华南师范大学学报(自然科学版)
Journal of South China Normal University(Natural Science Edition)
2018,50(5):110-119
doi:10.6054/j.jscnun.2018107

知识图谱及其在学术信息服务领域的应用

汤　庸[1*]，陈国华[2]，贺超波[3]，彭　博[1]
(1. 华南师范大学计算机学院，广州 510631; 2. 华南师范大学网络中心，广州 510631;
3. 仲恺农业工程学院信息科学与技术学院，广州 510225)

摘要: 介绍了知识图谱产生的历史背景和典型的面向通用领域和学术信息服务领域的知识图谱;综述了知识图谱涉及的关键技术,包括知识抽取、知识融合以及知识加工;并以学者社交网络 SCHOLAT 为背景,给出学术知识图谱的模式设计,讨论知识图谱在学术信息服务领域中的应用意义.

关键词: 大数据; 人工智能; 知识图谱; 学术信息服务; 学者网
中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1000-5463(2018)05-0110-10

Knowledge Graph and its Application in Academic Information Service

TANG Yong[1*], CHEN Guohua[2], HE Chaobo[3], PENG Bo[1]
(1. School of Computer Science, South China Normal University, Guangzhou 510631, China;
2. Network Center, South China Normal University, Guangzhou 510631, China;
3. School of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China)

Abstract: The history of knowledge graph is introduced and several typical knowledge graphs for the general field and academic information services are introduced. The key technologies of knowledge graphs are surveyed, including knowledge extraction, knowledge integration and knowledge processing. Finally, the applications of knowledge graph in the field of academic information service, SCHOLAT, and its corresponding design pattern are presented.
Key words: big data; artificial intelligence; knowledge graph; academic information service; SCHOLAT

2012 年,谷歌在其产品线中推出了知识图谱(Knowledge Graph),以此构建了智能搜索引擎,在传统的基于关键字搜索的基础上,提供了语义理解,带来了全新的搜索体检. 除智能搜索外,知识图谱还可用于自动问答、情报分析和个性化推荐等多个领域,受到了工业界和学术界的广泛关注[1-3]. 虽然知识图谱与早期流行的语义网络和语义网具有密切的联系. 1956 年,RICHENS[4] 为解决机器翻译问题,提出了“语义网络”(Semantic Network)的概念,目的是将语义网络作为各种自然语言之间的中间语言,在各种不同的语言间进行转换. 语义网络利用“图”这一数据结构来存储知识,图中的结点表示实体或者概念,而“边”表示结点之间存在的语义联系. 语义网络提出后,成为建设知识系统的有力工具,各个细分领域的语义网络也陆续出现,WordNet 是其中具有代表性的成果[5]. 另外,人们也开始研究逻辑与语义网络的关系,提出了多种逻辑体系来进行知识的推理,包括一阶谓词逻辑、术语逻辑和描述逻辑等[6-9]. 进入互联网时代后,网络中积累了海量数据. 大部分数据以网页的形式直接面向用户展示,除了网页间的链接关系外,没有明显的结构化特征. 2001 年,BERNERS-LEE[10] 提出了语义网(Semantic Web),利用 W3C 标准(XML、RDF、OWL 等)在 Web 上建立数据间的语义关系,形成统一的知识库,便于机器检索和处理. 知识图谱本质上也属于语义网络或者语义网,它包含真实世界或大规模 Web 中的各种实体、概念及其关系,目前也被用来泛指各种大规模的知识库.

收稿日期: 2018-06-04　　《华南师范大学学报(自然科学版)》网址: http://journal.scnu.edu.cn/n
基金项目: 国家自然科学基金项目(61772211);广东省科技计划项目(2017A040405057,2016B010124008);广州市产学研协同创新重大专项(201704020203)
* 通讯作者: 汤庸,教授,Email: ytang4@qq.com.

在当前大数据时代,海量的数据蕴藏了丰富的价值,知识图谱可以帮助建立从这些数据中挖掘和组织知识的技术构架,在实现人工智能的过程中可以发挥重要作用. 本文将首先分析当前典型的知识图谱,其中包括通用领域和学术信息服务领域的知识图谱,并综述知识图谱涉及的关键技术,最后讨论学术信息服务领域的知识图谱构建与模式设计.

1　典型知识图谱

1.1　通用领域的知识图谱

(1) Freebase[11]. Freebase 是一个大型的协作式知识库,由美国 Metaweb 公司于 2007 年推出. 它的数据大部分来源于 Wikipedia、名人库 NNDB 和音乐知识库 MusicBrainz 等,还有部分数据来源于个人用户的贡献. 它的数据可公开下载,并公布了 API 供用户获取数据. 2010 年,Metaweb 公司被 Google 收购,由其推出的 Freebase 后来也成为 Google 知识图谱的重要组成部分. 2014 年,Freebase 的数据被并入到 Wikidata 中. 2016 年,Freebase 的 API 服务宣布终止,但其数据依然可以公开下载. 目前,Freebase 包含了 19 亿条关系.

(2) NELL[12-13]. NELL(Never-Ending Language Learning)是卡内基·梅隆大学于 2010 年推出的语义学习计划,该计划的目标是设计一个系统,不间断地从互联网几十亿网页中挖掘知识,并已得到美国国防高等研究计划署(DARPA)、谷歌、美国国家科学基金会(NSF)、Yahoo! 等机构的支持. 目前,NELL 已通过学习的方式挖掘到 5 千万条知识,每条知识都设置了可信度,其中有超过 2 百万的知识是高可信的. 挖掘到的知识提供在线浏览,也可以离线下载.

(3) Google Knowledge Graph[14]. 如前所述,谷歌知识图谱(Google Knowledge Graph)是谷歌于 2012 年 5 月推出的产品,其初始目的是提供智能搜索以增强传统的搜索体检,后面也应用于 Google Now、Google Assistant 和 Google Home 等产品. 其数据部分来源于 Freebase、CIA World Factbook 和 Wikipedia 等,自发布后,数据量持续增长,截止至 2016 年 10 月,谷歌知识图谱收录的事实数量已超过了 700 亿条. Google 设计了 API,供开发者访问. 只要开发者在谷歌中注册了开发者账号,然后在 API 用户管理中心中启用谷歌知识图谱的 API 就可以正常使用.

