Web 智能与科学

Web Intelligence and Web Science

3

# 面向语义 Web 的知识管理技术

# Techniques of Knowledge Management for the Semantic Web

漆桂林　黄智生　杜剑峰　编著

高等教育出版社

Web 智能与科学
Web Intelligence and Web Science

3

MIANXIANG YUYI WEB DE ZHISHI GUANLI JISHU

# 面向语义 Web 的知识管理技术

# Techniques of Knowledge Management for the Semantic Web

漆桂林　黄智生　杜剑峰　编著

高等教育出版社·北京

内容提要

本书比较详细地介绍了描写语义 Web 的知识管理技术。全书共分为九章。在对语义 Web、语义技术以及基于本体（Ontology）的知识管理进行简要概述的基础上，介绍了本体生命周期的相关技术。首先介绍了 Web 本体语言 OWL 的逻辑基础；其次，对本体构建过程中出现的本体演化、本体诊断、本体融合、本体不一致性推理、本体版本化和本体模块化分别进行介绍；最后，对本体管理的应用做了一个简单介绍。本书除了介绍一些基本概念和技术以外，还重点概述了基于本体的知识管理的最新研究成果和相关的工具，并且结合实例，由浅入深地对各项技术进行讨论，有助于对基于本体的知识管理技术进行系统的学习和掌握。

本书可作为高等学校计算机科学专业、信息科学专业、自动化专业高年级本科生和研究生的教材或者教学参考书，也可供高等学校、研究所和互联网企业从事知识管理技术的科研人员学习参考。

**图书在版编目(CIP)数据**

面向语义 Web 的知识管理技术 / 漆桂林,黄智生,杜剑峰编著. -- 北京:高等教育出版社, 2015.10
(Web 智能与科学)
ISBN 978-7-04-043700-3

Ⅰ. ①面… Ⅱ. ①漆… ②黄… ③杜… Ⅲ. ①语义网络-计算机应用-知识管理 Ⅳ. ①G302-39

中国版本图书馆 CIP 数据核字(2015)第 184020 号

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 策划编辑 | 刘　英 | 责任编辑 | 刘　英 | 封面设计 | 王　洋 | 版式设计 | 王艳红 |
| 插图绘制 | 邓　超 | 责任校对 | 刘娟娟 | 责任印制 | 刘思涵 | | |

| | | | |
|---|---|---|---|
| 出版发行 | 高等教育出版社 | 咨询电话 | 400-810-0598 |
| 社　　址 | 北京市西城区德外大街 4 号 | 网　　址 | http://www.hep.edu.cn |
| 邮政编码 | 100120 | | http://www.hep.com.cn |
| 印　　刷 | 北京明月印务有限责任公司 | 网上订购 | http://www.landraco.com |
| 开　　本 | 787mm×1092mm　1/16 | | http://www.landraco.com.cn |
| 印　　张 | 12 | 版　　次 | 2015年10月第 1 版 |
| 字　　数 | 220 千字 | 印　　次 | 2015年10月第 1 次印刷 |
| 购书热线 | 010-58581118 | 定　　价 | 59.00 元 |


物 料 号　43700-00

# “Web 智能与科学丛书”编审委员会

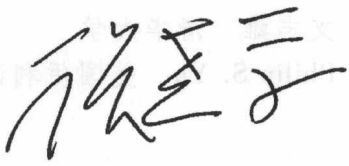
# 总　序

20世纪90年代以来，人类社会经历了一场新的科技革命，科技进步日新月异，国际竞争日趋激烈。在这场革命中，信息技术对全球社会经济的发展与进步起了巨大的推动作用，并极大地促进了世界经济结构的变革。网络化、智能化已成为当今世界科技革命的一个重要特征。在《国家中长期科学技术发展规划纲要(2006—2020)》中，已把与网络化、智能化相关的信息科学及技术纳入其中。因此，需要尽快与国际前沿接轨并达到国际领先水平，以提高我国在该领域的国际竞争力。

Internet和Web是发展最快的网络形式之一，也是最活跃的研究领域之一，网络智能化研究已成为国际人工智能研究领域的一个新趋势。钟宁教授、刘际明教授等学者顺应这一国际研究新趋势，率先对Web智能科学进行了系统的研究，获得了一系列的研究成果，在国际Web智能科学研究领域取得了领先地位。而国内对Web智能科学的研究起步相对较晚，研究成果较少，还没有引起国内人工智能研究领域及相关研究领域科研人员的广泛关注和足够重视。

