体进行抽取。该方法简单地使用ME进行命名实体识别,可能发生命名实体的不准确检测和错误分类。Liu等[27]在半监督学习框架下将K-最近邻(K-Nearest Neighbor,KNN)与线性条件随机场(Conditional Random Field,CRF)相结合,实现了对Twitter文本数据中实体的识别。上述方法需要足够数量的训练数据来获得良好的结果,当训练数据规模较小时,这些方法的有效性会受到影响[23]。

(3)面向开放域的抽取方法

面向开放域的实体抽取方法面向全网信息进行抽取。Whitelaw等[28]提出一种迭代扩展实体语料库的方法,该方法不仅可以识别地点和人物等类别,还可以识别更精细的类别。Jain等[29]采用无监督学习算法进行开放域实体的提取,并聚类查询日志,其实验结果优于现有的基于Web文档或查询日志的监督、半监督系统。对于上述方法,需要建立一个科学完整的命名实体分类体系[23]。

4.1.2 关系抽取

关系抽取指提取出实体间的关系或者实体与属性值之间的关系[4]。属性抽取也属于关系抽取的一部分,可以看作实体与属性值之间的一种关系,因此本文不对其单独进行分析。关系抽取方法可以分为:基于规则与模板的方法、面向开放域关系抽取、基于传统机器学习的方法和基于深度学习的方法。现有的主流的关系抽取技术是基于传统机器学习的方法和基于深度学习的方法。

(1)基于传统机器学习的方法

基于传统机器学习的方法可以分为:有监督的关系抽取、半监督的关系抽取和无监督的关系抽取。

1)有监督的关系抽取

有监督的关系抽取利用经过标注的样本数据集进行学习,形成了对关系的抽取。目前应用得较为广泛的方法包括基于特征向量的关系抽取和基于核函数的关系抽取。

基于特征向量的关系抽取主要从关系实例中提取一系列特征向量,其中包括词汇、句法和语义特征,并将上述特征用向量表示,再通过这些特征向量来训练关系分类器模型[30]。因此,如何选择具有代表性的特征来有效地反映待抽取关系,是基于特征向量的关系抽取方法的研究重点。Huang等[31]抽取词法、句法、实体3个层面的特征,再将训练语料向量化,训练支持向量机(Support Vector Machine,SVM)分类器来实现关系抽取。Suresh等[32]在TREC-QA2008评测语料库中,通过提取训练语料的语义、句法等特征,使用朴素贝叶斯分类器从非结构化文本中提取概念关系。上述方法比较简单且容易实现,但是存在特征维度过高、召回率下降的问题。

基于核函数的关系抽取主要通过核函数来计算实体间的距离,距离接近的两个实体可以看作有关系的实体对。Zelenko等[33]将设计的内核与SVM相结合,用于从文本中提取人员关系和组织位置关系。Bunescu等[34]基于依赖图中两个关系实体之间的最短路径,提出了一个用于关系提取的新内核。从ACE语料中提取关系的对比实验结果表明,新的最短路径依赖性内核优于基于依赖树的内核。

2)半监督的关系抽取

半监督的关系抽取利用少量标注数据进行学习[30]。Brin[35]考虑利用模式和关系集的二元性,自动从万维网中提取(作者,标题)对的关系。Agichtein等[36]提出Snowball系统,该系统在提取过程的每次迭代中,可以在没有人干预的情况下评估这些模式和元组的质量。

3)无监督的关系抽取

无监督的关系抽取利用未标记的样本数据集进行训练。Hasegawa等[37]提出一种基于大型语料库的无监督的关系发现方法,该方法根据命名实体的上下文词的相似性来聚类命名实体对。该方法不仅检测到了高召回率和高精确度的命名实体之间的关系,而且自动为关系提供合适的标签。Bollegala[38]提出一种联合聚类(Co-clustering)算法,该算法可以从未标记的数据中有效地提取大量关系。实验结果表明,该方法在精度和召回率方面优于开放域关系提取系统。

(2)基于深度学习的方法

随着近年来深度学习的崛起,深度学习逐渐被应用到实

体关系抽取任务中[39]。基于数据集标注量级的差异,基于深度学习的方法可以分为:有监督的关系抽取和远程监督的关系抽取。

1)有监督的关系抽取

有监督的实体关系抽取依靠人工标注的方法得到数据集,数据集的准确率、纯度较高,训练出的关系抽取模型的效果较好。目前应用得较为广泛的方法包括流水线方法和基于联合学习方法。

流水线学习方法指在实体识别已经完成的基础上直接进行实体之间关系的抽取。Socher 等[40]于 2012 年首次提出基于递归神经网络(Recurrent neural network,RNN)模型的关系抽取方法。Zeng 等[41]于 2014 年首次提出使用卷积神经网络(Convolutional Neural Network,CNN)进行关系抽取。Xu 等[42]于 2015 年提出基于长短期记忆(Long Short-Term Memory,LSTM)网络的方法进行关系抽取,该方法以句法依存分析树的最短路径为基础,融合单词特征、词性特征、语法关系特征、WordNet 上位词等,使用最大池化层和 Softmax 层进行关系分类。Zhang 等[43]提出一种基于修剪依存树的图卷积神经网络(Graph Convolutional Network,GCN)用于实体关系抽取。上述方法存在错误累积传播、忽视子任务间关系依赖、产生冗余实体等问题。

联合学习方法主要是基于神经网络的端到端模型,同时完成实体的识别和实体间关系的抽取。Miwa 等[44]于 2016 年首次利用循环神经网络、词序列以及依存树,将命名实体识别和关系抽取作为一个任务进行实验。Zheng 等[45]提出一种混合神经网络进行联合抽取。随后,注意力机制被引入联合学习任务中。Katiyar 等[46]提出将注意力机制与 Bi-LSTM 联合,进行实体识别和关系分类,从而获得更丰富的上下文信息。这些方法在实现实体关系抽取的过程中,仍将实体识别和关系分类两个子任务分开完成,仍然会产生没有关系的实体对这种冗余信息。为此,Zheng 等[47]于 2017 年提出一种新的标注方案和一个具有偏置目标函数的端到端模型来共同抽取实体及实体之间的关系。

2)远程监督的关系抽取

远程监督采用远程知识库对齐的方式自动标注数据,极大地减小了人力成本且领域迁移性较强。但远程监督自动标注得到的数据存在大量噪声,并且错误标签的误差会逐层传播,最终影响整个模型的效果[48]。针对错误标签这一问题,Surdeanu 等[49]提出多实例多标签学习方法。Lin 等[50]采用注意力机制进一步减弱了错误标注的示例语句产生的噪声。Ji 等[51]提出添加实体的描述信息来辅助学习实体的表示。为了增强关系提取效果,Ren 等[52]提出了 CoType 模型,其将实体信息、关系信息、文本特征和类型标签共同嵌入到有意义的表示中。Huang 等[53]提出深度残差网络来解决深层网络增大噪声的问题。

4.1.3 事件抽取

事件抽取指从非结构化信息中抽取出事件信息,并以结构化形式呈现给用户[54]。在事件抽取的过程中,一个事件往往被更形式化地定义为事件触发器、事件类型、事件元素和事件元素角色,因此事件抽取就是识别出上述事件要素并进行结构化组织。事件抽取方法可以分为:基于模式匹配的方法、基于机器学习的方法和混合事件抽取方法。

(1)基于模式匹配的方法

基于模式匹配的方法是将待抽取的句子与已有的事件模板进行匹配。目前应用得较为广泛的方法包括基于词汇句法模式和基于词汇语义模式。

1)基于词汇句法模式

词汇句法模式利用词汇特征和句法特征,从文本中自动获取词汇之间的关系[55]。Nishihara 等[56]提出一个支持系统,使用词汇句法模式分割句子,从博客中提取描述事件的 3 个关键词(地点、对象和动作),这 3 个关键词引导人们撰写关于个人经历的博客事件。该支持系统使用与提取的关键词相关的 3 张图片来表示事件,这些图片可以帮助用户判断是否将个人经历写在博客中。Xu 等[57]提出一种通过学习提及事件的模式来自动检测自然语言文本中的事件的方法,可以将事件类型解释为关系,用词汇句法模式初始化种子事件实例,实现自动识别事件范围和事件。上述方法由于缺乏模式表达性,因此没有合理利用具有特定含义的概念和关系。

