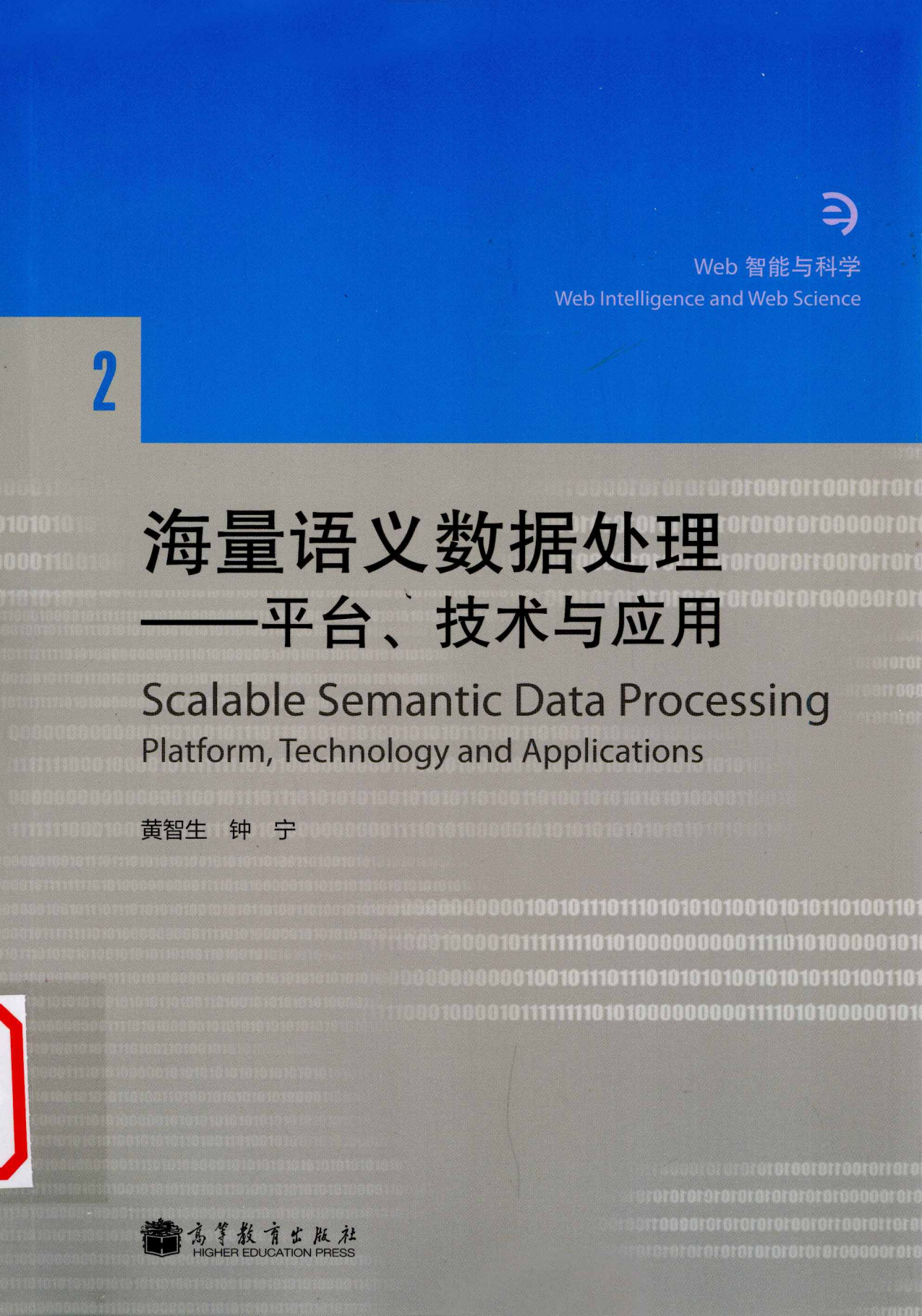

Web 智能与科学
Web Intelligence and Web Science
2
海量语义数据处理
——平台、技术与应用
Scalable Semantic Data Processing
Platform, Technology and Applications
黄智生 钟 宁
高等教育出版社
HIGHER EDUCATION PRESS

