知识图谱：方法、工具与案例

[奥] 迪特·芬塞尔(Dieter Fensel) 等著
郭 涛 译

清華大學出版社
北 京

北京市版权局著作权合同登记号　图字：01-2023-1868

图书在版编目(CIP)数据

知识图谱：方法、工具与案例 / (奥) 迪特·芬塞尔 (Dieter Fensel) 等著；郭涛译. —北京：清华大学出版社，2023.6
书名原文：Knowledge Graphs
ISBN 978-7-302-63463-8

I. ①知… II. ①迪… ②郭… III. ①知识管理－研究 IV. ①G302

中国国家版本馆 CIP 数据核字(2023)第 084357 号

责任编辑：王　军
装帧设计：孔祥峰
责任校对：成凤进
责任印制：朱雨萌

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市春园印刷有限公司
经　　销：全国新华书店
开　　本：148mm×210mm　　**印　　张**：4.75　　**字　　数**：138 千字
版　　次：2023 年 7 月第 1 版　　**印　　次**：2023 年 7 月第 1 次印刷
定　　价：59.80 元

产品编号：099591-01

译者简介

郭涛，主要从事模式识别与人工智能、智能机器人、软件工程、地理人工智能(GeoAI)和时空大数据挖掘与分析等前沿交叉研究。担任《深度强化学习图解》《AI 可解释性(Python 语言版)》《概率图模型及计算机视觉应用》等畅销书的译者。

译 者 序

知识图谱是一种图数据模型，以图的形式展现实体、实体属性以及二者之间的关系。知识图谱是一个典型的交叉技术领域，涉及人工智能、机器学习、计算机视觉、数据库、自然语言处理、物联网和互联网等技术。知识图谱并不是一个新的领域，而是一个由语义网络、本体论和知识工程与专家系统等随着时代变迁衍生出的领域，是人工智能符号主义流派的主要代表，主张智能的实现能模拟人类的心智，使用计算机符号记录人脑的记忆。2012 年，Google 在解决搜索引擎时正式推出知识图谱；知识图谱天生是科学和技术的产物，含有系统工程的基因。该领域的技术要素主要涉及知识图谱的表示、存储与查询、获取与构建、融合、推理、问答和分析等方面。在学术界和工业界，基于此技术要素组合而成的知识图谱项目井喷式增长，相关企业纷纷建立了自己的知识图谱课题和项目。

知识图谱开创了人工智能的新范式，以数据驱动和知识驱动相结合，开启了下一代人工智能，实现了人与人、人与机器、机器与机器的协同协作。此外，知识图谱突破了传统的人工智能研究领域，从广泛的文本、结构化、视觉和时序等多模型数据中提取知识已成为知识图谱发展的主要方向之一，多模态知识图谱的构建可深度融合并灵活运用显式符号知识和隐式数据知识。将深度学习、图深度学习、迁移学习与元学习深度融合是知识图谱的发展趋势，可用于全类型、高涵盖的大规模知识图谱构建，实现更精深的知识推理，是通往鲁棒、可解释的人工智能之路。

Dieter Fensel 是语义网络研究的先驱之一，本书是其团队在知

识图谱领域的主要成果之一。本书共 5 章，主要讨论了知识图谱的整个生命周期，知识图谱的概念、构建、实现、维护和部署、技术架构和未来工作的方向，可作为知识图谱、模式识别与人工智能和计算机视觉等方面的科学家、工程师的参考用书。

在翻译本书的过程中，我得到了很多人的帮助。对外经济贸易大学英语学院研究生许瀚、吉林大学外国语学院研究生吴禹林和吉林财经大学外国语学院研究生张煜琪对本书进行了审阅。最后，感谢清华大学出版社的编辑，他们进行了大量的编辑与校对工作，保证了本书的质量，使其符合出版要求。

由于本书涉及的广度和深度较大，加上译者翻译水平有限，在翻译过程中难免有不足之处。若各位读者在阅读过程中发现问题，欢迎批评指正。

郭 涛

推　荐　序

为帮助读者更好地理解知识图谱的作用，下面讲述一个故事，故事的主人公是 Manuel Sahli。

2005 年，Manuel Sahli 成为维基百科的撰稿人，并作出了重要贡献。他情系家乡——美丽的瑞士城市温特图尔。多年来，Manuel 精心编辑并持续更新一篇介绍温特图尔的文章，友好城市列表就是他的贡献之一。

大约在同一时间，另一位撰稿人在介绍加利福尼亚安大略省的文章中将温特图尔加入友好城市列表。安大略省位于洛杉矶以东，距洛杉矶约一小时车程。虽然友好城市理应反映的是某两个城市之间的相互关系，但这一信息从未出现在 Manuel 介绍温特图尔的文章中。