(4) DBpedia[15]. DBpedia 是由德国柏林自由大学、莱比锡大学及 OpenLink 软件公司合作发起的一个项目,旨在将 Wikipedia 中的数据转化为结构化的知识. Wikipedia 中的文档大部分是自由的自然语言文本,但也包含了部分结构化信息,如它的“infobox”、类别、图像、与外部网页的链接等. DBpedia 设计了一定的模式,将这些信息抽取出来,组织成为结构化的多语言的知识库,目前共包含 127 种语言版本. 第一个可用的数据库于 2007 年发布. 据最新发布的官方统计报告,截止至 2016 年 4 月,DBpedia 共包含了 0.28 亿个各种语言的实体,各类型的声明(statements)合计 2.39 亿条. 据评测,对事实样本进行抽样得到的准确率达到了 88%. DBpedia 的独特价值在于它作为开放数据链接(Linked Open Data)的中心,与多种知识库(如 Freebase、GeoNames、MusicBrainz、DBLP、CIA World Fact Book 及 Flickr 等)都建立了实体间的关联[15]. 根据 Linked Open Data 提供的数据显示,截止至 2018 年 4 月,DBpedia 与外部知识库关联的实体数量达到了 3 108 万个.

(5) Wikidata[16]. Wikidata 是维基媒体基金会主持的协作式编辑知识库. 其目的是为旗下的项目(包括维基百科、维基导游、维基物种及维基文库等)提供公共的数据来源. Wikidata 是一个面向文档的数据库,每个文档就是一个词条,由以字母“Q”开始的 ID 来唯一标识. 文档中包含了简介、别名以及一个或多个声明,其中,知识以键值对的方式存储到文档的声明列表中. 到目前为止,Wikidata 已包含了 4 700万篇文档,支持超过 350 种语言,支持以 JSON、XML 和 RDF 等多种格式的下载.

(6) Yago[17]. Yago 是由德国 Max Planck Institute for Informatics 及法国 Telecom ParisTech University 联合开发的一个大规模知识库,于 2008 年发布了第一个版本,将 WorkNet 中的本体层次结构体系与维基百科中丰富的信息进行了整合,将实体分类到 35 万余种类别中. 后来又结合 GeoNames,为很多实体和事件标定了时间和空间属性,完成了 Yago2 的建设[18-19];之后,利用维基百科的多语言版本,构建完成了 Yago3[20]. 目前,Yago 包含超过 1 千万个实体和 1.2 亿条事实. 经手工抽样验证,其准确率达到了 95.03%. Yago 的数据可以公开下载,其源代码也分享在 Github 上.

(7) ConceptNet[21]. ConceptNet 是一个多语言知识库,起源于 1999 年麻省理工大学媒体实验室发起的“Open Mind Common Sense”众包项目,旨在使得机器理解人类的语言,并推动人类常识的计算. ConceptNet 的数据来源广泛,包括维基词典、Open Mind Common Sense、特别设计的游戏(Verbosity、nadya.jp)以及

专家构建的 WordNet 和 JMDict. 目前,ConceptNet 支持 304 种语言,收录了2 800万条声明和超过 8 百万个实体. 实体之间的关系是固定的,包括 related-to、is-a、form-of、part-of、has-a 和 used-for 等 36 类关系.

(8) Linked Open Data Cloud. Linked Open Data Cloud 是一个开放数据联盟,旨在将现有的开放数据集或知识图谱链接起来,从而构成知识图谱云. 截止至 2018 年 4 月,Linked Open Data Cloud 已包含了 1 184 个数据集,并在数据集间建立了 15 993 条关联.

1.2 学术信息领域的知识图谱

(1) 微软学术图谱[23]. 微软学术搜索从 Bing 爬虫抓取的网页中智能分析出学术实体及它们之间的关系,构建出微软学术图谱(Microsoft Academic Graph,MAG). 微软学术图谱包含 6 种类型的实体,分别是作者、科研机构、论文、期刊或会议等论文出版发行商、研究领域及事件,其中事件指的是具体的某年在某地召开的学术会议,研究领域部分是利用文献[23]提到的方法构建而成. 目前,微软学术图谱已收录了 1.7 亿篇学术文献,提供了 2.1 亿位作者、4.7 万份期刊和 4 千余种会议实体.

(2) AMiner 科学知识图谱[24]. AMiner 是清华大学唐杰研究团队研发的国际化的学术文献服务平台,收录的文献大部分面向计算机领域,数量有 2.3 亿,并以此构建了科学知识图谱(Science Knowledge Graph,SciKG). SciKG 的实体包含 3 种类型: 概念、专家和论文,其中的概念即研究领域,有着上下级的包含与被包含关系,是从 ACM Computing Classification System[25] 中提取的. 目前,SciKG 共包含 908 个概念、20 万个专家及 51 万个论文实体.

(3) SCHOLAT 学者知识图谱[26]. 华南师范大学汤庸研究团队基于学者社交网络 SCHOLAT 上研发的学术信息服务平台构建了学者网知识图谱(Scholat Knowledge Graph,SKG). 截止至 2018 年 10 月,SCHOLAT 已收录 1 亿余条覆盖多个学科的学术论文、科研项目和知识产权信息,实名制注册的科研人员约 10 万,积累了大量学术信息服务领域相关数据,包括学术档案、学术文献以及合著关系数据等. 目前,SKG 共包含学者、学术刊物、学术组织及学科领域等实体 61 万个,实体间的联系数量为 3.9 亿个. 通过使用 SKG,学者网为注册用户提供了更准确的学术文献推荐、学者推荐及科研团队推荐等服务[27-28].