2)基于词汇语义模式

词汇语义模式除了包含词汇表征信息,还包含语义信息[58]。Cohen 等[59]考虑到领域概念的语义,将事件检测、参数识别、否定和推测检测等问题作为概念识别和分析任务。在生物领域采用概念识别器,从语料库中提取医学事件。Capet 等[60]提出一种旨在为自动预警系统提取事件的方法,以尽早发现新出现风险的信号。他们使用词汇语义模式进行概念匹配,使用特定域的词汇资源进行增强,以便自动分析识别事件类型。

(2)基于机器学习的方法

基于机器学习的方法将事件类别及事件元素的识别转换为分类问题[61]。目前应用得较为广泛的方法包括基于事件元素驱动、基于事件触发词驱动和基于事件实例驱动。

基于事件元素驱动的方法判断句子的词表示何种事件元素,并将这些元素进行分类识别。基于事件触发词驱动的方法判断句子的词是否是事件的触发词。基于事件实例驱动的方法将事件中的每个句子进行聚类。Zhao 等[62]对事件类别识别和事件元素识别两项关键技术进行了研究。其在事件类别识别阶段,将触发词扩展和二元分类方法相结合;在事件元素识别阶段,采用基于最大熵的多元分类。实验结果表明,该方法解决了训练过程中数据的不平衡问题和数据稀疏问题。Liu 等[63]利用建模实体和新闻文档作为加权无向二分图,从每日的网络新闻中提取关键实体和重要事件。Llorens 等[64]提出一种识别和分类 TimeML 事件的方法,该方法通过 CRF 模型进行语义角色标注,CRF 模型包含形态句法特征、额外的语义信息和语义角色标记。

(3)混合事件抽取方法

混合事件抽取方法是基于模式匹配的方法和基于机器学习的方法的结合。Jungermann 等[65]提出一个事件框架,将信息检索、机器学习、预处理技术相结合,用于命名实体识别,以便从大型文档集中提取事件。Piskorski 等[66]提出一种事件提取系统 NEXUS,重点研究了弱监督采集提取模式,它会自

动从在线新闻文章中提取与安全有关的事实。

4.2 知识表示

知识表示的有效性直接影响到了知识图谱构建的质量和效率。在知识表示的演化过程中,最主要的变化是从基于数理逻辑的知识表示过渡到基于向量空间学习的分布式知识表示。

基于数理逻辑的知识表示是以符号逻辑为基础的表示方法,这些方法易于表达显性、离散的知识,但在计算效率、数据稀疏性等方面存在着一些问题。基于向量空间学习的分布式知识表示将知识图谱中的实体和关系嵌入到低维连续的向量空间,并且在该向量空间中完成语义计算。这种表示方法可以有效地挖掘隐形知识,缓解数据稀疏性带来的问题,对知识库的构建、推理、融合以及应用具有重要意义[67]。

4.2.1 代表模型

按照模型提出的时间先后顺序,知识表示学习的代表模型主要包括距离模型、矩阵分解模型、单层神经网络模型、TransE 模型和双线性模型。

(1)距离模型

Bordes 等[68]提出一种基于神经网络架构的学习过程,将知识库中的实体和关系嵌入到连续向量空间中。在该空间中可以通过计算向量之间的距离来表示实体之间的相关度,距离越小说明两个实体的语义相关度越高,存在某种语义关系。这种嵌入式的表示学习方法可用于实现实体预测和信息检索。

(2)矩阵分解模型

Nickel 等[69]提出的 RESCAL 模型是一种关系潜在特征模型,该模型通过张量分解来考虑二元关系数据的固有结构,用于预测两个实体之间的关系。RESACL 模型将知识库中的三元组表示为一个张量。如果该三元组在知识图谱中存在,则张量中对应位置的元素置 1,否则置 0。

(3)单层神经网络模型

单层神经网络模型[70]通过一个标准的单层神经网络,采用非线性操作隐式连接实体向量,用于解决距离模型无法精准描述实体与关系的语义联系的问题。虽然这是对距离模型的改进,但单层神经网络的非线性操作只提供两个实体向量之间的弱相互作用,增加了计算开销,并且引入了更高的计算复杂度。

(4)TransE 模型

TransE 模型是由 Bordes 等[71]于 2013 年提出的,它在低维向量空间中嵌入实体和多关系数据,利用较少量的参数训练一个规范模型并将其扩展到大规模数据库。对于每一个三元组(h,r,t),将其关系 r 看作从头实体 h 到尾实体 t 的翻译。TransE 被成功应用在大规模数据集中,但无法表达复杂关系。

(5)双线性模型

双线性模型通过实体间关系的双线性变换来刻画实体和关系的语义联系。Yang 等[72]提出一个学习框架,其中实体是从神经网络中学习到的低维向量,关系是双线性或线性映射函数,该框架可以从双线性目标中学习嵌入表示关系语义。

4.2.2 复杂关系模型

复杂关系模型主要针对 1-N,N-1 和 N-N 这 3 种复杂关系,其代表包括 TransH[73]模型、TransR[74]模型、CTransR[74]模型、TransD[75]模型、TransG[76]模型。知识表示中的复杂关系模型如表 3 所列。

表 3 知识表示中的复杂关系分类模型分类汇总

Table 3 Summary of complex relation models in knowledge representation

模型	主要特点	优点	缺点
TransH	在 TransE 基础上增加关系超平面,保证一个实体在涉及不同的关系时有不同的表示	利用关系超平面增强模型的灵活性,解决 TransE 一对多、多对一、多对多建模的难题	在一个共同的语义空间中表示实体和关系
TransR	在两个不同的语言空间建模实体和关系	保证相同关系的头尾实体在嵌入空间中接近	关系关联的头尾实体共享相同的投影矩阵,未考虑头尾实体类型的差异
CTransR	在 TransR 基础上对每一组实体对学习一个关系向量	将关系细分成子关系,为每一个子关系学习向量表示	相比 TransR,模型的计算复杂度更高
TransD	分别定义头实体和尾实体在关系空间上的投影矩阵	解决原来 TransR 模型参数过多的问题	未考虑不同关系的复杂程度差异
TransG	一种关系对应多种语义,每一种语义可以用高斯分布表示	通过考虑关系的不同语义,形成多个高斯分布,具有较高的区分度	未考虑知识库中的关系和实体的语义本身的不确定性

然而,现有的复杂关系模型对所有的三元组都一视同仁,均将三元组视为关系三元组,不能有效实现实体到属性值的翻译。Lin 等[77]提出的表示学习模型 KR-EAR(Knowledge Representation model with Entities, Attributes and Relations),不仅可以表示实体以及实体之间的关系,还可以表示实体的属性,将实体、关系、属性映射到低维空间中,有效提高了知识表示的精度。

4.3 知识融合

经过知识抽取阶段,抽取来的知识可能存在冲突或重叠。因此,有必要应用知识融合技术对知识进行处理,提升知识图谱的质量,丰富知识的存量[78]。知识融合可以分为知识评估和知识扩充。

4.3.1 知识评估

知识评估是知识融合的第一步,对验证为正确的知识进行融合计算才有意义。知识评估算法可以分为:基于贝叶斯模型的方法、基于 D-S 证据理论的方法、基于模糊集理论的方法和基于图模型的方法。

(1)基于贝叶斯模型的方法

基于贝叶斯模型的方法在知识为真时的先验概率和从数据源观察到的条件概率都已知的情况下,求出知识为真的后验概率。后验概率最大时对应的知识就是我们要找的正确知识。然而,该方法需要满足如下条件:不同来源的知识之间的观测是相互独立的,而且这些知识的先验概率是可预知的。

(2)基于 D-S 证据理论的方法

D-S 证据理论的方法是融合不同观测结果的信任函数,得到基础概率分配后,再选择最大支持度的假设作为最优判断,从而选择认为正确的知识。该方法也存在一些问题:知识源冲突较大时,会产生相悖的结论,同时该方法的时间复杂度也会增大。

(3)基于模糊集理论的方法

模糊集理论的方法在 D-S 证据理论的基础上,进一步放宽了贝叶斯模型的限制条件[79-80]。目前应用得较为广泛的方法基于模糊积分的方法[81]。模糊积分是一个非线性函数,可以完成质量评估,找到置信度最高的知识作为正确的知识。然而,该方法需要凭经验设置知识的模糊规则,不适用于不同知识源类型的知识评估。

