金 海 袁平鹏 著

语义网数据管理技术及应用

科学出版社
www.sciencep.com

语义网数据管理技术及应用

金　海　袁平鹏　著

科学出版社
北京

内 容 简 介

万维网已经深刻影响到社会生活的各个方面。本书围绕 Web 科学的核心问题之一——数据管理技术进行阐述。首先，介绍了语义网的意义及语义网的标准规范；然后，研究了语义数据模型及存储技术，以实现语义数据的高效存储、快速查询和动态添加。由于现实中的数据通常为半结构化或非结构化数据，为了从中获取语义数据，本书介绍了信息抽取技术、数据分类技术。在此基础上，介绍了语义数据检索、语义数据可视化技术，以便为用户提供高质量的检索结果，并向用户可视化地呈现丰富的数据间语义关联。最后，介绍了一些语义网应用系统。

本书可供计算机科学与技术、模式识别等相关专业的研究人员、工程技术人员、教师、研究生和本科生学习参考。

图书在版编目(CIP)数据

语义网数据管理技术及应用/金海，袁平鹏著. —北京：科学出版社，2010.2

ISBN 978-7-03-026468-8

Ⅰ.①语… Ⅱ.①金…②袁… Ⅲ.①语义网络-应用-数据管理-研究 Ⅳ.①TP18②TP274

中国版本图书馆 CIP 数据核字(2010)第 012346 号

责任编辑：魏英杰 王志欣 / 责任校对：桂伟利

责任印制：赵 博 / 封面设计：嘉华永盛

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

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2010 年 2 月第 一 版 开本：B5(720×1000)

2010 年 2 月第一次印刷 印张：20 1/4

印数：1—3 500 字数：385 000

定价：60.00 元

(如有印装质量问题，我社负责调换)

序

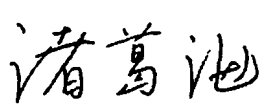
国家重点基础研究发展计划项目“语义网格的基础理论、模型与方法研究”从2003年开始围绕“规范重构，语义互联，智能聚融”这三个科学问题系统地探索了未来互联环境的语义基础问题，取得了一系列成果，本书是其中的重要成果之一。

自从计算机诞生以来，数据管理一直是计算机科学的核心问题之一。特别是随着信息技术的快速发展与普及，个人与社会每时每刻都在大量产生数据。人类将生活在一个充满数据的世界里。如何对海量数据进行有效管理就成为一个重要的研究问题。

要实现海量数据的有效管理首先要解决语义问题。有效利用海量数据更加依赖语义问题的解决。计算机科学乃至信息科学的许多问题在本质上都可以归结为语义问题。虽然人们自觉不自觉地都在使用语义，但不同的领域对语义有不同的认识。例如，语义网研究领域希望对万维网中的数据增加机器可理解的语义，以使机器能快速、准确地从万维网上获取全面有用的信息并进行自动处理。但大量加上标注的数据如何进行管理就成为了一个新的问题。传统的数据库模型难以适应如今的网络数据管理，从而研究规范组织的语义网数据模型就成为问题的核心。

该书结合作者的研究体会，吸收了新的技术和方法，大跨度地阐述了语义数据管理的技术和研究问题，涉及诸如数据模型、抽取、分类、存储、检索、呈现等。适应了语义网数据管理开发、应用和研究的需要。

该书内容丰富，希望读者能从中获得对语义数据管理的知识，对相关研究有所启示。故向读者郑重推荐。



2009年11月29日于北京

前　言

现代社会数据增长速度惊人。每隔18个月，整个数据量会翻一番。IDC统计2007年产生了281EB(2810亿GB)数据。并且IDC估计从2008年至2011年，数据会以每年60%的速度增长。尽管从2008年下半年开始全球处于经济危机之中，但这丝毫不影响数据的增长速度。IDC的报告显示2008年实际产生的新信息量为487EB，超过之前预计的3%。如果这些数据转化成文本并汇编成书，这些书的页面可以从地球铺到太阳系边缘的冥王星大约10个来回。用这些页面大约可以覆盖8次整个地球。根据数据产生的趋势，IDC估计到2012年产生的数据量将是2008年的5倍。

