Ontology and Its Application

本体方法及其应用

甘健侯　姜　跃　夏幼明◎著

科学出版社

云南师范大学学术文库

Ontology and Its Application

本体方法及其应用

甘健侯　姜　跃　夏幼明◎著

科学出版社
北京

研究基地

民族教育信息化教育部重点实验室
中国科学院计算机网络信息中心
云南省高校智能信息处理重点实验室
昆明理工大学冶金与能源工程学院

前言

本体论原本是一个哲学概念，是表达哲学理论的一个术语。20世纪90年代初，本体概念被广泛地引用到计算机领域特别是人工智能和知识工程研究领域。本体已经成为知识工程、自然语言处理、协同信息系统、智能信息集成、Internet智能信息获取、知识管理等各方面普遍研究的热点。

首先，本书介绍本体的概念、本体的组成、建立本体的原则和一般方法、本体的常用关系、本体常用开发工具以及典型本体等。其次，本书介绍本体描述语言，包括XML、RDF/RDFS、OWL等。

本书结合旅游信息资源本体、高校就业管理领域本体、常用软件本体构建介绍本体的开发步骤及开发过程中存在的问题，使读者对本体及其构建有一个较清晰的认识，为其深入的研究和应用打下一个良好的基础。

本体映射是解决本体异构问题的手段之一，概念语义相似度和相关度计算对实现本体集成和信息的语义检索起重要的作用。本书介绍本体映射相关概念、方法和典型系统，并对基于本体的概念语义相似度和相关度进行探讨和研究。

描述逻辑是基于对象知识表示的形式化，依据提供的构造器，在简单的概念和关系上构造出复杂的概念和关系。本书在基本描述逻辑ALC的基础上添加构造器——最大数量约束（$\leqslant nR.C$）、最小数量约束（$\geqslant nR.C$）、传递关系（R^+）、反关系（R^-）、关系并（$R_1 \sqcup R_2$）、关系复合（$R_1 \circ R_2$）、个体实例集（$\{a\}$），提出扩展描述逻辑ALC^+，通过证明得到ALC^+的一些性质。

在以上研究的基础上，本书以“软件开发”领域为例，设计并开发一个基于Lucene和本体的语义检索原型系统；基于科学家资源服务领域，对已有数据进行完善、更新和管理，构建科学家资源领域本体，并介绍基于本体的科学家资源服务平台框架。

从2003年开始，在云南师范大学林毓材教授的指导和帮助下，作者在语义Web领域进行系统研究。在本书出版之际，向林毓材教授表示衷心感谢。

感谢中国科学院计算机网络信息中心阎保平研究员、昆明有色冶金设计研究院谢刚教授，两位恩师在百忙中审阅本书初稿，并给出修改意见和建议，让作者受益匪浅。

云南师范大学徐天伟、文斌、李金绪、刘江涛、张妹老师，保山学院段寿建老师，中国人民银行昆明中心支行胡绍波等参加了课题的研究，并做出了很好的科研成果。

本书得到国家自然科学基金项目、中国科学院 CNIC 创新基金项目、云南省自然科学基金项目、云南省教育厅自然科学基金重点项目、云南师范大学学术著作出版基金的资助。在此，一并表示感谢。

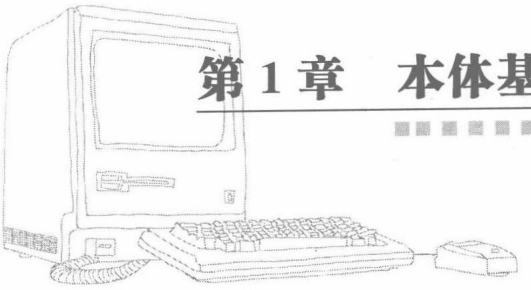
由于作者水平有限，在完整性、准确性等方面难免存在问题。希望得到广大读者特别是专家和同行的指点，共同探讨有关问题，交流研究经验，使研究工作取得更大的进步。

2011 年 5 月 10 日于中国科学院

目 录

第一篇

本体与本体描述语言



第1章　本体基础

1.1　本体概述

20世纪90年代初，本体概念被广泛地引用到计算机领域特别是人工智能（AI）和知识工程研究中，因为AI和知识工程需要开发一个领域共享的、公共的概念，实现知识共享和重用。在AI领域，本体通常被称为领域模型（Domain Model）或概念模型（Conceptual Model），是关于特定知识领域内各种对象、对象特性以及对象之间可能存在关系的理论。通过对应用领域的概念和术语进行抽象，本体形成了应用领域中共享和公共的领域概念，可以描述应用领域的知识或建立一种关于知识的描述。本体的抽象可能是很高层次的抽象，也可能是针对特定领域的概念抽象。本体已经成为知识工程、自然语言处理、协同信息系统、智能信息集成、Internet智能信息获取、知识管理等各方面普遍研究的热点。因此，随着高度结构化的知识库在AI和面向对象系统中的出现，对于实际应用和理论研究，本体变得日益重要。

