知识工程技术及其应用

Knowledge Engineering Technology and Its Application

张春霞 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

知识工程技术及其应用

Knowledge Engineering Technology and Its Application

张春霞 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

内 容 简 介

知识工程是人工智能的重要分支领域。本书围绕知识工程技术及其应用，对领域本体、时间本体与时间信息抽取、实体识别、实体关系获取、实体属性知识获取、描述流抽取以及知识评估进行了详细的论述。全书共分 8 章，沿着从知识表示到知识抽取、知识评估这一主线逐步展开，围绕相关应用，由浅入深，阐明相关概念和核心方法，为知识工程领域的从业人员了解相关知识表示和知识获取技术提供参考。本书可作为知识工程、人工智能等专业本科生及研究生的教学参考用书，也可供研究院所从事知识工程、行业信息化领域的研发人员阅读参考。

图书在版编目（CIP）数据

知识工程技术及其应用 / 张春霞著. --北京 : 北京理工大学出版社, 2023.3

ISBN 978-7-5763-2237-8

Ⅰ. ①知… Ⅱ. ①张… Ⅲ. ①知识工程 Ⅳ. ①TP182

中国国家版本馆 CIP 数据核字（2023）第 056471 号

出版发行 / 北京理工大学出版社有限责任公司
社　　址 / 北京市海淀区中关村南大街 5 号
邮　　编 / 100081
电　　话 / （010）68914775（总编室）
　　（010）82562903（教材售后服务热线）
　　（010）68944723（其他图书服务热线）
网　　址 / http://www.bitpress.com.cn
经　　销 / 全国各地新华书店
印　　刷 / 保定市中画美凯印刷有限公司
开　　本 / 710 毫米 × 1000 毫米　1/16
印　　张 / 17
字　　数 / 294 千字
版　　次 / 2023 年 3 月第 1 版　2023 年 3 月第 1 次印刷
定　　价 / 98.00 元

责任编辑 / 多海鹏
文案编辑 / 把明宇
责任校对 / 刘亚男
责任印制 / 李志强

前　言

知识工程是一门以知识为研究对象的学科，是人工智能领域的重要研究内容。知识工程技术是利用知识工程原理与方法构建知识嵌入的人工智能系统的技术体系。知识工程技术主要研究知识表示、知识获取、知识推理、知识评估以及知识应用等核心技术和智能应用系统构建方法，是实现知识嵌入的人工智能的重要知识基础和技术支撑。开展知识工程技术研究，旨在利用现代计算技术为人们的生产生活提供精准的知识服务、提高知识获取的效率和质量，为各行业领域的信息化建设、智能化建设提供技术支撑。

本书作者在长期进行相关研究的基础上，对知识工程中的知识表示、知识获取、知识评估方法以及相关应用等进行全面论述。撰写本书的目的是为了呈现相关概念、技术思想、核心方法和相关应用，从而为语义检索、问答系统、信息推荐等下游任务提供知识支撑，为“知识”在多种人工智能系统中的应用提供方法论和技术支持。

全书共分 8 章。第 1 章主要介绍知识工程技术的研究背景，阐述知识工程技术的研究内容以及面临的挑战，并对本书内容进行梳理，给出本书的组织结构。

第 2 章以领域本体为切入点，从领域本体角度，阐述形式领域本体；从知识获取角度，介绍领域知识获取本体和模式本体；从领域本体应用角度，作为应用案例，论述考古学领域本体和数学课程本体。

第 3 章讨论时间本体和时间信息抽取。从本体表示角度，阐述时间本体；从时间信息抽取角度，论述时间实体识别方法；从本体应用角度，介绍时间本

体在问答系统中的应用。

第4章介绍实体识别。从知识获取对象角度，分别阐述领域概念获取方法、术语定义抽取方法以及领域术语抽取方法。

第5章论述实体关系知识获取。阐述实体上下位关系抽取方法，以及实体对齐关系识别方法。

第6章讨论实体属性知识获取。对于实体属性知识，介绍实体的显式槽和隐式槽的属性知识获取方法。进一步，介绍作者身份识别应用，从三个方面开展相关论述。阐述非结构化文本的作者身份属性识别方法，博客作者身份属性识别方法，源代码作者身份属性识别方法。

第7章介绍描述流抽取。讨论描述流的表示和结构，给出描述流的形式分析、定性分析和定量分析，并阐述领域本体驱动的描述流抽取方法。

第8章论述知识评估。首先，介绍概念分类层次知识的评估方法；然后，阐述实体属性知识评估方法，包括单种属性关系的属性知识评估、多种属性关系的属性知识评估以及相关应用方法。