上面所提及的数据，涵盖人类社会生活，包括科研、经济、政治、地理等各个方面，比较抽象，通常大家感觉不到。我们可以看一看另一个与现代社会密切相关的数据增长的具体例子，即Web上网页数量呈指数飞速增长。虽然很难准确地统计出Web上网页数量，但可以根据Google公司公布的数据一窥Web上网页数量。1998年Google大约索引了260M个网页。到2000年，Google索引了10亿以上网页。据2008年7月Google官方Blog所给出的数据显示：Web上存在1T(1 000 000 000 000)个不同的URL。虽然URL不一定等同于网页，但是仍然可以从中推断出网页的数量及其增长速度同样是惊人的。

海量数据蕴含着海量有价值信息，然而信息不都是直观容易发现的，必须对数据有效处理以后才能挖掘出来有价值信息。正如1991年诺贝尔经济学奖获得者Coase所说："If you torture the data long enough, Nature will confess"。然而，如何有效地管理和利用这些海量数据信息是目前面临的一项巨大挑战。依赖人工处理海量数据显然不现实，因此需要借助于机器来处理。由于这些数据绝大部分都是半结构化/非结构化数据，没有采用形式化表示，缺乏明确的语义信息。计算机无法实现这些数据的自动处理。为便于机器高效地处理数据，需要对数据增加语义标记，以便于机器更好地理解数据的含义，快速和准确地提供给用户有价值的信息。

为了使计算机能够准确理解和快速处理半结构化/非结构化数据，1998年万维网的创始人Berners-Lee提出了语义网(Semantic Web)概念，其中数据管理是语义网的核心问题。随着对Web及语义Web研究的深入，2006年11月美国麻省理工学院和英国南安普敦大学公布了"Web科学研究计划"(web science research initiative，WSRI)。该计划的目的是研究Web的不断普及带来的社会和科

技方面的影响。这标志着一门新兴学科——Web 科学诞生了。Web 科学是一门跨计算机科学、数学、社会科学、心理学、生命科学等多门学科的综合学科。Web 科学关注的核心问题之一同样是数据管理。这需要对 Web 中的数据进行管理，增加机器可理解的语义，形成语义网，以方便机器能够自动处理，从而方便人们快速、准确获取全面的信息。语义网所发展的技术不仅仅针对 Web 上存在的大量数据处理需要。在科研领域，如生物、化学、医学等；在国民经济领域，如地理信息系统、电子政务、电子商务等许多应用同样存在大量数据需要处理同样可以应用语义网技术。

语义网的愿景虽然很好，但由于语义网尚面临诸多技术及其他问题，导致了语义网迟迟得不到大规模应用。数据、智能及安全是语义网面临的几个主要问题，其中数据管理是最主要的问题之一。数据管理涉及数据的表达、存储、分类、检索以及呈现等方面。关于数据表达，W3C 成立了一些工作组来制定规范。目前语义网所发展的一些规范，如 RDF 等已经成为了这些领域进行知识和元数据表达的基本规范。如在化学里，用 RDF 表达化学元素、分子等的元数据；在地理信息系统里，用 RDF 表达地貌、城市、交通等的元数据。同样，在语义数据的存储、检索也开展了一些研究，但仍存在诸多问题。

基于此，本书围绕 Web 科学的核心问题之一——数据管理技术进行阐述。首先，作为铺垫，本书介绍了语义网的意义及语义网的标准规范。然后研究语义数据模型及存储技术，以实现语义数据高效存储、快速查询和语义数据的动态添加。本书提出了一种主动语义数据模型 ASDM。ASDM 包括两个子模型：语义实体关联模型和 FEM 事件模型。同时，给出了面向在线分析的语义网数据存储系统 DBLink。其次，由于现实中的数据通常为半结构或非结构化数据，为了从中获取语义数据，本书介绍了语义数据抽取技术、数据分类技术等。针对语义数据分类，提出了 SVM 与 kNN 相结合的多类别单标签分类方法 MSVM-kNN 和多类别多标签分类方法。在此基础上，本书针对目前语义数据检索技术存在的一些问题提出了一些方法：针对目前语义数据检索缺乏语义关联分析和知识评价的问题，提出了支持知识评价的语义关联数据模型 RSS；针对当前语义数据检索主要采用复杂的本体查询语言，从而造成用户认知困难的问题，提出了一种基于关键字的语义检索机制。该方法是 RSS 模型中非概念约束检索机制的扩展。针对数据的语义模糊特性及用户的搜索偏好，提出了一种在 RSS 模型中支持模糊语义的检索机制；针对基于对等网络的信息检索缺乏语义的问题，提出了一种基于语义小世界的数据检索机制。由于语义数据蕴含丰富的信息，文本方式不直观，承载信息量少。因此，需要可视化呈现这些语义信息。本书介绍了语义数据可视化技术，以便向用户快速、高效、可视化地呈现丰富的数据间语义关联。最后介绍了一些语义网应用系统。

