张德海 著

本体学习的认知模型

A Cognitive Model for Ontology Learning

科学出版社

张德海 著

本体学习的认知模型

A Cognitive Model for Ontology Learning

科学出版社
北京

内 容 简 介

本体是具有不同知识表示的智能系统之间进行知识交换和知识共享的基础结构。这种基础结构，实际上就是一种对于客观世界的共识。随着本体在知识工程、语义网、非结构化数据分析等众多领域的应用，本体的自动获取显得尤为重要。

本书选择这一问题加以研究，具有较强的理论意义和应用价值。借鉴认知科学的理论研究成果，对本体学习过程中的认知状态及其变化进行深入研究，提出本体学习的认知模型，给出该模型的形式化表示及实现方法。此外，本书还对本体学习研究的发展方向进行展望。

本书适合于信息管理、知识管理、计算机应用及相关专业的高年级本科生和研究生作为教材，也适合于从事信息技术、知识处理、应急管理等相关技术专业人员作为参考书。

图书在版编目（CIP）数据

本体学习的认知模型 / 张德海著. —北京：科学出版社，2016.11
ISBN 978-7-03-050316-9

Ⅰ. ①本… Ⅱ. ①张… Ⅲ. ①智能技术-研究 Ⅳ. ①TP18

中国版本图书馆 CIP 数据核字（2016）第 258092 号

责任编辑：付 艳 苏利德 / 责任校对：郑金红
责任印制：张欣秀 / 整体设计：楠竹文化
编辑部电话：010-64033934
E-mail: edu-psy@mail.sciencep.com

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com
北京中石油彩色印刷有限责任公司 印刷
科学出版社发行 各地新华书店经销
*
2017 年 6 月第 一 版 开本：720×1000 B5
2017 年 6 月第一次印刷 印张：14 3/4
字数：225 000
定价：79.00 元
（如有印装质量问题，我社负责调换）

张德海

博士，云南大学副教授，云南大学中青年骨干教师，云南省青年骨干教师。IEEE会员，中国人工智能学会（CAAI）会员，国际计算机科学与信息技术学会（IACSIT）会员。主要从事人工智能、知识工程、系统分析与集成及大数据分析的研究。

主持国家基金、省基金及横向项目20余项，发表论文30余篇，公开申请国际发明专利及国家发明专利15项，软件著作权12项。参与云南省政府政策研究室项目，主持或参与多个智慧城市、领域大数据应用项目。

内容简介

本体是具有不同知识表示的智能系统之间进行知识交换和知识共享的基础结构。这种基础结构，实际上就是一种对于客观世界的共识。随着本体在知识工程、语义网、非结构化数据分析等众多领域的应用，本体的自动获取显得尤为重要。

本书选择这一问题加以研究，具有较强的理论意义和应用价值。借鉴认知科学的理论研究成果，对本体学习过程中的认知状态及其变化进行深入研究，提出本体学习的认知模型，给出该模型的形式化表示及实现方法。此外，本书还对本体学习研究的发展方向进行展望。

本书适合于信息管理、知识管理、计算机应用及相关专业的高年级本科生和研究生作为教材，也适合于从事信息技术、知识处理、应急管理等相关技术专业人员作为参考书。

责任编辑：付　艳　苏利德

整体设计： 楠竹文化
010-68964483

研 究 基 地

云南省软件工程重点实验室

云南省云计算工程中心

民族教育信息化教育部重点实验室（云南师范大学）

云南省高校民族教育与文化数字化支撑技术工程研究中心

序　言

在 20 世纪 90 年代初，本体从哲学领域进入计算机相关领域，其主要目的是实现不同计算机系统之间的知识共享、语义互操作等问题。在任何一个论域中，当我们确定了其中的概念及其属性、概念之间的关系、概念及其关系的性质等语义内涵，并且这种确定被大家接受之时，知识共享和语义互操作就有了保障，计算机之间的“语义鸿沟”、人与人之间的“语义鸿沟”、人与计算机之间的“语义鸿沟”就能缩小或消失。

