国家科学技术学术著作出版基金资助项目

The Semantic Grid: Fundamental Issues, Methodologies and Applications

语义网格：模型、方法与应用

吴朝晖 陈华钧 著
袁保宗 主审

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社
Springer

国家科学技术学术著作出版基金资助项目

语义网格：模型、方法与应用

吴朝晖　陈华钧　著
袁保宗　主审

浙江大学出版社

内容提要

语义网格综合语义 Web 的语义表达技术和网格计算的分布式体系架构技术，为下一代互联网技术的发展提供新的思路、方法和技术规范。本书在对这一国际前沿领域进行分析和综述的基础上，围绕语义网格的知识表达、信任计算、复杂问题求解、分布式数据集成、复杂流程与服务的管理、分布式的数据挖掘与知识发现等核心问题展开了系统性的论述。此外，还结合生命科学和智能交通系统两个应用领域的特点，介绍了语义网格一些相关核心技术的实际应用，展示相关技术的潜在应用前景。

本书可以作为从事网格计算、语义 Web、语义网格研究和开发工作的人员以及大专院校有关专业师生的参考书。

图书在版编目（CIP）数据

语义网格：模型、方法与应用 / 吴朝晖，陈华钧著.
—杭州：浙江大学出版社，2008.2
ISBN 978-7-308-05681-6

Ⅰ. 语.… Ⅱ. ①吴…②陈… Ⅲ. 语义网络—研究 Ⅳ. TP18

中国版本图书馆 CIP 数据核字（2007）第 196794 号

语义网格：模型、方法与应用

吴朝晖　陈华钧 著
袁保宗　主审

策　　划	希　言
责任编辑	黄娟琴　邹小宁
封面设计	刘依群
出版发行	浙江大学出版社 （杭州天目山路 148 号　邮政编码 310028） （E-mail：zupress@mail.hz.zj.cn） （网址：http://www.zjupress.com http://www.press.zju.edu.cn） 电话：0571－88925592，88273066（传真）
排　　版	浙江大学出版社电脑排版中心
印　　刷	杭州余杭人民印刷有限公司
开　　本	787mm×1092mm　1/16
印　　张	15.75
字　　数	327 千
版 印 次	2008 年 2 月第 1 版　2008 年 2 月第 1 次印刷
书　　号	ISBN 978-7-308-05681-6
定　　价	48.00 元

前 言

随着互联网的发展与普及，以智能分布式的方式协同共享与集成管理海量的网络信息资源，已经成为一个十分突出的难题和亟待解决的问题。语义网格作为网格计算的一个分支，用一种包含语义的方式描述各类网格资源，使之更加易于被发现、聚合和连接。这种资源描述通常是基于语义Web的关键技术如RDF和OWL来实现的。语义网格将语义Web为代表的语义技术和以网格计算为代表的体系架构技术结合起来，为下一代互联网提供了开放、安全、有序、可扩展的管理体系架构，并支持解决和实现复杂网络环境下跨多个机构的大规模分布式协同计算和信息共享问题。

语义网格作为网格计算和语义Web的交叉技术，为未来互联网环境的发展提供了新的思路、技术和方法。本书针对这一发展趋势进行了有益的探讨与探索，并对语义网格所涉及的一些主要问题展开了广泛的讨论。具体而言，重点针对语义网格中的语义表达和知识表示方法、语义网格中的数据集成和管理、基于语义的流程组合与服务拼接、语义网格中的信任管理与问题求解，以及基于语义网格的数据挖掘与知识发现等问题，进行了分析和探讨。最后，综合理论探索和应用研究探讨了语义网格中的各种技术如何被实际应用到诸如医学信息学和智能交通管理系统。

本书第1章介绍了语义网格的概念和发展历程。第2、3、4章从理论方法和模型的角度探讨了语义网格相关的核心问题，比如第2章从人工智能知识表达领域出发，阐明了知识表达技术是语义网格的核心技术之一，是人工智能技术在互联网领域的重要应用。第3章探讨了语义网格环境下的复杂问题求解。第4章提出了一个语义网格中的分布式信任计算模型。第5、6、7、8章从关键技术的角度对语义网格进行了介绍。第5、6章介绍了基于语义的数据库网格技术，即如何在语义网格中整合和管理跨多个机构的数据库资源。其中的一些关键技术包括语义映射技术、语义注册技