(4)基于图模型的方法

基于图模型的方法使用从其他类型的数据中获得的先验知识,为知识分配一个概率。简单来说,就是根据图上的一组现有的边,预测其他边存在的可能性。Lao 等[82]提出了一种基于路径的排序算法(Path Ranking Algorithm,PRA),该算法通过已有知识之间的关系预测它们之间可能产生的隐含知识,再与数据源中抽取到的知识进行对比,识别不同来源知识中可能的真值。

上述 4 种知识评估方法都在一定程度上提高了知识的可靠性和置信度。然而,这些评估都适用于静态知识。目前,对于具有动态演化性的知识,缺乏直接的评估方法。另外,由于缺乏对知识获取渠道和获取方式的建模,因此难以从不可靠的知识获取方式中区分不可靠的数据源[22]。

4.3.2 *知识扩充*

知识扩充指将验证正确的知识扩充到知识库的方法。受篇幅的限制,本文只考虑在知识库中存在与文本实体映射的实体,并且在知识库中存在与文本实体关系映射的实体关系。知识扩充方法可以分为:实体对齐、实体链接和关系对齐。

(1)实体对齐

实体对齐指在异构数据中判断两个实体是否指向同一对象,可以消除实体冲突、指向不明等不一致性问题。实体对齐算法可以分为:成对实体对齐和集体实体对齐。

1)成对实体对齐

成对实体对齐考虑实体及其属性的相似程度。目前应用得较为广泛的方法包括传统概率模型的实体对齐方法和基于机器学习的实体对齐方法。

传统概率模型的实体对齐方法是一种基于属性相似的成对比较的方法,该方法不考虑匹配实体之间的关系[83]。Newcombe 等[84]和 Fellegi 等[85]通过比较实体的特征和值,来判断实体是否表示相同的事物。Winkler 等[86]将待匹配属性值出现的频率代入属性相似性计算中,通过期望最大化(Expectation-maximization,EM)算法自动估计参数,结合字符串比较器度量属性的权重,并认为出现频度越高的属性对实体对的匹配贡献越低。

基于机器学习的实体对齐方法可以分为有监督学习、无监督学习两类,其主要的区别在于是否使用了有标注的数据。目前主流的有监督学习方法有决策树、SVM、集成学习和主动学习等。

决策树方法通过训练数据迭代生成一棵规则树,其内部节点为判断规则,叶子节点表示分类的结果。Elfeky 等[87]在 TAILOR 工具包中实现了一种 ID3 决策树算法,用户可以调整系统参数和插入工具来构建自己的实体对齐模型。实验结果表明,该算法的匹配效果高于传统的概率模型方法。

SVM 方法通过训练数据集在高维空间中产生一个分类超平面来对真实数据进行分类,使得匹配实体和不匹配实体的间距尽可能大。Christen 等[88]提出通过自动选择训练样例来训练一种迭代的 SVM 二元分类器,实验结果表明,其匹配效果远高于 TAILOR 中的混合算法。

集成学习是一种通过结合多种基本的实体对齐系统来提高对齐质量的方法。Chen 等[89]使用两种组合方法将多个实体对齐系统的结果与上下文特征相结合,形成统一的聚类决策模型。实验结果表明,集成学习框架在不同领域的应用上,实现了更高的匹配质量。

主动学习通过初始少量的训练数据集和人机交互算法迭代式地训练分类模型,不断提升分类效果。Sarawagi 等[90]构建的 ALIAS 系统,基于人机交互来完成实体记录链接和重复数据删除任务。系统通过 3 棵分类树构建一种集成式的分类模型,那些在 3 棵分类树模型下结果不同的比较向量则由用户手工指定,再将分类结果代入模型进行迭代。实验结果表明,主动学习显著减少了实现高精度所需要的实例数。

另外,当缺乏足够的训练数据时,可以通过无监督学习完成实体对齐。无监督学习利用聚类算法将相似实体聚集到同一类。Verykios 等[91]提出一种基于聚类的无监督学习匹配模型,其基本思想是基于交互训练,通过少量的已标记数据的匹配情况去推理聚类中所有数据的匹配情况。

2)集体实体对齐

集体实体对齐在成对实体对齐的基础上考虑实例之间的相互关系,用于计算实体相似度[77]。目前应用较为广泛的方法包括局部集体实体对齐和全局集体实体对齐。

局部集体实体对齐在计算实体相似性时将实体的关联实体属性纳入计算。全局集体实体对齐将实体之间的相互关联关系,作为实体相似性的计算依据,并依此来调整实体之间的相似度。Bhattacharya 等[92]提出一种将属性相似性和关系信息相结合的关系聚类算法。实验结果表明,当数据中存在模糊引用时,关系聚类算法的效果优于属性相似度。Lacoste-Julien 等[93]在 Bhattacharya 的基础上进一步提出了 SiGMa(Simple Greedy Matching)算法,用于将知识库与数百万个实体进行对齐。SiGMa 算法是一种迭代传播算法,它将大规模知识库实体对齐问题看作一个全局匹配评分目标函数的贪婪优化问题。Domingos 等[94]使用多关系链接方法,对所有候选对执行同时推断。通过基于条件随机场模型,允许使用图切割算法在多项式中找到最优解,同时使用投票的感知器算法学习参数。Wick 等[95]提出一种基于条件随机场的判别式层次模型,将实体划分为树结构,这些树形成了一个高度紧凑、信息丰富的结构,以实现实体的高效推理。实验结果表明,这种判别式层次模型相比成对实体对齐,对齐速度快出了几个数量级。

(2)实体链接

实体链接指将实体对齐后的实体与知识库中的实体进行

链接,以补充知识图谱的现有内容。实体链接方法可以分为:基于实体属性的实体链接、基于实体流行度的实体链接、基于上下文的实体链接和基于外部证据的实体链接。

1)基于实体属性的实体链接

基于实体属性的实体链接通过计算描述实体的属性的相似度来判断实体是否相同。早期的基于实体属性的实体链接的研究工作主要通过计算编辑距离、Jaccard 系数等方式来计算实体相似度和属性相似度。但是,在计算相似度时,经常存在语义异构的问题。为了解决上述问题,需要引入实体语义相似度来辅助实体相似度的计算。Chen 等[96]提出一种结合 WordNet 和模糊形式概念分析的方法来生成模糊本体,新的模糊本体比一般本体更为灵活,该方法将实体的字符串相似度和语义相似度进行加权平均,并将结果作为实体相似度的度量。上述方法在属性信息丰富、没有噪音的情况下是有效的。但是,从网络数据中获取的属性难以保证完全没有噪声。另外,上述方法假定所有属性的权重是相同的,忽视了有些属性可能更加典型。

2)基于实体流行度的实体链接

基于实体流行度的实体链接认为对于给定的实体指代,与其对应的链接实体可能是流行度最大的实体,即在线百科锚文本中出现频率最大的实体。Ratinov 等[97]认为维基百科的全面性使得其成为了消除歧义的主要数据来源。但是,这种方法可能忽视实体的上下文,没有考虑实体的歧义问题。

3)基于上下文的实体链接

基于上下文的实体链接通过计算给定实体上下文之间的相似性来判断两个实体是否为同一个实体。Bunescu 等[98]提出一种命名实体检测和消歧方法,利用在线百科对命名实体进行消歧。该方法通过实体指代的上下文与所指候选实体的维基百科页面的内容之间的余弦相似度来度量实体相似度。选择上下文相似度得分最高的候选实体作为实体指代对应的映射实体。Cucerzan[99]提出一个大规模的命名实体消歧系统,该系统可以分析任何 Web 页面或客户端文本文档,在大规模的实体提取和消歧系统中进行应用。上述方法弥补了基于实体流行度的缺点,但是必须具备一个约束条件,即两个用于比较的文本之间存在词重叠,这在实际情况下很难满足。另外,实体上下文信息可能出现稀疏或噪音,这也会影响实体链接的准确率。