与通用领域的知识图谱相比,学术信息服务领域的知识图谱具有如下特点:

(1) 实体和实体关系的类型是固定的. 由于学术活动的纯粹性,学术涉及的实体及他们之间的关系基本不变. 实体基本由学者、科研机构、论文、期刊会议、论文发表机构及论文出版机构等组成.

(2) 研究领域本体是动态改变的. 随着学术研究活动的发展,将不断有新兴的研究领域出现. 因此,在构建研究领域本体时,应特别注意发现新的研究领域并能够动态扩展相应的本体库.

(3) 学者实体容易重名. 在学术文献库中,学者占据了核心位置,准确识别学者实体也是构建学术知识图谱的关键. 由于学术文献中的学者姓名非常容易出现重名问题,这给构建学者实体带来了很大的挑战,需要首先通过人名消歧方法进行学者身份的唯一性识别.

2 知识图谱关键技术

互联网积累海量数据. 这些数据复杂多样,既有大量非结构化的自然语言文本,也有半结构化的网页,还有少量结构化的数据. 如何从这些数据中提取出知识,组织为可用的知识库,是知识图谱面临的主要问题. 知识图谱涉及的技术非常多,整体上可分为三大部分: 第一部分是知识抽取,即从非结构化、半结构化或结构化的数据中提取出实体、实体属性以及实体之间的关系; 第二部分是知识融合,即当新发现的知识需要整合到现有的知识库时,有时同一个称谓会指向不同的实体,同样的实体也可能有不同的称谓,而且句子中会经常出现代词,这就需要对实体进行匹配和指代消解; 第三部分是知识的加工、推理及应用.

2.1 知识抽取

知识图谱的构建方式通常有 2 种: 一种是自顶向下的方式构建,从高质量的数据中提取本体和模式信息加入到知识库中,建成知识库的骨架,再从其他信息源中抽取知识,逐步丰富知识库的内容; 一种是自底往上的构建方式,设计高准确率的算法,从丰富的自然语言文本中提取知识,经人工审核后加入到知识库. 在这 2 种构建方式中,知识抽取是非常重要的第一步,所要解决的问题是如何从无结构或半结构化的数据中,抽取出实体以及实体与实体之间的关系. 第 2 种构建方式,对知识抽取的准确度要求更高.

2.1.1 实体抽取 实体抽取是利用命名实体识别(Named Entity Recognition,NER)技术[29],将实体从海量的自然语言文本中挖掘出来. 实体是知识图谱

的最基本元素,实体抽取算法的准确率直接决定着知识图谱构建的质量.

早期的命名实体识别系统依赖于专家手工编制的启发式规则或者字典,根据句子里出现的句法特征进行实体的识别[30],常见的句法特征有标点符号、关键词和位置词等信息. 这些方法的优势是简单、运行速度快,但仅能处理狭窄领域的命名实体识别任务,需耗费大量的人工,通用性和扩展性较差. 针对以上弊端,研究人员开始尝试利用机器学习的相关理论来解决命名实体识别问题. 例如,COLLIER 等[31]利用隐马尔可夫模型,从 MEDLINE 数据库文献的摘要和正文中提取基因名称; WANG 等[32]为将传统中医开具的手写诊疗记录转化为结构化的数据,综合利用隐马尔可夫、最大熵马尔可夫、条件随机域模型对最为关注的症状名称进行抽取; LIU 等[33]利用 K-最近邻算法和条件随机域,从 Twitter 文本中进行实体识别; SAHA 等[34]采用支持向量机,综合利用字符特征和层次聚类信息构造了核函数,在 JNLPBA 2004 生物医学数据集中进行了基因分类; LIN 等[35]首先利用最大熵识别文本中的生物名称,然后利用词典和规则修正前一阶段的错误,提高了系统性能.

传统的监督式机器学习算法需要大量经过标注的样本数据,但在实际情况中,这往往需要耗费大量的人力. CHEN 等[36]针对这一问题,提出采用主动学习框架,训练时每轮迭代只选择一小部分样本进行标注,然后再利用 CRF 模型(Conditional Random Field)进行训练. 经过在医学数据集验证,该方法可有效降低对标注样本量的需求. 为解决通用领域的命名实体识别问题,SEKINE 等[37]提出建立一个命名实体的分类体系,将网络中的所有实体划分成 150 个分类,并构成层次结构. LING 和 WELD[38]借鉴 Freebase 的分类方法,设计了 112 种类别,利用条件随机域和自适应感知机算法,实现了通用领域中的实体自动识别和分类,与 Stanford NER 的命名识别系统[39]相比,效果要更优. HABIBI 等[40]采用深度学习框架长短期记忆网络,结合 CRF 模型,实现了领域无关的 NER 模型,在进行特定领域优化的情况下,大幅提高了召回率. LIN 等[41]采用双向的 LSTM-CRF,在没有针对领域手工构建特征的情况下,仅利用字符和句法信息,就很好地完成了在带噪声的用户生成文本中发现新的命名实体的任务.

2.1.2　实体关系提取　在知识库的构建过程中,实体关系提取与实体提取处于同样重要的地位,它要解决的是实体间的语义关联问题. 与实体提取一样,传统的实体关系提取也是基于启发式规则的,通过人工构造关系模板提取特定的关系文本; 后来,利用统计机器学习,结合实体关系的上下文进行实体关系提取成为研究的共识. 例如,在早期,KAMBHATLA[42]借助上下文中的词法、名法和语义特征进行实体关系建模,利用最大熵方法成功抽取了实体关系; ZHOU 等[43]在文献[42]的基础上,加入了实体类别信息,结合 WordNet,使用 SVM 作为分类器,使得实体关系识别准确率达到了 55%. 在最近一段时期,深度学习技术在图像、视频、语音领域取得了极大的成功,并开始向智能文本处理领域扩展[44-46]. HASHIMOTO 等[47]利用词嵌入(Word Embedding)的方法,从标注的语料库中学习特定名词对的上下文特征,然后利用神经网络分类器进行分类,在 SemEval-2010 task 8 的数据集上,其 F1 值达到了 82%; LI 和 JI[48]提出了一种联合抽取方案,同时提取实体和实体关系,与此同时,加入全局特征作为软约束,避免了错误累积. 以上统计机器学习及深度学习模型都属于监督学习范畴,需要使用大量手工标注的样本去训练模型. 为解决这一问题,BRIN[49]利用 Bootstrap 方法构建了 DIPRE 系统,首先加入少量手工设计的种子模板,利用模板进行学习,然后不断从网络中大量非标注的文本中提取新的模板,加入到模板集里面,从而解决需花费大量人力标注数据的问题.