21 世纪以来，本体在智能信息集成、Internet 信息检索、知识管理、语义 Web、数字图书馆、自然语言处理、机器翻译等应用研究方向得到逐步重视，在生物、地理、医学、宗教、军事等学科中也得到广泛研究。

与此同时，国内外出现了大量的本体建设语言和工具，如 Ontolingua、OntoSaurus、WebOnto、Protégé、WebODE、OilEd、OntoEdit 及 KAON 等。这些工具提供了友好的图形化界面和一致性检查机制，在一定程度上方便了本体的构建，但是提供的仅仅是本体编辑功能，支持的仍然是手工构建本体的方式。

由于手工建设本体的代价很大，所以利用机器学习技术提高本体构建效率的本体学习（ontology learning）技术应运而生，大量本体学习工具也不断涌现，国外有代表性的本体学习工具

有 Text2Onto、Hasti、OntoLearn、OntoBuilder、OntoLiFT 及 NELL 等，国内的有 OnShpere、SOAT、WebOntLearn 等。这些工具在运行过程中或多或少地需要人工干预。

张德海博士从 2000 年起从事知识获取和本体构建的研究，是我国最早研究本体的学者之一。在该书中，他总结了本体学习的研究现状，认为目前的本体学习研究主要集中在概念获取、公理获取、关系获取等方面，在本体形成阶段的工作仍然需要人工介入，其主要原因在于对本体学习过程中本体的扩展、修正、更新等过程缺乏深入的研究。他进一步认为，要深入地研究本体学习，就要研究学习者的认知状态及其在学习过程中的变化，把机器学习回归到学习的本质，也就是人工智能研究的起点，即模仿人类的学习方式，从而把本体学习变成学习者（或智能体）的自主知识构建过程。鉴于此，张德海博士结合现代认知主义学习理论的成果，提出了本体学习的认知模型，用来描述本体学习过程中学习者的认知状态及其变化，并用形式化的方法来表示该模型中的知识，对已学习的知识进行归纳泛化、对矛盾知识进行修正、对过时的知识进行更新等认知策略的具体操作，从而进一步建立本体自主学习的认知模型。这些都是对本体学习技术的一种有益的新探索。

今天，人类已进入大数据时代，海量非结构化数据的语义分析对大规模本体的需求更加迫切。大规模语义网的构建依赖于本体学习技术的进步。虽然本书的研究成果还不足以引领一个新的研究方向，但已让我们看到了一个模仿人类认知过程的从理论到应用的本体学习方法是可行的，作者的很多创见将为我们提供有益的参考。

曹存根

2016 年 8 月于北京

前　言

近年来，“本体”这个概念在智能信息集成、Internet 信息检索、知识管理、语义 Web、数字图书馆等众多领域广为流传。本体能够流行的一个主要原因是本体提供了人与计算机之间以及机器与机器之间对领域知识的共享和共同的理解。于是，用于领域本体获取的机器学习模型或称“本体学习”近年来成为一个非常热门的研究方向。

从认知的观点看，目前的本体学习主要集中在概念获取、公理获取、关系获取等层面。确切地说，上述本体学习算法应该被称为本体挖掘算法，因为在本体形成阶段的工作仍然是手工的或半自动的，需要人利用本体构建工具参与本体形成的过程，其主要原因在于对本体学习过程中学习者的认知状态及其变化过程缺乏深入的研究。

要深入地研究本体学习，就要研究学习者的认知状态及其在学习过程中的变化。认知科学研究的一些成果说明和解释了人在完成认知活动时是如何进行信息加工的。鉴于此，结合现代认知主义学习理论的成果，本书提出本体学习的认知模型，用来描述本体学习过程中学习者的认知状态及其变化，并用形式化的方法来表示该模型中的知识，对已学习的知识进行归纳泛化，对矛盾知识的修正以及对过时的知识进行更新等认

