语义网信息组织技术与方法

戴维民 等著

YUYIWANG
XINXI ZUZHI
JISHU YU FANGFA

学林出版社

语义网信息组织技术与方法

戴维民 等著

YUYIWANG
XINXI ZUZHI
JISHU YU FANGFA

学林出版社

图书在版编目(CIP)数据

语义网信息组织技术与方法/戴维民等著. —上海:学林出版社,2008.12

ISBN 978-7-80730-724-2

Ⅰ.语… Ⅱ.戴… Ⅲ.语义网络—信息管理 Ⅳ.TP18

中国版本图书馆 CIP 数据核字(2008)第 171500 号

语义网信息组织技术与方法

作　　者——戴维民等
责任编辑——乐惟清
特约编辑——陈金生
封面设计——魏　来
出　　版——上海世纪出版股份有限公司
学林出版社(上海钦州南路 81 号 3 楼)
电话:64515005　传真:64515005
发　　行——新华书店上海发行所
学林图书发行部(上海钦州南路 81 号 1 楼)
电话:64515012　传真:64844088
印　　刷——上海展强印刷有限公司
开　　本——640×965　1/16
印　　张——20.75
字　　数——27 万
版　　次——2008 年 12 月第 1 版
2008 年 12 月第 1 次印刷
印　　数——2300 册
书　　号——ISBN 978-7-80730-724-2/D·29
定　　价——35.00 元

前言

在网络环境下，信息揭示和组织的方式已经具有和信息内容的创作同等分量的重要性。信息组织工作的时代意义不言而喻。信息组织也自然成为当代学术活动中最具有现实价值、最令人瞩目也最具活力的研究主题之一。该领域涌现了大量研究成果，并极大地推动了网络信息组织的实践。然而，客观地讲，网络信息组织的现状仍然不尽如人意。人们逐渐认识到，网络信息组织面貌的根本性改观，不可能仅仅通过一些局部改良性质的方法、途径来实现，比如对传统信息组织方法如分类法、主题法等的网络化改造及应用，再比如对当下网络上的主要信息组织方式即搜索引擎的改进等。究其本质，在于万维网初始设计只是为了便捷人际间交流与合作，而这一主导思想决定了其技术路线的选择。相应地，万维网主流语言 HTML 仅是一种面向人类的格式显示标记，万维网上的信息对于计算机来讲，只是通过超链接简单关联起来的海量堆砌的字符串，计算机不能从它们中间发现任何语义关联。网络检索总是带来“望文生义”之憾的原因也即在此。

简言之，只有对网络的根本设计理念实现革命性变革，网络信息组织才有可能进入理想的信息存取新境界。

语义网正是这样一种对下一代万维网形态的新设想。该理念于 1998 年由万维网的发明者、W3C 主席蒂姆·伯纳斯-李(Tim Berners-Lee)首次提出，其基本思想是扩展当前的万维网，使其能够表达可以被机器(即计算机)所“理解”的语义，以便于人

和机器、机器和机器之间的交流与合作。2000年，伯纳斯-李又提出了以XML(S)、RDF(S)和ONTOLOGY三大技术为核心的语义网标准体系结构，在国际上引发了一轮语义网研发热潮。

从信息组织的视角来看，语义网理念的提出，既是新挑战，更是新机遇。它既对网络信息组织提出了一系列新要求、新任务，更为信息组织的变革提供了一系列新的有利条件，为信息组织的发展揭示了新方向。因此，摸索新形势下网络信息组织的新特点、新规律，并据此提出新理论、新技术、新方法，就成为极具现实意义和价值的重大课题。然而，当前国内对语义网进行论述的著作尚且不多，集中以信息组织为主题，系统、深入地进行专门研究的著作就更是几乎难觅一本了。本书的问世，即是希望弥补这一空白。

具体而言，本书的价值主要体现在以下三个方面：(1)通过对语义网信息组织原理与机制的剖析，通过对本体语言理论与方法的系统研究，发展了检索语言的传统方法，进一步丰富和完善了网络环境中信息组织学科的理论体系，为信息组织研究向知识组织研究的转型打下了较为坚实的基础；(2)系统建立了本体语言的理论、技术与方法，为我国发展本体语言奠定了理论基础，提供了技术保障，其中的内容可直接进入图情档专业以及其他信息管理专业的教学内容；(3)编制规范可作为国家标准的基础，促使尽早出台国家标准，从一开始就对本体语言编制进行规范。

本书共八章、一个附录，分别从理论、技术及应用层面研究和探讨语义网信息组织。

在理论层面，提出了本体开发与应用的四个方面的理论模型，具体包括：(1)本体语言的构成模型。引入时间性作为理论架构的思考基点，以共时和历时为基本维度，提出了本体语言构成的基本理论模型：首先是本体语言构成的静态理论模型，主要研究本体的结构及其构造的方法、规则；其次是本体语言构成的动态理论模型，主要研究了本体进化的基本原理与方法，研究了进化管理的技术机制及版本管理问题。(2)基于本体的信息组织模型。常见的信息组织模型都有一些共同的不足，如对概念之间关