最近十年以来，各种研究机构和知识工程研究者提出了多种面向AI、具有细微差别的本体定义。

（1）一个本体是一个非形式的概念化系统；

（2）一个本体是由一个逻辑理论表示的概念系统；

（3）一个本体定义了组成主题领域词汇的基本术语和关系以及用于组合术语和关系以定义词汇的外延规则；

（4）本体是概念模型的明确规范说明。

其中，Tom Gruber 的定义被引用最多，“本体是概念模型的明确规范说明”。Studer 等总结认为：“本体是共享概念模型明确的形式化规范说明。”

这包含 4 层含义：概念模型（Conceptualization）、明确（Explicit）、形式化（Formal）和共享（Share）。

“概念模型”指通过抽象出客观世界中一些现象的相关概念而得到的模型，概念模型所表现的含义独立于具体的环境状态；“明确”是指所使用的概念及使用这些概念的约束都有明确的定义；“形式化”指本体是计算机可读的（即能被计算机处理）；“共享”指本体中体现的是共同认可的知识，反映的是相关领域中公认的概念集，即本体针对的是团体的共识。

根本上，本体的作用是为了构建领域模型。例如，在知识工程过程中，一个本体提供了关于术语概念和关系的词汇集，通过该词汇集可以对一个领域进行建模。虽然不同的本体之间存在一些差异，但它们之间存在普遍的一致性。针对应用领域中一些特殊的任务，知识表达可能还需要一种在很高的普遍性层次上的本体抽象概念。

当语义通过一定的形式添加到网络资源上之后，下一步工作就是如何使得这些资源被理解和共享。对于 Web 上不同的数据资源，它们对同一个概念可能采用不同的标识符。如对于“下载工具”这个概念，可以使用〈DownloadTools〉，也可以使用〈Download _ tools〉。为了识别这些标记所代表的概念，将本体论的方法引入到语义 Web 中来。

在语义 Web 中，本体具有非常重要的地位，它是解决语义层次上 Web 信息共享和交换的基础。就 Web 而言，本体可以应用在如下方面：①提高搜索引擎的精确度，它只需根据元数据查找网页，而不会像现在用语义含糊的关键词进行全文搜索；②利用本体从 Web 页面到相应的知识结构和推理规则建立关系；③本体还可用于电子商务网站，使买卖双方可以基于机器进行交流等。

1.2 本体的组成

一个本体，一般首先给出一组概念的层次性结构，概念间的包含关系、组成关系、划分关系等。

（1）分类层次结构：如常用软件、旅游信息资源中的分类层次。

（2）按 IS-A 和 Part-of 关系组织、组成概念结构。

（3）概念的语义描述，应该不局限于静态结构。其他如“先后关系”、“因果关系”以及语义复杂的“参照关系”，具有丰富语义的关系往往无法清楚地表达出它们的语义来。

一般来说，一个本体可由概念类、关系、函数、公理和实例等 5 种元素组成。

（1）本体中的概念是广义上的概念，它既可以是一般意义上的概念，也可以是任务、功能、行为、策略、推理过程等。本体中的这些概念通常构成一个分类层次。

（2）本体中的关系表示概念之间的一类关联，典型的二元关联如子类关系形成概念类的层次结构，一般情况下用 $R: C_1 \times C_2 \times \cdots \times C_n$ 表示概念类 C_1，C_2，…，C_n 之间存在 n 元关系 R。

（3）函数是一种特殊的关系。其中，第 n 个元素相对于前 $n-1$ 个元素是唯一的，一般情况下，函数用 $F: C_1 \times C_2 \times \cdots \times C_{n-1} \rightarrow C_n$ 表示。

（4）公理用于表示一些永真式。更具体地，在许多领域中，函数之间或关联之间也存在着关联或约束。

（5）实例是指属于某概念类的基本元素，即某概念类所指的具体实体，特定领域的所有实例构成的领域概念类在该领域中的指称域。

1.3　本体建立的原则

Tom Gruber 给出了 5 条设计本体的基本准则：

（1）明确性和客观性：本体应该有效地传达所定义的术语内涵。

（2）一致性：一个本体应该前后一致，即由它推断出来的概念定义应该与本体中的概念定义一致。

（3）可扩展性：可扩展性是指一个本体提供一个共享的词汇，它应该在预期的任务范围内提供概念的基础，同时，它的表示应该使得人们能够单调地扩展和专门化说明这个词汇，即人们应该能够在不改变原有定义的前提下，以这组存在的词汇为基础定义新的术语。

（4）最小编码偏差：本体应该处于知识的层次，而与特写的符号级编码无关。

（5）最小本体承诺：一个本体应该在提供必需的共享知识条件下，要求有最小的本体承诺。

除了上述原则外，J. Arpirez 等提出本体设计应该坚持如下几点原则：

（1）尽可能使用标准术语；

（2）同层次概念之间保持最小的语义距离；

（3）可以使用多种概念层次，采用多重继承机制来增加表达能力。

1.4　本体建立的一般方法

常用的本体开发方法有：

Uschold 和 King 的“骨架法”：由英国爱丁堡大学 AI 应用研究所基于开发