中国人工智能学会推荐
“十三五”国家重点图书出版规划

人工智能丛书

知识图谱

Knowledge Graph

赵军 主编
赵军 刘康 何世柱 陈玉博 编著

高等教育出版社

中国人工智能学会推荐
“十三五”国家重点图书出版规划

人工智能丛书

知识图谱

Knowledge Graph

赵军 主编
赵军 刘康 何世柱 陈玉博 编著

高等教育出版社·北京

内容简介

本书聚焦于知识图谱，分十个章节围绕知识建模、知识获取、知识融合、存储和检索、知识推理以及知识服务等知识图谱生命周期各个主要环节展开介绍。每章以任务为导引，引出任务描述、难点问题、基本方法、研究现状和存在的问题，并从多个相关的研究方向对各个任务的发展进程进行系统的、多维度的梳理，注重介绍传统知识工程的思想和理论以及机器学习和深度学习在知识图谱各个环节中应用的技术和方法，从而使读者能够了解发展脉络，激发研究兴趣，思考核心问题，领悟发展方向。

本书可以作为自然语言处理、知识工程、人工智能等相关课程的研究生教材，也可供计算机科学技术领域相关工程技术人员学习参考。

图书在版编目（CIP）数据

知识图谱 / 赵军主编；赵军等编著. -- 北京：高等教育出版社，2018.12 (2019.4重印)

（人工智能丛书）

ISBN 978-7-04-050984-7

Ⅰ. ①知… Ⅱ. ①赵… Ⅲ. ①知识管理 Ⅳ. ① G302

中国版本图书馆 CIP 数据核字（2018）第 258451 号

策划编辑	张江漫	责任编辑	张江漫	封面设计	赵 阳	版式设计	徐艳妮
插图绘制	于 博	责任校对	刘丽娴	责任印制	陈伟光		

出版发行	高等教育出版社	网 址	http://www.hep.edu.cn
社 址	北京市西城区德外大街4号		http://www.hep.com.cn
邮政编码	100120	网上订购	http://www.hepmall.com.cn
印 刷	北京市联华印刷厂		http://www.hepmall.com
开 本	787mm × 960mm 1/16		http://www.hepmall.cn
印 张	19.5		
字 数	350 千字	版 次	2018 年 12 月第 1 版
购书热线	010-58581118	印 次	2019 年 4 月第 3 次印刷
咨询电话	400-810-0598	定 价	49.90 元

物 料 号 50984-00

赵 军

中国科学院自动化研究所模式识别国家重点实验室，研究员，博士生导师；中国科学院大学人工智能学院岗位教授。研究领域为自然语言处理、知识图谱、信息抽取、问答系统等。作为项目负责人承担国家自然科学基金重点项目等多项国家级重要科研项目以及企业应用项目。在ACL、IJCAI、SIGIR、AAAI、COLING、EMNLP、TKDE等顶级国际会议和重要学术期刊上发表论文80余篇。曾获第25届国际计算语言学大会COLING 2014最佳论文奖，主持研发的“大规模开放域文本知识获取与应用平台”获得2018年中国中文信息学会“钱伟长中文信息处理科学技术奖”一等奖。兼任中国中文信息学会常务理事，语言与知识计算专业委员会副主任，计算语言学专业委员会副主任，《中文信息学报》编委，ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)副主编等学术职务。在中国科学院大学主讲“知识图谱导论”等课程。

刘 康

博士，中国科学院自动化研究所模式识别国家重点实验室副研究员，西安电子科技大学客座教授。研究领域包括信息抽取、网络挖掘、问答系统等，同时也涉及模式识别与机器学习方面的基础研究。在人工智能、自然语言处理、知识工程等领域国际重要会议和期刊发表论文60余篇，曾获第25届国际计算语言学大会COLING 2014最佳论文奖、中国中文信息学会“钱伟长中文信息处理科学技术奖——汉王青年创新奖”，2018年获得中国中文信息学会“钱伟长中文信息处理科学技术奖”一等奖。兼任中国中文信息学会青年工作委员会主任、语言与知识计算专业委员会秘书长等学术职务。