本书共八章，其中各章的主要内容安排如下：

第一章是绪论。本章介绍了语义网的目的与面临的挑战，并分析了语义网与 Web 1.0 和 Web 2.0 的区别。然后介绍语义网的体系结构，并重点介绍语义网几个重要的协议规范。

第二章是介绍语义数据模型。本章提出了一种主动语义数据模型 ASDM。ASDM 包括两个子模型：语义实体关联模型和 FEM 事件模型。语义实体关联模型是一种图模型，主要描述语义数据的组织及表达。本模型适应 RDF 广泛接受为数据表达的基础而提出的。当对语义数据进行操作时，会产生一些事件。对事件的响应，会引发在语义数据图上进行新的操作。FEM 事件模型规定如何描述事件、事件如何响应和事件如何产生等。这种模型同前人提出的事件模型相比，在复合事件上作了简化，同时并没有减少事件模型的表达能力。

第三章介绍语义数据抽取技术。本章就语义数据的抽取进行了论述，对信息抽取方法的发展及其技术手段加以讨论。以科技文献为例，分析了文献信息中提取技术及系统。

第四章介绍数据分类技术。本章首先简述自动分类的研究背景，然后介绍自动分类的意义及国内外研究现状。在此基础上研究了多类别文献自动分类方法，提出了 SVM 与 kNN 相结合的多类别单标签分类方法 MSVM-kNN 和多类别多标签分类方法。对多类别自动分类系统的主要实现技术，如文本表示、特征降维、权重计算及分类器的设计等进行了介绍。最后，对所介绍分类方法的准确率、召回率和 F-measure 值进行了测试和分析。

第五章介绍语义网数据存储技术。本章首先对语义网数据存储技术进行介绍，随后分析语义网数据管理系统的设计思想，并以我们研发的 DBLink 语义数据管理系统为例，具体阐述了语义网数据管理系统的设计和实现。

第六章介绍语义网数据检索技术。目前 Web 上存在着大量的数据资源，如何从海量信息中获取有用的知识成为亟待解决的难题这涉及许多问题，如目前的语义网数据检索缺乏语义关联分析和知识评价；当前语义网数据检索主要采用复杂的本体查询语言，从而造成用户认知困难；由于数据的语义模糊特性及用户的搜索偏好，如何更好地满足用户检索请求的检索机制仍然是具有挑战性的问题；在对等网络环境下，大规模的基于对等网络的信息检索/共享系统越来越受到人们关注。然而基于对等网络的信息检索缺乏语义的问题。本章将针对这些问题进行研究。

第七章介绍语义网数据检索可视化技术。本章介绍了数据可视化的基本概念，说明了现在数据可视化的基本技术，并举例阐述了常见的几种可视化方法。针对语义数据的特殊性，重点说明基于语义关联的数据可视化策略。该策略利用数据之间的关联关系，提出关联信息的双缓存算法。该策略更能高效的显示数据之间的关联性，具有比较好的实用性、准确性，同时具有较强的展现力和较高的交互

实时性。

第八章介绍语义网数据应用系统。本章所介绍语义网数据应用系统主要集中在基于语义的文献管理和检索应用系统。这些系统既包括已取得广泛应用的CiteSeer、GoogleScholar、Libra、Scirus 等系统，也包括由大学开发的系统，如 Bibster、SemreX 等。通过介绍这些语义网数据应用系统，了解该领域的发展现状，分析存在的问题，以及未来的发展趋势。