术、语义查询重写技术等。第7章介绍了基于语义的服务流程发现与组合技术,即如何在语义网格中整合跨多个机构的流程服务资源。其中的一些关键技术包括服务流程的语义组合、基于语义的服务发现等。第8章介绍了语义网格中的数据挖掘与知识发现。第9章从数据管理、流程组合和消息中间件三个方面介绍了一个称为DartGrid的语义网格平台系统。第10章、第11章分别介绍了语义网格的中医药应用和在ITS智能交通系统中的应用。从应用的角度具体介绍了语义网格的相关关键技术的潜在应用价值。

本书是经过浙江大学CCNT实验室的众多科研人员多年学习、研究和工程实践沉淀的成果。参与本书工作的人员包括:郑骁庆、毛郁欣、周春英、张宇、邓水光、封毅、于彤、郑国轴、施伟、吴健等。在此对他们表示衷心的感谢。

本书所介绍的工作主要得到了国家自然科学基金杰出青年基金"智能空间的语义模型与行为感知认证"(No. NSFC60533040)、国家"973"计划语义网格专项子项目"语义网格在中医药知识共享与服务中的应用"(No. 2003CB317006)、教育部新世纪人才计划"面向普适计算的嵌入式语义网格及若干关键技术研究"(No. NCET-04-0545)和现代服务业服务基础技术研究(2006BAH02A01)项目的资助;此外,参与本书相关项目的其他研究人员还得到了长江学者和创新团队发展计划资助(IRT0652)、国家"863"高科技发展计划(No. 2006AA01A122)、国家自然科学基金项目(NSF60503018, NSF60603025)、国防预研(No. 060651306030101, No. 9140A06060307JW0403, No. A1420060153)等的资助;本书的出版得到"国家科学技术学术著作出版基金"的资助。在此,一并表示感谢。

语义网格是当前处于科学前沿的论题,许多理论和思想还处于探索阶段,由于作者的水平和经验有限,错误和不妥之处在所难免,恳请读者给予批评指正,共同推进下一代互联网技术研究的进步与发展。

作　者

2007年9月1日

目 录

第 1 章　语义网格概述 …… (1)

1.1　概　述 …… (1)
1.1.1　网格计算 …… (2)
1.1.2　语义 Web …… (5)
1.2　语义网格主要研究概况 …… (8)
1.2.1　国外研究现状 …… (8)
1.2.2　国内研究现状 …… (10)
1.3　语义网格的基本概念 …… (11)
1.3.1　语义网格的定义 …… (11)
1.3.2　语义网格需要探索的关键问题 …… (11)
1.3.3　语义网格需要实现的关键技术 …… (12)
1.3.4　语义网格的层次模型 …… (13)
1.4　语义网格的典型应用举例 …… (14)
1.4.1　语义网格的医学应用 …… (14)
1.4.2　语义网格的电子商务应用 …… (15)
1.5　小　结 …… (15)
参考文献 …… (16)

第 2 章　语义网格与知识表达 …… (18)

2.1　概　述 …… (18)
2.2　语义网格知识表示的理论基础 …… (19)
2.2.1　逻辑形式系统 …… (20)
2.2.2　语义网络 …… (21)
2.2.3　框架系统 …… (22)
2.2.4　本体论 …… (23)
2.2.5　描述逻辑 …… (24)

2.3 语义 Web 的知识表示框架 …… (25)
2.3.1 语义 Web 的技术层次结构 …… (25)
2.3.2 XML 与 XML Schema …… (27)
2.3.3 资源描述框架 RDF 与 RDFS …… (27)
2.3.4 本体语言 OWL …… (28)
2.3.5 比较研究 …… (29)
2.4 一个本体开发的实际案例 …… (33)
2.4.1 开发本体的关键步骤 …… (33)
2.4.2 中医药本体设计与开发 …… (34)
2.4.3 中医药本体开发的成果 …… (37)
2.5 小 结 …… (42)
参考文献 …… (43)

第 3 章 面向语义网格的问题求解 …… (46)