Web 智能与科学
Web Intelligence and Web Science

2

HAILIANG YUYI SHUJU CHULI

海量语义数据处理
——平台、技术与应用

Scalable Semantic Data Processing
Platform, Technology and Applications

黄智生　钟　宁

高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING

图书在版编目(CIP)数据

海量语义数据处理：平台、技术与应用/黄智生，钟宁编著. —北京：高等教育出版社，2012.10
（Web智能与科学）
ISBN 978-7-04-036246-6

Ⅰ. ①海… Ⅱ. ①黄… ②钟… Ⅲ. ①语义网络-数据处理 Ⅳ. ①TP18②TP274

中国版本图书馆CIP数据核字（2012）第226416号

策划编辑 刘 英	责任编辑 冯 英	封面设计 张 楠	版式设计 马敬茹
插图绘制 尹文军	责任校对 殷 然	责任印制 朱学忠	

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	涿州市星河印刷有限公司	网上订购	http://www.landraco.com
开 本	787mm×1092mm 1/16		http://www.landraco.com.cn
印 张	17.25	版 次	2012年10月第1版
字 数	320千字	印 次	2012年10月第1次印刷
购书热线	010-58581118	定 价	59.00元

物料号 36246-00

《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。在这场竞争中，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年）》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。

在这个时代，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，网络智能化研究已经成为当今国际人工智能研究领域的一个新趋势。以钟宁教授、刘际明教授等为代表的一批学者顺应这一国际研究新趋势，率先对Web智能科学领域进行了系统的研究，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领先地位。但是，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，钟宁、刘际明提议并任主编，建立了“Web智能与科学”系列丛书，以解决国内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、最新的研究成果和学术交流等情况。

在此，衷心希望通过该系列丛书的出版，为国内研究者构建一个Web智能科学研究领域交流互动的窗口，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，为把我国建设成自主创新的科技强国做出应有贡献。

2010年12月

序

语义技术经过十多年的研究和发展，正在越来越多地被许多应用系统采用。目前，许多行业的大公司都部署使用语义技术，如传媒业的英国广播公司（British Broadcasting Corporation，BBC），搜索引擎领域的谷歌和雅虎，新闻业的纽约时报，等等。而且在许多其他领域都可以看到类似的情况。同样，世界上许多国家的政府部门、地区政府部门及其城市主管部门，也越来越多地使用语义技术，为公民提供相关信息。这一切都使得可以获得的语义网数据急剧增长。据最新估算，LOD（Linked Open Data）云系统上可通过语义网免费获得的事实和知识描述条目已达数百亿。

但是，成功经常会带来一系列新的问题。在语义技术的发展过程中，海量性首次成为一个实实在在的问题。我们如何处理以前从未遇到过的高达数十亿事实描述的数据规模。而且，数据规模不仅巨大，这些数据还是分布式的、异构的、不完整的，甚至常常是部分不正确的。

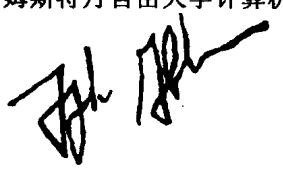
上述观察（即语义网成功之后所带来的海量性问题）是我们着手构建大型知识对撞机（Large Knowledge Collider，LarKC）的动机。在 LarKC 中，我们结合并行计算、知识表示、机器学习和自然语言处理等技术，建立了一个拥有超大规模语义数据的推理平台。LarKC 平台已被成功地用于构建许多应用系统，包括意大利米兰和韩国首尔的智能城市应用、关联生命数据集（Liked Life Data）以及用于发现基因和疾病之间的关系。

LarKC 是一个真正由国际化的联合体建设而成的，参加者横跨三大洲，分别来自欧洲国家、美国、韩国和中国。对我个人而言，我们同中国合作伙伴的合作一直在丰富着我们的经验。我对来自于钟宁教授所领导的实验室的研究同事所表现的知识和技能印象深刻，他们的贡献对于 LarKC 的成功至关重要。

因此，我非常高兴这本多方位介绍 LarKC 平台的书能够出版。黄智生教授和钟宁教授把相关内容汇印成册，做了一件令人钦佩的工作，这将进一步加强国际上不同语义研究群体之间的联系。

LarKC 技术总裁

荷兰阿姆斯特丹自由大学计算机系教授



Frank van Harmelen

2012 年 3 月 14 日

Preface

After more than a decade of research and development, Semantic Web technologies are now increasingly being adopted in many applications. Very large industrial players are now deploying Semantic Web technology: in the media industry (BBC), in the search-engine business (Google, Yahoo), in the news business (New York Times), and in many other places. The same holds for governments around the world at both national, regional and city level, who are increasingly using Semantic Web technologies to make their data available to their citizens. All this has caused a very large increase in the size of Semantic Web data that is currently available. Latest counts on the size of the Linked Open Data cloud range in the tens of billions of facts and knowledge items that are freely available on the Semantic Web.