本书的撰写凝聚了华中科技大学服务计算技术与系统教育部重点实验室暨集群与网格计算湖北省重点实验室的语义网与知识管理研究组全体人员的智慧和辛勤劳动。具体写作分工如下：第一章、第二章由袁平鹏副教授撰写；第三章由黄莉、黄泽武、郭志鑫、袁平鹏撰写；第四章由陈玉芹、袁平鹏撰写；第五章由常冰琳、马忠宝、袁平鹏撰写；第六章由宁小敏博士、袁平鹏副教授撰写；第七章由黄莉、陈羚、袁平鹏、李毅撰写；第八章由余一娇博士撰写。全书的统稿和审定工作由金海教授主持完成。参与本书撰写工作的还有褚帆、倪雯、吴德龙等。同时感谢吴步文、严奉伟、方飞同学协助校对书稿。感谢项目组成员章勤教授、吕泽华博士的支持。

本书是作者在华中科技大学服务计算技术与系统教育部重点实验室暨集群与网格计算湖北省重点实验室长期从事语义网及其数据管理技术研究基础上形成的。我们在撰写本书的过程中力图做得完美。由于 Web 科学是一门新兴学科，有许多问题有待进一步研究，再加上我们的学识有限，书中不妥之处在所难免，敬请各位读者批评指正。

本书的出版得到了国家重点基础研究发展计划（973 计划）课题“基于语义网格的语义关联存贮模型及管理和通信平台”（课题编号：2003CB317003）及湖北省科技基础条件平台建设项目的资助，在此我们表示衷心的感谢！

感谢所有关心和支持我们的同仁！

作　者

2009 年 11 月于瑜珈山

目　录

序
前言
第一章　绪论 …… 1
1.1　语义网及其面临的挑战 …… 2
1.2　与 Web 1.0 和 Web 2.0 的区别 …… 5
1.3　语义网体系结构 …… 7
1.4　语义网规范 …… 10
1.4.1　资源描述框架 …… 10
1.4.2　RDFS/OWL …… 12
1.4.3　SPARQL …… 17
1.4.4　GRDDL …… 19
1.4.5　RDFa …… 21
1.5　本章小结 …… 23
参考文献 …… 23
第二章　语义数据模型 …… 25
2.1　数据模型简介 …… 25
2.1.1　语义及数据模型 …… 26
2.1.2　语义数据模型特征 …… 27
2.1.3　语义数据模型分类 …… 28
2.2　逻辑及物理数据模型 …… 29
2.2.1　层次和网状数据模型 …… 29
2.2.2　关系数据模型 …… 30
2.2.3　面向对象的数据模型 …… 31
2.2.4　模糊语义模型 …… 33
2.2.5　图模型 …… 35
2.3　概念数据模型 …… 35
2.3.1　实体-关系模型 …… 35
2.3.2　RM/T …… 36
2.3.3　SDM …… 37
2.3.4　资源空间模型及语义链网络 …… 39

2.4 主动语义数据模型 …… 40
2.4.1 主动语义数据模型概述 …… 40
2.4.2 事件模型研究现状 …… 43
2.4.3 FEM …… 46
2.4.4 事件探测 …… 58
2.5 本章小结 …… 71
参考文献 …… 72
第三章 语义数据抽取 …… 75
3.1 信息抽取技术简介 …… 75
3.1.1 国内外研究现状 …… 75
3.1.2 信息抽取技术分类 …… 78
3.2 科技文献元数据抽取 …… 84
3.2.1 科技文献特征 …… 85
3.2.2 信息抽取预处理 …… 88
3.2.3 基于模板匹配的头部信息抽取 …… 93
3.2.4 基于统计的尾部信息抽取 …… 98
3.2.5 信息封装 …… 104
3.3 科技文献信息抽取系统及其性能评价 …… 105
3.3.1 科技文献信息抽取系统 …… 105
3.3.2 评价指标 …… 107
3.3.3 实验结果及分析 …… 109
3.4 语义数据清理 …… 114
3.4.1 语义数据清理概述 …… 114
3.4.2 歧义实体的识别 …… 115
3.4.3 重复数据的发现 …… 119
3.5 本章小结 …… 126
参考文献 …… 126
第四章 数据分类 …… 131
4.1 数据分类意义 …… 131
4.2 研究现状 …… 132
4.2.1 传统文本分类 …… 133
4.2.2 层次分类 …… 134
4.2.3 基于知识的分类 …… 135
4.3 多类别文本自动分类方法 …… 136
4.3.1 支持向量机分类方法 …… 136