3.1 概 述 …… (46)
3.1.1 问题求解 …… (46)
3.1.2 分布式协同问题求解 …… (47)
3.1.3 多智能体系统 …… (48)
3.2 基于语义网格的问题求解 …… (49)
3.2.1 网格与问题求解 …… (49)
3.2.2 面向语义网格的问题求解 …… (50)
3.3 支持问题求解的本体网格 …… (51)
3.3.1 基于网格的本体管理架构 …… (52)
3.3.2 本体网格节点 …… (52)
3.3.3 语义视图 …… (56)
3.4 支持问题求解的本体重用 …… (58)
3.4.1 动态存储模型 …… (58)
3.4.2 基于案例的本体重用模型 …… (59)
3.5 子本体进化的问题求解 …… (62)
3.5.1 子本体操作 …… (62)
3.5.2 基本概念 …… (64)
3.5.3 问题求解环境 …… (64)
3.5.4 基于子本体进化的问题求解 …… (66)
3.6 问题求解与语义网格的关系 …… (68)
3.7 相关工作 …… (70)
3.8 小 结 …… (71)

参考文献 …… (72)

第 4 章 面向语义网格的信任计算 …… (74)

4.1 概 述 …… (74)
4.2 语义网格与信任计算 …… (75)
4.3 信任的特点和涉及的因素 …… (76)
4.3.1 信任的特点 …… (76)
4.3.2 信任涉及的因素 …… (77)
4.4 信任信息交互的语义 …… (77)
4.5 “封闭”信任计算模型 …… (79)
4.6 “开放”信任计算模型 …… (82)
4.7 实验和结果分析 …… (84)
4.8 相关工作 …… (87)
4.9 小 结 …… (87)
参考文献 …… (88)

第 5 章 数据库语义网格 …… (90)

5.1 概 述 …… (90)
5.2 数据库语义网格模型与体系架构 …… (91)
5.2.1 数据库语义网格的虚拟组织 …… (91)
5.2.2 数据库语义网格虚拟组织的构成 …… (92)
5.2.3 数据库语义网格虚拟组织角色描述 …… (93)
5.2.4 数据库语义网格的核心组件 …… (94)
5.3 数据库资源访问管理协议 …… (96)
5.3.1 数据库元信息访问协议 …… (96)
5.3.2 数据库会话管理协议 …… (98)
5.3.3 数据库断言控制协议 …… (100)
5.3.4 数据库模式管理协议 …… (107)
5.3.5 数据库授权管理协议 …… (109)
5.4 数据库语义解析协议族 …… (113)
5.4.1 本体查询协议 …… (113)
5.4.2 语义注册管理协议 …… (115)
5.5 小 结 …… (118)
参考文献 …… (118)

第 6 章 语义映射与语义查询 …… (120)

6.1 概 述 …… (120)

6.2 背景知识 …… (122)
6.2.1 SHIQ 描述逻辑 …… (122)
6.2.2 基于视图的查询和问答 …… (123)
6.3 语义映射系统 …… (124)
6.4 语义查询处理 …… (126)
6.4.1 语义查询 …… (126)
6.4.2 语义查询重写 …… (128)
6.5 小 结 …… (133)
参考文献 …… (134)

第 7 章 语义网格中的服务流程管理 …… (137)

7.1 概 述 …… (137)
7.2 服务流程管理的技术基础 …… (138)
7.2.1 工作流管理技术 …… (138)
7.2.2 Web 服务技术 …… (140)
7.2.3 语义 Web 服务技术 …… (142)
7.3 服务流程管理的现状分析 …… (144)
7.3.1 服务流程管理的研究框架 …… (144)
7.3.2 Web 服务组合 …… (145)
7.3.3 服务流程的表达语言 …… (150)
7.3.4 服务流程的验证方法 …… (152)
7.3.5 服务流程管理相关平台 …… (154)
7.4 服务流程管理的关键方法 …… (156)
7.4.1 基于接口依赖的语义 Web 服务发现方法 …… (156)
7.4.2 基于柔性工作流的服务组合方法 …… (159)
7.4.3 基于 π 演算的服务兼容性验证方法 …… (161)
7.5 小 结 …… (164)
参考文献 …… (165)

第 8 章 语义网格中的数据挖掘与知识发现 …… (168)

8.1 概 述 …… (168)
8.2 KDD 系统体系结构的发展 …… (169)
8.2.1 单机体系结构 …… (169)
8.2.2 并行体系结构 …… (170)
8.2.3 分布式体系结构 …… (171)
8.2.4 基于网格的体系结构 …… (171)

