计算机科学学术丛书

基于语义的XML信息集成技术

顾进广　陈莘萌　著

WUHAN UNIVERSITY PRESS
武汉大学出版社

中国博士后科研基金（20060400277
湖北省自然科学基金（2007ABA29
江苏省博士后科研基金（0601009
湖北省教育厅科学研究项目（Q2007110
软件工程国家重点实验室（武汉大学）开放基金(SKLSE05-03)
武汉科技大学高层次引进人才科研专项资助项目
武汉科技大学计算机科学与技术学院学科建设基金

计 算 机 科 学 学 术 丛 书

基于语义的XML信息集成技术

顾进广 陈莘萌 著

图书在版编目(CIP)数据

基于语义的XML信息集成技术/顾进广，陈莘萌著．—武汉：武汉大学出版社，2007.10

计算机科学学术丛书

ISBN 978-7-307-05843-9

Ⅰ.基… Ⅱ.①顾… ②陈… Ⅲ.可扩充语言，XML—程序设计 Ⅳ.TP312

中国版本图书馆CIP数据核字(2007)第147397号

责任编辑：黄金文　　责任校对：刘　欣　　版式设计：支　笛

出版发行：**武汉大学出版社**　(430072　武昌　珞珈山)

(电子邮件：wdp4@whu.edu.cn 网址：www.wdp.whu.edu.cn)

印刷：湖北新华印务有限公司

开本：787×1092　1/16　印张：8.875　字数：208千字　插页：1

版次：2007年10月第1版　　2007年10月第1次印刷

ISBN 978-7-307-05843-9/TP·274　　定价：18.00元

顾进广，博士，副教授，硕士生导师。男，1974年11月出生。主要研究方向为分布式计算、智能信息处理、语义Web及软件工程。现为IEEE、IEEE-CS和ACM会员，中国计算机学会和中国电子学会高级会员。1995年和1997年分别于武汉科技大学信息科学与工程学院获工学学士学位和工学硕士学位，2005年于武汉大学计算机学院获工学博士学位。现为武汉科技大学计算机科学与技术学院教师，并在东南大学计算机科学与技术博士后流动站从事智能信息处理、软件工程等方面的研究工作。在国内外专业期刊和会议上发表学术论文二十余篇，其中SCI-E、EI和ISTP收录二十余次。主持中国博士后科研基金项目、江苏省博士后科研基金项目和湖北省教育厅科学研究项目各一项，参与湖北省自然科学基金项目、湖北省教育厅科学研究重点项目多项，主持和参与横向项目多项。

陈莘萌，教授，博士生导师。男，1939年9月出生。1958年8月毕业于武汉大学数学系，1981～1983年在日本京都大学作访问学者，从事并行处理研究。历任中国计算机学会委员、中国计算机学会体系结构专委会委员、信息存储专委会委员、中国计算机学会体系结构专委会主任、湖北省微机领导小组成员、湖北省财会电算化软件评审专家组组长、武汉大学计算机科学研究所副所长、国务院电子信息系统推广应用办公室“电子计算机应用系列教材”常务主编等职。早年从事计算机系统结构与计算机应用的研究，主持开发了GNB-Ⅰ型通用计算机系统等填补多项国内技术空白的大型系统。1979年以来，主要从事分布并行处理研究。20世纪80年代初，主持开发了国内第一个分布式计算机系统Wudp-80，随后又主持开发了Wudp-85、Wudp-88、Wudp-91等分布式并行计算机系统，并在分布并行算法方面进行了卓有成效的研究。1989年，在人工智能和智能系统新原理研讨会上提出“多时空思维”的新学术观点，受到学术界的普遍关注。20世纪90年代以来，主持多项国家自然科学基金课题、国家攀登计划课题和863基础研究课题。近几年发表论文30余篇。

内容简介

基于XML的半结构化信息集成技术成为当前信息技术十分活跃的前沿研究领域之一。本书系统介绍了分布式环境下基于语义的XML信息集成的原理、方法，技术及原型系统。并总结了作者在该领域的研究成果和国内外同行的研究工作。本书较系统和全面地介绍了基于语义的XML信息集成技术的各种背景知识和相关的新思路、新观点和新成果，可以作为计算机科学与技术和信息技术专业高年级本科生、研究生教学用书，也可供从事这方面研究和开发工作的科技人员参考。

前 言

信息系统的广泛应用和互联网技术的发展，促进了人们对完整获取分布、异质信息的需求。然而，由于分布式环境下半结构化信息和非结构化信息在结构上和语义上的异构性，实现信息的共享、交换和互操作往往十分困难。主要表现在以下方面：

（1）在互联网应用领域，由于大部分信息资源采用基于 HTML 的语言进行表示和存储，虽然方便了人们之间的信息交流，但由于 HTML 语言本身的限制，计算机之间无法识别相互表示的信息资源，造成互联网资源利用率过低，大量的资源被浪费和闲置。

（2）在企业信息化建设、电子商务和电子政务方面，由于各个系统均采用不同的信息表示机制，造成企业内部的“信息孤岛”，无法建立统一的信息访问机制，从而充分利用各信息系统之间的关联信息进行分析、统计，造成企业内部资源被浪费，或者降低了信息处理工作的效率。