Manuel 虽然对有关温特图尔的所有事物都感兴趣，但在近十年，他都不曾知道这一增添内容。

2012 年，维基百科的姊妹项目 Wikidata 推出，这是一个任何人都可以编辑和使用的知识图谱。一段时间后，一位撰稿人使用维基百科上的安大略文章，在 Wikidata 上添加了安大略省的友好城市列表。

此后，Manuel Sahli 被选为温特图尔市议会议员。Manuel 想用 Wikidata 制作一张温特图尔的友好城市的地图，而 Wikidata 有一个强大的查询界面，允许在知识图谱上显示查询结果。令 Manuel 大吃一惊的是，地图上呈现了一个他不熟悉的地方：安大略省。

起初，Manuel 认为这是一个尚未被发现的破坏行为，但这一事实却源自安大略省政府的官方网站。Manuel 也的确在该网站上找到

了温特图尔！作为市议会议员，他询问温特图尔市政府是否知道这一信息，对方都给出了否定答案。

Manuel 给安大略市政府写了一封正式信函，说明了相关情况。其后，出乎意料的是，安大略市政府给出了双方签署的文件，证明两市于 1982 年建立了友好城市关系。温特图尔市档案馆现已根据这些文件和准确日期找到相关文件。多亏了 Wikidata，温特图尔发现了它失散多年的第五个友好城市；从此，两个城市再续前缘，重新建立了友好关系。

没有人试图向 Manuel 或任何人隐瞒这一事实。温特图尔市档案馆存有相应文件(只是起初没有找到)；维基百科公布这一信息已长达近十年的时间；安大略省的官方网站上也有这一信息。然而即使 Manuel 对他的家乡和维基百科投入了很多精力，也并不了解此事。只有当信息被添加到 Manuel 可访问且知道如何查询的知识图谱时，真相才浮出水面。

将知识隐藏在文档和大量自然语言中太容易了，而在适当时候将其公开却很难。使用知识图谱，将能更方便地索引、处理和查找事实。

然而，Manuel 不仅有足够的兴趣发现这一事实，且能使用 SPARQL 查询语言查询 Wikidata，这完全是侥幸。大多数人其实都不知道该怎么做。

我们正在见证计算机使用方式的范式转变：通过自然语言与计算机交互的新生能力。这一转变的重要性不亚于以往分时系统、图形用户界面、Web 和智能手机的出现。对话界面已经存在了很长时间，就像任何新技术在被广泛采用之前都会先应用于专业领域一样。然而，所谓的智能助手正在迅速普及和发展，从手表到汽车，从电视到耳机，越来越多的设备都具备了听从命令和回答问题的能力。

越来越多的人使用智能助手获取知识来完成日常任务，且人数正在以惊人的速度增长。无论问题并不重要还是足以改变生活，为解决当务之急，我们将不再需要办公室、学校或图书馆中的传统计算机。没有人需要学习如何使用最新的智能手机功能，如何使用鼠

标，甚至如何阅读和打字，因此，一个更具包容性的世界展现在人们眼前。

这种新的范式需要新的方式使人们加入新的数据空间，使潜在客户和消费者访问数据和服务，使人们发现并享受新方式所提供的东西。如果越来越多的人不使用应用程序和网站预订酒店、租车、购买机票、订午餐和查询信息，我们就需要确保了解用户的意图，并以允许集成和组合的方式提供数据和服务。这不是对未来的展望，而是我们已经建立的世界。不计其数的功能和服务已集成到 Google Assistant、Apple Siri、Amazon Alexa、开源 Mycroft、Microsoft Cortana、Samsung Bixby 中。

将理解用户查询所需的自然语言技术和在大型知识图谱中构建事实的能力相结合，我们就能在终端用户需要时为其传递知识。人们将能探询世界，能追随好奇心，能够学习。企业能将其服务和产品扩展到许多新客户。

本书对架构以及了解智能助手世界的步骤进行实践性概述，指导我们构建、发展并维护知识图谱。本书讨论如何用知识图谱表示组织中的重要事实以及如何使这些事实对其他供应商可用。

Dieter Fensel 是很早就认识到语义技术的潜力以及将服务组合成新颖界面的必要性的人士之一。他和他的研究小组在服务的语义描述、链接数据应用、本体工程以及学习方法和工具等领域处于领先地位。本书由 Dieter Fensel 和其团队研究人员共同编写，总结了他们多年的研究经验。