系揭示不足、语义表达不充分、没有形式化的描述语言等。因此本书中提出了基于本体的信息组织模型，主要包括知识模型、方法模型和存储模型。(3)基于本体的信息检索模型。本书中基于本体技术构建了一个可以进行族式返回的检索模型，通过本体表达和处理检索系统中的语义，通过族式返回输出信息完全的检索结果。(4)本体标注模型。本体标注也称语义标注，是指利用本体中定义的词汇显式地揭示和表达网络中的语义。本体标注面临两大任务：一是如何将当前不计其数的普通 Web 页面转换为富含语义信息的语义页面；二是如何发布语义页面。针对这两类任务，本书中分别提出了基于本体的语义标注模型。

在技术层面，研究和开发了本体开发与应用的具体技术方法和工具，从本体的构建、存储与查询、标注技术、RDF 存储与查询以及规则描述与推理等方面详细讨论了用于语义网信息组织的具体技术方法，主要包括：(1)利用分类法、主题词表及分类主题词表构建本体。这三类“表”一般由国家或相关专业领域的权威机构公开出版发行，具有很强的权威性，是领域内公认的知识表示集合。本研究认为，利用“三表”构建本体不仅可行，而且还具有很大的成本优势。(2)使用 Protégé 构建本体。本研究认为本体开发工具 Protégé 是一款非常优秀的本体开发工具，在开发中文网络本体语言方面具有明显的优势，并给出了具体使用指导，并对其进行了汉化处理。(3)提出了以公理为中心的本体存储方案。(4)分析了如何运用 Protégé 进行规则描述和推理等。

在应用层面，取得了两项成果，即开发了一个导弹领域本体，并提出了一个基于本体的搜索引擎设计方案。

本书附录中提供的导弹领域本体中共定义了包括子概念在内的 612 个概念，134 个属性，并利用属性定义描述了概念间的逻辑关系。这些逻辑关系纵横交错形成了一个立体的、直观的语义联系网。该本体已尽可能详尽囊括了导弹领域的所有相关概念，已达到中型本体的规模，并已通过推理器 racer 的逻辑推理检验。

针对目前搜索引擎发展的技术瓶颈，结合本体在信息组织、语义揭示、知识表现方面的优势，应用前述基于本体的信息组织

模型、基于本体的信息检索模型，本书提出了一个基于本体的搜索引擎设计方案，具体包括搜索引擎的基础设计、接口设计、核心设计和优化设计四个方面，并对其中的查询请求获取、族式返回检索和排序算法等进行了深入的讨论。

本书是国家社科基金项目“面向网络信息组织的中文网络本体语言研究”（项目编号：04BTQ026）的最终研究成果，我作为课题负责人，得到了罗昊、田春虎、王梅等主要成员的大力支持，正是他们潜心和深入的研究，使我们比较顺利地完成了课题研究任务。本书是由我们4人集体撰写的。在课题研究的不同阶段，刘永丹、孙瑾、包冬梅等也参与了研究工作。同时，在研究和评审阶段我们还广泛听取了一些专家的宝贵意见，并参阅了国内外同行的大量相关成果，在此付梓之际，一并表示衷心的感谢。

本书出版得到了国家社会科学基金以及南京政治学院上海分院重点学科学术著作出版基金的资助。

本书作为国内从信息组织角度系统研究语义网的引玉之作，希望能引发大家对语义网信息组织的兴趣和关注，推动相关研究和实践更好更快发展。

戴维民

二〇〇八年十二月

目录

CONTENTS

第1章 语义网基础

1.1 语义网概述

1.1.1 语义网的概念

微软公司董事长比尔·盖茨那幢坐落在西雅图、被喻为未来生活预言的科技住宅，无疑是当今世界上最现代化的豪华住宅之一，堪称智能建筑的经典之作。住宅内共铺设各种电缆52英里，将所有的电器设备相互连接成了一个智能网络。主人即使出门在外，也可以通过网络遥控家中的任何一件电器。

在住宅内还装有气象感知器，它可以根据各项气象指标，控制室内的温度和通风情况。走进大厅时，空调系统会将室温调整至你感觉最舒适的温度，音响系统也会针对你的喜好播放音乐，灯光系统自动调整照明颜色与强度，就连墙上的LCD显示屏，也会自动显示你喜爱的世界名画或播放你上次只看到一半的影片。

在住宅内四处随意走动时，地板能在6英寸的范围内跟踪到人的足迹，在有人经过时会自动打开照明，离去时自动关闭。

每个房间的温度、照明、音响等都将随不同的设定自动调整。即使是在水池中，也会从池底“冒”出如影随形的音乐。尤其有意思的是，比尔·盖茨非常喜欢车道旁边一棵 140 岁的老枫树，所以就通过专门的监视系统对其进行 24 小时的全方位监控，一旦监视系统发现它有任何干燥的迹象，灌溉系统就会启动浇水程序。

人们不禁要问，是什么样的尖端科技系统使得盖茨先生的豪宅拥有如此高的智能化水平？你可能不曾想到，通过扩展今天的万维网(World Wide Web，即 www)就完全可以使这一梦想变为现实。这种扩展后的万维网称为语义网(Semantic Web)。