2.2　知识融合

知识融合是指将抽取到的知识与知识库已有的知识整合起来. 在这一过程中经常遇到的问题是: 很多实体具有歧义性. 这里的歧义有两方面: 一是相同的实体具有不同的名字; 二是相同的名字指向不同的实体. 解决这一问题需要用到的技术是实体匹配(Entity Matching),亦称为实体链接(Entity Linking)、实体对齐(Entity Alignment)、实体解析(Entity Resolution). 按采用模型的不同,实体匹配算法可分为基于概率模型的实体匹配算法和基于机器学习的实体匹配算法. 在基于概率模型的实体匹配算法中,主要考虑 2 个实体各自属性的相似性. NEWCOMBE 等[50]将实体匹配问题转化为分类问题,根据属性的相似度建立了概率模型. 在该模型中,所有的属性都具有同等的重要性. HERZOG 等[51]为不同的属性分配了不同的权重,准确率有所提高. 基于机器学习的实体匹配算法可以将实体匹配问题当作二分类问题,在属性值之间建立相似度函数,就可以利用常见的机器学习算法进行求解,如决策树、支持向量机和集成学习等[52-53]. 例如,

CHRISTEN[54]提出了无监督式的二阶段分析模型,首先按照一定的规则把最相似的或者相似度超过一度阈值的数据聚在一起,作为训练样本,然后在第二阶段利用SVM来训练二类的分类器,避免了手工标注训练样本,算法的效果也优于K-means的效果;COHEN和RICHMAN[55]将实体匹配视为聚类问题,将尽量相似的实体聚到一起再进行匹配,提出了一种自适应距离函数,具有较强的扩展性.

2.3 知识加工

经过知识抽取与知识融合后,知识库中就包含了一系列事实,但要形成可用的知识图谱,还需要对知识进行进一步的加工处理,这一过程主要有本体构建和知识推理.

2.3.1 **本体构建** 本体描述了概念之间的关系,是对客观世界的建模. 本体的一个特点是共享,是同一领域的不同主体之间进行交流的语义基础[56]. 本体可由领域专家手工构建,但由于工作量巨大,仅适用于特定的细分领域. 目前通用领域的本体库产品,都是由数据驱动自动构建后由人工审核进行修正的. 例如,微软的Probase利用Bootstrap模型,使用语法和语义结合迭代的方法,首先利用少量精确的Hearst模板[23],从海量的自然语言文本中提取出高质量的"is-a"实体关系,构成初始的知识库,然后利用知识库中的语义知识作为语义种子,不断进行迭代扩展新的语义知识[57]. 例如,对于句子中一个片断: "…domestic animals other than dogs such as cats…",如果仅仅从语法方面去解析,会得到2种可能的解读"cat is-a dog"和"cat is-a domestic animal". 如果不确定cat、dog与domestic animal的关系(如在第一轮迭代时),这个句子只能被舍弃掉. 而在下一轮迭代的时候,如果知识库已有了知识"domestic animal is-a animal"和"cat is-a animal",那么就可以在这2个解读中选择可信度更高的那一个. 利用这一方法,可以在保证高准确率的同时提高召回率. 到目前为止,Probase已从16亿网页中抽取到了265万个概念,包含了超过2千万个的"is-a"概念对,准确率达到92.8%,在规模和准确率上都居前列.

2.3.2 **知识推理** 知识推理是从知识库中已有的实体关系出发,推理出新的关系,在丰富和扩展知识库方面有重要作用. 知识推理的方法可分为两大类:基于逻辑的推理和基于图的推理.

(1) 基于逻辑的知识推理. 基于逻辑的推理主要包括一阶谓词逻辑和描述逻辑[58-59]. 一阶谓词逻辑建立在命题的基础上,命题分为个体和谓词两部分,可分别对应知识库中的实体和实体关系. 设定好逻辑推理的规则和约束条件,就可以对知识进行简单的推理. 一阶谓词逻辑适合于简单的人物关系推理. 例如,可以设计同事关系的逻辑推理规则如下:

IF (B workIn A) and (C workIn A)
and (B is A person) and (C is A person)
and (A is A organization)
THEN →(B colleagueWith C)

描述逻辑可以用来进行复杂的实体关系推理,是一种基于对象知识的形式化表示,由个体、概念和属性3个元素构成,其中,个体对应于知识库中具体的实体,概念是实体的集合,相当于实体的上位词,而属性代表实体之间的二元关系. 描述逻辑具有很强的表达能力,而且是可判定的,可以保证推理算法总能终止. 基于描述逻辑的知识库,一般都设计有2种公理集:TBox(Terminology Box)和ABox(Assertion Box). 其中,TBox声明了知识库中上下位实体间的包含关系,ABox声明了实体之间的关系断言. 借助这2个工具,可以利用Tableau算法,将基于描述逻辑的推理归结为ABox的一致性检测问题,从而简化并最终实现关系推理[60-61].

(2) 基于图的知识推理. 基于图的知识推理主要使用各种图算法进行关联推理,例如,LAO和COHEN[62]提出了Constrained Random Walk模型和Path Ranking算法,利用关系路径中蕴含的信息,通过图中2个实体间的多步路径来预测它们的语义关联,该方法被用于包含约50万关系的NELL知识库中,取得了良好的效果,与NELL所采用的Horn子句学习和归纳算法[13]相比,准确率提高了近一倍.