享受阅读这本书的过程吧！本书将帮助你更好地理解知识图谱的新领域以及使用方法。本书将使你和你的组织进入一个新世界。在这个世界中，你的档案和系统中已有的知识不会再被长久隐藏，你和用户都将能使用已有的知识，用户将获得更大的能力，你将获得新的收入来源，且再也不会失去一个“友好城市”。

Denny Vrandečić

作于美国加利福尼亚州旧金山

致　　谢

感谢 Andreas Harth 和 Aidan Hogan 对本书进行了富有成效、大有裨益的讨论和评论。感谢 MindLab 项目的所有参与者，本书总结了他们工作的精髓。[1]此外，我们与 2019 年 STI 知识图谱峰会的与会者(尤其是 Mark Musen、Juan Sequeda、Rudi Studer 和 Sung-Kook Han)进行了讨论，获得了极大帮助，受益匪浅。[2]最后，感谢 Andreas Lackner 作为探路者，帮我们在(蒂洛尔)电子旅游迷宫中寻找出路。

1 见[1]。
2 见[2]。

前　言

Alexa 和 Google Home 等智能音箱将人工智能(Artificial Intelligence，AI)引入了数百万乃至数十亿个家庭，使人工智能渗透到人们的日常生活。现在，人们足不出户，或不使用电脑，也可查找信息、订购产品和服务。我们只需要和一个盒子说话，这个盒子会执行所需的任务，非常方便。这些新的沟通渠道为成功的电子营销和电子商务带来了新挑战。仅运行一个带有许多彩色图片的传统网站已不是最先进的技术了。如今的 Web 甚至也在通过应用 schema.org 进行自我改造。数据、内容和服务都有了语义标注，这使得软件智能体(即所谓的机器人)通过 Web 搜索了解其内容。需要人类浏览大量网站、手动提取并解释信息的时代已经过去。如今，用户通过他们的个人机器人来查找、聚合和个性化信息，还能预留、预订或购买产品和服务。因此，对于信息、产品和服务的供应商而言，在这些新的在线渠道中占据地位变得越来越重要，这可帮助他们确保未来可持续获益。本书讨论了有助于实现这些目标的方法和工具，核心是应用和开发内容、数据及服务的语义标注(可由机器处理)，以及它们在大型知识图谱中的聚合。只有这样，机器人才能不仅可理解问题，而且拥有渊博的知识以回答问题。

这些知识图谱，尤其是基于 schema.org 的知识图谱，在基于 Internet 的信息搜索中发挥着越来越重要的作用，已成为成功电子商务和电子营销的关键技术，并对与客户在线互动的经济部门的价值分配产生关键影响。知识图谱是另一种大规模的可扩展数据集成方法，而且很可能不是解决这一难题的最后一种方法。然而，这是我

们第一次在全局范围内研究这个问题。在本书中，我们描述了一些方法和工具，这些方法和工具使信息供应商能构建和维护此类知识图谱。具体将介绍以下几个方面：

- 用于手动、半自动、自动构建和验证语义标注，并将其集成到知识图谱的方法和工具。
- 用于实施知识图谱的方法和工具。
- 基于生命周期的半自动和自动管理知识图谱的方法，包括评估、纠正和利用其他静态和动态资源蕴含的知识的方法。

仅是知识本身还不够，还必须可作为问题的潜在答案和对话的指导。

- 电子营销：通过推理方法和工具，我们可从知识图谱中为特定任务和领域得到基于对话的机器人。
- 电子商务：基于服务和产品的语义描述，可设计一个面向目标的对话，改进预留、租用、预订或购买商品和服务的过程。

为说明这些方法的实际用途，我们讨论了几个试点，重点是电子旅游领域。旅游业是全球最大的垂直行业之一，具有巨大的增长潜力。此外，旅游业也是欧洲前途良好的垂直领域之一，价值分配关键取决于在电子营销和电子商务方面的适当能力。潜在客户遍布世界各地，服务供应商也是分散的，大多数是小型企业(例如，蒂罗尔州的数万家小型家庭旅馆)。总的来说，我们关注以下几个方面：

- 集成来自开放、专有、异构和分布式源的内容、数据和服务描述。
- 高效地维护数据上下文(如出处、地理有效性和时间有效性)。
- 使用知识图谱引导对话。
- 集成静态源和动态源。
- 集成语义网络服务以促进操作和自动服务调用。

本书结构如下所述。第 1 章提供知识图谱的定义。我们的目标

不是追求数学精度，而是试图涵盖有关其影响的各种方法。第 2 章详细介绍如何构建、实施、维护和部署知识图谱。第 3 章介绍可使用此类知识图谱构建的相关应用层。解释如何使用推理来定义此类图谱上的视图，使其能在开放和面向服务的对话系统中发挥作用。实践是最好的检验标准。第 4 章阐述知识图谱在电子旅游领域的应用以及在其他垂直领域的用例和试点。第 5 章对全书进行总结，并指出未来的研究方向。附录 A 为领域规范引入一种抽象语法和语义，用于使 schema.org 适应特定的领域和任务。