何世柱

博士，中国科学院自动化研究所模式识别国家重点实验室副研究员，2016年获得中国科学院大学工学博士学位。研究方向为问答系统、对话系统和自然语言处理。主持国家自然科学基金青年科学基金项目“知识问答中的自然答案生成关键技术研究”，2018年获得中国中文信息学会“钱伟长中文信息处理科学技术奖”一等奖。

陈玉博

博士，中国科学院自动化研究所模式识别国家重点实验室助理研究员，2017年获得中国科学院大学工学博士学位。研究方向为信息抽取、知识图谱和自然语言处理。主持国家自然科学基金青年科学基金项目“面向非结构化文本的大规模事件信息抽取关键技术研究”，2018年获得中国中文信息学会“钱伟长中文信息处理科学技术奖”一等奖。

知识图谱

赵军

1 计算机访问http://abook.hep.com.cn/1253771，或手机扫描二维码、下载并安装Abook应用。

2 注册并登录，进入“我的课程”。

3 输入封底数字课程账号（20位密码，刮开涂层可见），或通过Abook应用扫描封底数字课程账号二维码，完成课程绑定。

4 单击“进入课程”按钮，开始本数字课程的学习。

课程绑定后一年为数字课程使用有效期。受硬件限制，部分内容无法在手机端显示，请按提示通过计算机访问学习。

如有使用问题，请发邮件至abook@hep.com.cn。

http://abook.hep.com.cn/1253771

人工智能丛书编委会

序

近几年来知识图谱无疑是一个火热的名字。从 2016 年秋天接受高等教育出版社的邀请着手撰写本书开始我一直在思考：什么是知识图谱？为什么它受到学术界和产业界如此的关注？带着这个疑问，书稿断断续续写了一年多，到今天十章全部起草修改完毕后，心中逐渐有了一个相对清晰的认识。

狭义地讲，知识图谱是由谷歌公司首先提出，被互联网公司用来从语义角度组织网络数据，从而提供智能搜索服务的大型知识库。形式上，它是一个用图数据结构表示的知识载体，描述客观世界的事物及其关系，其中节点代表客观世界的事物，边代表事物之间的关系。在具体实现上，知识图谱用语义网（Semantic Web）中的资源描述框架（Resource Description Framework, RDF）对知识体系和实例数据两个层面的内容进行统一表示，共同构成一个完整的知识系统。扩展开来，知识描述、实例数据及其相关的配套标准、技术、工具及应用系统构成了广义的知识图谱。

让我们从人工智能和语义网两个领域追溯知识图谱的发展轨迹。在人工智能发展的第一个十年，科学家们的目标集中在如何构建推理模型进行问题求解，后来逐渐认识到领域知识在问题求解过程中占有不可或缺的地位，于是在 20 世纪六七十年代陆续提出了语义网络、框架、脚本等一系列知识描述理论，其中语义网络（Semantic Network）是一个通过语义关系连接的概念网络，节点代表概念，边则表示概念之间的语义关系。这种图结构知识表示方法与知识图谱一脉相承。基于这些知识描述理论，领域专家开始利用人工方法编写实例数据建立知识库（例如 Cyc 项目），但是规模非常有限，只在某些很受限的领域取得成功。互联网时代，人类在与自然和社会的交互中产生了异常庞大的数据，它们以文字、图片、音频、视频等各种模态存在，如何让计算机自动阅读、分析、理解这些海量、繁杂乃至泛滥的数据，从中挖掘有价值的信息，为用户提供精准的信息服务，是构建下一代信息服务的核心目标之一。为了实现这一目标，万维网之父 Tim Berners-Lee 于 2001 年提出了语义网（Semantic Web）的概念，其本质是：定义一种对客观世界进行描述的概念化规范（即本体），基于本体并通过一套统一的元数据规范对网络内容进行详细的语义标记，从而赋予万维网信息以含义，将网页互联的万维网转化为内容互联的语义网。维基百