8.2.5 KDD 体系结构发展总结 …… (173)
8.3 基于语义网格的知识发现 …… (174)
8.3.1 基于语义网格的知识发现的虚拟组织模型 …… (174)
8.3.2 基于语义网格的知识发现的体系架构和组件 …… (175)
8.3.3 基于语义网格的知识发现的特点 …… (177)
8.4 中医药方剂挖掘原型系统 …… (178)
8.4.1 平台体系结构 …… (178)
8.4.2 平台技术实现 …… (179)
8.4.3 平台应用实例:方剂药对挖掘 …… (180)
8.5 小 结 …… (182)
参考文献 …… (182)

第 9 章 DartGrid 语义网格平台 …… (185)

9.1 概 述 …… (185)
9.2 基于语义的数据库集成 …… (186)
9.2.1 系统体系架构 …… (186)
9.2.2 DartMapping:可视化语义映射编辑工具 …… (187)
9.2.3 DartQuery:语义浏览和查询工具 …… (188)
9.2.4 DartSearch:基于语义的搜索 …… (189)
9.3 基于语义的服务流程管理 …… (191)
9.3.1 系统体系架构 …… (191)
9.3.2 DartFlow 的主要功能 …… (193)
9.3.3 DartFlow 的主要特色 …… (195)
9.4 基于语义的消息中间件 …… (196)
9.4.1 系统概述 …… (196)
9.4.2 系统体系架构 …… (198)
9.4.3 基于语义的消息征订 …… (199)
9.5 小 结 …… (202)
参考文献 …… (203)

第 10 章 语义网格与中医药 e-Science 环境 …… (204)

10.1 概 述 …… (204)
10.1.1 中医药信息化的现状 …… (204)
10.1.2 中医药信息化与共享的主要问题 …… (205)
10.2 中医药 e-Science 体系结构 …… (206)
10.2.1 体系结构概述 …… (207)

10.2.2 中医药 e-Science 环境的应用平台 …… (208)
10.3 中医药语义网格应用平台 …… (209)
10.3.1 中医药本体工程 …… (209)
10.3.2 基于语义的中医药数据共享平台 …… (212)
10.3.3 中医药数据加工与共建平台 …… (213)
10.3.4 中医药数据挖掘与知识发现平台 …… (215)
10.4 小 结 …… (218)
参考文献 …… (218)

第 11 章 语义网格在智能交通系统中的应用 …… (219)

11.1 概 述 …… (219)
11.1.1 背 景 …… (219)
11.1.2 网格技术与智能交通系统 …… (221)
11.1.3 交通本体库 …… (222)
11.2 智能交通信息服务共享平台的功能需求 …… (223)
11.2.1 信息服务共享平台的应用场景 …… (223)
11.2.2 信息服务共享平台的目标功能 …… (226)
11.3 智能交通语义网格的服务体系框架 …… (227)
11.3.1 资源层 …… (228)
11.3.2 服务层 …… (229)
11.3.3 ITS 服务层 …… (229)
11.3.4 ITS 子系统层 …… (230)
11.3.5 ITS 应用层 …… (230)
11.4 智能交通语义网格原型系统的设计与实现 …… (230)
11.4.1 系统体系架构 …… (230)
11.4.2 智能交通语义本体的设计 …… (231)
11.4.3 基于智能交通本体库的交通信息查询门户 …… (234)
11.4.4 杭州城市交通诱导服务 …… (235)
11.5 小 结 …… (235)
参考文献 …… (237)

第 1 章

语义网格概述

【摘要】 语义网格是一种基于语义的分布式计算技术，它建立在语义 Web (Semantic Web)及网格计算(Grid)相关技术规范基础之上，通过规范化语义来表达信息和描述资源，通过开放和有序的管理体系架构来解决和实现在复杂网络环境下跨多个机构的大规模分布式协同计算和信息共享。本章对语义网格的研究背景、发展历史和国内外研究进展进行了综述。

1.1 概 述

随着信息化和网络化的飞速发展与深入应用，互联网逐渐成为人们日常生活与工作中的信息共享空间和协同工作平台。一方面，各个领域都产生了极为巨大的海量信息，例如在生物信息学、医学以及天文气象学等领域。这些海量信息的典型特征是广泛分布、深度异构、分散自治。如何提供一个易于扩展、容错的分布式计算基础设施，支持这些海量信息的协同共享、语义集成以及综合管理，成为一个亟待解决的难题。另一方面，跨多个机构的大规模协同工作，例如全球协同科学研究、跨企业的电子商务等，日益需要处理复杂流程和服务的灵活组合和动态集成。由此，如何提供一个强有力的分布式计算平台，支持更加灵活和易于扩展的大规模协同工作，也是一个极大的挑战。