本书可作为知识工程、人工智能等专业本科生及研究生的教学参考用书，也可供研究院所从事知识工程、行业信息化领域的研发人员阅读参考。

本书凝聚了作者同事、朋友和研究生的心血。在本书的撰写过程中，参阅并引用了诸多文献和部分国内外相关研究成果，敬致感谢。在本书写作和出版过程中，北京理工大学出版社给予了大力帮助，特此表示感谢。

知识工程发展迅速，呈现与自然科学、社会科学交叉融合的态势，相关理论与技术尚处于不断完善、探索和发展之中。由于作者水平所限，书中难免有疏漏和不妥之处，欢迎专家和读者提出宝贵意见，给予批评指正，激励我们不断完善和提高，将不胜感激。

目　录

第1章

引　言

随着人工智能和大数据技术的迅猛发展，知识工程技术获得飞速发展和广泛应用。作为一门学科，知识工程的研究对象是知识。知识工程技术是利用知识工程原理与方法构建知识嵌入的人工智能系统的技术体系。知识工程技术主要研究知识表示、知识获取、知识推理、知识评估以及知识应用等核心技术和智能应用系统构建方法，是实现知识嵌入的人工智能的重要知识基础和技术支撑。本章介绍知识工程技术的研究背景、研究内容和未来挑战。

1.1 研究背景

当今诸多行业领域面临杂乱信息泛滥、精准知识匮乏的问题。在社会生产和日常生活中，随着人工智能技术的发展与深度应用，人们冀望获得高质量的精准知识服务。

人工智能的目标是通过计算机来模拟和延伸人类智能行为。当前，深度学习方法的广泛应用迅猛地推动着专用人工智能技术的发展。作为一种数据驱动的方法，深度学习在计算机视觉、自然语言处理、语音识别等领域获得了极大的成功。但是，随着开放环境的泛在以及复杂智能感知与决策任务需求的增长，纯数据驱动的人工智能技术发展到了极致阶段，技术上难以实现大超越。技术上，“知识”嵌入是实现高阶人工智能的必要基础。

知识工程是人工智能学科中的重要内容。作为知识工程的核心内容，知识表示、知识获取、知识推理的相关方法与技术已经广泛应用于诸多智能系统之中[1,2,3]。在诸多行业领域中，领域知识是广泛存在的。在方法论层面，领域知识描述特定领域的概念内涵与外延、领域实体关系以及领域实体属性知识等。将领域知识与计算机技术有机融合可有效地处理专业领域应用任务[4,5]。比如，在医学领域、金融领域以及气象领域中，现有人工智能系统充分利用了领域知识，提升了相关人工智能产品的性能和服务质量。另外，问答系统、信息推荐和知识融合等知识工程应用技术所取得的突破性进展，同样离不开对相关领域知识或常识知识的深度应用。知识工程技术主要包括知识表示、知识获

取、知识验证或评估、知识推理等内容。知识表示是指对各种类型的知识进行表示与组织，为知识工程的众多下游任务提供知识支撑。知识获取是指从互联网、书籍、期刊以及专家等来源中获取知识。其中，涉及结构化数据、半结构化数据和非结构化数据。知识类型包括领域知识、常识知识，或事实性知识、程序性知识、过程性知识以及意见性知识。知识验证是指对于以自动方式、半自动方式或人工方式获取的知识，验证其一致性、正确性和完全性。知识推理是指根据已有知识推理出未知的新知识。

鉴于知识工程的相关内容十分广泛，本书主要聚焦于知识表示、知识获取和知识验证及应用四个层面。具体涉及以下内容：

（1）在知识表示方面，主要包括形式领域本体、领域知识获取本体、模式本体、课程本体、时间本体。

（2）在知识获取方面，主要包括时间信息抽取、实体识别、术语定义抽取、实体上下位关系抽取、实体对齐关系识别、领域实体属性知识获取、非结构化文本作者属性识别、博客作者属性识别、源代码作者属性识别、文本描述流抽取。

（3）在知识验证方面，主要包括概念分类层次知识评估或验证、实体属性知识评估或验证。

（4）在知识应用方面，主要呈现时间本体在问答系统中的应用，以及概念分类层次知识评估和实体属性知识评估方法在多领域知识验证中的应用。

1.2 知识工程挑战

第三代人工智能通过“知识、数据、算法和算力”四要素，实现能够模拟人类认知、思维和决策等的人工智能技术和应用[6]。知识工程的研究内容主要包含知识处理的理论、方法，以及技术和应用。知识工程涉及自然语言处理、人工智能、机器学习、大数据分析和数据挖掘等技术[7]。