4.3.2 k近邻分类方法 …… 138
4.3.3 多类别单标签分类方法 MSVM-kNN …… 139
4.3.4 多类别多标签分类方法 …… 143
4.4 多类别文本分类实现技术 …… 146
4.4.1 文献预处理 …… 147
4.4.2 特征降维 …… 147
4.4.3 权重计算 …… 148
4.4.4 分类器设计 …… 151
4.5 多类别分类方法性能评价 …… 153
4.6 本章小结 …… 157
参考文献 …… 157
第五章 语义网数据存储技术 …… 162
5.1 语义网数据存储技术现状 …… 162
5.1.1 RDF数据模型 …… 162
5.1.2 RDF存储结构 …… 164
5.1.3 RDF查询语言 …… 165
5.1.4 语义数据索引技术 …… 168
5.1.5 语义网数据存储系统 …… 171
5.2 语义网数据管理系统 DBLink …… 174
5.2.1 DBLink的体系结构 …… 175
5.2.2 基于主存的面向对象存储模型 …… 178
5.2.3 URI映射与内置数据类型 …… 179
5.2.4 语义网数据分割与映射 …… 180
5.2.5 语义网数据的压缩存储 …… 182
5.2.6 模式空间与实例空间分离 …… 183
5.3 DBLink性能评价 …… 187
5.3.1 测试数据集 …… 187
5.3.2 实验环境 …… 189
5.3.3 测试查询集 …… 189
5.3.4 实验结果 …… 191
5.4 本章小结 …… 193
参考文献 …… 194
第六章 语义数据检索技术 …… 197
6.1 语义数据检索的分类 …… 197
6.1.1 传统数据/信息检索 …… 198

6.1.2 语义网数据检索 …… 200
6.1.3 语义网格/知识网格中的数据检索 …… 201
6.1.4 对等网络中的信息检索 …… 202
6.2 语义关联数据模型及其检索的关键问题 …… 203
6.2.1 支持知识评价的语义关联数据模型 …… 203
6.2.2 实用且有效的语义数据检索机制 …… 203
6.2.3 模糊语义数据的表达及检索 …… 203
6.2.4 基于语义的对等网数据组织及其检索 …… 204
6.3 支持知识评价的语义关联数据模型 …… 204
6.3.1 语义关联数据模型 …… 206
6.3.2 数据的检索及回答 …… 210
6.4 基于关键字的语义数据检索 …… 213
6.4.1 问题描述 …… 215
6.4.2 近似搜索算法及分析 …… 217
6.4.3 相关问题讨论 …… 221
6.5 基于语义模糊性和用户偏好的检索机制 …… 221
6.5.1 后台知识库的语义模糊性 …… 224
6.5.2 用户检索请求的规范化 …… 229
6.5.3 检索结果的排序 …… 231
6.6 基于语义小世界的数据检索 …… 233
6.6.1 语义小世界的构建 …… 234
6.6.2 数据检索算法 …… 237
6.7 本章小结 …… 238
参考文献 …… 239
第七章 语义数据检索可视化 …… 246
7.1 信息检索可视化 …… 246
7.2 研究现状 …… 247
7.2.1 可视化技术概述 …… 247
7.2.2 基于分类的文档簇法 …… 248
7.2.3 基于超链接法 …… 249
7.2.4 基于文档内容的信息可视化方法 …… 249
7.2.5 网络知识发现的可视化方法 …… 250
7.2.6 分类信息检索可视化方法 …… 251
7.2.7 其他信息可视化系统 …… 252
7.3 基于关联的语义网情景检索 …… 253