下载 Reference 文件以及 Links 文件

在阅读本书时，会不时看到诸如“(Newell 1982)”的引用内容；你可扫描封底二维码，下载 Reference 文件，从 Reference 文件查看引用详情。

本书脚注中多处涉及网页链接，形式是[*]，即方括号中间加数字编号。你可扫描封底二维码，下载 Links 文件，从 Links 文件查找与[*]对应的网页链接。

目　　录

以下内容可扫描封底二维码下载

第 1 章
引言：什么是知识图谱?

对于当今的许多企业而言，知识图谱至关重要：它提供结构化数据和事实知识，驱动许多产品发展并使其更加智能神奇(Noy et al. 2019)。

摘要 自从知识图谱由 Google 创立以来，这一术语近年来逐渐被广泛使用，却没有明确的定义。本章试图通过汇编文献中的现有定义，并结合先前为解决我们今天面临的数据集成挑战所付出的努力的独特性，来得出知识图谱的定义。我们将通过实证来调查现有知识图谱。本章提供了对某些概念和构建知识图谱动机的共同理解，为本书的其余部分奠定了基础。

1.1 引言

Alexa 和 Google Home 等智能音箱将基于人工智能(AI)的通信方式引入了数百万乃至数十亿的家庭，如今的 Web 甚至也在通过应用 schema.org[1]进行自我改造(Guha et al. 2016)。数据、内容和服务都有了语义标注，这使得软件智能体(即所谓的机器人)[2]通过 Web 搜索了解其内容。因此，对于信息、服务和产品供应商而言，在这

1 见[1]。
2 见[2]。

些新的在线渠道中占据有利地位变得越来越重要，这帮助供应商确保未来可以持续获益。

开发这种自动语音识别方法是开发自动对话系统的重要前提[3]。在自动语言理解方面的突破是依靠大数据[4]和机器学习(Goodfellow et al. 2016)完成的。但回答查询或运行面向目标的对话则需要更多技术。为给出有意义的答案，智能体需要知识。因此，Google 于 2012 年开始开发所谓的知识图谱[5]，这一知识图谱应该包含在 Web 或其他数据源中被语义标注的重要人类知识。同时出现了与知识图谱相关的炒作[6]。因此，我们有必要更好地理解知识图谱。我们使用互补方法处理这个问题。首先，我们尝试通过分析知识图谱的基本原理给出一个概念性答案。其次，我们对现有知识图谱进行了实证调查。

1.2 知识图谱的概念性定义

Ehrlinger 和 Wöß(2016)对知识图谱的潜在定义进行了有效简明的概述，说明了定义的变体。他们还基于本体和推理机获取新知识，为知识图谱添加了一个新定义。从我们的角度看，这个定义兼容性较差，过于关注特定方法。我们应该从知识图谱的潜在定义开始。从解释学上讲，可首先区分构成这个概念的两个术语，因为两个术语完全不相关。

“图是一种结构，相当于一组对象，其中一些对象对在某种意义上是相关的”[7]。严格地说，我们需要稍微将此定义扩展到多集，因为同一个对象在语法和语义上可在图上多次出现。规范化可解决这个问题，但这意味着要使用某些特定的处理技术。这个简单的定

3 见[3]。

4 见[4]。

5 A.辛格尔：介绍知识图谱、事物，而不是字符串。博客文章位于[5]。

6 仅以几本书为例(Chen et al.2016;Croitoru et al.2018; d’Amato 和 Theobald 2018; Ehrig et al.2015; Li et al.2017; Pan et al.2017a, b; Qi et al., 2020; Qi et al. 2013; Van Erp et al. 2017)。另见 Bonatti et al.(2019)。

7 见[7]。

义可向各个方向扩展，我们最终得到许多种图：简单图、无向图、有向图、混合图、多重图、Quiver、加权图、半边图、松边图、有限图与无限图等[8]。这令人想起 20 世纪信息系统领域的情况，当时“每个”新博士都为 Petri-Net 引入了一个新变体[9]：标记、着色、分级等。引用维基百科的介绍：“Petri 网还有更多扩展，但重要的是要记住，随着 Web 在扩展属性方面的复杂性增加，使用标准工具评价 Web 的某些属性就越来越难。因此，对于给定的建模任务，最好使用最简单的 Web 类型。”[10]在语义网络领域，通常使用 RDF 表示知识图谱[11]。但仔细研究后会发现，通常情况下，RDF[12]更多地用于语法方面而非认识论或语义方面。例如，RDF 仅提供全局属性定义，人们通常希望将属性定义为具有特定范围的属性来定义它们的概念[13]。同理，RDF/OWL 语义将范围限制解释为一种推断新知识的方法，而不是使用它们检测类型错误。RDF/OWL 语义不一定是 schema.org[14,15]中范围定义的预期含义。