在高等教育出版社的支持下，由钟宁教授、刘际明教授提议并任主编，建立了“Web智能与科学”系列丛书，将解决国内Web智能科学研究领域中最新资源短缺的状况。该系列丛书主要向国内读者介绍最新的Web智能科学领域的学术动态、最新的研究成果以及学术交流等内容。

在此，衷心地希望通过该系列丛书的出版，为国内相关领域研究人员构建一个Web智能科学研究交流互动的窗口，为国内该领域的研究提供最新的信息和学术资源。希望该系列丛书的出版成为我国Web智能科学研究进一步发展的新起点，进而带动该领域的水平提升，为我国的科技进步、自主创新做出应有贡献。



2010年12月

# 作者简介

**漆桂林** 东南大学计算机科学与工程学院教授，获英国贝尔法斯特女皇大学计算机科学博士学位。曾在德国 Karlsruhe 大学 AIFB 研究所从事博士后研究，指导教授是语义 Web 创始人之一 Rudi Studer 教授。长期从事人工智能中知识表示和推理、语义 Web 和本体管理技术的教学和科研工作，主持两项国家自然科学基金项目和多项省部级项目，主持欧盟第七研究框架计划 SemData 中的高效推理工作组的研究。在国际期刊和会议上发表论文百余篇。担任多个国际重要学术会议（例如 IJCAI、AAAI、WWW、ISWC）的程序委员会委员和负责多个国际学术会议（包括 ISWC、EKAW、JIST）的组织工作。

**黄智生** 荷兰阿姆斯特丹自由大学计算机系高级研究员，北京工业大学 WIC 国际研究院等多个大学兼职教授。主持欧盟第七研究框架重大研究计划 LarKC 项目中的推理工作组的研究。在国际期刊和会议上发表学术论文近 200 篇。担任 80 余个国际学术会议的组织委员会或程序委员会的委员或联合主席。

**杜剑峰** 广东外语外贸大学电子商务系系主任、副教授，获中国科学院软件研究所计算机科学与技术博士学位。曾经主持两项本体推理领域的国家自然科学基金课题。在权威国际学术期刊和国际学术会议（例如 KAIS、IJCAI、AAAI、WWW、CIKM、ISWC）上发表过多篇论文，曾担任多个国际学术会议程序委员会成员。

# 前　言

近年来，随着语义 Web 的兴起以及各方面的应用需求，本体技术受到了广泛关注。很多大型跨国公司都已经投入研究本体技术并开发对应的应用。例如，谷歌于 2012 年提出了知识图谱的项目，旨在利用本体技术来提高搜索的精度和实现更智能化的知识浏览；本体技术在 IBM 的著名问答系统 Watson 中发挥了重要的作用；微软提出了 Probase 项目，旨在通过爬取网页中的信息来构建大规模的本体；IBM 利用语义 Web 技术来处理异构医疗数据的整合以及提供更准确的查询回答；Oracle 实现了一个强大的语义数据推理和索引系统。国内的互联网公司，如百度和搜狗，也已经开展了这方面的项目。本体技术还受到欧美政府的支持。例如，英国政府发起了 Data.gov.uk 项目，把很多政府网站的信息都以本体的形式发布；美国政府也有类似的项目。学术界对本体的研究有很多成果，特别是在计算机科学领域，开发了很多实用的技术。例如，欧盟在最近 5 年投入大量科研经费（累计数亿欧元）用于本体相关的研究；中国的自然科学基金也有数亿人民币投入到与本体相关的项目。目前，本体技术已经趋向成熟，但是全面介绍面向语义 Web 的本体技术方面的书籍比较匮乏，特别是缺少这方面的中文书籍，本专著的编写将弥补这一空缺。

本书将系统化地介绍基于 OWL（Ontology Web Language）本体知识管理方面的技术，其中 OWL 是由 W3C 推荐的标准本体语言。由于 OWL 是基于描述逻辑（Description Logic），并且具有严格的逻辑语义，所以受到了学术界的广泛关注以及工业界的认可。基于 OWL 的本体管理技术得到了快速的发展，这些技术包括描述逻辑的推理技术、基于描述逻辑的本体诊断技术等。

本书分为九章。第一章为导论；第二章介绍本体语言与逻辑基础，重点介绍 Web 本体语言 OWL 和描述逻辑；第三章介绍本体调试的基本概念和算法；第四章介绍本体演化的主要过程以及基于描述逻辑的本体修正方法；第五章介绍本体融合相关的理论和技术，特别是本体映射以及本体映射的修复；第六章介绍本体不一致性推理的几种主要方法；第七章介绍本体版本化的相关技术；第八章介绍本体模块化的相关技术；第九章介绍基于本体的知识管理的应用。本书第三～第六章由漆桂林教授执笔，第一章、第七章、第九章由黄智生教授执笔，第二章、第八章由杜剑峰副教授执笔。本书肯定存在许多不足之处，希望能够得到同行的批评指正。