知策略的具体操作，从而进一步建立本体自动学习的认知模型，探索本体自动构建的关键过程和技术。

本书的主要研究结果如下。

（1）在综合机器学习理论与认知主义学习理论的基础上，提出本体学习的认知模型，并给出其形式化表示。在该模型中，把本体学习过程看作一个认知过程，通过 7 种不同的认知策略来实现，即断言获得、本体扩展、本体缩减、本体归纳、本体演绎、本体修正和本体更新。本书讨论了除断言获得策略以外的 6 种策略。

（2）为进一步讨论本体学习认知模型的实现，基于本体系统的本体论假定，建立本体知识表示模型，给出了本体的形式定义，提出了本体系统的描述语言和推理规则，定义了本体的协调性及其逻辑闭包的协调性。

（3）在本体知识表示模型的基础上，建立了基于公理化认知模型的本体学习系统。用公理化的方法为学习模型中的 5 个认知操作：本体扩展“+”、本体缩减“-”、本体归纳“$\prec$”、本体修正“$\circ$”和本体更新“$\diamond$”建立了相关的假定、公理体系，给出了操作的具体算子，并证明这些算子满足所建立的公理系统。

（4）讨论了本体学习的认知模型中归纳算子、修正算子和更新算子中存在的问题及可能的解决方法。从修正与更新的最小改变原则的要求出发，提出了从类结构差异和类的差异两个层次进行本体比较的算法。

（5）给出了基于本体学习认知模型的本体自动构建工具原型系统，并对该系统的应用案例作了分析。

（6）总结展望了本体学习技术的发展方向。

自 2000 年以来，作者在中国科学院计算技术研究所曹存根研究员、眭跃飞研究员和云南师范大学林毓材教授的指导下，开始接触知识工程的研究，后来在从事本体工程、本体学习方面的研究中一直都得到几位老师的悉心指导，在攻读博士学位期间还得到了云南大学曹克非教授的指导，在此特向各位恩师表示衷心的感谢。

云南师范大学甘健侯教授、云南大学岳昆教授对本书的出版进行了指导并提供帮助，云南大学研究生王乃尧、王斌、杨忠昊、赵航等在本书出版过程中

提供了校对、排版等帮助，特表示感谢。

本书得到国家自然科学基金项目（61263043）、云南省教育厅自然科学基金重点项目（2011Z020）、云南大学“海量数据分析与服务”创新团队培育计划（XT412011）、云南大学中青年骨干教师培养计划（XT412003）的资助，在此一并表示感谢。

由于作者水平有限，部分个人观点可能不一定准确，敬请广大读者批评指正。

张德海

2016年8月于昆明

目　录

第一章 绪 论

人工智能（artificial intelligence，AI）被定义为“对使计算机变得具有智能的学说的研究”（Winston，1984），以及“对智能行为的研究”（Genesereth & Nilsson，1987）。计算机系统是被制造出来展示智能行为的，但是这些行为又是如何被加以分类或者被认可为具有智能的呢？换句话说，究竟什么是智能？

对智能的一种定义是“获取和应用知识的能力”。而人工智能研究的众多学派中，以 McCarthy、Nilsson 为代表的逻辑学派和以 Simon、Minsky、Newell 为代表的认知学派均认为“知识与概念化是人工智能的核心”（史忠植，2006）。

McCarthy 指出人工智能的长期目标是实现人类水平的人工智能（McCarthy，2005）。而机器学习就是要让机器具备学习的能力，获取知识是机器学习的一个重要方向。一个机器学习系统如何获取新的知识、知识如何编辑，以及如何学习到新的知识是知识工程研究中的难点。在获取新知识的过程中要对知识的完整性、新旧知识的一致性进行检查与处理。

20 世纪 90 年代初，本体在知识工程、自然语言处理和知识表示等人工智能领域引起了越来越多研究人员的兴趣。近年来“本体”这个概念在如智能信息集成、Internet 信息检索、知识管理、语义 Web、数字图书馆等很多领域广为流传。本体能够流行的一个主要原因是由于本体提供了人和计算机之间以及机器与机器之间对领域知识的共享和共同的理解。于是，用于领域本体获取的机