But as so often, success brings with it a new set of problems: for the first time in the development of semantic technologies, scalability is becoming a real problem. How to deal with data of a size that we had never seen before (billions of facts), and that is not only large, but also distributed, heterogeneous, incomplete, and often even partially incorrect?

This observation (that scaling problems were the price of the success of the Semantic Web) was the motivation to start building the Large Knowledge Collider, or LarKC for short. In LarKC we have combined technologies from parallel computing, from knowledge representation, from machine learning and from natural language processing to build a platform for reasoning with very large volumes of Semantic Web information. The LarKC platform has been successfully used to build applications for smart city applications in Milan, Italy and Seoul, Korea, for a large biomedical knowledgebase (Linked Life Data) and for discovering the relation between genes and diseases.

LarKC was built by a truly international consortium, spanning three continents, with participants from Europe, U.S.A, Korea and China. For me personally, the collaboration with our Chinese partners has been a very enriching experience. I have been impressed by the skills, knowledge and motivation of my research colleagues from the lab of Prof. Ning Zhong, and their contribution was crucial to the success of

LarKC.

I am therefore extremely pleased with the publication of this book that discusses many different aspects of the LarKC platform. Both Prof. Zhisheng Huang and Prof. Ning Zhong have done an admirable job in putting together this volume, which will serve to further strengthen the links between the different international semantic communities.

Frank van Harmelen
LarKC Scientific Director
Mar.14, 2012

前　言

语义网所面临的一个重大问题就是如何处理海量语义数据，它一直被认为是制约语义技术发展的瓶颈问题。在这样一个技术背景下，欧盟第七研究框架开展了语义网 LarKC 重大项目的研究。其研究目标在于开发海量语义数据处理平台，并通过一系列应用开发为海量语义数据处理的技术研究提供科学实验证据。经过三年多的努力，LarKC 项目在海量语义数据技术方面取得了一系列重大进展。本书系统地介绍了海量语义数据处理的最新技术和进展，特别是通过介绍 LarKC 项目所开发的海量语义处理平台及其应用，来阐述海量语义数据处理技术的基本原理、实现方法和应用开发等一系列关键问题。

本书由 13 个章节构成，分为上、下两篇。上篇为技术篇，内容涵盖海量语义数据处理基本原理、海量语义数据处理平台体系结构、识别与选择技术、抽象与转换技术、推理与决策技术、LarKC 平台应用开发技术等。下篇为应用篇，内容涉及 LarKC 平台现有开发的一系列应用系统，包括关联生命数据集（Linked Life Data）、基于语义技术的生物医学文献检索技术、语义技术在生命科学上的应用、语义技术与城市计算（Urban Computing）、语义技术在智能交通上的应用等。

北京工业大学国际 WIC 研究院作为 LarKC 联合体的研究团队之一，参与了欧盟第七框架语义网重大项目的研究。多位中国学者在 LarKC 项目的研究中作出了重要的贡献。可以说，在 LarKC 项目的各个研究环节，都有中国学者的参与，这就为本书的系统性介绍及其科学性描述提供了基本保证。

本书由 LarKC 项目的推理技术总负责人黄智生教授和 LarKC 项目中国团队总负责人、北京工业大学国际 WIC 研究院的钟宁教授担任主编，由 LarKC 项目中多位资深中国学者参与撰写对应的章节。各个章节的撰写人分别是：第 1 章（导论）黄智生、钟宁，第 2 章（LarKC 海量语义数据处理平台）方俊，第 3 章（识别与选择）曾毅、Danica Damljanovic、任旭、王岩，第 4 章（抽象与转换）黄异，第 5 章（推理与决策）黄智生，第 6 章（非常规语义推理）方俊、黄智生，第 7 章（LarKC 系统与应用开发）方俊，第 8 章（关联生命数据集）黄智生，第 9 章（生物医学文献语义检索）黄智生，第 10 章（海量语义数据处理与基因研究）黄智生，第 11 章（城市计算Ⅰ）黄异，第 12 章（城市计算Ⅱ）黄智生，第 13 章（海量语义数据处理技术展望）黄智生、荷如和钟宁。这些作者分别来自荷兰阿姆斯特丹自由大学、北京工业大学国际 WIC 研究院、德国西门子公司、