本书不仅是一本系统了解本体相关技术的重要参考书，同时对基于本体的应用的研究人员和本体技术的应用开发人员均具有重要的参考价值。本书的

读者群包括高校的教师、博士生、研究生、高年级本科生以及企业中从事本体技术的科研人员和开发人员。

漆桂林　黄智生　杜剑峰

2015 年 8 月

# 目　录

# 第一章 导 论

## 1.1 语义万维网与语义技术

人类社会已经进入了大数据(Big Data)时代。在这个以万维网为主要特征的大数据时代里,人们处于一个全新的信息环境之中。万维网上的信息为人类社会提供了极其丰富的信息资源,成为人类获取信息的主要途径之一。但是,从万维网的诞生到现在,仅仅在20多年的时间里,人类已经面临着万维网上信息急剧增长而不能有效处理这样一个迫切的问题。人们每天要花费大量的时间从事在万维网上人工进行搜索信息、筛选信息以获得自己真正需要的信息这样一个枯燥而繁重的智力劳动。人类需要寻找一个全新的方式来描述万维网上的信息,以便能够通过自动化的手段更有效更精准地获得自己想要的信息。

为了应对这样的技术挑战,万维网之父 Tim Berners-Lee 在10多年前提出了语义万维网(Semantic Web,简称语义 Web 或语义网)的思想[1,2],其目的就是让计算机能够自动处理万维网上的信息。语义网的主要特征是采用逻辑的手段来描述数据,从而能够在语义层面上更精确地刻画数据内容中所包含的意义(即语义),使得人们可以通过其对应的推理机或者是其他自动处理工具有效地分析数据内容,为知识管理提供了基本的技术手段。10多年来,语义网及其相关技术已经取到了巨大的发展[3],其标志性的发展是国际万维网组织出台了一系列语义数据描述语言标准,其中包括用于描述网络信息资源的 RDF/RDFS (Resource Description Framework/Resource Description Framework Schema)、网络本体语言 OWL (Ontology Web Language)、语义数据查询语言 SPARQL (SPARQL Protocol and RDF Query Language)、规则交换框架 RIF (Rule Interchange Format)等。这些国际规范的语义描述及查询语言的出台,为人们提供了统一的数据描述格式,为数据的互操作提供了共同的基础。

图1.1是著名的语义网技术层次图,它由七层结构组成。从图中我们可以看出,统一资源标识符(Uniform Resource Identifier, URI)、国际资源标识符(Internationalized Resource Identifier, IRI),以及 Unicode 技术构成了语义网的最底层基础。因为 URI 实现了网络信息资源的唯一标识,IRI 作为对 URI 的补充,允许使用 Unicode 来标识网络资源, Unicode 实现了不同语言的符号集的单一表示,从而实现了在信息的数字化表达层面的最基础的技术统一。可扩充标记

语言 XML 的引入使得我们能够通过标签(Tag)来表达半结构化的数据,名字空间 NS(Name Spaces)的使用使得概念的 URI 前缀表达有了更方便的表达方式。在这基础上引入 XML 框架模式(XML Schema)及 XML 查询语言,使得我们可以用它来表达许多不同领域的元数据。RDF/RDFS 模式所提供的一系列技术使我们能够方便地描述和询问网络数据资源。

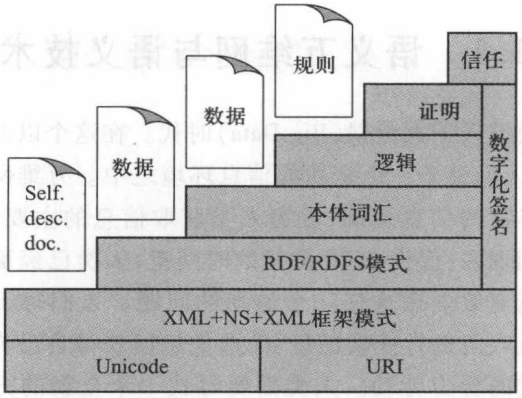



图 1.1 语义网技术层次图

RDF/RDFS 语言被看成是一种规范的元数据描述语言,用于描述万维网信息资源及其提供的推理支持。元数据(Meta Data)指的是那些用于描述数据的数据。网络本体语言 OWL 是在 RDF/RDFS 基础上的一些扩展和改造,主要是增加了更强的逻辑描述能力。我们把这种采用比元数据更强的逻辑关系描述的特定领域的知识称为本体(Ontology)。本体可以被理解成特定领域规范概念集及其逻辑关系的描述。本体为特定领域中的信息提供了一个基本的分类框架,同时也为特定领域中信息之间的关联性提供了一定程度的逻辑描述,使得特定领域中的信息资源能够在本体描述的框架上组织成一个有机的整体。我们把这个基于语义网思想所发展的信息及其本体处理技术称为语义网技术(Semantic Web Technology),简称为语义技术(Semantic Technology)。