语义网的概念由万维网的发明者、现任万维网联盟(World Wide Web Consortium，即 W3C)主席蒂姆·伯纳斯-李(Tim Berners-Lee)于 1998 年首次提出。在他看来，“语义网并非是另外一个独立的万维网，而是现在的万维网的一个延伸。在其中，所有的信息都具有定义完好的含义，更利于人与机器之间的合作。”①

现在万维网上的大部分内容都是设计给人阅读的，而不是让计算机程序按网页的内容自动进行处理。语义网的目的就是扩展当前的万维网，使其能够表达可以被机器(计算机)所“理解”的语义，以便于人和机器(计算机)以及机器和机器之间的交互与合作。

语义网中的“语义”，实际上就是指文本的含义。例如对于文本“计算机”，当我们看到这三个字时，我们首先会想到这是一个与我们日常学习、工作和生活息息相关的具有实物形态的工具，甚至就是我们手头正在使用的“东西”；更进一步，我们知道“计算机”这一文本既不代表数字、也不表示某一动作，而是一个名词概念；并且就此概念我们可以说出一系列与之相关的信息，如它的

① Tim Berners-Lee, James Hendler, Ora Lassila. The Semantic Web. Scientific American. [2006-10-5].
http://www. ryerson. ca/~ dgrimsha/courses/cps720 _ 02/resources/Scientific%20American%20The%20Semantic%20Web. htm

基本组成、配置和功能等。因此，我们（或是说“人”）可以很好地“理解”“计算机”这一文本所代表的含义。而对于机器而言，事情就没有那么简单了。它非但不知道“计算机”这一文本所代表的一系列“含义”，甚至连这一文本代表的是数字还是字符都不知道。因此语义网的主要任务之一，就是要以某种方式或方法明确地“告诉”计算机某一文本的具体含义。

语义网目前是 W3C 的主要研究目标之一，也是国际国内研究的热点。它通过表达充分、完备的可以为计算机“理解”的语义，使现有的万维网具有一定的推理和自动处理能力，从而彻底改变现有万维网的工作方式以及人们的生活方式。因此可以说，语义网就是未来的万维网，是下一代的互联网络。

1.1.2 语义网与万维网的区别

语义网是万维网的延伸，但却与万维网有着很大的不同，主要表现在：

- **面向的对象不同**

目前的万维网主要使用 HTML 表达网页内容。使用 HTML 标记的网页的确可以表达一些控制网页显示格式之类的信息，从而使人们认为计算机真的可以“理解”我们的意图。但实际上 HTML 仅注重文本的表现形式，如字体颜色、大小、类型等，而不考虑文本的具体内容与含义。虽然万维网上有一些自动的脚本程序可以帮助人们实现一部分功能，但在开放式的网络环境中，它们并不能很好地用于计算机之间的交互。因此目前我们所使用的万维网主要是供“人”阅读和使用的。

而语义网则是要在万维网之上加入一些可以被计算机“理解”的语义信息，它在方便人们阅读和使用的同时，也方便计算机之间的相互交流与合作。因此，万维网面向的对象主要是“人”，而语义网面向的对象则主要是“机器”。

- **信息组织方式不同**

由于两者面向的对象不同，因此在信息组织方式上自然会存在很大的差异。万维网在组织信息资源时主要以“人”为中心，按

照人们的思维习惯和方便性组织网络信息资源。语义网在组织信息资源时则必须兼顾计算机对文本内容的“理解”以及它们之间的相互交流和沟通。

● 信息表现的侧重点不同

万维网侧重于信息的显示格式和样式，而不关心所要显示的内容。例如对于比较重要的信息，万维网可能会在其显示上以大字体或颜色鲜明的字体表示；而语义网则更加侧重于信息的语义内容，对具有特定意义的文本必须进行一定的标注或解释。

● 主要任务不同

万维网主要是供人阅读、交流和使用的，其主要任务就是信息发布与获取。通过在网络上发布或获取信息来达到共享和交流的目的。语义网的主要任务则是计算机之间的相互交流和共享，从而使计算机可以代替人们完成一部分工作，使网络应用更加智能化、自动化和人性化。

● 工作方式不同

语义网与万维网面向的对象不同，它们的工作方式自然也有所不同。万维网主要面向“人”，因此其大部分工作都是由人来完成的，包括信息的收集、检索、整理、排序和分析等等。而语义网通过加入一些可以被计算机“理解”的语义信息，则可以把人从上述各类繁琐的工作中解脱出来，利用“智能代理”帮助完成上述的大部分工作。一个典型的例子就是信息检索，利用智能搜索代理，语义网将提供给人们真正需要的信息内容，而不像现在的搜索引擎那样输出数以万计的无用的搜索结果。

1.1.3 语义网的优点

语义网的最大优点就是对网络信息的“理解和处理”能力。通过加入可以被计算机“理解”的语义，从而使得对文本含义的理解并不再是人的专利，利用计算机同样也可以完成相同的工作。

例如，对于网上书店关于某本书的介绍，我们可以很容易地