“知识”的概念有点模糊。我们回到 Newell(1982)所说的知识水平。假设智能体遵循理性原则；该原则后来细化为有限理性概念(Simon，1957)，包括“最优”决策的成本。基于该假设，我们认可知识，理解知识为实现某些目标而采取的行动。从这个意义上讲，知识由观察者从外部分配给这一智能体。在内部，知识在符号级别被编码。

知识层之下是符号层。知识层是面向世界的，即知识层涉及智

8 见[8]。参见 Angles 和 Gutiérrez (2005)关于图和模型的综述。

9 见[9]。另见 Reisig(2013)。

10 见[10]。

11 请参阅 Angles 和 Gutiérrez (2008)有关图数据库的概述。

12 见[12]。

13 Hayes (1981)和 Patel-Schneider (2014)。Guha et al. (2016)已通过多态性扩展了 RDF，schema.org 中的大量子属性试图避开 RDF 的这一瓶颈。另见 Patel-Schneider(2014)。

14 见[14]。

15 假设 human 类和 birthPlace 属性具有范围 City。如果 RDF/OWL 推理机发现语句{birthPlace(domain:Human,range:City, human (Hans),birthPlace(Hans) “Austria”}，则推断 Austria 是一个城市，而不会出现识别范围错误。另见 De Bruijn et al.(2005)、Patel-Schneider 和 Horrocks (2006)。

能体操作所在的环境，而符号层是面向系统的，因为符号层包括智能体可用于操作的机制。知识层使智能体的行为合理化，而符号层使智能体行为机械化(Newell 1982)。

可用类似的思路解释知识图谱。智能体通过解释图来拥有/生成知识，即将其元素与所谓的现实世界对象和动作相关联。此外，图是一种特定的编码形式。为进一步完善定义，我们可能希望将图置于逻辑或认识论层面，而不是实现层面(Brachman 1979)。在实现层面，我们掌握基于图的数据库等方法。

语义网络是 20 世纪 60 年代和 70 年代流行的知识表示形式(Brachman 1990；Sowa 1992)。一方面，将知识图谱与语义网络区分开来并不简单。[16]另一方面，二者明显的区别在于它们的实际规模以及知识图谱的潜在影响。规模很重要，数量也可产生新品质，即使没有人能确定新品质产生时的确切数字(Hegel 1812)也同样如此。与此相关的一个问题是区分知识库和知识图谱。根据 Akerkar 和 Sajja(2010)所言，基于知识的系统由两部分组成：包含知识的知识库，和可用于推理新事实或在知识库上回答问题的推理引擎[17]。知识库的另一个重要特征是 ABox 和 TBox 之间的区别，参见 Brachman 和 Schmolze(1985)。断言框(assertion box，ABox)包含一组断言/事实陈述。严格分离的术语框(terminological box，TBox)定义了 ABox 语句使用的术语，并使用该术语添加了更通用的逻辑公式。TBox 提供了无限多潜在附加事实的内涵定义(由推理引擎得出)。知识图谱可能具有完全不同的架构和结构。逻辑公式缺失，术语“知识”与“断言”被存储在同一层。也就是说，只是几个额外断言，有些图甚至在没有这些额外断言的情况下工作。这将导致两个后果：

(1) 知识图谱几乎无法用于理解其他信息。在基于 schema.org 的知识图谱中，通常是一个简单工具(如 Google 的结构化数据测试

16 见[16]。

17 这些事实并非真正意义上的“新”，而是通过使用逻辑公式从内涵定义推导出来的，其中逻辑公式隐含内涵定义。

工具[18])用于强制约束有效图谱。

(2) 可使用严格定义的模式来定义用户界面，以确保数据质量(正确性和完整性)，并允许优化存储、查询、维护和交易。这些是关系数据库成功的关键因素(Codd 1970)。然而，当需要集成来自各种半结构化、异构和动态源的数据时，这种严格的模式定义也会带来严重的瓶颈。严格定义的结构化、同质和稳定模式的假设被破坏，使此类数据集成无效且不可扩展。因此，知识图谱在这方面没有传统数据库或知识库严格，这不是一种错误，而是知识图谱的一个特点。