4)基于外部证据的链接

目前应用得较为广泛的方法包括采用话题连贯性和借助在线百科的结构信息。Han 等[100]根据同一文本的话题连贯性提出一种推断,认为同一文本中的实体不是独立的,它们之间存在语义相关性。这种语义关系可以促进实体链接的准确率的提升。Shen 等[101]利用维基百科中嵌入的丰富语义知识,通过图表上的随机漫游完成实体链接任务。

(3)关系对齐

关系对齐指对两个实体之间可能存在的命名不同但含义相同的关系进行归类和融合[83]。关系对齐方法可以分为:基于语义的方法和基于嵌入学习的方法。

1)基于语义的方法

基于语义的方法通过对比关系词汇之间的语义相似度来验证其是否是同一种关系。目前应用得较为广泛的方法包括基于语义词典的方法和基于语料库的方法。

基于语义词典的方法通过计算两个词汇在词典上的距离,来度量词汇之间的语义相似度。最早的工作是由 Resnik[102]提出的基于两个词汇在分类树中最小公共祖先节点的信息度量相似度的方法。该方法为了计算每个词汇的信息量,需要在一个大规模的文本语料中获取词的共现信息。该方法的局限在于在分类体系中相同概念下的所有词汇的相似度是一样的。针对上述问题,Patwardhan 等[103]将 WordNet 的内容以及基于语料库的数据相结合,引入一种新的语义相关度量。上述方法虽然很简单,但是词典无法提供足够的词语覆盖度,原因在于这些词典大多数是基于人工方式构建的,来不及进行实时更新。因此,当词典中出现词语缺失时,上述方法无法有效工作。

基于语料库的方法利用语料库中词汇的上下文分布来计算语义相似度。Chen 等[104]提出一种 double checking 模型,将网络作为实时语料进行探索,利用 Web 搜索引擎返回的文本片段计算词之间的语义相似度。Skip-gram 模型是学习分布式向量的有效方法,可以捕获大量精确的句法和语义单词顺序。Mikolov 等[105-106]提出了基于连续词袋模型和 Skip-gram 模型的 word2vec 方法,将词表征为实数值向量,通过计算向量空间上的相似度来表示词汇语义上的相似度。上述方法可以有效改进语义相似度度量的准确性。然而,上述方法也面临一些局限性:首先这种度量方式存在偏差,这是因搜索引擎使用的索引和排序机制导致的;其次,有些搜索结果需要与搜索引擎进行交互,这导致通信开销和索引成本大大增加,无法适用于在线应用。

2)基于嵌入学习的方法

基于嵌入学习的方法在嵌入关系中进行实体的嵌入表示,利用这种表示表达实体之间的关系,并判断两个实体的关系是否表达同一种关系[22]。现有的嵌入学习方法将实体 h 和 t 映射到一个语义空间,通过打分函数度量 (h,r,t) 在嵌入空间中的合理性。r 在嵌入空间中通过 h 和 t 来表示。典型的基于嵌入学习的工作主要是考虑关系的表示学习,对描述实体关系的上下文、类型和时间信息等缺乏有效的表示建模。针对上述问题,目前有些研究工作已经利用 CNN 模型建立了结合上下文信息的关系嵌入学习模型,提升了嵌入学习方法的表示能力,提高了关系扩充的准确率。

Zeng 等[107]利用分段卷积神经网络和多实例学习进行远程监督关系提取。其使用分段最大池化层来自动学习特征,并结合多实例学习来解决错误的标签问题。Santos 等[108]提出一种排名分类模型 CR-CNN,使用单词嵌入作为输入要素,利用卷积神经网络来处理关系分类任务。使用新的成对排名损失函数,可以有效地减小人工分类的影响。

4.4 知识推理

知识推理方法可以分为基于逻辑的推理、基于统计的推理和基于图的推理。

4.4.1 基于逻辑的推理

基于逻辑的推理指通过已有规则,从已有图谱中推理出新的实体或关系,还可以对知识进行冲突检测,验证已有知识

的正确性。基于逻辑的推理可以分为:基于一阶谓词逻辑推理、基于描述逻辑推理和基于规则推理。

(1)基于一阶谓词逻辑推理

一阶谓词逻辑推理是以命题为基础进行的推理。Allen等[109]指出帧表示语言和语义网络语言是一阶谓词演算的句法变体。以一阶谓词演算的符号表示作为一种表示语言,有助于设计语义网络的检索器,从而赋予语义网路形式化语义。

(2)基于描述逻辑推理

描述逻辑推理是在命题逻辑与一阶谓词逻辑的基础上发展而来的,是一种基于对象的知识表示的形式化工具,描述逻辑也是万维网联盟(W3C)推荐的 Web 本体语言 OWL 的逻辑基础[110]。Horrocks 等[111]开发了 FaCT 系统,用于处理一个比较大的医疗术语本体 GALEN,其性能优于其他类似的推理机。

(3)基于规则推理

基于规则推理通过添加实体隐含关系进行推理。Lu 等[112]提出一种基于 WordNet、本体和 SWRL 规则匹配 Web 服务的机制。它可以使用语义网络技术来提高发现 Web 服务的质量和精度,并且请求服务者还可以使用 WordNet 服务获得更多同义词。

4.4.2 基于统计的推理

基于统计的推理指通过统计规律从知识图谱中学习到隐含的实体关系。基于统计的推理可以分为:基于实体关系的推理、基于类型的推理和模式归纳方法。

(1)基于实体关系的推理

基于实体关系的推理通过统计方法学习知识图谱中实例和实例之间的隐含关系。Nickel 等[69]提出的 RESACL 模型是一种关系潜在特征模型,用于预测两个实体之间的关系。该模型可以将知识库中的三元组表示为一个三阶张量。如果该三元组存在,则张量中对应位置的元素置 1,否则置 0。

(2)基于类型的推理

基于类型的推理学习知识图谱中实体和概念之间的 is-a 关系。Paulheim 等[113]提出基于启发式链接的类型推理机制 SDType,它可以处理噪声数据和错误数据,用于知识图谱的推理。但是,该机制无法做到跨数据集的类型推理。

(3)模式归纳方法

模式归纳方法主要学习概念之间的关系。目前应用得较为广泛的方法包括基于归纳逻辑编程(Inductive Logic Programming,ILP)的方法和基于关联规则挖掘(Association Rule Mining,ARM)的方法。

1)基于归纳逻辑编程的方法

ILP 结合机器学习和逻辑编程技术,使得人们可以从实例和背景知识中获得逻辑结论。Lehmann 等[114]提出一个用于学习描述逻辑和 OWL 的框架 DL-Learner,DL-Learner 包含几种学习算法,支持不同的 OWL 格式、推理器接口和学习问题。

2)基于关联规则挖掘的方法

ARM 从知识图谱的信息中找出规则,这些规则可以直接转换成本体中的公理[110]。Völker 等[115]介绍了两种学习类别不相交公理的方法 GoldMiner 和 LeDA。GoldMiner 可以看作无监督的自下而上的方法,它依赖于数据集中的实例,并使用关联规则来发现模式。LeDA 可以看作有监督的自上而下的方法,它的结果是基于类的逻辑和词汇描述。为了评估这两种方法,他们建立了基于 DBpedia 本体的类别不相交公理的标准,并提出了一种基于特征选择的转移学习方法,该方法大大提高了产生不相交公理的准确性。

4.4.3 基于图的推理

基于图的推理方式指利用从知识图谱中观察到的三元组的边的特征来预测一条可能存在的边[110]。基于图的推理可以分为:基于神经网络模型的推理和 Path Ranking 算法。

(1)基于神经网络模型的推理

基于神经网络模型的推理利用神经网络对知识图谱中的三元组进行建模,获取三元组中的向量表示,将其与预测元素的向量进行比较,神经网络的输出值就是最终得分。目前应用得较为广泛的方法包括基于语义的推理、基于结构的推理和基于辅助存储的推理。

1)基于语义的推理

基于语义的推理建立在挖掘和利用语义信息的基础上,挖掘信息之间存在着语义关联。基于语义的推理模型包括 DKRL[116]、ProjE[117]、MT-KGNN[118]、ConMask[119] 和 HNM[120]。