知识工程技术面临的主要挑战包括：

（1）海量知识的多源异构性。

在知识工程中，知识来源呈现多样性和分散性特点。知识来源包括书籍、领域专家和互联网等。互联网主要包含新闻网站、百科网站、社交平台等。从数据类型看，主要包括文本、图片、音频和视频等非结构化数据。文本语言进

一步包括书面语、口头语和网络用语等。文本包括短文本和长文本。从知识来源的模态看，知识来源包括单种模态数据（如文本或图像）和多种模态数据（如文本－图像对）。总之，知识工程所处理的数据多源异构，具有不同的模态特性、不同的模态规模、不同的时空粒度、不同的表征结构、不同的语义信息、不同的度量特性。如何实现高效的知识表示、知识获取，突破模态平衡的陷阱和粒度鸿沟，是其中的一个挑战性问题。

（2）知识的动态更新性。

近年来，社交平台和社交软件呈持续增长的态势，知识来源的种类不断增加，数据获取的渠道不断丰富。同时，各领域知识处于不断发展和更新之中。结构化数据、半结构化数据和非结构化数据呈现几何级增长态势。总之，不同软件系统或知识库中的知识表示不断更新，需要构建可自适应转换不同表示语言的知识表示方法，不断动态更新知识库，为知识融合、语义互操作、知识共享和重用开辟新的技术途径，是一个持续面临的技术挑战问题。

（3）多粒度知识融合。

知识工程涉及文本、图片、视频和音频等多种模态数据。在知识融合的内容方面，需要融合同一实体的相关知识、同一概念的相关知识、具有相同属性或相同关系的知识、隶属于相同类别的实体知识等。在知识融合过程中，还需要充分考虑不同类型的知识载体，包含来自不同专家的知识和来自不同软件系统的知识。因此，如何高效融合具有不同知识来源、不同模态特征、不同表示语言、不同自然语言形态、不同表示粒度的知识，是需要解决的技术问题。

（4）多类型知识验证。

在知识工程中，“知识”往往呈现出不同的性质。在知识内容方面，需要验证概念层面的知识、实体层面的知识。概念层面的知识包括概念的含义、概念之间的语义关系、概念分类层次知识等。实体层面的知识即知识图谱，包括实体之间的语义关系、实体的属性知识、实体与概念的隶属关系等。在知识验证的时间维度方面，需要验证不同时间的知识正确性和一致性。在知识验证的知识粒度方面，需要验证不同粒度的知识，包括同一概念同一属性的不同粒度的相关知识、同一实体和同一属性的不同粒度的相关知识。因此，知识验证需要充分解决多类型知识中可能存在的语义歧义、矛盾以及异构性，确保知识的一致性、正确性和完全性。

1.3 本书组织结构

本书共包括8章内容，主要包括知识表示、知识获取，以及知识评估等内容。

第1章，引言，介绍知识工程的研究背景、研究内容及其挑战性问题。

第2章，领域本体，阐述形式领域本体、领域知识获取本体、模式本体、课程本体的构成和表示。

第3章，时间本体和时间信息抽取，阐述时间本体的构成框架、基础时间本体、扩展时间本体以及时间信息抽取方法。

第4章，实体识别，论述概念抽取方法、术语定义抽取方法等。

第5章，实体关系知识获取，阐述上下位关系抽取方法、实体对齐关系识别方法。

第6章，实体属性知识获取，阐述领域实体属性知识获取方法、非结构化文本作者属性识别方法、博客作者属性识别方法、源代码作者属性识别方法。

第7章，描述流抽取，阐述描述流的表示和结构、描述流的定性分析和定量分析以及描述流抽取方法。

第8章，知识评估，阐述概念分类层次知识评估方法、实体属性知识评估方法。

第2章

领域本体

目前，互联网载体信息具有海量繁杂和多源异构特点。多源异构知识的异质性和分散性严重地阻碍了知识在多主体和软件实体之间的语义互操作、共享和重用。从技术发展趋势来看，形式本体已被认为是很有前途的解决方法。为此，本章首先介绍形式领域本体、领域知识获取本体；其次，阐述模式本体和考古学领域本体；最后，论述数学课程本体。