OWL 已经成为基于语义网的本体描述语言的国际标准,也成为各类语义知识库的常用描述语言。如著名的英文电子词典 Wordnet 有对应的 OWL 格式表达(http://www.w3.org/TR/wordnet-rdf/),Cyc 大型知识库系统也有对应的 OWL 文件。OWL 语言根据其不同的应用需求分为不同的子语言,如 OWL Lite、OWL DLP、OWL DL 和 OWL Full 等。最常用的是 OWL DL 语言,它建立在描述逻辑(Description Logics,DLs)[4]的逻辑体系之上,是对应着一阶谓词逻辑中可判定的一个子语言[5]。OWL2 是 OWL 基本版本的进一步扩充,已经成为了当

前网络本体语言的推荐标准①。

使用标准的网络本体语言来描述信息仅仅完成了对本体最基本的性质描述,更多的本体数据性质可能需要更强的本体描述能力,也就是通过规则(Rule)来描述。除此之外,还需要用本体查询(Query)管理语言对本体数据进行询问和基本推理处理。这就是图 1.1 中所展现的规则与查询层(Rules/Query)所要求的内容。

近 10 多年来,在计算机科学家、领域的科学家以及工程人员的共同努力之下,在许多领域都已经创建有对应的元数据与本体,这为语义技术的发展与普及提供了必不可少的数据基础。使用这些特定领域的元数据与本体,我们就可以对万维网上现有的许多信息资源采用手工、半自动化或自动化的手段进行语义标注(Semantic Annotation)。这样,我们就可以通过针对特定领域的语义搜索引擎有效地、更精准地提供人们所需要的信息资源。由于在语义网上这些信息资源(包括元数据与本体)都是采用规范的语义描述,使得计算机能够有效地理解和处理这些数据所包含的内容,这也为信息搜索的自动化处理提供了基本的技术环境。

RDF 和 RDFS 的数据表达可以采用不同的数据格式,如基于 XML 标记格式、Ntriple 表达格式等。但是,RDF/RDFS 的语义模型在逻辑层面上都是通过"主语(Subject)-谓语(Predicate)-宾语(Object)"这样一个三元组结构来表达的。虽然这种三元组结构看起来比较简单,但它们的组合实际上可以表达极其丰富的数据内容。OWL 是 RDF/RDFS 语言在表达能力上的进一步扩充,所以在语义网与本体技术研究领域,语义数据的规模通常都是以三元组(Triple)的数量来度量的。

由于许多领域知识之间都有一定的关联性,某个元数据或本体中的一些概念可能等价于其他一些元数据或本体中的另外一些概念,故这些特定领域的本体与元数据都存在着一定的语义关联性,这种关联性可以通过其关联描述来刻画。于是,关联语义数据集为我们提供了跨学科跨领域的语义数据的整合体。在过去的几年中,越来越多的数据提供者和互联网应用开发者将他们各自的数据发布到互联网上,并且与其他数据源关联在一起,构成巨大的数据资源网(Web of Data),这就形成了著名的关联开放数据云图(Linked Open Data, LOD)。

2007 年关联开放数据项目的设立极大地促进了关联语义数据集的发展,关联数据应用范围的云图不断增大,关联的开放数据呈几何级数飞速增长。截至 2011 年 9 月 19 日,由 295 个数据集构成的关联数据网络中包含了 310 亿条 RDF 语句,这些语句被 5.04 亿个 RDF 链接,而且这些链接正在持续地增加。

① http://www.w3.org/TR/2008/WD-owl2-syntax-20080411/。

云图内容也逐步扩展，从早期的地理信息、生命科学数据、百科词条等，发展到目前涉及媒体、出版、政府信息、图形图像等，几乎无所不包。值得注意的是，关联开数据云图中的大多数语义数据集，正如其“开放”两字所表达的那样，都是免费公开用于共享的数据，这对语义技术的普及及其应用技术的推广提供了良好的数据环境和技术基础。

近10多年来，语义技术已经在许多领域如生命科学与医学、智慧城市（智能交通、智慧医疗与健康、智慧楼宇与智慧社区等）、工程领域、人文科学研究等领域都得到了广泛的应用。2011年6月搜索引擎巨头谷歌、雅虎和微软必应共同宣布新的语义搜索的技术标准 schema. org。2012年5月谷歌搜索引擎推出基于语义技术的知识图谱（Knowledge Graph）服务。由于这些重量级的公司的介入，语义万维网的研究和开发开始步入了一个新纪元。