3 学术信息服务知识图谱构建

本节以学者网为应用背景,讨论学术知识图谱的模式设计. 学者网是以学者为中心的社交网络,目前已有10余万个实名注册的学者用户,形成了活跃的学术社区. 这些学术社区涉及了多种不同的类型,如包含几十名甚至上百名学者的大中型实验室、少数几人组成的学术讨论组、几名老师构成的教学团队、老师与学生之间的课程等等. 因此,除了传统的学术搜索中的学术实体之外,学者网中的学术知识图谱还需要对学术社交活动中涉及的各种实体进行建模.

3.1 学术实体类型

本文梳理了学术社交活动中涉及的各种学术实体及它们之间的相互关系. 表1给出了在学者网的学术知识图谱中涉及的各种学术实体类型.

表 1　学者网学术实体类型表

Table 1　Academic entity types in SCHOLAT

实体名称	描述
学术成果(Publication)	包含论文、图书、图书章节、图书系列和专利等各种形式的学术成果.
学者(Scholar)	一切与学术活动相关的个人都可被称为学者.
团队(Team)	包括大中型实验室团队、小型学术讨论组和教学团队等.
课程(Course)	教师开设的教学课程.
论文发表机构(Avenue)	期刊、会议等.
单位(Affiliation)	院校、研究所等科研机构.
基金(Funding)	支持学术活动的基金项目.
研究领域(Field)	与微软学术图谱一样,将研究领域也构建为一个实体,并构建研究领域的上下级包含关系,形成领域本体.

学术实体之间存在相互关系,大部分学术活动具有主客体之间的区别,因此,学术实体之间的关系也是单向的,由此构成的学术图谱属于有向图. 图 1 给出了学者网中学术实体之间不同关系的图形化表示. 为了简化图形的表示,图中 2 个实体间的双向关系用无向边表示,用"/"分隔; 单向关系用有向边表示.

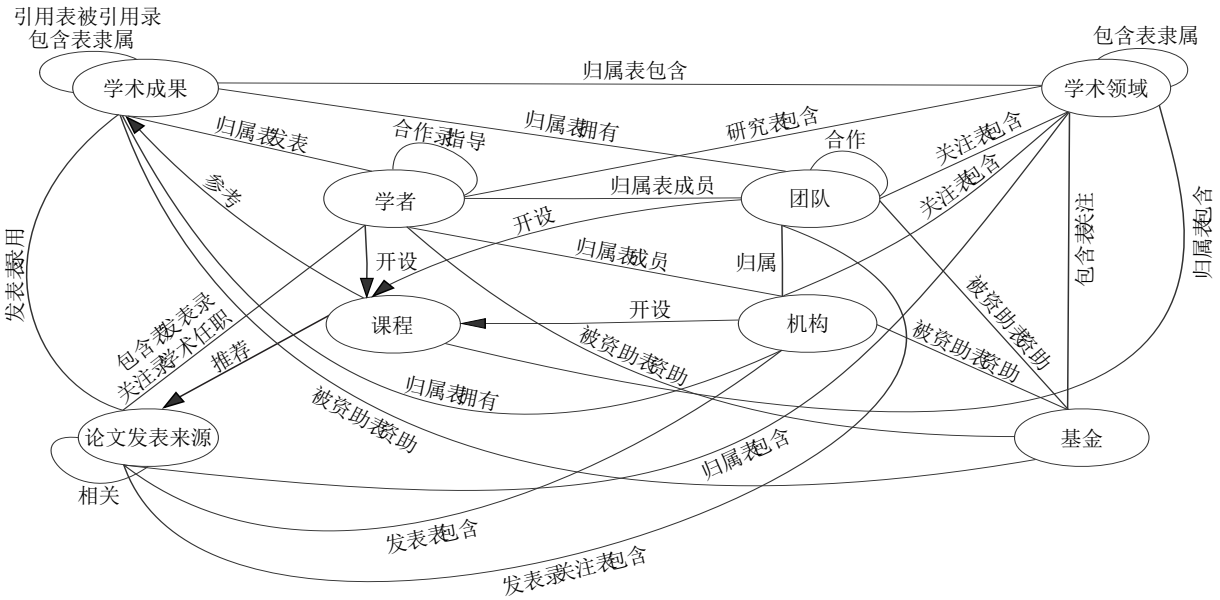

图 1　学者网学术实体关系类型

Figure 1　Academic entity relationships in SCHOLAT

3.2　学术信息获取

学者网的学术信息包含两部分: 一部分是由注册用户产生的个人学术信息,包括个人学术资料、开设的课程、创建的团队、与好友的互动等,该部分信息是用户产生的,保存在学者网网站中,无需到外部去获取; 另一部分是收录的站外的学术文献资料,提供给学者检索学术信息时使用,该部分信息需要到各类学术资源网站去抓取. 本小节主要介绍站外学术信息的抓取方案.

为实现对站外学术资源的抓取,学者网团队先后设计并实施了 2 套爬虫系统: 其一是基于 Nutch 的爬虫,其二是基于浏览器插件的爬虫. 两者各有优劣: 前者基于开源系统 Nutch,架构设计合理,只需要针对不同的页面设计相应的页面解析器即可,适应于页面结构较为简洁的学术网站,学术信息原生显示在页面中,不存在过多的 Ajax 调用; 后者可直接运行在用户浏览器上,与用户的行为完全一致,运行过程直接显示在浏览器的页面上,抓取到的学术信息所见即所得,非常直观,也便于调试. 由于浏览器对 Ajax 过程原生的支持,因此,浏览器爬虫非常适合页面结构复杂、Ajax 调用频繁的场景.

由浏览器爬虫抓取 Link Springer 学术论文时的运行效果(图 2) 可知: 学术爬虫通过对页面元素的解析,直接完成了对学术信息的结构化提取,这对之后的学术信息处理提供了非常大的便利.