2012 年，Google 创建了知识图谱(Knowledge Graph)一词，从而为世界构建一个模型。同时，知识图谱也已成为产品和服务行业广泛讨论的术语。一般来说，旅游业[19]未必是最具创新性的领域，但其中每个主要参与者都有一个知识图谱，成千上万的参与者(例如目的地管理组织)都需要或想要一个知识图谱。这是因为成功的电子营销和电子商务对于旅游业等领域的价值分配变得越来越重要。一般来说，当前知识图谱的动态来自经济部门，而非学术界。因此，一个简单的知识图谱定义可能是：知识图谱是一个用于描述和指导当前的数据(和服务)集成问题的流行术语。这不是解决“知识图谱定义”难题的第一种方法，也不会是最后一种。然而，这是我们第一次在全球范围内解决这个问题。

知识获取瓶颈(参见 Feigenbaum 1984；Hoekstra 2010)导致了大约 40 年前的 AI 寒冬[20]。看到数百万人破解知识图谱(即语义网络)中的数十亿条语句来解决 AI 寒冬问题是一种有趣的经历(Fensel 和 Musen 2001)。如果我们在 1980 年前后提出这个解决方案，人们会立即将我们赶出大学。即使是 Douglas Lena(Lenat 和 Guha 1989)提出的依靠 50 名员工生成解决方案的愿景，与这个解决方案相比也显

18 见[18]。

19 请注意，旅游业是全球范围内最重要的经济垂直领域之一，占 2017 年全球 GDP 和总就业人数的 10%左右(WTTC 2018)。

20 见[20]。

得微不足道。同时，我们有一个基于 schema.org 语义网络，托管超过 380 亿条语义语句，用于超过 12 亿个 Web 站点。[21]

尤其令人惊奇的是，这些操作是在事实层面上完成的，而非通过捷径使用一阶逻辑表达式表达潜在数十亿条语句完成的。知识库是通过写下数十亿个事实建立的。列举事实的方式就像扩展描述一个集合，而非采用集中方式。就像用归纳法来代替证明，先证明 1，然后证明 2、3，以此类推。也许这只是由于知识图谱处于早期发展阶段？也许将由机器接管这项任务。不幸的是，机器可能不会使用逻辑，并可能产生我们不再能理解的捷径。同样，机器和深度学习方法将越来越多地用于知识图谱的构建、改进和丰富(Paulheim 2018a)。考虑到规模、异质性、半结构化特征和速度，LarKC 项目旨在语义网络上应用推理，该项目已深入研究了各种组合归纳和演绎技术的方法；可参见 Fensel 和 van Harmelen(2007)、Fensel et al.(2008)。传统逻辑的基本假设具有小公理集、100%正确性、100%完整性以及知识静态特征，在 Web 或大型知识图谱规模下被打破。事实上，Wahlster 在他的最近一次演讲中指出，如何适当结合归纳方法、演绎方法和基于知识的方法是未来人工智能研究的一个关键挑战。[22]这一点尤其适用于物联网的研究。物联网是由物理设备、车辆、家用电器等其他物品组成的网络，这些物品嵌入电子设备、软件、传感器、制动器以及连接性，连接性使这些物品能连接、收集和交换数据，并开始在虚拟世界以外发挥作用。[23]

总之，知识图谱是一个非常大的语义网络，集成了各种不同的信息源，以表示关于某些领域的知识。下面将讨论如何构建和使用

21 Web 数据共享——RDFa、微数据、嵌入式 JSON-LD 和微格式数据集，2017 年 11 月。见[21]。Guha et al.(2016)称，在 100 亿个 Web 样本中，有 31%的 Web 正在使用 schema.org。最新的抓取报告显示，超过 300 亿个 quad 发现超过 37%的语义标注网站，包含大约 32 个付费级别的域。见[21]。

22 Wahlster 教授在 30 Jahre DFKI—KI für den Menschen 活动上的主题演讲：30 Jahre DFKI—Von der Idee zum Markterfolg，柏林，2018 年 10 月 17 日。

23 见[23]。

知识图谱。[24]

1.3　知识图谱的实证定义

下面概括介绍开放和专有知识图谱的定义和重要特征图谱(参见 Paulheim 2017)。[25]方法按照其首次发布年份排序。

1.3.1　开放知识图谱

对语义网络和链接数据的研究导致许多开放数据集最终组成了链接开放数据云(参见 2.5 节)。现在，这些数据集大多被重新命名为知识图谱。这些知识图谱通常是跨领域的。许多开放知识图谱都来自维基百科，因为维基百科涵盖了多个领域中的大量事实知识。一些知识图谱在构建过程中也受益于非结构化语料库、词典、本体和众包。在本节中，将介绍一些著名的开放知识图谱。