7.3.1 基本概念 …… 254
7.3.2 情景值计算 …… 255
7.3.3 情景可视化 …… 256
7.3.4 实体推荐 …… 257
7.4 语义网情景检索的实现 …… 258
7.4.1 关联信息可视化体系结构 …… 259
7.4.2 关联信息双缓存 …… 262
7.4.3 关联信息可视化策略 …… 264
7.5 本章小结 …… 271
参考文献 …… 271
第八章 语义网数据应用系统 …… 273
8.1 CiteSeer …… 273
8.1.1 CiteSeer 的设计理念和功能 …… 273
8.1.2 软件系统分析 …… 275
8.1.3 自动引用索引 …… 277
8.1.4 论文相似度计算方法 …… 278
8.1.5 CiteSeer^x …… 279
8.1.6 CiteSeer 的意义 …… 281
8.2 Google Scholar …… 282
8.2.1 系统功能和服务 …… 283
8.2.2 特征分析 …… 283
8.2.3 对数字图书馆的影响 …… 287
8.3 Libra …… 287
8.3.1 Libra 应用状况 …… 288
8.3.2 对象级检索 …… 289
8.4 Bibster …… 292
8.4.1 Bibster 的特点 …… 292
8.4.2 Bibster 的软件结构 …… 293
8.4.3 Bibster 中的本体 …… 294
8.5 Scirus …… 296
8.6 基于语义关联的科研文献共享系统 SemreX …… 298
8.6.1 SemreX 的软件结构 …… 298
8.6.2 文献共享系统的比较 …… 301
8.7 语义网数据应用现状与趋势 …… 303
参考文献 …… 304

第一章 绪 论

万维网已经成为人们获取信息的主要渠道之一，深刻影响到人类社会生活的各个方面：人们在 Web 上浏览新闻、搜索信息、买卖商品及服务。然而，目前的万维网是面向人而不是面向机器的。换句话说，人可以理解万维网上的内容，而机器则不能理解。随着万维网上的内容越来越多（据 Google 的官方 Blog 公布的数据，目前 Google 索引了 1T 个 URL），人们准确、快速、全面获取到信息也越来越难。

对此，WWW、URIs、HTTP 和 HTML 的发明人 Berners-Lee 提出了语义网概念。他认为“The Semantic Web will bring structure to the meaningful content of web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users”自从提出语义网之后，语义网得到了学术界和工业界的重视。Gartner 认为企业语义网是 3 个具有影响力的万维网技术之一。许多语义网公司，如 Siderean、SandPiper、Intellidimension 等在兴起。一些大公司，如 Adobe、HP、IBM、Nokia、Oracle、Vodaphone 等纷纷对语义网技术进行研发。一些国家和国际组织，如英国、欧盟、美国也投入资金资助相关研究。

语义网是关于万维网的元数据网。数据一直是语义网的核心问题，Berners-Lee 曾经公开表示，他认为对语义网更准确的名称应该是数据网（data web）。语义网存在及互联的目的是使得计算机能够自动处理、集成来自不同数据源的数据，而不仅仅是为了让计算机能够正确显示。为了实现让机器或设备能够自动识别和处理网上数据，需要在内容中加入标记，即需要采用标记语言。考虑到适应不同领域的标记需要，标记语言必须是灵活、可扩展的。为了实现此目标，研究和制定在万维网上发布能够帮助计算机理解内容的设施、技术和规范成为迫切需要解决的问题。

对此，研究人员、标准化组织包括互联网工程任务组和 W3C 等都致力于增强、扩展语义网能力，在研究、制定以及部署能够共享语义的语言以及工具方面进行了大量工作。这些语言为语义互操作提供了基础。1997 年，W3C 定义了资源描述框架规范 RDF。此后，相继提出了 RDFS/OWL、GRDDL、RDFa 等规范或建议规范。研究人员也开发了许多工具，如 Kowari、RDFLib、Jena、Sesame、Protégé、SWOOP 等。W3C 将其中的部分技术标准化，形成了一些标准，包括资源描述框架 RDF、RDFS/OWL 等。

虽然语义网的研究正在蓬勃开展，但语义网仍面临着一些问题。其中一个问

题就是数据问题。本章将首先介绍语义网面临的挑战，语义网与 Web2.0 的区别以及语义网的体系结构，然后介绍目前 W3C 制定的一些规范以及学术界、工业界提交的一些建议规范。

1.1 语义网及其面临的挑战

从 Web 诞生经历发展至今，Web 上网页数量呈指数飞速增长。虽然很难准确地统计出 Web 上的网页数量，但我们可以根据 Google 公布的数据一窥 Web 上的网页数量。据 2008 年 7 月 Google 的官方 Blog 给出的数据：Google 数据显示 Web 上存在 1T (1 000 000 000 000)个不同的 URL。而 1998 年 Google 索引了 260 000 000个网页。到 2000 年，Google 索引了 10 亿以上网页[1]。这些 Web 网页承载着海量的数据。Web 上不仅仅存在着大量的网页及数据，网页之间也存在超链接等，这都使信息变得更加复杂化。