科（Wikipedia）无疑是语义网的一种杀手级应用，它以一种亿万网民协同构建的方式建立百科全书，极大地促进了知识资源的快速成长，在知识类型、覆盖范围和规模上都达到前所未有的程度，为知识图谱的诞生起到了决定性作用并为它奠定了雄厚的资源基础。2012 年 5 月，谷歌首次提出知识图谱的概念，很快互联网巨头们纷纷跟进，构建了自己的知识图谱，包括微软 Probase、百度知心、搜狗知立方等。各个行业也在探索建立垂直领域的知识图谱，以知识赋能提升金融、医疗、司法、教育、出版等各个行业业务的智能化水平。学术界则致力于研究各种知识图谱构建和应用的自动化方法。回顾知识图谱的发展过程，以语义网络为代表的知识表示理论研究的积淀，互联网智能化信息处理的迫切需求，语义网在标准、技术、工程和应用方面的实践，以及以 Wikipedia（维基百科）为代表的网络协同构建知识资源的迅猛发展，共同推动知识图谱成为新一代人工智能的一个极具代表性的奠基性工作。

总结而言，知识图谱之所以成为学术界和产业界共同关注的热点，主要是由于它的以下几个特点：

（1）知识图谱是人工智能应用不可或缺的基础资源。知识图谱在语义搜索、问答系统、智能客服、个性化推荐等互联网应用中占有重要地位，在金融智能、商业智能、智慧医疗、智慧司法等领域具有广阔的应用前景。

（2）语义表达能力丰富，能够支持很多知识服务应用任务。知识图谱源于语义网络，是一阶谓词逻辑的简化形式，并在实际应用中通过定义大量的概念和关系类型丰富了语义网络的内涵。它既能够描述概念、事实、规则等各个层次的认知知识，也能够有效组织和描述人类在自然环境和社会活动中形成的海量数据，从而为各类人工智能应用系统奠定了知识基础。

（3）描述形式统一，便于不同类型知识的集成与融合。本体（Ontology）和分类系统（Taxonomy）是典型的知识体系载体，数据库是典型的实例数据载体，它们的描述形式截然不同。知识图谱以资源描述框架（RDF）的形式对知识体系和实例数据进行统一表示，并可以通过对齐、匹配等操作对异构知识进行集成和融合，从而支撑更丰富、更灵活的知识服务。

（4）表示方法对人类友好，给以众包等方式编辑和构建知识提供了便利。传统知识表示方法和描述语言需要知识工程师具备一定的专业知识和技能，普通人群难以操作。知识图谱以实体和实体关系为基础的表示形式，无论是专家还是普通民众都能够接受，这给以众包等方式编辑和构建知识提供了便利，为大众参与大规模知识构建提供了低认知成本的保证。

（5）二元关系为基础的描述形式，便于知识的自动获取。知识图谱对各种类型的知识采取统一的二元关系进行定义和描述，给基于自然语言处理和机器

学习方法进行知识的自动获取提供便利，从而为大规模、跨领域、高覆盖的知识采集提供了技术保障。

（6）表示方法对计算机友好，支持高效推理。推理是知识表示的重要目标，传统方法在进行知识推理时复杂度很高，难以快速有效地处理。知识图谱的表示形式以图结构为基础，结合图论相关算法的前沿技术，利用对节点和路径的遍历搜索，可以有效提高推理效率，极大降低计算机处理成本。

（7）基于图结构的数据格式，便于计算机系统的存储与检索。知识图谱以三元组为基础，这使得它在数据的标准化方面更容易推广、相应的工具更便于统一。结合图数据库技术以及语义网描述体系、标准和工具，知识图谱为计算机系统对大规模知识系统的存储与检索提供技术保障。

当然，我们应该冷静地看到：知识图谱只是知识工程发展进程中的一个节点，知识工程还有很长的路要走。目前所定义的知识图谱只能表示事实性的知识，或者称为以实体为核心的结构化知识。知识还有很多类型：常识知识、场景知识、事务知识、情感知识等，这些知识如何表示、构建与应用还处在不断的探索和研究当中，还需要我们付出更加艰辛的努力。