DBpedia[26](Auer et al. 2007；Lehmann et al. 2015)是在 2007 年首次发布的知识图谱，是语义网络真正的中心数据集(也称为“核”)，因为它链接到其他许多数据集。该知识图谱主要通过提取器从维基百科页面(主要是信息框)上的结构化数据中提取，可调整提取器以提取不同类型的数据。DBpedia 知识图谱建立在众包维护的 DBpedia 本体(在 OWL 中指定)之上，该本体从维基百科的信息框元数据映射而来。该知识图谱按照语义网络标准，作为 RDF 转储和 SPARQL 端点发布。DBpedia 是定期发布的，但也提供与维基百科同步的实时端点。[27]2016 年 10 月发布的版本包含 130 亿个 RDF 三元组。DBpedia 本体包含 760 个类和大约 3000 个属性。[28]DBpedia 可应用

24　也可将其称为 Fensel et al.(1997)提出的知识网络。

25　可参见 Noy et al.(2019)了解有关 Bing (Microsoft)、eBay、Facebook、Google 和 IBM Watson 的知识图谱的详细信息。

26　见[26]。

27　见[27]。

28　见[28]。

于许多领域，如自然语言处理以及知识探索和丰富。

Freebase[29](Bollacker et al. 2008)是 2007 年推出的协作知识库。运营 Freebase 的公司在 2010 年被 Google 收购，此后，该知识库改进了 Google 的知识图谱。2016 年，Google 关闭了 Freebase，其知识已逐渐纳入 Wikidata。最新的转储包含 19 亿个事实。[30]Freebase 具有自定义数据模型，支持局部属性。

YAGO[31](Suchanek et al. 2007；Hoffart et al. 2013；Mahdisoltani et al. 2015)是另一个基于维基百科内容构建的知识图谱，于 2008 年首次发布。YAGO 融合了从维基百科文章中提取的实体与 WordNet 同义词集，以丰富类型层次结构。YAGO 和 DBpedia 之间的主要区别在于，YAGO 本体仅提取少量关系，并专注于保持知识图谱高度紧凑、准确和一致。YAGO 的形式使用 RDFS 的扩展版本。YAGO2 使用地理空间和时间信息改进初始知识库，YAGO3 专注于多语言，并集成了来自 Wikidata 的数据。YAGO 目前包含 1.20 亿个事实和 35 万个类。[32]

NELL[33](Carlson et al. 2010；Mitchell et al. 2018)基于 5 亿个 Web 站点，利用机器学习方法构建跨领域知识库。它使用预定义的初始本体来确定需要从 Web 中提取的关系类型。NELL 智能体持续运行，并通过创建新事实和删除过时错误的事实来不断改进知识库。NELL 智能体从 2010 年开始运行，知识库目前包含超过 270 万个事实。[34]NELL 知识库使用简单的基于框架的形式，可用作一个大的制表符分隔值文件。有研究人员将 NELL 数据模型映射到 RDF 和 OWL(Giménez-García 2018)。

29 见[29]。
30 见[30]。
31 见[31]。
32 见[32]。
33 见[33]。
34 最后访问时间：2019 年 8 月。NELL 提取了数百万个观点，但“事实”只是置信度值高于 0.9 的观点。

Wikidata[35](Vrandečić 和 Krötzsch 2014)是一个由社区策划的知识图谱，是维基百科的姊妹项目，于 2012 年启动。Wikidata 与 DBpedia 相似，都以维基百科的知识为基础，但 Wikidata 也可由社区成员编辑。事实上，Wikidata 和维基百科可能具有双向关系，这意味着来自 Wikidata 的事实也可用来丰富维基百科中的文章。Wikidata 的主要特点是其关注数据的出处和上下文。例如，一个城市的人口不仅以二元关系给出，而且带有上下文信息，例如“根据某个统计机构的测定结果”。Wikidata 有一个自定义的数据模型，支持限定符和上下文信息，但也提供到 RDF 和 OWL 的映射(Erxleben et al. 2014)。为访问知识图谱，Wikidata 提供了 RDF 转储和 SPARQL 端点；但由于数据模型支持 n 元关系，因此查询更复杂[36]。知识图谱包含超过 6500 万个实体。[37] RDF 表示在 2019 年 8 月包含超过 70 亿个三元组。[38]

KBpedia[39](Bergman 2018)知识库包含到维基百科、Wikidata、schema.org、DBpedia、GeoNames、OpenCyc 和 UMBEL 的映射。它的主要目的是通过为机器学习启用训练集生成等方式支持人工智能应用程序[40]。该知识库于 2016 年启动，于 2018 年开源。KBpedia 使用 OWL 2 形式发布，在其基础模型中包含 55 000 个概念、5000 个属性和 70 个基本分离的类型，以简化知识库的模块化。该知识库声称拥有 98%的 Wikidata 涵盖率，因此包含 4500 万个实例。[41]