2)基于结构的推理

基于结构的推理利用知识图谱中三元组内部或相互之间的结构联系进行推理。Neelakantan 等[121]提出一种多跳关系推理方法,使用组合向量空间模型支持任意长度路径上的推理链。其通过 RNN 将连续多元路径上的多跳关系信息合为整体,推理出一个连接首尾实体“合并关系”的路径。Das 等[122]将内容丰富的符号逻辑推理与泛化能力较强的神经网络相结合,训练出了一个单一的高能力 RNN。该方法在多跳关系推理的基础上进行改进,综合考虑了多条路径上的关系和中间实体信息,使预测结果更加准确。

3)基于辅助存储的推理

基于辅助存储的推理利用共享记忆组件或外部存储矩阵来存储推理所需要的中间结果,便于获取隐含信息,提高推理效率。基于辅助存储的推理模型包括 DNC[123] 和 IRN[124]。

(2)Path Ranking 算法

Path Ranking 算法主要用于判断源节点和目标节点之间是否存在关系,其主要依据是能否找到一条从源节点到目标节点的路径,而节点可以看作知识图谱的实体,边可以看作知识图谱上实体之间的关系[125]。Gardner 等[126]探讨了使用随机游走推理方法进行知识图谱补全的方法,指出 PRA 就是以两实体间的连通路径作为特征,来学习目标关系的分类器,据此判断这两个实体是否属于目标关系。他们在 PRA 的基础上,定义了一个子图特征提取算法,可以提取到比两个节点之间的路径更丰富的特征。然而,Path Ranking 算法的计算复杂度较高,无法满足大规模知识图谱的应用需求[127]。

基于上述问题,Nickel 等[128]回顾了在大型知识图谱上如何进行知识推理的两种方法,一种是基于潜在特征模型的,另一种是基于图模型的。他们指出将两种方法相结合,可以降低计算成本,提高建模能力,并进一步探索了路径排序算法中不同的特征抽取和特征值计算策略对整体效率以及性能的影响。

4.5 知识存储

知识存储可以分为基于 RDF 的存储与基于图的存储。

4.5.1 基于 RDF 的存储

基于 RDF 的存储使用唯一的 URI 标识一个资源，以三元组的方式来存储数据。每个三元组可以表示为(s,p,o)，其中 s 是主语，p 是谓语，o 是宾语。(s,p,o)表示 s 与 o 之间具有关系 p，或者 s 的属性 p 取值为 o。三元组模式的查询、归并和连接都非常高效，数据易于发布和共享[129]。但是，因其自身索引方式的问题，更新维护的代价较大。各种 RDF 数据模式近几年的发展情况如图 2 所示，其中横坐标表示时间，竖轴表示流行度。可以看出，基于 RDF 的存储中数据库 MarkLogic 比较领先，是 RDF 存储领域中目前最流行的存储框架。

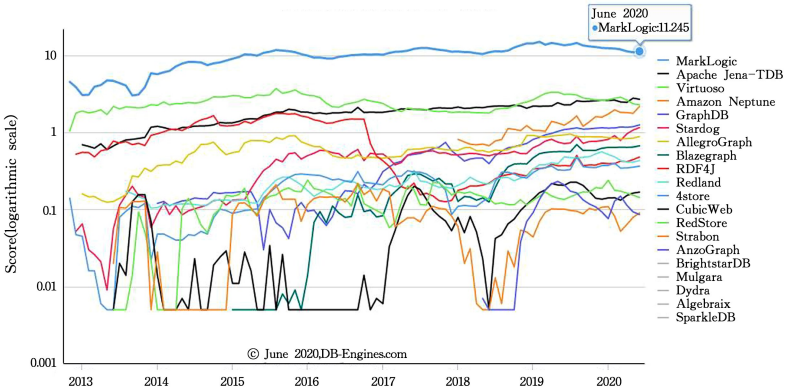

图 2 RDF 数据模式的排名[130]

Fig. 2 DB-Engines ranking of RDF stores[130]

4.5.2 基于图的存储

基于图的存储以属性图为基本的表示形式，图数据库可以提供完善的查询语言，有利于实现在图上的高效查询和搜索，更容易表达现实的业务场景。但是，图数据库的分布式存储使得数据更新缓慢，整体的实现代价高。各种图数据模式的发展情况如图 3 所示，其中横轴表示时间，竖轴表示流行度。可以看出，基于图的存储的数据库 Neo4j 比较领先，是图存储领域最流行的存储框架。目前，主流的图数据库有：gStore[132]，Virtuoso[133]，Stardog[134]，AllegroGraph[135]，Titan[136]，OrientDB[137]，Neo4j[138]等。常用的图数据库如表 4 所列。

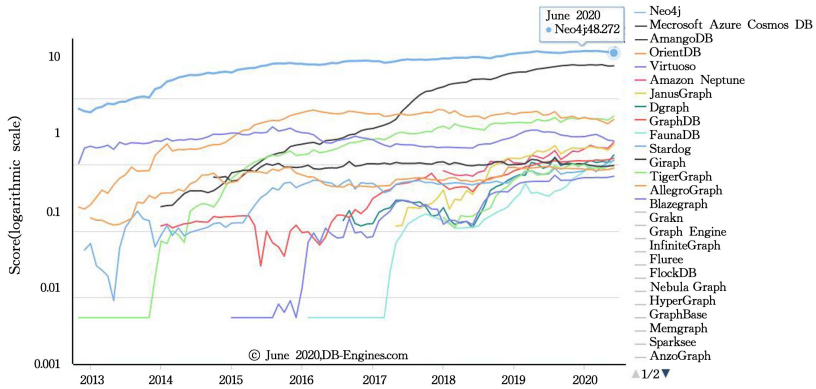

图 3 图数据模式的排名[131]

Fig. 3 DB-Engines ranking of graph stores[131]

表 4 图数据库汇总表

Table 4 Summary table of graph database

图数据库	种类	查询语言	底层存储	特点
gStore	开源图数据库	SPARQL	RDF 图/VS 树	可以支持 10 亿三元组规模的 RDF 知识图谱的数据管理任务，使用基于位串图存储
Virtuoso	商业图数据库	SPARQL/SQL	RDF 图/RDB 存储	支持多种数据模型的混合数据库管理系统，其基础源自于传统关系型数据库管理系统
Stardog	商业图数据集	SPARQL	属性图存储	能够良好地支持 OWL2 推理机制，能够支持不同编程语言和 Web 访问接口
Allgrograph	商业图数据库	SPARQL	RDF 图/三元组索引	支持永久存储的图数据库，具有强大的加载速度、查询速度和高性能
Titan	原生图数据库	SPARQL，Gremlin	Cassandra，HBase	在服务器集群搭建的分布式的图形数据库中，可以存储和处理大规模图形
OrientDB	原生图数据库	SQL/Gremlin	属性图/原生图存储	支持分布式、多模型的原生图数据管理系统，灵活支持各种数据模式
Neo4j	原生图数据库	Cypher	属性图/原生存储	高性能的 NoSQL 图形数据库，可以将结构化数据存储在网络上，系统本身的查询效率高，不支持分布式

4.5.3 基于RDF的存储和基于图的存储的区别

基于RDF的存储和基于图的存储,可以从图谱的规模、操作复杂度和模型的结构3个角度进行区分。

(1)从图谱的规模来看,小规模知识图谱可以采用基于RDF的存储;如果图谱达到数亿节点的规模,可以采用基于图的存储。

(2)从图谱的操作复杂度来看,如果图谱上的操作简单,则可以采用基于RDF的存储;如果图谱上的操作复杂,涉及到多步遍历,则可以采用基于图的存储。

(3)从模型的结构来看,基于RDF的存储表达能力强于基于图的存储。这主要是因为RDF中三元组的谓语可以在另一条三元组中做主语或谓语,其数据模型特性相对完善,但正因为理论性过强,影响了其在工业界的推广[139]。随着图数据库的应用,基于图的存储获得了较强的用户认可度[126]。

5 未来研究方向

根据上文对国内外研究现状的分析,本节给出了知识图谱构建在未来的研究方向。

(1)时空多元关系抽取。现阶段的知识抽取往往针对的是二元关系抽取,然而二元关系很难表达实体关系的时空特性。具有时空特性的多元关系抽取,是未来的研究方向之一。

(2)动态知识的获取与表示。在知识图谱构建的过程中,领域应用不仅要关联到静态知识,还需要包括一些非结构化的动态知识。需要研究动态知识获取与表示方法,将这些动态知识添加到知识图谱中。但是,依靠动态知识的添加来完全代替领域专家的工作,仍然是十分困难的,只能一定程度地降低对专家的依赖,实现简单的知识工作自动化。