英国谢菲尔德大学和西北工业大学等。我们非常感谢上述各位学者在 LarKC 项目研究过程中的真诚合作以及在本书撰写过程中的辛勤工作。感谢 LarKC 技术总裁、荷兰阿姆斯特丹自由大学计算机系的 Frank van Harmelen 教授为本书作序。特别感谢高等教育出版社刘英编辑在本书撰写和出版过程中提供的大力支持和帮助。

本书不仅是一本系统介绍海量语义数据处理处理 LarKC 平台及其技术与应用的参考书，同时对语义数据处理技术的研究人员、语义数据处理平台实现的技术人员以及语义技术的应用开发人员均具有一定的参考价值。我们热烈欢迎各位读者把对本书的意见和建议积极地反馈给我们，并对本书的纰漏与瑕疵提出批评。

黄智生　钟宁

2012 年 3 月

目　录

第 1 章　导论 …… 1
1.1　语义技术概述 …… 2
1.2　海量语义数据处理 …… 4
1.3　LarKC 概述 …… 5
1.3.1　LarKC 项目 …… 5
1.3.2　LarKC 技术概述 …… 8
1.3.3　LarKC 海量语义数据处理平台概述 …… 10
1.3.4　LarKC 应用技术开发概述 …… 11
1.4　本章小结 …… 12
参考文献 …… 12

第一部分　技　术　篇

第 2 章　LarKC 海量语义数据处理平台 …… 17
2.1　LarKC 体系结构 …… 17
2.2　LarKC 平台的安装与使用 …… 19
2.2.1　获取 LarKC …… 19
2.2.2　运行 LarKC …… 20
2.2.3　一个简单工作流实例 …… 20
2.3　工作流设计器 …… 22
2.3.1　工作流设计器概览 …… 22
2.3.2　安装及主要操作 …… 23
2.4　LarKC 插件概述 …… 24
2.5　用户支持和版权信息 …… 26
2.6　本章小结 …… 27
参考文献 …… 27

第 3 章　识别与选择 …… 29
3.1　识别方法与识别插件 …… 29
3.2　基于兴趣的选择方法与插件实现 …… 31
3.2.1　基本原理与基本算法 …… 31

3.2.2 方法的可扩展性与效率比较 34
3.2.3 基于兴趣的选择插件设计与实现 37
3.3 随机索引选择方法与插件实现 40
3.3.1 语义索引 40
3.3.2 基于随机索引与 Lucene 索引的检索比较 43
3.4 选择方法与选择插件的应用 47
3.4.1 基于兴趣的选择插件应用示例 48
3.4.2 随机索引选择插件应用示例 51
3.5 本章小结 53
附录 第 3.3.2 节相关询问 53
参考文献 57

第 4 章 抽象与转换 59
4.1 机器学习 59
4.1.1 SUNS 60
4.1.2 机器学习插件 63
4.2 数据流 66
4.2.1 C-SPARQL 67
4.2.2 数据流插件 73
4.3 归纳与演绎结合的数据流推理 77
4.3.1 动机 77
4.3.2 数据流推理的结构框架 78
4.4 本章小结 79
参考文献 79

第 5 章 推理与决策 82
5.1 LarKC 推理与决策插件 83
5.2 常规语义推理 86
5.2.1 OWLAPI 推理机 86
5.2.2 SPARQL DL 推理机 88
5.3 并行与分布式推理 92
5.3.1 采用 MapReduce 技术的海量分布性推理 92
5.3.2 采用 WebPIE 进行 OWL 分布性推理 96
5.4 基于规则的推理 98
5.5 本章小结 100
参考文献 101