互联网技术的飞速发展为解决这些难题提供了可能。以语义 Web 为代表的语义技术，以其严格的逻辑理论基础和标准化的技术路径，正逐渐成为构建未来互联网系统的一项关键性支撑技术。事实上，缺乏统一的资源语义表达模型，是造成在分布式系统中资源难于被发现和集成，难以建立资源之间的逻辑连通性的本质原因之一。语义技术通过明确的、规范化的描述信息资源的语义试图解决互联网系统中资源的自动发现、数据的直接交换与服务的无缝集成，并希望通过缩小人的认知域与计算机的处理域之间的距离来支持人们用直观的语义对信息资源在概念层次进行直观的操作。同时，以网格技术为代表的体系架构技术通过在不同层面定义标准和规范，为下

一代互联网提供更加适应变化、自主容错、动态可扩展的分布式计算体系架构。并通过研究如何在现有的互联网上实现一个支持共享的分布式基础设施,来支持面向共享的应用程序的开发、运行和管理,支持跨地域、跨领域、跨单位的虚拟组织的动态生成和有效管理。

语义网格,将以语义 Web 为代表的语义技术和以网格计算为代表的体系架构技术结合起来,通过规范化描述和明确表达包括计算、存储、数据库、服务等各种信息资源的内涵语义,提供开放、安全、有序、可扩展的管理体系架构来解决和实现复杂网络环境下跨多个机构的大规模分布式协同计算和信息共享问题。

1.1.1 网格计算

1. 发展简述

网格(Grid)这个词来源于电力网格(Power Grid),最初兴起于高性能计算领域,其简单的理想是利用互联网把分散在不同地理位置的电脑组织成一台"虚拟化的超级计算机",并像电力网透明地对不同的终端用户提供一致的电力服务一样,向计算用户提供透明的计算力服务。一些早期的典型应用都来源于计算密集型的科学研究应用,如在高能物理、生命科学、天文学等领域的高端计算应用,都是将分布在全球的闲置计算机或位于不同地域的高性能计算机聚合起来提供可靠并且廉价的计算服务。例如,Folding@Home[①]就是一个研究蛋白质折叠及由此引起的相关疾病的分布式计算工程,它通过使用联网式的计算方式和大量的分布式计算能力来模拟蛋白质折叠的过程,从 2000 年 10 月 1 日起,世界各地有近百万个 CPU 参加了该项目。

随着计算网格的成功,网格的研究范畴迅速扩充到对各种信息资源包括计算、存储、数据、软件、设备等的大规模协同共享。在 2001 年,被尊称为网格之父的I. Foster 先生就对网格计算的概念进行了详细的界定,并定义其主要目标是"在动态变化的多个组织之间和谐地共享各种软硬件信息资源,支持协同解决问题"[1~4],并把其领导开发的网格软件平台 Globus[②] 界定为"面向共享的分布式计算基础设施",强调其主要目的是要为构建跨多个机构的虚拟组织提供一个可靠、一致的开发和运行平台。从这个角度讲,网格计算是一种高度融合的"协同计算"。在这种"共享协同计算"环境中,用户可以从中享受一体化的、动态变化的、可灵活控制的协作式信息服务。正因为如此,网格技术日益受到关注,目前已经从科学研究应用起步,进入到制造业信息化、电子政务、企业协同、教育信息化、娱乐空间等多种应用领域。

早在 2000 年,在美国、欧洲、亚太地区的网格研究者的大力推动下,正式成立了

① Folding@Home 项目官方网站:http://folding.stanford.edu/

② Globus 项目官方网站:http://www.globus.org

全球网格论坛 Global Grid Forum[①](简称 GGF),并全面推动网格计算相关规范和标准的制定工作。2003 年,在 IBM、Oracle 等企业的推动下,又成立了企业网格联盟 EGA(Enterprise Grid Alliance)。在随后的几年中,全球各个国家和地区都纷纷启动了大量的网格项目。如美国 TeraGrid、GriPhyN、Grid3,欧盟的 EuroGrid、DataGrid、MyGrid、CrossGrid,中国的 CNGrid、ChinaGrid 等。同时,网格计算的研究也引起了商业界的极大重视,商业界各大型企业纷纷投入巨资大力开发网格产品、技术和服务,典型的比如 IBM 的 Grid Toolbox、Oracle10g 数据库网格等系统与工程,都试图在以网格为核心的软件技术的竞争中取得主导地位。