(3)融合先验知识的实体链接。之前的统计模型已经证明先验知识对于实体链接任务的有效性,在深度学习模型中融合先验知识进行实体链接是提升现有深度模型的有效手段之一。

(4)高效的推理。随着通信宽带、GPU、内存等硬件性能的提高,以及多核、多处理技术的提出,采用共享内存模型和分布式推理技术来提升推理效率,这是突破大数据处理界限并实现高效推理的一种有效途径[110]。

(5)基于图数据的混合存储。相比传统的单一化的存储方式,混合存储能够针对知识图谱中不同的数据类型,选取合适的数据结构进行存储,保证数据的高效存储与管理。

(6)多模态知识图谱。多模态知识图谱将多模态知识(文本、视频、图片)进行整合,可以为用户提供多个不同维度的知识,还可以实现不同模态数据之间的跨模态交互[140]。其核心挑战在于,如何更好地利用模态内部信息和模态之间的交互信息,如何进行复杂的多模态关系的挖掘和推理,如何对多模态知识图谱进行增量更新。

结束语 本文以知识图谱构建为主线,对目前知识图谱构建关键技术的研究现状进行了全面调研和深入分析。首先介绍目前主流的两种知识图谱,并且描述两者在构建过程中的区别;其次介绍知识图谱构建工作面临的关键问题和重要挑战;然后分类描述知识抽取、知识表示、知识融合、知识推理、知识存储5个层面的解决方法和策略;最后,展望知识图谱构建过程的未来研究方向。本文将帮助研究人员在研究知识图谱构建平台时,理清知识图谱构建体系,针对知识图谱构建过程中的问题和挑战,选择适当的方法和算法,避免重复或冗余的工作。

参 考 文 献

[1] AMIT S. Introducing the knowledge graph:things,not things [EB/OL]. (2012-05-16) [2020-06-25]. https://blog.google/products/search/introducing-knowledge-graph-things-not/.

[2] JI S,PAN S,CAMBRIA E,et al. A survey on knowledge graphs:Representation,acquisition and applications[J]. arXiv:2002.00388,2020.

[3] XU Z L,SHENG Y P,HE L R,et al. Review on knowledge graph techniques[J]. Journal of University of Electronic Science and Technology of China,2016,45(4):589-606.

[4] LIU Q,LI Y,DUAN H,et al. Knowledge graph construction techniques[J]. Journal of Computer Research and Development,2016,53(3):582-600.

[5] HUANG H Q,YU J,MIAO X,et al. Review on knowledge graphs[J]. Journal of Computer System Application,2019,28(6):1-12.

[6] 搜狗知立方[DB/OL]. (2017-03-06) [2020-06-25]. http://baike.sogou.com/h66616234.htm.

[7] 百度知心[DB/OL]. (2013-20-23) [2020-06-25]. http://baike.baidu.com/view/10972128.htm.

[8] NIU X,SUN X,WANG H,et al. Zhishi.me-weaving Chinese linking open data[C]// Proceedings of the International Semantic Web Conference. 2011:205-220.

[9] JIA Y T,WANG Y Z,CHENG X Q,et al. OpenKN:an open knowledge computational engine for network big data[C]// Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2014:657-664.

[10] XU B,XU Y,LIANG J Q,et al. CN-DBpedia:A never-ending Chinese knowledge extraction system[C]// Proceedings of the International Conference on Industrial,Engineering and Other Applications of Applied Intelligent Systems. 2017:428-438.

[11] MILLER G A. WordNet:a lexical database for English [J]. Communications of the ACM,1995,38(11):39-41.

[12] AUER S,BIZER C,KOBILAROV G,et al. DBpedia:A nucleus for a web of open data[M]// The semantic web. Berlin:Springer,2007:722-735.

[13] BOLLACKER K,EVANS C,PARITOSH P,et al. Freebase:a collaboratively created graph database for structuring human knowledge[C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. 2008:1247-1250.

[14] 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.

[15] MAHDISOLTANI F,BIEGA J,SUCHANEK F M. YAGO3:A knowledge base from multilingual wikipedias[C]// Proceedings of the CIDR. 2013.

[16] WU W T,LI H S,WANG H X,et al. Probase:A probabilistic taxonomy for text understanding[C]// Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. 2012:481-492.

[17] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: A web-scale approach to probabilistic knowledge fusion [C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 601-610.

[18] IMDB Official. IMDB [EB/OL]. [2020-06-25]. http://www.imda.com.

[19] DrugBank Official. DrugBank [EB/OL]. [2020-06-25]. http://www.drugbank.ca.

[20] ROSPOCHER M, VAN ERP M, VOSSEN P, et al. Building event-centric knowledge graphs from news[J]. Journal of Web Semantics, 2016, 37: 132-151.

[21] AcemapOfficial. Acemap[EB/OL]. [2020-06-25]. https://www.acemap.info/.

[22] LIN H L, WANG Y Z, JIA Y T, et al. Network big data-oriented knowledge fusion methods: a survey[J]. Chinese Journal of Computer, 2016, 40(1): 3-29.

[23] SUN Z, WANG H L. Overview on the advance of the research on named entity recognition[J]. New Technology of Library and Information Service, 2010, 26(6): 42-47.

[24] RAUL L F. Extracting company names from text[C]// Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications. 1991: 29-32.

[25] CHINCHOR N, MARSH E. Muc-7 information extraction task definition[C]// Proceedings of the 7th Message Understanding Conference. 1998: 359-367.

[26] LIN Y F, TSAI T, CHOU W C, et al. A maximum entropy approach to biomedical named entity recognition[C]// Proceedings of the 4th International Conference on Data Mining in Bioinformatics. 2004: 56-61.

[27] LIU X H, ZHANG S D, WEI F R, et al. Recognizing named entities in tweets[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011: 359-367.

[28] WHITELAW C, KEHLENBECK A, PETROVIC N, et al. Web-scale named entity recognition[C]// Proceedings of the 17th ACM Conference on Information and Knowledge Management. 2008: 123-132.

[29] JAIN A, PENNACCHIOTTI M. Open entity extraction from web search query logs[C]// Proceedings of the 23rd International Conference on Computational Linguistics. 2010: 510-518.

[30] LI D M, ZHANG Y, LI D Y, et al. Review of entity relation extraction methods[J]. Journal of Computer Research and Development, 2020, 57(7): 1424-1448.

[31] HUANG X, ZHU Q M, QIAN L H. Chinese entity relation extraction based on features combination[J]. Microelectronics & Computer, 2010, 27(4): 198-200.

[32] SURESH KUMAR G, ZAYARAZ G. Concept relation extraction using naïve Bayes classifier for ontology-based question answering systems[J]. Journal of King Saud University-Computer and Information Sciences, 2015, 27(1): 13-24.

[33] ZELENKO D, AONE C, RICHARDELLA A. Kernel methods for relation extraction[J]. Journal of Machine Learning Research, 2003, 3(2): 1083-1106.

[34] BUNESCU R C, MOONEY R J. A shortest path dependency kernel for relation extraction[C]// Proceedings of the Conference on human language technology and empirical methods in natural language processing. 2005: 724-731.

[35] BRIN S. Extracting Patterns and relations from the world wide web[C]// Proceedings of the International Workshop on the World Wide Web and Databases. 1998: 172-183.

[36] AGICHTEIN E, GRAVANO L. Extracting relations from large plain-text collections[C]// Proceedings of the ACM Conference on Digital Libraries. 2000: 85-94.

[37] HASEGAWA T, SEKINE S, GRISHMAN R. Discovering relations among named entities from large corpora[C]// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. 2004: 415-422.

[38] BOLLEGALA D T, MATSUO Y, ISHIZUKA M. Relational duality: unsupervised extraction of semantic relations between entities on the web[C]// Proceedings of the International Conference on World Wide Web. 2010: 151-160.

[39] KUMAR S. A survey of deep learning methods for relation extraction[J]. arXiv preprint arXiv: 1705.03645, 2017.

[40] SOCHER R, HUVAL B, MANNING C D, et al. Semantic compositionality through recursive matrix-vector spaces[C]// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational natural Language Learning. 2012: 1201-1211.

[41] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network[C]// Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 2014: 2335-2344.