第 6 章 非常规语义推理 102
6.1 不一致本体的推理 102

6.1.1 语义网与不一致性 102
6.1.2 基本方法 104
6.1.3 LarKC 平台下的 PION 系统 108
6.2 转折推理 111
6.2.1 基本定义 111
6.2.2 计算方法和实现 112
6.2.3 转折推理插件 113
6.3 嘈杂语义数据的推理 116
6.3.1 基本定义 116
6.3.2 韩国首尔 RSM 系统示例 117
6.4 本章小结 122
参考文献 123

第 7 章 LarKC 系统与应用开发 124
7.1 LarKC 工作流开发 124
7.1.1 工作流图 124
7.1.2 工作流描述 125
7.1.3 更复杂的一个示例 127
7.2 LarKC 插件开发 130
7.2.1 LarKC Maven 原型的使用 130
7.2.2 插件代码编写 135
7.2.3 整合插件到 LarKC 平台 138
7.3 相关的开发工具 140
7.3.1 集成开发环境 Eclipse 140
7.3.2 项目管理 Maven 140
7.3.3 单元测试 JUnit 141
7.3.4 版本控制 SVN 141
7.4 本章小结 142
附录 复杂的工作流描述示例 142
参考文献 144

第二部分 应 用 篇

第 8 章 关联生命数据集 147
8.1 概况 147
8.2 关联生命数据组成 147
8.3 语义关联构造 150

8.4 关联生命数据集的使用 …… 152
8.4.1 关联生命数据集关键词查询 …… 153
8.4.2 关联生命数据集 SPARQL 语义查询 …… 155
8.5 本章小结 …… 164
参考文献 …… 164

第 9 章 生物医学文献语义检索 …… 166
9.1 需求分析 …… 166
9.2 通过 LLD 进行医学文献检索 …… 169
9.3 医学文献语义标注 …… 178
9.4 LarKC 医学文献语义检索插件 …… 180
9.5 本章小结 …… 181
参考文献 …… 182

第 10 章 海量语义数据处理与基因研究 …… 183
10.1 概述 …… 183
10.2 基因研究与语义数据 …… 184
10.3 LarKC 海量语义数据处理平台用于 GWAS 研究 …… 189
10.3.1 LarKC 的 GWAS 插件 …… 189
10.3.2 关键词扩展推理器 …… 189
10.3.3 GWAS 识别器 …… 197
10.3.4 GWAS 工作流 …… 199
10.4 本章小结 …… 204
参考文献 …… 205

第 11 章 城市计算 I：交通与社交媒体 …… 207
11.1 交通路线规划 …… 207
11.1.1 框架结构 …… 208
11.1.2 交通预测 …… 209
11.1.3 语义交通路线规划 …… 211
11.1.4 评价 …… 213
11.1.5 小结 …… 215
11.2 社交媒体分析 …… 216
11.2.1 概况 …… 216
11.2.2 框架结构 …… 218
11.2.3 BOTTARI 的 LarKC 工作流 …… 220
11.3 本章小结 …… 229

参考文献 229

第 12 章 城市计算Ⅱ：路标管理 231

12.1 基本思想 231

12.2 RSM 数据集与数据整合 234

12.2.1 RSM 数据集 234

12.2.2 数据整合 236

12.2.3 路标的有效性审核 238

12.3 RSM 语义数据处理 239

12.3.1 系统结构 239

12.3.2 RSM 工作流 240

12.3.3 RSM 查询与推理 241

12.4 RSM 系统用户界面 243

12.5 本章小结 246

参考文献 246

第 13 章 海量语义数据处理技术展望 248

13.1 市场分析 248

13.1.1 语义技术的市场观察和潜力 248

13.1.2 市场分析的结论 249

13.2 LarKC 海量语义数据平台应用展望 250

13.2.1 药物研发海量语义数据处理 250

13.2.2 语义技术用于政治文化分析 251

13.2.3 智能交通系统 251

13.2.4 基于语义技术的电子病历 252

13.3 海量语义数据处理研究展望 253

13.4 结束语 255

参考文献 255

第1章 导 论

万维网对人类社会产生了不可估量的影响，彻底地改变了人类社会的信息环境。通过万维网来获取信息已经成了人类社会的生活方式之一，也成为了科学研究的主要信息处理手段之一。万维网已经对各个领域的科学研究环境产生了重大影响。