图2 基于浏览器插件的学术爬虫

Figure 2 Academic crawler based on browser plugin

3.3 学术实体匹配

抓取到结构化的学术信息后,一个非常重要的工作就是对学术信息进行实体匹配. 在学术文献信息中,论文、作者、论文来源、作者单位是彼此相关的实体. 其中,论文来源、作者单位较为明确,表述也较为规范,几乎没有歧义现象; 少数期刊、会议存在简写现象,这通过建立别名字典即可解决. 本小节主要讨论论文和作者实体化方法.

3.3.1 论文实体化 由于一篇同样的学术论文可能会在多个不同的学术网站出现,如果没有恰当的实体化方案,会导致学者网重复收录该论文,一方面造成搜索结果的冗余,另一方面也会影响论文被引用情况的统计,影响论文、作者和作者单位等相关学术实体的影响力分析.

在构建学者网学术知识图谱时,通过对论文的关键字段求哈希值作为论文的ID来实现论文的实体匹配. 论文的DOI编号可以确定一篇论文的唯一性,因此,如果存在DOI字段,则系统可利用它来计算论文ID; 如果不存在DOI字段,则采用最容易获取的、也最具备区分性的字段来生成论文ID. 经实验,系统中最终采用的关键字段是论文题目、前3位作者名字以及论文的年份.

3.3.2 作者实体化 在所有实体匹配中,作者的重名、别名现象最为突出,因此,作者的实体匹配也最具有挑战性,作者重名判别问题也得到了广泛的研究. 目前主流的方法可分为分类法[63]、聚类法[64-65]及概率模型法[66].

在本系统中,因为需要处理上亿级的文献数量,常规的基于机器学习的重名判别算法无法胜任,所以,设计实现了基于MapReduce的两段式作者重名判别算法[67]. 与题目、关键词、摘要、单位和发表来源等论文的内容语义特征相比,合作者信息最容易获取,计算简便而且区分度非常好,因此,该算法仅采用了合作者特征. 共同合作者越多,则2个姓名指向同一作者的概率越大. 由此,根据“先易后难”原则,在第一阶段对最确定的姓名对进行合并,以尽可能高的准确率构建出“原子聚类”; 在第二阶段,根据其他作者的原子聚类情况进行扩展,提高召回率. 经实验验证,该方案可以有效地在上亿级的数据库中实现作者实体匹配.

3.4 学术知识图谱的应用意义

在一篇学术文献中蕴含了丰富的信息,如作者之间的合作关系网络、作者在某时段的工作单位、受基金项目资助情况、作者的研究主题、刊物的覆盖内容和学术会议召开的时间及地点等. 对这些信息进行深入分析,可构建关系丰富的学术知识图谱. 此外,学术知识图谱中的内容还可以由学术社区用户自发生成. 由于其内容是学者贡献的,涉及个人学术声誉,因此,在内容的准确性和及时性上,相比自动化分析的手段更有优势. 学术知识图谱可在多方面对学术社区起到促进作用[27-28]:

(1) 促进学术社区用户之间的协作水平和协作效率. 学术知识图谱对作者、论文和科研项目等进行了建模,并建立了各种学术实体间的关联关系,因此,学术社区用户可以非常方便地获取合作者的相关科研成果,实现“一人录入,好友共享”,提高协作效率.

(2) 提高学术搜索的准确率. 当前学术搜索大

多基于传统的文档关键字匹配模型,为用户提供符合搜索条件的论文列表. 而借助于知识图谱,学术搜索可以做到在用户键入查询词的同时理解用户的搜索意图,并根据查询意图做出特定的优化,从而提高学术搜索的准确率,提升搜索体验.

(3) 便于建立学术评价机制. 学术知识图谱将相关的信息全部实体化并进行了关联,因此,可以非常方便地提取出某个作者、研究机构或者刊物所发表的论文,并根据论文的被引用情况对作者、研究机构或刊物进行学术评价. 在建立论文、专利与基金项目的关联后,基金管理部门还可以借此考核基金的资助效果.

4　总结与展望

随着技术的发展,机器变得越来越"智能",使用知识图谱可以为实现人工智能搭建基本的技术理论框架,因此,知识图谱是现代人工智能与智能信息服务的重要前沿研究课题和关键技术,具有重要的学术研究意义和广泛的应用价值.

目前,虽然工业界和学术界对知识图谱的研究与应用已经取得了一定的成果,但是在以下方面仍然需要进一步研究:

(1) 目前互联网的信息处于爆炸性增长的状态,信息提取和加工处理的效率受到严峻的考验,如何从快速增长的海量数据中高效提取知识图谱所需要的语义信息,同时保持较高的召回率,仍然是一个需要持续研究的问题.

(2) 目前对图像、音频及视频等类型数据的语义理解已经借助深度学习实现了巨大的进步,但文本类型数据的语义分析和准确理解方面仍有不足,而有效的文本分析在知识图谱构建中具有重要作用. 因此,如何借助深度学习进一步提升文本数据的语义分析性能将是一个重要的研究课题.

(3) 在学术信息服务领域中,目前知识图谱还仅限于文献检索、论文获取及学术资源推荐等较为基础的应用,在支持更高级的智能应用方面仍然很薄弱. 例如,通过知识图谱实现对研究领域发展脉络的精准呈现、预测下一个可能存在的研究热点、分析论文被引用模式及帮助学者的论文被更广泛的引用等将是更有价值的应用.

参考文献:

[1] SAWANT U, CHAKRABARTI S, RAMAKRISHNAN G. "Open - domain question answering using a knowledge graph and web corpus" by Uma Sawant, Soumen Chakrabarti and Ga-nesh Ramakrishnan with Martin Vesely as coordinator [J]. ACM SIGWEB Newsletter, 2018(Winter): 4/1-8.