[42] XU Y, MOU L, LI G, et al. Classifying relations via long short-term memory networks along shortest dependency paths[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1785-1794.

[43] ZHANG Y, QI P, MANNING C. Graph convolution over pruned dependency trees improves relation extraction[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 2205-2215.

[44] MIWA M, BANSAL M. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 1105-1116.

[45] ZHENG S, HAO Y, LU D, et al. Joint entity and relation extraction based on a hybrid neural network[J]. Neurocomputing, 2017, 257: 59-66.

[46] KATIYAR A, CARDIE C. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 917-928.

[47] ZHENG S, WANG F, BAO H, et al. Joint extraction of entities and relations based on a novel tagging scheme[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1227-1236.

[48] E H H, ZHANG W J, XIAO S Q, et al. Survey of entity relationship extraction based on deep learning[J]. Journal of Software, 2019(6): 1793-1818.

[49] SURDEANU M, TIBSHIRANI J, NALLAPATI R, et al. Multi-instance multi-label learning for relation extraction[C]// Proceedings of the 2012 Joint Conference on Empirical Methods in

Natural Language Processing and Computational Natural Language Learning. 2012:455-465.

[50] LIN Y,SHEN S,LIU Z,et al. Neural relation extraction with selective attention over instances[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016:2124-2133.

[51] JI G L,LIU K,HE S Z,et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions[C]// Proceedings of the 31st Conference on Artificial Intelligence. 2017:3060-3066.

[52] REN X,WU Z,HE W,et al. CoType:Joint extraction of typed entities and relations with knowledge bases[C]// Proceedings of the 26th International Conference on World Wide Web. 2017:1015-1024.

[53] HUANG Y Y,WANG W Y. Deep residual learning for weakly-supervised relation extraction[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017:1803-1807.

[54] GUO X Y,HE T T. Survey about research on information extraction[J]. Computer Science,2015,42(2):14-17.

[55] HEARST M A. Automatic acquisition of hyponyms from large text corpora[C]// Proceedings of the 14th Conference on Computational Linguistics. 1992:539-545.

[56] NISHIHARA Y,SATO K,SUNAYAMA W. Event extraction and visualization for obtaining personal experiences from blogs [C]// Proceedings of the Symposium on Human Interface. 2009:315-324.

[57] XU F Y,USZKOREIT H,LI H. Automatic event and relation detection with seeds of varying complexity[C]// Proceedings of the AAAI Workshop on Event Extraction and Synthesis. 2006:12-17.

[58] BORSJE J,HOGENBOOM F,FRASINCAR F. Semi-automatic financial events discovery based on lexicon-semantic patterns [J]. International Journal of Web Engineering and Technology,2010,6(2):115-140.

[59] COHEN K B,VERSPOOR K,JOHNSON H L,et al. High-precision biological event extraction with a concept recognizer[C]// Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing:Shared Task. 2009:50-58.

[60] CAPET P,DELAVALLADE T,NAKAMURA T,et al. A risk assessment system with automatic extraction of event types [C]// Proceedings of the International Conference on Intelligent Information Processing. 2008:220-229.

[61] DING X,SONG F,QIN B,et al. Research on typical event extraction method in the field of music[J]. Journal of Chinese Information Processing,2011,25(2):15-20.

[62] ZHAO Y Y,QIN B,CHE W X,et al. Research on Chinese event extraction[J]. Journal of Chinese Information Processing,2008,22(1):3-8.

[63] LIU M,LIU Y,XIANG L,et al. Extracting key entities and significant events from online daily news[C]// Proceedings of the 9th International Conference on Intelligent Data Engineering and Automated Learning. 2008:201-209.

[64] LLORENS H,SAQUETE E,NAVARRO-COLORADO B. TimeML events recognition and classification: learning CRF models with semantic roles[C]// Proceedings of the 23rd International Conference on Computational Linguistics. 2010:725-733.

[65] JUNGERMANN F,MORIK K. Enhanced services for targeted information retrieval by event extraction and data mining[C]// Proceedings of the 13th International Conference on Application of Natural Language and Information Systems. 2008:335-336.

[66] PISKORSKI J,TANEV H,WENNERBERG P O. Extracting violent events from on-line news for ontology population[C]// Proceedings of the 10th International Conference on Business Information Systems. 2007:287-300.

[67] LIU Z Y,SUN M L,LIN Y K,et al. Knowledge representation learning:A review[J]. Journal of Computer Research and Development,2016,53(2):247-261.

[68] BORDES A,WESTON J,COLLOBERT R,et al. Learning structured embeddings for knowledge bases[C]// Proceedings of the 25th AAAI Conference on Artificial Intelligence. 2011:301-306.

[69] NICKEL M,TRESP V,KRIEGEL H. A three-way model for collective learning on multi-relational data[C]// Proceedings of ICML. 2011:809-816.

[70] SOCHER R,CHEN D,MANNING C D,et al. Reasoning with neural tensor networks for knowledge base completion[C]// Proceedings of the International Conference on Neural Information Processing Systems. 2013:926-934.

[71] BORDES A,USUNIER N,GARCIA-DURAN A. Translating embeddings for modeling multi-relational data[C]// Proceedings of the International Conference on Neural Information Processing Systems. 2013:2787-2795.

[72] YANG B,YIH W,HE X,et al. Embedding entities and relations for learning and inference in knowledge bases[C]// Proceedings of International Conference on Learning Representations. 2015.

[73] WANG Z,ZHANG J W,FENG J L,et al. Knowledge graph embedding by translating on hyperplanes[C]// Proceedings of the 28th AAAI Conference on Artificial Intelligence. 2014:1112-1119.

[74] LIN Y,LIU Z,SUN M,et al. Learning entity and relation embeddings for knowledge graph completion[C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence. 2015:2181-2187.

[75] JI G L,HE S Z,XU L H,et al. Knowledge graph embedding via dynamic mapping matrix[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015:687-696.

[76] XIAO H,HUANG M L,HAO Y,et al. TransG:a generative mixture model for knowledge graph embedding[C]// Proceedings of the Computation and Language. 2015.

[77] LIN Y K,LIU Z Y,SUN M S. Knowledge representation learning with entities attributes and relations[J]. IEEE Signal Processing Letters,2016,23(4):41-52.

[78] MA Z G,NI R Y,YU K H. Recent advances,key techniques and future challenges of knowledge graph[J]. Chinese Journal of Engineering,2020,42(10):1254-1266.

[79] Zadeh L A. Fuzzy sets[J]. Information and Control,1965,8(3):338-353.

[80] ABDULGHAFOUR M,CHANDRA T,ABIDI M A. Data fu-

sion through fuzzy logic applied to feature extraction from multisensory images[C]// Proceedings of the IEEE International Conference on Robotics and Automation. 1993:359-366.

[81] GRABISCH M, SUGENO M, MUROFUSHI T. Fuzzy measures and integrals: theory and applications[M]. Heidelberg: Physica, 2010.

[82] LAO N, MITCHELL T, COHEN W. Random walk inference and learning in a large-scale knowledge base[C]// Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011:529-539.

[83] ZHUANG Y, LI G L, FENG J H. A survey on entity alignment of knowledge base[J]. Journal of Computer Research and Development, 2016, 53(1):165-192.

[84] NEWCOMBE H B, KENNEDY J M, AXFORD S J, et al. Automatic linkage of vital records[J]. Science, 1959, 130(3381):954-959.

[85] FELLEGI I P, SUNTER A B. A theory for record linkage[J]. Journal of the American Statistical Association, 1969, 64(328):1183-1210.

[86] WINKLER W E, THIBAUDEAU Y. An application of the Fellegi-Sunter model of record linkage to the 1990 US decennial census[M]// Washington. DC: US Bureau of the Census, 1991.

[87] ELFEKY M G, VERYKIOS V S, ELMAGARMID A K. TAILOR: a record linkage toolbox[C]// Proceedings of the 18th International Conference on Data Engineering. 2002:17-28.

[88] CHRISTEN P. Automatic training example selection for scalable unsupervised record linkage[C]// Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2008:511-518.

[89] CHEN Z Q, KALASHNIKOV D V, MEHROTRA S. Exploiting context analysis for combining multiple entity resolution systems[C]// Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. 2009:207-218.