虽然 Web 上存在海量的信息，但是目前 Web 实际上只是一种面向人的存储和共享信息的媒介。这是因为：首先，Web 内容主要是提供给人来理解和浏览的。由于 Web 内容没有采用形式化表示，缺乏明确的语义信息，故而计算机“看到的”Web 内容只是二进制数据，对其内容无法进行识别。这使得计算机不能理解网页内容的语义，无法实现 Web 内容的自动处理。其次，Web 是按 URL 而非内容来定位信息资源。网页之间通过超链接关联，但超链接的含义是模糊的，因此网页所承载的数据及其之间的关系在语义上是孤立的或是缺乏丰富的细粒度语义关联，无法检索到语义关联信息。这导致精确查找所需的信息，开发复杂的互联网应用非常困难。最后，随着 IT 技术的发展，互联网将不仅仅互联计算机，会有更多的智能机器、手持终端及家用电器等设备互联。为了方便人们的日常生活，这些设备之间需要进行智能交互，需要获取和处理 Web 上的信息。当前的互联网技术尚不能支持这一点。

考虑到目前 Web 存在的问题，为了使计算机能够理解和处理网页内容，迅速准确地从海量网页中查找出所需要的内容，1998 年 Berners-Lee 提出了语义网(Semantic Web)[2]。顾名思义，语义网是对现有 Web 增加语义支持，它是现有万维网的变革和延伸，其目标是帮助机器在一定程度上理解信息的含义，使得高效的信息共享和机器智能协同成为可能。这将使网络有能力提供动态与主动的服务，从而更便于人与机器、机器与机器之间的对话和协同工作。简单地说，语义网是以数据的内容，即数据的语义为核心，用机器能够理解和处理的方式链接起来的海量分布式数据库。

语义网最大的好处是可以让计算机具有对网络空间储存的数据进行智能评估的能力。这样，计算机就可以像人脑一样理解信息的含义，完成智能代理的功能，

使用语义网搜索引擎检索的结果也会比万维网更为精确。然而，语义网的愿景虽然很好，但是由于语义网尚面临着诸多问题，导致了其迟迟得不到大规模应用。数据问题、智能问题及安全问题是语义网目前面临的主要问题。

1. 数据问题

数据一直是语义网的核心问题。Berners-Lee 曾经认为，将这种智能网络称为语义网或许不够准确，更准确的名字应该是数据网。在语义网中，为实现让机器或设备能够自动识别和处理网上数据，需要在内容中加入标记，即需要采用标记语言。考虑到适应不同领域标记的需要，标记语言必须是灵活、可扩展的。目前 W3C(world wide web consortium)定义了一些语义网规范，如 RDF(resource description framework)[3]、RDFS(RDF schema)[4]/OWL(web ontology language)[5]等。同时，互联网上绝大部分内容尚未加上符合语义网规范的标记。因此，如何对现有内容自动加上符合语义网规范的标记，将是语义网快速走向实用化的关键之一。这涉及一系列技术，如信息抽取、分类、表达、存储、检索等。

针对数据问题，W3C 的 Semantic Web Education and Outreach 组织的互联开放数据(linking open data)项目通过发布开放的 RDF 数据集并互联来自多个数据源的数据，使得有更多符合语义网规范的数据可用。目前，互联数据(linked data)项目用 Web 去链接互联以前未链接的相关数据。图 1.1 是到 2009 年 3 月为止已互联的数据云图。同 2008 年 10 月相比，2009 年 5 月互联数据的规模已经达到 47 亿个三元组，通过 1.42 亿链接互联起来[6,7]。尽管符合语义网标准的数据在增多，但相对于人类社会所产生的海量数据来说，仍显得微不足道。研究人员仍需发展高效的方法从半结构化/非结构化数据中获取语义信息，并对其进行高效管理。

2. 智能问题

语义网面临的另一个重要的技术难题是如何能够让机器或设备进行“思考”和“推断”，这涉及本体、逻辑和规则等若干方面。例如，对于本体来说，尽管一些概念比较持久，但实际上他们并不是始终保持不变。一个例子是关于学科分类本体演变：随着科学技术的发展，不断有新的学科出现，或同其他学科合并，从而学科的分类会发生变化。因此，必须把本体当成可以演变的。这就产生了若干问题，例如如何将变化前后的本体对应起来；如何解决变化的本体可能导致知识库不一致的问题等。其次，本体依赖于所采用的本体语言。描述逻辑语言是语义网的逻辑基础，然而描述逻辑在表达能力上存在一定的不足。某些应用可能需要研究更强表达能力的描述逻辑。实际应用中，本体数目一般都有上百个，有的甚至达到上万个。对于这么大规模的本体即使在现有本体语言上进行推理，也是非常具有挑战性的。