2006 年,全球网格论坛 GGF 与企业网格联盟 EGA 通过重新组合相关工作组以后,成立了开放网格论坛 OGF[②](Open Grid Forum)。OGF 充分吸收两个国际技术标准组织的优势和特点,并极大地扩展了网格计算在企业应用领域的标准化工作,并明确提出了企业网格计算(Enterprise Grid Computing)的概念,强调企业网格的主要目的是为企业提供基于 SOA (Service-Oriented Architecture)的解决方案,支持更加灵活、自动、易于扩展的方式帮助企业管理日益庞大的 IT 基础设施,这既包括硬件也包括软件,既包括计算资源,也包括数据库、应用程序等各个方面的资源。

2. 网格基本概念及关键技术

网格计算的核心概念之一是"虚拟组织"[1]。一个虚拟组织由来自于地理上分布、逻辑上独立的多个机构为解决一个共同的问题或实现共同的目标,动态组成的跨多个管理域的虚拟计算组织。一个虚拟组织既包括通过网络互联起来的各种信息资源,又包括通过信息资源接入网络的各种设备仪器,也包括管理和使用信息资源的人。在虚拟组织中,各种异质异构资源都以网格服务的方式发布[5],并通过网格层的协议进行聚合、集成和互操作。

"体系结构"是网格计算另一重要概念。网格计算实现的主要途径是从开放系统的体系架构的角度出发,通过在分布式系统的不同层面定义标准和规范,以及研究如何在现有的互联网上实现一个支持共享的分布式基础设施,支持跨地域、跨领域、跨单位的虚拟组织的动态生成和有效管理。一个开放的网格体系架构标准和基础设施规范对于实现网格所提出的理想至关重要。概括起来讲,网格体系结构就是关于如何建造网格的技术。它给出了网格的基本组成与功能,描述了网格各组成部分的关系以及它们集成的方式或方法,刻画了支持网格有效运转的机制。

网格研究者在网格体系架构方面已经进行了非常深入的工作。最早由 I. Foster 提出的五层沙漏模型主要是面向科学研究领域高端设备的管理和共享的分层体系架构[1]。在该体系结构中,把网格系统刻画分为典型的五层:构造层(Fabric)对应于具

① Global Grid Forum: http://www.ggf.org

② Open Grid Forum: http://www.ogf.org

◇ **跨多个管理域的安全机制**。由于网格中的资源通常属于不同的机构或个人，并通常要求具有局部的自治性。所以，要在多个管理域之间进行资源共享和协同，必须提供一种有效的方式解决跨多个管理域的用户授权、安全认证等问题。网格一方面在标准上定义了跨管理域的安全规范，另一方面在基础设施上实现了多种适合网格特征的安全机制，比如代理认证(Delegation)、单点登录(Single Sign On)等。

1.1.2 语义 Web

1. 发展简述

传统的信息获取技术和搜索引擎技术通常采用自然语言处理、数据挖掘和统计分析等方法对 Web 文档进行处理。这些技术的瓶颈是无法完整、有效、精确地提取出蕴涵于文档中的信息语义。1994 年，Web 创建者 T. Berners-Lee 在年度国际万维网会议上指出："Web 文档本身描述的是现实世界中的对象、概念和它们之间的关系，但这些信息都是用自由文本描述的，这虽然方便人们浏览，但机器却无法自动提取与理解 Web 文档中所蕴涵的语义。"这是导致当前搜索引擎无法对信息进行精确搜索的根本原因。随后不少研究人员致力于通过在 Web 网页中增加元数据来提高信息的易检索性和易处理性。1996 年，美国马里兰大学J. Hendler领导发起了一个名为 SHOE 的项目(Simple HTML Ontology Extension)[7]，其主要目的是通过在 Web 网页中增加语义本体描述来实现基于语义的搜索，这个项目为语义 Web 概念的形成提供了基础。