[90] SARAWAGI S, BHAMIDIPATY A. Interactive deduplication using active learning[C]// Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002:269-278.

[91] VERYKIOS V S, ELMAGARMID A K, HOUSTIS E N. Automating the approximate record-matching process[J]. Information Sciences, 2000, 126(1-4):83-98.

[92] BHATTACHARYA I, GETOOR L. Collective entity resolution in relational data[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1):5-40.

[93] LACOSTE-JULIEN S, PALLA K, DAVIES A, et al. Sigma: simple greedy matching for aligning large knowledge bases[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013:572-580.

[94] DOMINGOS P. Multi-relational record linkage [C]// Proceedings of the KDD-2004 Workshop on Multi-Relational Data Mining. 2004:31-48.

[95] WICK M, SINGH S, MCCALLUM A. A discriminative hierarchical model for fast coreference at large scale[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. 2012:379-388.

[96] CHEN R C, BAU C T, YEH C J. Merging domain ontologies based on the WordNet system and fuzzy formal concept analysis techniques[J]. Applied Soft Computing, 2011, 11(2):1908-1923.

[97] RATINOV L, ROTH D, DOWNEY D, et al. Local and global algorithms for disambiguation to Wikipedia[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011:1375-1384.

[98] BUNESCU R, PAŞCA M. Using encyclopedic knowledge for named entity disambiguation[C]// Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. 2006.

[99] CUCERZAN S. Large-scale named entity disambiguation based on Wikipedia data[C]// Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007:708-716.

[100] HAN X, SUN L, ZHAO J. Collective entity linking in web text: a graph-based method[C]// Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2011:765-774.

[101] SHEN W, WANG J, LUO P, et al. A graph-based approach for ontology population with named entities[C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. 2012:345-354.

[102] RESNIK P. Using information content to evaluate semantic similarity in a taxonomy[C]// Proceedings of the 14th International Joint Conference on Artificial Intelligence. 1995.

[103] PATWARDHAN S, PEDERSEN T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts [C]// Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together. 2006.

[104] CHEN H H, LIN M S, WEI Y C. Novel association measures using web search with double checking[C]// Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. 2006:1009-1016.

[105] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Proceedings of the Advances in Neural Information Processing Systems. 2013:3111-3119.

[106] MIKOLOV T, YIH W, ZWEIG G. Linguistic regularities in continuous space word representations [C]// Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013:746-751.

[107] ZENG D J, LIU K, CHEN Y B, et al. Distant supervision for relation extraction via piecewise convolutional neural networks [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015:1753-1762.

[108] SANTOS C N, XIANG B, ZHOU B. Classifying relations by ranking with convolutional neural networks[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. 2015.

[109] ALLEN J F, FRISCH A M. What's in a semantic network? [C]// Proceedings of the 20th Annual Meeting on Association for Computational Linguistics. 1982:19-27.

[110]QI G L，GAO H，WU T X. The research advances of knowledge grape[J]. Technology in Intelligence Engineering，2017，3(1)：4-25.

[111]HORROCKS I. Using an expressive description logic：FaCT or fiction? [J]. KR，1998，98：636-645.

[112]LU S Y，HSU K H，KUO L J. A semantic service match approach based on wordnet and SWRL rules[C]// Proceedings of the IEEE 10th International Conference on e-Business Engineering. 2013：419-422.

[113]PAULHEIM H，BIZER C. Type inference on noisy RDF data [C]// International Semantic Web-ISWC. 2013：510-525.

[114]LEHMANN J. Dl-learner：learning concepts in description logics [J]. Journal of Machine Learning Research，2009，10(6)：2639-2642.

[115]VÖLKER J，FLEISCHHACKER D，STUCKENSCHMIDT H. Automatic acquisition of class disjointness[J]. Journal of Web Semantics，2015，35：124-139.

[116]XIE R B，LIU Z，JIA J，et al. Representation learning of knowledge graphs with entity descriptions[C]// Proceedings of the 30th AAAI Conference on Artificial Intelligence. 2016：2659-2665.

[117]SHI B，WENINGER T. ProjE：Embedding projection for knowledge graph completion[C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence. 2017：1236-1242.

[118]TAY Y，TUAN L A，PHAN M C，et al. Multi-task neural network for non-discrete attribute prediction in knowledge graphs [C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017：1029-1038.

[119]SHI B，WENINGER T. Open-world knowledge graph completion[J]. Proceedings of the 30th AAAI Conference on Artificial Intelligence，2018：1957-1964.

[120]DENIS L，ASJA F，JENS L，et al. Neural network-based question answering over knowledge graphs on word and character level[C]// Proceedings of the International World Wide Web Conference Committee. 2017：1211-1220.

[121]NEELAKANTAN A，ROTH B，MCCALLUM A. Compositional vector space models for knowledge base completion[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. 2015：1-16.

[122]DAS R，NEELAKANTAN A，BELANGER D，et al. Chains of reasoning over entities，relations，and text using recurrent neural networks[C]// Proceedings of the European Chapter of the Association for Computational Linguistics. 2017.

[123]GRAVES A，WAYNE G，REYNOLDS M，et al. Hybrid computing using a neural network with dynamic external memory[J]. Nature，2016，538(7626)：471-476.

[124]SHEN Y，HUANG P S，CHANG M W，et al. Modeling large-scale structured relationships with shared memory for knowledge base completion[C]// Proceedings of the 2nd Workshop on Representation Learning for NLP. 2017：57-68.

[125]GUAN S P，JIN X L，JIA Y T，et al. Knowledge reasoning over knowledge graph：a survey [J]. Journal of Software，2018，29(10)：74-102.

[126]GARDNER M，MITCHELL T. Efficient and expressive knowledge base completion using subgraph feature extraction[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015：1488-1498.

[127]WANG S，DU Z J，MENG X F. Research progress of large-scale knowledge graph completion technology[J]. Scientia Sinica Informationis，2020，50(4)：551-575.

[128]NICKEL M，MURPHY K，TRESP V，et al. A review of relational machine learning for knowledge graphs[J]. Proceedings of the IEEE，2015，104(1)：11-33.

[129]WANG X，ZOU L，WANG C K，et al. Research on knowledge graph data management：a survey[J]. Journal of Software，2019，30：1-38.

[130]DB-Engines. DB-Engines Ranking of RDF Stores [EB/OL]. (2020-01) [2020-06-25]. https://db-engines.com/en/ranking_trend/rdf+store.

[131]DB-Engines. DB-Engines Ranking of Graph DBMS. [EB/OL]. (2020-01) [2020-06-25]. https://dbengines.com/en/ranking_trend/graph+dbms.

[132]ZOU L，ÖZSU M T，CHEN L. gStore：a graph-based SPARQL query engine[J]. The VLDB journal，2014，23(4)：565-590.

[133]OpenLink Software [EB/OL]. (2015-01-01) [2020-06-25]. OpenLink Virtuoso. https://virtuoso.openlinksw.com/.

[134]Stardog Union. Stardog - The Knowledge Graph Platform for the Enterprise [EB/OL]. (2017-02-03) [2020-06-25]. http://www.stardog.com/.

[135]Franz Inc. Allegro Graph [EB/OL]. (2018-05-07) [2020-06-25]. https://franz.com/agraph/allegrograph/.

[136]Spmallette. Titan-Distributed Graph Database [EB/OL]. (2017-01-01) [2020-06-25]. http://titan.thinkaurelius.com/.

[137]Callidus Software Inc. OrientDB - Multi-Model Database [EB/OL]. (2017-12-20) [2020-06-25]. http://orientdb.com/.

[138]The Neo4j Team. The Neo4j Manual v3.4[EB/OL]. (2018-05-16) [2020-06-25]. https://neo4j.com/docs/developer-manual/current/.

[139]HARTIG O. Reconciliation of RDF * and property graphs[J]. arXiv：1409.3288，2014.

[140]WANG K，HE R，WANG L，et al. Joint feature selection and subspace learning for cross-modal retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2015，38(10)：2010-2023.

HANG Ting-ting，born in 1986，Ph. D，lecturer，is a member of China Computer Federation. Her main research interests include domain knowledge graph construction and information extraction.

FENG Jun，born in 1969，Ph.D，professor，Ph.D supervisor，is a professional member of China Computer Federation. Her main research interests include data management，domain knowledge discovery research，and water conservancy informatization.