万维网为人类社会及其科学研究提供了浩瀚的信息和知识资源。由万维网之父 Tim Berners-Lee 所提出的语义万维网（the Semantic Web，简称语义网）为处理万维网上浩瀚的信息资源提供了一个全新的技术手段。我们把这个基于语义网思想所发展的信息和知识处理技术，统称为语义技术（Semantic Technology）。近 10 多年来，语义技术已经在许多领域（如医学与生命科学、软件工程与网络技术、智能交通与城市管理、工程技术、农业与食品、社会科学研究等）得到了广泛的应用[1]。

在语义技术的发展过程中，通过计算机领域与各个应用领域的科学家及研究人员的努力和紧密合作，已经产生了极其庞大的语义数据集。这些语义数据集之所以不同于现有万维网上的信息资源，在于前者使用国际规范的语义描述语言，如 RDF/RDFS 和 OWL 等来描述数据，使得人们能够方便地根据其数据内容更精准地获得数据。因此，如何有效地处理海量语义数据已经成为了语义技术研究的核心课题之一[2]。

本书的目的就是系统化地研究和阐述海量语义数据处理的最新技术，特别是通过介绍欧盟第七研究框架重大语义技术 LarKC 项目（http：//www.larkc.eu）开发的 LarKC 平台[3]和提出的一系列技术方法，以及所实现的一系列具体的应用系统，系统而全面地阐述海量语义数据处理的最新思想和技术进展。本书的所有作者都是亲身参与 LarKC 项目的资深研究人员。本书除第 1 章导论外，全书其他部分由技术篇和应用篇组成。技术篇介绍 LarKC 海量语义数据处理平台技术研究的各个层面，包括平台体系结构、插件与工作流开发、对应的各种标识、转换、学习、选择以及推理与决策等一系列技术。应用篇介绍 LarKC 团队所开发的各种应用系统，包括基于语义技术的智能交通系统、生命科学语义数据集的开发、医学文献语义检索技术和语义技术用于生命科学研究（包括癌症的基因关联性分析研究）等。

为了使读者能够更方便而有效地了解本书后面要介绍的各个技术内容，本章首先从分析语义技术的主要思想与发展趋势出发，提供一个语义技术的概述。在此之上，阐述海量语义数据技术的主要思想和技术方法。同时，对 LarKC 海量

语义数据处理平台以及技术开发和应用系统提供一个概述性的介绍。

1.1　语义技术概述

从万维网的诞生到现在的20多年时间里，人类已经面临着万维网上信息急剧增长的问题。人们每天要花费大量的时间在万维网上进行人工搜索信息，筛选出自己真正需要的内容，这是一个枯燥而繁重的智力劳动。人类需要寻找一种全新的方式来描述网络上的信息，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。

于是，万维网之父Tim Berners-Lee在2000年前后提出了语义万维网的概念，其核心思想就是让计算机能够自动处理网络上的信息内容。10多年来，语义技术已经得到了巨大的发展。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及其推理技术的RDF/RDFS、网络本体语言OWL、语义数据查询语言SPARQL、规则交换框架RIF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。语义技术的主要特征是采用逻辑的手段来描述数据，使得人们可以通过其对应的推理机有效地分析数据内容，为知识管理提供基本技术手段。

用于描述网络信息资源及其提供推理支持的RDF/RDFS语言被看成是元数据描述语言。所谓的元数据（Meta Data）顾名思义就是用于描述数据的数据。网络本体语言OWL可以被看成是在RDF/RDFS基础上的扩展和改造，主要是增加了更强的逻辑描述能力（如引入逻辑否定、逻辑合取等）。所谓本体（ontology）可以被理解成特定领域规范概念集及其关系的描述。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。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。

近10多年信息领域最主要的成果之一就是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基础条件。使用这些特定领域的元数据与本体，可以对万维网上的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垂直搜索引擎（更准确地讲，也就是语义搜索引擎）有效地寻找到对应的信息或知识。这同时也为信息搜索的自动化提供了基本的技术环境。图1-1提供了一个网络信息资源结构图，说明了各种网络资源及其与语义数据的关系。RDF/RDFS的语义模型是基于三元组结构的，OWL是RDF/RDFS语言的进一步扩充，所以在语义网与本体技术研究领域，语义数据的规模通常是以三元组（Triple）的数量来度量的。