Datacommons.org[42]是 Google 于 2018 年推出的开放知识图谱，集成了多个公共资源，主要包含关于地理和行政区域、人口统计数

35 见[35]。

36 可通过在“真实图”上编写查询来解决复杂性；见[36]中的示例。

37 见[37]。

38 COUNT SPARQL 查询的结果见[38]。关于语句数量的统计显示，2018 年 4 月，Wikidata 包含超过 4 亿个语句(Malyshev 2018)。这种巨大差异是因为 Wikidata 中的一个语句可用多个 RDF 三元组表示，因为需要具体化。

39 见[39]。

40 见[40]。

41 2.10 版本。

42 见[42]。

据以及其他公开可用数据(如天气和房地产)的知识。关于该知识图谱形式和大小的信息并不公开；但事实以三元组形式表示，每个事实都附有出处值。另一个显示其形式的迹象是，它使用 schema.org 词汇表来描述事物，并略微扩展了词汇表。该知识图谱不提供转储，但可通过浏览器界面和 Python API 访问知识图谱。

1.3.2 专有知识图谱

多个公司已经开发了各种知识图谱来支持客户的应用程序。在本节中，我们将简要介绍其中一些知识图谱及其用途。

Cyc[43](Lenat 和 Guha 1989；Lenat 1995)是持续时间最长的 AI 项目之一，也是一个常识性知识库，最初发布于 1984 年。Cyc 知识库包含 150 万个概念、2000 万个通用公理，以及针对医疗保健、交通和金融服务等领域的扩展。Cyc 的内容由 Cyccorp 策划，但在必要时会采用工具和方法从外部资源中检索知识。知识库用 CycL 语言表示，这是一种基于框架的语言，其表达能力超越一阶逻辑。[44] Cyc 借助微观理论(Guha 1991)，可进行有效推理，并避免大型知识库中可能出现的不一致。知识库包含从一般知识(如“原因在其影响出现时或出现前开始”)到高度领域特定知识(如股票价格)的各种知识。Cyc 知识库和支持应用程序是专有的；但可获得非商业用途的研究许可。2017 年前，Cyc 的一个片段以 RDF 格式发布作为 OpenCyc 供公众使用，但 2017 年后已不再发布。

Facebook 的 Entities Graph[45]由 Facebook 维护，并用于内部支持图搜索功能。Entities Graph 最初于 2010 年推出，包含有关 Facebook 用户的知识，即用户个人资料信息、兴趣和人际关系。该知识图谱可通过 Facebook Graph API 访问，包含 5 亿个事实(Noy et al. 2019)。

43 见[43]。
44 2019 年 2 月的技术报告，见[44]。
45 见[45]。

Google 知识图谱[46]最初于 2012 年推出，旨在改善 Google 搜索引擎结果，有效地将 Google 转变为问答引擎。2016 年 10 月，Google 宣布知识图谱拥有超过 700 亿个事实(Noy et al. 2019)[47]。有关该知识图谱的基础技术和形式的文档并不公开，但众所周知，标准 schema.org 类型可用于描述图中的事物。Google 在其知识图谱中集成了多个来源的数据，如维基百科、World Bank、Eurostat 等，还利用了 schema.org 的标注和来自 Web 的数据。Google 知识图谱目前也支持 Google Assistant。该知识图谱可通过 Google Knowledge Graph API 访问。

Yahoo!知识图谱[48](Blanco et al. 2013)于 2013 年推出。该知识图谱从异构源获取数据，并将其融合到一个通用的 OWL 本体下。融合过程包括知识的融合和清洗。知识图谱的核心是来自维基百科数据，然后通过 Music Brainz 和商业数据供应商等资源来充实数据。在知识图谱构建过程中，使用了机器学习方法和基于模板的方法。通用本体包含 300 个类和 1300 个属性。Yahoo!利用知识图谱来完成搜索、关系发现和自然语言处理等任务。

Knowledge Vault[49](Dong et al. 2014a)是 Google 收购的一个研究项目，旨在从不同类型的 Web 内容和数据中提取一个大型概率知识库。Knowledge Vault 融合了来自四种不同来源的知识，如非结构化文本、HTML DOM 树、HTML 表格和 schema.org 标注。提取的三元组针对现有的知识图谱(如 Freebase)进行验证，以提高事实的可靠性。在这项研究(Dong et al. 2014a)发表时，Knowledge Vault 包含超过 2.7 亿个事实。

46 见[46]。
47 见[47]。
48 见[48]。
49 见[49]。