[2] OSTERMANN S, ROTH M, MODI A, et al. SemEval-2018 task 11: machine comprehension using commonsense knowledge [C] // Proceedings of the 12th International Workshop on Semantic Evaluation. New Orleans, Louisiana: Association of Computational Linguistics, 2018: 747-757.

[3] PAULHEIM H. Knowledge graph refinement: a survey of approaches and evaluation methods [J]. Semantic Web, 2017, 8(3): 489-508.

[4] RICHENS R H. Interlingual machine translation [J]. The Computer Journal, 1958, 1(3): 144-147.

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

[6] SIMMONS R F, BRUCE B C. Some relations between predicate calculus and semantic net representations of discourse [C] // Proceedings of the 2nd International Joint Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers Inc, 1971: 524-530.

[7] SCHUBERT L K. Extending the expressive power of semantic network [J]. Artificial Intelligence, 1975, 7(2): 158-164.

[8] JIANG Y C, WANG J, TANG S Q, et al. Reasoning with rough description logics: an approximate concepts approach [J]. Information Sciences, 2009, 179(5): 600-612.

[9] JIANG Y C, TAN H Y. Very expressive fuzzy description logics over lattices for the Semantic Web [J]. Journal of University of Electronic Science and Technology of China, 2012, 41(3): 322-335.

[10] BERNERS-LEE T, HENDLER J, LASSILA O. The Semantic Web [J]. Scientific American, 2001, 284(5): 34-43.

[11] BOLLACKER K, COOK R, TUFTS P. Freebase: a shared database of structured general human knowledge [C] // Proceedings of the 22nd National Conference on Artificial Intelligence. Palo Alto, California: AAAI, 2007: 1962-1963.

[12] MITCHELL T, COHEN W, HRUSCHKA E, et al. Never-ending learning [J]. Communications of the ACM, 2018, 61(5): 103-115.

[13] CARLSON A, BETTERIDGE J, KISIEL B, et al. Toward an architecture for never-ending language learning [C] // Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence. Palo Alto, California: AAAI, 2010: 1306-1313.

[14] VINCENT J. Apple boasts about sales; Google boasts about how good its AI is [EB/OL]. (2016-10-04) [2018-05-10]. https: // www. theverge. com/2016/10/4/13122406/google-phone-event-stats.

[15] BIZER C, LEHMANN J, KOBILAROV G, et al. DBpedia-a crystallization point for the web of data [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2009, 7(3): 154-165.

[16] TANON P T,VRANDEČIĆ D,SCHAFFERT S,et al. From Freebase to Wikidata: the great migration [C] // Proceedings of the 25th International Conference on World Wide Web. Montreal,Canada: International World Wide Web Conferences Steering Committee,2016: 1419-1428.

[17] SUCHANEK F M,KASNECI G,WEIKUMG. Yago: a core of semantic knowledge [C] // Proceedings of the 16th International Conference on World Wide Web. New York: ACM,2007: 697-706.

[18] 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.

[19] HOFFART J,SUCHANEK F M,BERBERICH K,et al. Yago2: exploring and querying world knowledge in time,space,context,and many languages [C] // Proceedings of the 20th International Conference Companion on World Wide Web. New York: ACM,2011: 229-232.

[20] MAHDISOLTANI F,BIEGA J,SUCHANEK F M. Yago3: a knowledge base from multilingual Wikipedias [C] // Proceedings of 7th Biennial Conference on Innovative Data Systems Research. Asilomar,California: CIDR,2015.

[21] SPEER R,HAVASI C. ConceptNet 5: a large semantic network for relational knowledge [M] // GUREVYCH I,KIM J. The People's Web Meets NLP. Berlin: Springer,2013: 161-176.

[22] Microsoft Research. Microsoft Academic Graph [Z/OL]. (2015-06-05) [2018-05-03]. https://www.microsoft.com/en-us/research/project/microsoft-academic-graph.

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

[24] TANG J. Science knowledge graph [Z/OL]. (2009-12-12) [2018-05-03]. https://www.aminer.cn/scikg.

[25] ACM. The ACM computing classification systems [S/OL]. [2018-05-03]. https://dl.acm.org/ccs/ccs_flat.cfm.

[26] TANG F Y,ZHU J,HE C B,et al. SCHOLAT: an innovative academic information service platform [M] // CHEEMA M,ZHANG W,CHANG L. Databases Theory and Applications. Berlin: Springer,2016: 453-456.

[27] 贺超波,汤庸,刘海,等. 一种集成链接和属性信息的社区挖掘方法 [J]. 计算机学报,2017,40(3): 601-616.
HE C B,TANG Y,LIU H,et al. Method for community mining integrating link and attribute information [J]. Chinese Journal of Computers,2017,40(3): 601-616.

[28] 李春英,汤庸,林海,等. 基于标签传播的可并行复杂网络重叠社区发现算法 [J]. 中国科学: 信息科学,2016,46(2): 212-227.
LI C Y,TANG Y,LIN H,et al. Parallel overlapping community detection algorithm in complex networks based on label propagation [J]. Scientia Sinica Informationis,2016,46(2): 212-227.

[29] MOHIT B. Named entity recognition [M] // ZITOUNI I. Natural Language Processing of Semitic Languages. Berlin: Springer,2014: 221-245.

[30] RAU L F. Extracting company names from text [C] // Proceedings of the Seventh IEEE Conference on Artificial Intelligence Applications. Miami Beach,FL: IEEE,1991: 29-32.

[31] COLLIER N,NOBATA C,TSUJII J. Extracting the names of genes and gene products with a hidden Markov model [C] // Proceedings of the 18th Conference on Computational Linguistics. Stroudsburg,PA: Association for Computational Linguistics,2000: 201-207.

[32] WANG Y Q,YU Z H,CHEN L,et al. Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: an empirical study [J]. Journal of Biomedical Informatics,2014,47: 91-104.

[33] 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. Stroudsburg,PA: Association for Computational Linguistics,2011: 359-367.