## 1.2 本体与本体工程

如上所述，本体是特定领域规范概念集及其逻辑关系的描述。本体为特定领域中的信息提供了一个基本的分类框架，同时也为特定领域中的信息之间的关联性提供了一定程度的逻辑描述，使得特定领域中的信息资源能够在本体描述的框架上有机地组织起来。

本体是语义网的核心。按照 Gruber 的定义，本体是感兴趣领域共享的概念化的显式规约[6]。作为一个规约，本体需要通过某种语言表达。在一般意义上，我们可以不严格区分元数据与本体，故语义网中的本体一般都是指使用 RDFS 或 OWL 语言构建的本体。本体通常被看成知识库中满足共同约定的常识部分。因此，特定领域的本体的构造成为了该领域语义信息检索的主要基础工作之一。

一个本体自从被构造生成之后存在着一系列本体生命周期（Life Cycle）。我们把围绕着本体生命周期的相关技术统称为本体工程（Ontology Engineering），包括下列主要环节。

（1）本体构造：研究如何构造和生成本体。本体构造与生成主要是通过下列几个技术途径：

① 改造生成：通过改造现有特定领域的叙词表或受控词汇表，采用现代化的本体描述语言来表达，实现表达格式上的转换，并根据本体描述语言所提供的更强的逻辑描述能力，增加其对应领域知识的描述。

② 多源融合：通过融合集成现有的元数据集合或电子词汇表等，采用现代化的元数据或本体描述语言来实现其领域知识的融合与表达格式的转换。

③ 自动生成：通过自动化（如通过机器学习或文本挖掘）或半自动化的手段，从特定领域的大量自然语言文本语料中，采用自然语言处理的技术，构造出

一个对应该领域的初步的本体，再通过人工审核与修改的办法提高所生产的本体的质量，以生成一个新的本体。

④ 人工生成：通过领域专家或研究人员的人工生成的手段，采用可视化的本体编辑工具，如 Protege 等，构造生成对应的本体。这种完全通过人工手段来构造本体的方法只适用于小规模的经常是供学习用的本体。

(2) 本体学习(Ontology Learning)：正如上面所介绍的，这是一个通过机器学习、数据挖掘、文本挖掘等手段来获得对应的领域知识本体描述的手段。这种本体学习的技术手段，不仅可用于本体的生成与构造，还可以用于本体后续的生命周期环节，包括本体知识的扩展、与其他本体的关联映射等。

(3) 本体映射(Ontology Mapping)：正如我们上面关于关联数据所介绍的那样，一个本体中的一些实体(包括概念、属性与个体)可能语义等价于或语义相关(如子概念关联关系等)于其他一些本体中的另外一些实体。本体映射就是来建立不同本体之间的这种语义关联性映射。

(4) 本体管理(Ontology Management)：一个本体在生成之后，根据应用发展的需求，总是处于不断地发展和变化之中，这就需要对之进行有效的管理。本体管理包括：

① 本体演化(Ontology Evolution)：研究本体的发展过程中的变化规律及其管理与维护的相关技术。

② 本体融合(Ontology Integration)：研究如何从多个本体中集成一个新的本体。

③ 本体诊断(Ontology Diagnosis)：由于本体是采用逻辑表达形式，为了使之能够采用规范的推理机进行有效的推理与管理，我们通常要求本体本身在逻辑上是一致的(即没有包含任何直接的或者是间接的逻辑矛盾)。对本体进行诊断就是对本体进行检测看看是否满足一致性。如果一个本体是不一致的，则有两种做法：一种是通过诊断与调试的办法，发现其引起不一致的部分，对之进行修改，使之满足一致性。另一种办法，就是不进行任何修改，而是采用非规范的推理机进行推理，以获得有意义的结论，这就是下面要介绍的基于不一致本体的推理。

④ 不一致性本体的推理(Reasoning with Inconsistent Ontologies)：研究如何针对不一致的本体进行有效推理的相关技术。

⑤ 本体版本化(Ontology Versioning)：研究如何维护与管理本体的演化过程中所生成的不同版本的本体的相关技术。

⑥ 本体模块化(Ontology Modularization)：研究如何把大规模的本体切分成小的模块，并对之进行有效管理与推理的相关技术。

在本体工程领域，我们通常把那些负责本体构造及本体管理与维护的工程人员称为知识工程师(Knowledge Engineer)。知识工程师的主要工作就是负责