我从2002年开始指导研究生开展信息抽取和问答系统方向的研究工作，在实体识别、实体消歧、关系抽取、事件抽取、知识表示学习、知识推理、问答系统、对话系统等多个方向做了较为系统的研究，同时，通过与多家应用单位合作，在百科知识图谱和行业知识图谱构建及应用方面进行了大量工程实践。本书是在这些研究成果的基础上，充分参考吸收国内外代表性的研究工作，以知识图谱为核心，以知识建模、知识构建、知识融合、存储和检索、知识推理以及知识服务等知识图谱生命周期的各个阶段为线索进行系统的梳理撰写而成。本书属于教材性质，首先对基本问题、基本方法和典型工作进行系统的介绍，并配以大量实例使读者具有感性的认识。在此基础上，本书也属于导论性质，紧密联系自然语言处理、知识工程、语义网、机器学习等多个与知识图谱密切相关的研究方向，试图从历史发展的角度、以更宽广的视角去梳理相关内容，以激发读者的研究兴趣。

全书由赵军主持撰写，赵军、刘康、何世柱、陈玉博共同编写，写作过程中参考了我指导的多位博士研究生的学位论文，主要包括韩先培、何世柱、陈玉博、张元哲、刘树林、魏琢钰、曾道建、齐振宇、来斯惟、纪国良、刘洋、张涛、吴友政、刘非凡等。刘树林、魏琢钰、纪国良、曾道建、王雪鹏等参加了材料准备工作，李文婷、刘操、张翔、郭尚敏、曾祥荣、刘健、杨航、左新宇、白桂荣等参加了书稿校对工作。初稿完成后，在中国科学院大学讲授了“知识图谱导论”研究生课程，根据同学们的反馈意见对书稿进行了修改和完善，形

成了最终的版本。

在本书选题过程中微软亚洲研究院副院长周明博士提出了建设性意见；清华大学计算机系李涓子教授审阅了全书，提出了宝贵的修改意见。高等教育出版社在出版过程中给予了大力支持和帮助。在此，谨向他们表示诚挚的谢意。最后，衷心感谢多年来给予我支持和帮助的各位师长、同事、同仁和朋友们，恕不一一列举，本书的研究成果与他们的大力支持是分不开的。

知识图谱是一个新兴的研究方向，发展快，涉及面广。我水平有限，书中肯定有不少疏漏、不妥甚至是错误的地方，恳请读者批评指教。

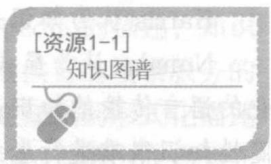
赵军

2018 年 3 月于北京中关村

目　录

第一章 概 述

1977 年，在第五届国际人工智能会议上，美国斯坦福大学计算机科学家爱德华·费根鲍姆发表了特约文章《人工智能的艺术：知识工程课题及实例研究》①，系统地阐述了“专家系统”的思想，并提出了“知识工程”的概念，从此知识工程成为人工智能的一个重要研究方向。在随后的 30 多年，研究人员提出一系列知识表示、构建和应用的理论和方法，基于人工构建知识库的专家系统在多个具体应用领域取得成功，稳步推动了知识工程学科方向的发展；另一方面，由于依赖人工构建知识库，知识工程也遇到知识瓶颈的严重挑战。2012 年 5 月，谷歌发布了新一代知识搜索引擎，可以展示与关键词所描述的实体或概念相关的人物、地点和事件等信息[资源 1-1]。例如：对于关键词“中科院自动化所”，知识搜索引擎将它理解成一个实体“中国科学院自动化研究所”，并展示成立时间、地理位置、主管部门等基本信息，以及“中国科学院计算技术研究所”“中国科学院软件研究所”“中国科学院信息工程研究所”等相关实体，从而让使用者更加便捷地发现新知识。支持这种知识搜索功能的是一个称为知识图谱（Knowledge Graph）的基础设施，它是从 Wikipedia（维基百科）抽取出来的、规模巨大的、以相互关联的实体及其属性为核心的知识网络。知识图谱丰富的语义表达能力和开放互联能力，为计算机理解万维网的内容以及万维网知识互联打下了坚实的基础。一时间互联网巨头们纷纷跟进，构建了自己的知识图谱，包括微软 Probase、百度知心、搜狗知立方等，而学术界也倾力研究各种知识图谱构建和应用的方法，知识工程再次成为人工智能领域的研究热点。



① “The Art of Artificial Intelligence: Themes and case studies of Knowledge Engineering”.