1998 年，T. Berners-Lee 与 J. Hendler 等首次明确提出了语义 Web(Semantic Web)的概念[8~10]，并将其定义为"是现有 Web 的扩展，并通过在 Web 中增加机器可理解的语义来更好地使机器与人之间进行互操作"。随后全球万维网标准化组织 W3C① 正式成立了语义 Web 工作组②，并陆续开展了一系列的相关标准化工作。

2001 年，W3C 完成了 RDF(Resource Description Framework)的标准化工作。RDF 是有关 Web 语义表达的最基础的规范，它规定了如何对 Web 资源的语义进行规范化、明确化描述的基本方法和框架。RDF 吸取了关系数据库模型的经验，借鉴了人工智能领域的知识表达研究成果，并针对互联网的本质特征而设计，是专门面向信息集成的数据模型。它比起 XML，语义描述更加简单直观，语义约束更加明确和规范。从某种角度来讲，XML 是资源描述的语法规范，而 RDF 是在 XML 之上，用于资源描述的语义规范。

但由于 RDF 在本体(Ontology)和术语规范(Terminology)定义上的表达能力不

① W3C 标准化组织：http://www.w3.org/

② W3C 语义 Web 工作组：http://www.w3.org/2001/sw/

足,W3C 又启动了 OWL(Web Ontology Language)[①]的规范化工作。OWL 是具有更强表达能力的资源语义表达规范,它充分借鉴了知识表达领域有关语义网络、框架系统和描述逻辑的研究,能用于定义复杂的领域本体和术语规范,比如医学术语等。2004 年,W3C 完成了有关 OWL 的规范化工作。但是要实现语义 Web 的目标,仍然还有大量的规范化和标准化工作需要做。比如,W3C 于 2005 年启动了 SPARQL[②],2006 年启动了 GRDDL[③]、RDFa[④]、RIF[⑤] 等一系列的标准化。其中,SPARQL 是标准化的语义查询语言,GRDDL 主要用于从 XML/HTML 中抽取 RDF 语义信息,RDFa 用于在 XML/HTML 中嵌入基于 RDF 的语义微内容,RIF 是标准的规则交换语言(Rule Interchange Format)。此外,W3C 还成立很多兴趣小组支持一些相关应用领域的研发工作,比如针对生命科学领域应用的兴趣小组,旨在推动语义技术在生命科学领域的应用[⑥]。

当前,国内外不少大学与研究机构都参与了语义 Web 的研发。比如美国的 MIT、斯坦福大学、卡内基梅隆大学、马里兰大学,欧洲的曼彻斯特大学、DFKI、ERCIM 等。特别是欧盟,最近几年陆续启动和完成了大量的有关语义 Web 的项目和工作,如 SWAD-Euro[⑦]。在企业方面,惠普研究院开发的 Jena 工具包[⑧]已经被大多数语义 Web 相关项目所使用;Nokia 公司开发了语义 Web 服务器[⑨];Adobe 公司的 XMP 已经支持对 PDF 文档添加语义描述信息;开发了 XMLSpy 的著名的 Altova 公司也推出了专门用于语义编辑和本体建模的工具 SemanticWorks;Oracle公司在其 10.2 的版本中已经加入了 RDF 存储与管理功能。另外,围绕语义 Web 还诞生了一系列新公司,如 SemaGix、NetworkInference、RDF Gateway 等。

语义 Web 的最终目标是要通过语义把各种数据(Data)和程序(Program)互联起来,综合利用知识的方法解决信息资源的语义问题,进而解决资源的共享问题;使 Web 成为一个能提供知识服务的巨大知识库。

2. 语义 Web 的基本概念和核心技术

语义 Web 的一个核心概念是语义本体(Semantic & Ontology)。本体内涵是指世界的概念模型,外延又包括比如元数据、数据库中的 E-R 模型、软件工程领域的 UML 所描述的模型、搜索引擎的目录、产品分类表、某个领域的词汇表等,这些都可

① W3C Web Ontology 语言工作组:http://www.w3.org/2004/OWL/
② W3C Data Access 工作组:http://www.w3.org/2001/sw/DataAccess/
③ GRDDL: http://www.w3.org/TR/grddl/
④ RDFa:http://www.w3.org/TR/xhtml-rdfa-primer/
⑤ RIF:http://www.w3.org/2005/rules/wg
⑥ W3C 语义 Web 与生命科学兴趣小组:http://www.w3.org/2001/sw/hcls/
⑦ 欧盟的语义 Web 高级研发项目:http://www.w3.org/2001/sw/Europe/
⑧ 惠普 Jena 工具包:http://jena.sourceforge.net/
⑨ Nokia 的语义 Web 服务器:http://sw.nokia.com/