器学习模型或称“本体学习”近年来成为一个非常热门的研究方向。

机器学习的最终目标是模拟人的学习行为。但从认知的观点看，目前的本体学习主要集中在概念获取、公理获取、关系获取等层面，而对获取到的每一条知识如何加入到已有的本体中几乎没有深入的研究。确切地说，上述本体学习算法应该被称为本体挖掘算法，在本体形成阶段的工作仍然是手工的或半自动的，需要人利用本体构建工具参与本体形成的过程。其主要原因在于对本体学习过程中学习者的认知状态及其变化过程缺乏深入的研究。因为除了找到一个适当的知识表示形式化系统来表示已获取的概念、关系和公理外，在知识获取的过程中还存在其他问题。首先，由于对世界的认识是不完全的，所以我们通常没有足够多的知识来得出正确的结论或者区分相似的事物；另外一个问题是关于不正确的知识。不正确的知识可能是由一次错误的观察或是通信中的一次失误，或者在信息源的部分出现的欺骗行为所引起的。最后，当我们没有观察到一个改变世界的行为或者事件；或者我们不知道这个行为或者事件将会导致什么样的结果时，以前曾经正确的知识在改变了的现实世界中进行使用可能是不适当的。于是，随着我们从或多或少可以信赖的信息源获取新的知识，我们关于世界的知识也必须不断地改变。因此，我们需要进行知识改变。同时，对知识进行表示的本体也需要不断地改变。

对学习的整个认知过程的研究，目前的论文主要集中 agent 的信念修正（belief revision）、信念更新（belief update）上，即：当 agent 获取到新的知识后，如何添加到已有的知识库中，对已有的知识进行扩充；当前后的知识出现矛盾的时候，如何对其进行修正；当外部世界的知识发生变化的时候，如何更新已有的知识。

要深入地研究本体学习，就要研究学习者的认知状态和在学习过程中学习者认知状态的改变过程。而认知科学研究的一些成果说明和解释了人在完成认知活动时是如何进行信息加工的，鉴于此，本书提出本体学习的认知模型，用来描述在本体学习的过程中学习者的认知状态及其变化，并用形式化的方法来描述该模型中的知识表示、学习的归纳泛化、矛盾知识的修正策略以及知识更新的策略及算法。

本章的具体安排如下。首先，第一节介绍本体的概念及本体知识表示模型，第二节介绍本体学习研究的进展，第三节介绍目前认知科学领域的认知主义学习理论，第四节介绍本体学习的认知模型，第五节给出本书的主要结果，第六节介绍本书的组织结构。

第一节 本体及基于本体的知识表示模型

一、本体论和本体

本体论（ontology）原是一个哲学名词。它最早来源于亚里士多德对世界万物的分类，是关于存在及真实世界的任何领域中的对象、性质、事件、过程和关系的种类和结构的学说。作为一个抽象的哲学概念，1721 年，Merriam Webster 给出了本体的两个定义：①涉及存在的本质和关系的形而上学的一个分支；②一个关于存在的本质或者存在物的种类的特殊的理论。“本体论”通常被哲学家作为“形而上学”（metaphysics）的同义词来使用。Smith 指出，早在 18 世纪初形式本体的概念已经被哲学家胡塞尔从形式逻辑中分离出来（Smith，1998），“本体论”这个术语开始被哲学界广泛采用，并且在 20 世纪的分析哲学中，本体论正式成为研究存在性和存在本质等方面的通用理论（Smith，2001）。

任何一个基于知识的系统都是基于对世界的某种概念化（conceptualization），而 ontology 正是一种概念化的规范说明系统，又独立于具体的符号层表示方式，因而是具有不同知识表示的智能系统（或者说其中的智能主体）之间进行知识交换和知识共享的基础结构。这种基础结构实际上就是一种对客观世界的共识。

1998 年，Guarino（1998a）提出了一个应用本体论的领域集合，包括知识工程、知识表示、质量建模、语言工程、数据库设计、信息检索与抽取和知识管理与组织等领域。在今天看来，这个集合还应包括图书馆科学以及新兴的语义 Web、电子商务等领域。