[34] SAHA S K,NARAYAN S,SARKAR S,et al. A composite kernel for named entity recognition [J]. Pattern Recognition Letters,2010,31(12): 1591-1597.

[35] LIN Y F,TSAI T H,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. Berlin: Springer,2004: 56-61.

[36] CHEN Y K,LASKO T A,MEI Q Z,et al. A study of active learning methods for named entity recognition in clinical text [J]. Journal of biomedical informatics,2015,58: 11-18.

[37] SEKINE S,SUDO K,NOBATA C. Extended named entity hierarchy [C] // Proceedings of the 3rd International Conference on Language Resources and Evaluation. Las Palmas,Spain: European Language Resources Association,2002: 1818-1824.

[38] LING X,WELD D S. Fine-grained entity recognition [C] // Proceedings of the 26th AAAI Conference on Artificial Intelligence. Palo Alto,California: AAAI,2012: 94-100.

[39] FINKEL J R,GRENAGER T,MANNING C. Incorporating non-local information into information extraction systems by gibbs sampling [C] // Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg,PA: Association for Computational Linguistics,2005: 363-370.

[40] HABIBI M,WEBER L,NEVES M,et al. Deep learning with word embeddings improves biomedical named entity recognition [J]. Bioinformatics,2017,33(14): i37-i48.

[41] LIN B Y,XU F,LUO Z,et al. Multi-channel BiLSTM-CRF model for emerging named entity recognition in social media [C] // Proceedings of the 3rd Workshop on

Noisy User-Generated Text. Stroudsburg,PA: Association for Computational Linguistics,2017: 160-165.

[42] KAMBHATLA N. Combining lexical,syntactic,and semantic features with maximum entropy models for extracting relations [C] // Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. Stroudsburg,PA: Association for Computational Linguistics,2004: 1-22.

[43] ZHOU G D,SU J,ZHANG J,et al. Exploring various knowledge in relation extraction [C] // Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg,PA: Association for Computational Linguistics,2005: 427-434.

[44] YANN L,YOSHUA B,GEOFFERY H. Deep learning [J]. Nature,2015,521: 436-444.

[45] CHEN L H,LING Z H,LIU L J,et al. Voice conversion using deep neural networks with layer-wise generative training [J]. IEEE/ACM Transactions on Audio,Speech and Language Processing,2014,22(12): 1859-1872.

[46] HE K M,ZHANG X Y,REN S Q,et al. Deep residual learning for image recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,NV: IEEE,2016: 770-778.

[47] HASHIMOTO K,STENETORP P,MIWA M,et al. Task-oriented learning of word embeddings for semantic relation classification [J]. Proceedings of the Nineteenth Conference on Computational Natural Language Learning. Stroudsburg,PA: Association for Computational Linguistics,2015: 268-278.

[48] LI Q,JI H. Incremental joint extraction of entity mentions and relations [C] // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg,PA: Association for Computational Linguistics,2014: 402-412.

[49] BRIN S. Extracting patterns and relations from the World Wide Web [J]. Lecture Notes in Computer Science,1998,1590: 172-183.

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

[51] HERZOG T N,SCHEUREN F J,WINKLER W E. Data quality and record linkage techniques [M]. New York: Springer,2007.

[52] HAN J W,KAMBE M,PEI J. Data mining: concepts and techniques [M]. San Francisco,CA: Morgan Kaufmann,2006.

[53] VAPNIK V. The nature of statistical learning theory [M]. Berlin: Springer,2000.

[54] CHRISTEN P. Automatic training example selection for scalable unsupervised record linkage [C] // WASHIO T,SUZUKI E,TING K M,et al. Advances in Knowledge Discovery and Data Mining. Berlin: Springer,2008: 511-518.

[55] COHEN W W,RICHMAN J. Learning to match and cluster large high-dimensional data sets for data integration [C] // Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM,2002: 475-480.

[56] STUDER R,BENJAMINS V R,FENSEL D. Knowledge engineering: principles and methods [J]. Data & Knowledge Engineering,1998,25(1): 161-197.

[57] 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. New York: ACM,2012: 481-492.

[58] LINDSTRÖM P. First order predicate logic with generalized quantifiers [J]. Theoria,1966,32(3): 186-195.

[59] BAADER F,CALVANESE D,MCGUINNESS D,et al. The description logic handbook: theory,implementation and applications [M]. New York: Cambridge University Press,2003.

[60] JIANG Y C,TANG Y,CHEN Q M,et al. Semantic operations of multiple soft sets under conflict [J]. Computers & Mathematics with Applications,2011,62(4): 1923-1939.

[61] HORROCKS I. Description logics in ontology applications [C] // Proceedings of the 14th International Conference on Automated Reasoning with Analytic Tableaux and Related Methods. Berlin: Springer,2005: 2-13.

[62] LAO N,COHEN W W. Relational retrieval using a combination of path-constrained random walks [J]. Machine Learning,2010,81(1): 53-67.

[63] HAN H,GILES L,ZHA H Y,et al. Two supervised learning approaches for name disambiguation in author citations [C] // Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries. New York: ACM,2004: 296-305.

[64] TANG J,ZHANG J,YAO L M,et al. ArnetMiner: extraction and mining of academic social networks [C] // Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM,2018: 990-998.

[65] HAN H,ZHA H Y,GILES C L. Name disambiguation in author citations using a k-way spectral clustering method [C] // Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries. New York: ACM,2005: 334-343.

[66] ZHANG D,TANG J,LI J Z,et al. A constraint-based probabilistic framework for name disambiguation [C] // Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management. New York: ACM,2007: 1019-1022.

[67] ZHANG L,CHEN G H,TANG Y,et al. Disambiguating authors in academic search engines [C] // LIN X M,MANOLOPOULOS Y,SRIVASTAVA D,et al. Web Information Systems Engineering-WISE 2013. Berlin: Springer,2013: 511-514.

【责任编辑: 庄晓琼　责任校对: 庄晓琼　英文审校: 程杰】