以看成不同复杂程度和规范化程度上的语义本体。而语义 Web 的核心思想就是通过提升描述信息的语义本体以及规范化程度来支持更加方便、迅速和智能化的信息集成、聚合与融合。图 1.1 显示了语义 Web 的这一本质思想。可以看到，传统的 Web 资源以一种隐语义的方式存在，大量的数据逻辑以机器难以处理的自由文本存在，而资源之间则以一种隐语义的超链接互联，资源之间的语义关系也因为没有明确描述而不方便被机器处理。相反，如果明确表达与描述 Web 资源语义，并把资源之间的关系冠以某种特定的含义，在这种情况下，信息以语义良定义的形式存在，则将大大提高资源的共享能力。另外，需要指出的是，语义 Web 与传统知识库系统和数据库系统的显著不同之处是：它并不追求一个由专家或专门人员在有限范围内建立的全局统一的知识库或数据库，而是依赖 Web 的社会性特征和网络效益(Network Effect)实现语义 Web 的自我增长和演化；语义 Web 还允许冲突和不一致的存在，但最重要的是数据本身的语义互联，以及语义互联所带来的潜在价值和对搜索技术的巨大影响。

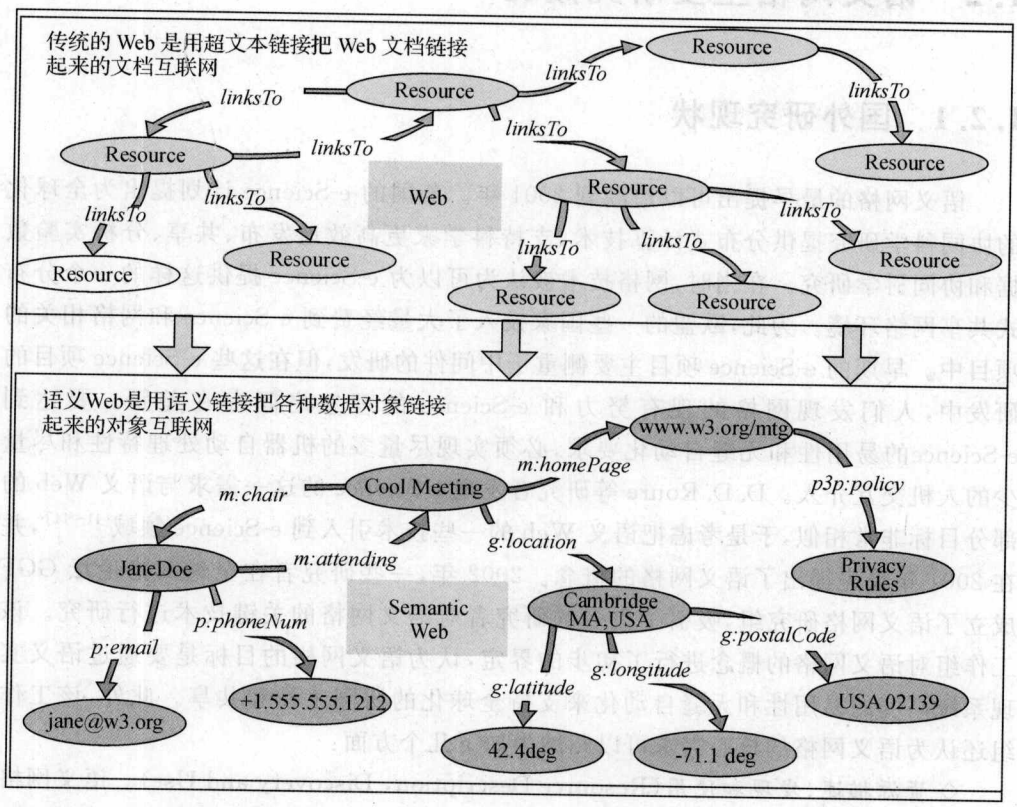

图 1.1　语义 Web 通过语义链接把数据互联起来

语义 Web 的核心技术建立在一系列技术标准和规范之上，其中 RDF 和 OWL 是最基本的技术标准。RDF 可用来描述任何 Web 资源，这包括物理上存在的网页、数