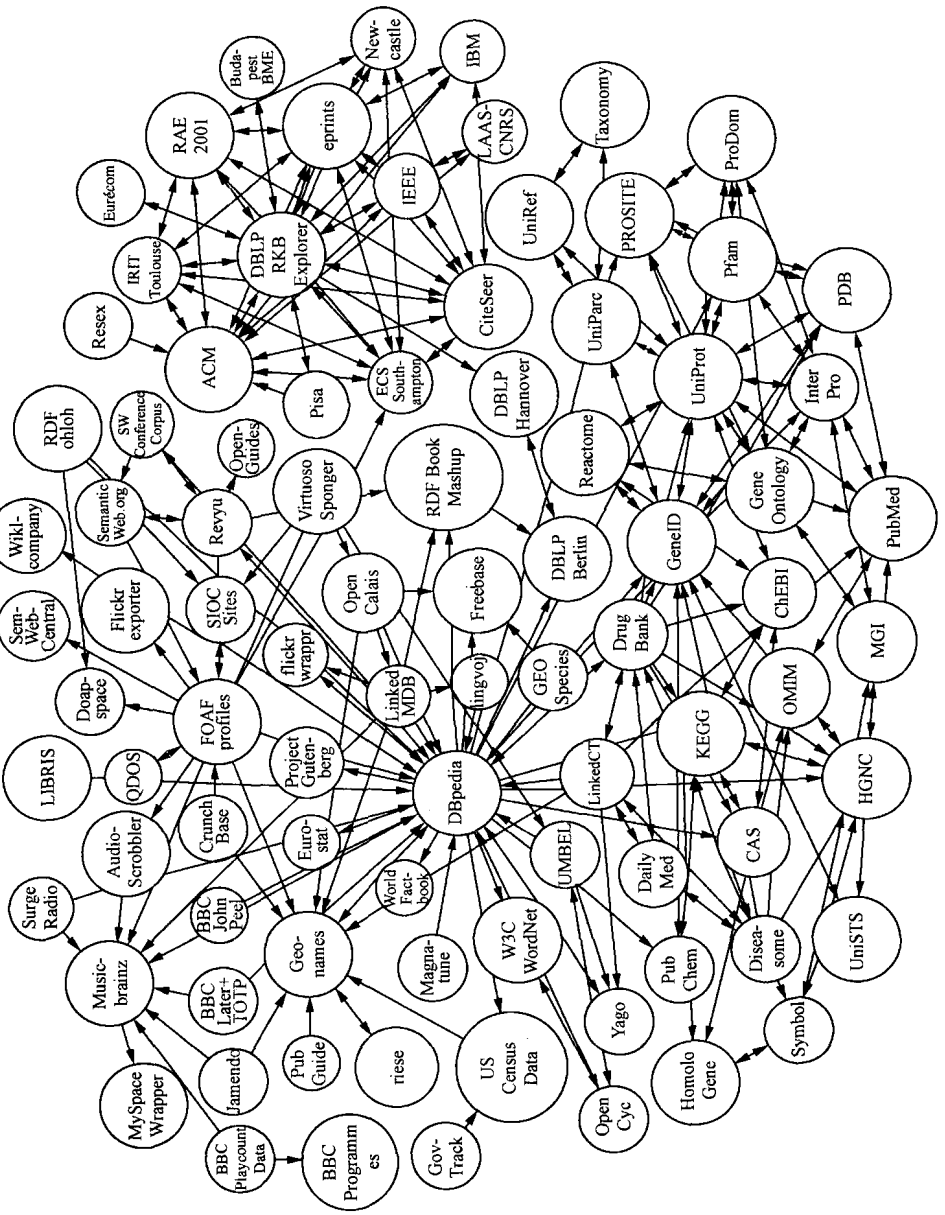


图 1.1 互联的数据(http://linkeddata.org/)