（3）在个人信息处理方面，传统的基于目录（Directory Oriented）的个人信息管理方式已经逐渐不能满足个人用户需求，每天人们不得不花费大量的时间从电子邮件、Word 文件及其他个人信息系统中寻找所需要的信息，并且需要花费大量的时间来对这些信息进行相应的格式变换处理以适应另一个业务系统的需求。

近年来，XML 信息表示技术和基于本体（Ontology）的知识表示技术取得了较大进展，为解决上述问题提供了相应的技术基础。如果利用 XML 优秀的信息表示能力表示某些领域的半结构化信息或者描述其关键信息点，并充分利用本体描述隐含于信息资源中的知识，实现各异构的信息资源在语义级的共享与处理将具有重要的意义，具体可以表现为：

（1）由于目前无法找到类似于关系代数的方式对半结构化数据进行形式化的描述，扩展现有数据与信息表示机制来支持对半结构化的数据处理是一种有益的探索。XML 作为一种文档结构描述语言，不具备语义表示的能力，采用基于本体的语义表示机制扩展 XML 表示的信息及其查询处理机制，以达到语义级别处理半结构化数据的能力。

（2）将极大地提高目前互联网资源的利用率，促进许多新的基于互联网的应用的发展。促进互联网逐渐向语义网和语义网格过渡。

（3）将为整个企业的信息资源提供统一的访问机制，实现企业信息资源的融合，实现基于语义的企业信息门户和知识共享平台。

（4）将改变目前个人信息的管理方式，消除应用程序和文档具体存放目录的差别，实现基于语义（Semantic Oriented）的个人信息管理方式，并可以进一步延伸到网格(Grid)或者 P2P 环境下的个人信息语义级别的管理。

采用基于 XML 的语言描述某些领域的半结构化信息是目前一种认可的方案。但需要指出的是 XML 毕竟只是一种定义文档结构的描述性语言，并且具有语法的多样性，它无法消除半结构化信息在语法和语义上的异构性。因此，如何充分利用本体来描述隐藏于非结构化

信息之中的语义信息及知识，并通过这些语义信息来克服不同节点之间的语义的异构性是一个值得深入研究的问题。另外，分布式信息集成是信息共享的一种主要方式，目前日益受到重视的网格（Grid）技术是信息集成的一种重要实现机制。如何在一个信息集成环境下提供一个一致的全局语义环境是信息集成技术目前急需要解决的问题，也是一个研究的热门话题。

基于上述目的，本书主要探讨分布式环境下基于语义的 XML 信息集成中一些需要解决的问题。

本书假定在分布式环境下各信息源可以采用或转化成 XML 表示的前提下，介绍了如何利用本体解决信息集成中的语义异构的问题。特别讨论了在 XML 信息集成环境下的语义处理问题。主要内容如下：

（1）介绍了一个分布式环境下基于 XML 的信息集成原型系统 OBSA，该系统采用了本体信息集成机制，利用 F-Logic 作为本体描述语言和表示机制，定义了一个信息表示机制的三级模型，并在此基础上设计了一个从本体到 XML Schema 的转化算法，以此为 XML 的数据处理提供一个语义环境。该系统采用语义适配器结构集成各种异质的半结构化信息资源，并利用一种基于本体扩展的 XML 查询语言 FL-Plus 实现对 XML 文档在语义级别的访问。

（2）针对现有的基于一对一本体映射机制的不足，分析了基于语义相似度的复杂本体映射机制，包括直接本体映射、包含本体映射、组合本体映射和分解本体映射等，定义了语义映射的特性，包括传递性、对称性和强映射特性。并在此基础上实现了基于复杂本体映射的本体集成，通过该集成机制，挖掘隐含于复杂映射中的概念及关系上的语义相似性。论述和提出了基于 Mediator-Wrapper 模式的本体集成机制的实现及相应的步骤和算法，包括四种本体映射机制的本体融合（Ontology Fusion）和根据本体映射机制的特性而实施的规范熔合（Canonical Fusion）。最终在一个集成信息环境下构建了一个全局共享的基于本体的语义环境。

（3）探讨了基于本体扩展的 XML 代数查询机制，克服了 XML 查询语言在语义级别处理的缺陷，解决了在一个集成环境下进行半结构化信息查询时的语义不完全或语义缺失问题，提高了查询精度，消除了查询过程中的冗余信息。并在此基础上讨论了如何利用集成的本体语义信息对查询进行重写，设计了相应的重写算法，制定了更为合理的查询规划。

（4）针对目前在数据网格环境下对基于信息集成研究存在的问题，提出一种基于 Mediator-Wrapper 的语义数据网格体系结构。它通过 Mediator 提供一个虚拟的数据源来兼容 OGSA-DAI 的数据网格标准，并在此基础上设计了一个基于 SOAP 的语义信息访问与处理的通信机制，实现了在网格环境下基于语义信息的处理。

（5）针对目前个人资源管理存在的问题，探讨了语义桌面（Semantic Desktop）的机制，设计了一个语义桌面原型系统 OntoBook。并讨论了扩展到语义 P2P 环境的方法。

全书是在陈莘萌教授的指导下，由顾进广具体负责编写，其中，本书第五章主要内容在杨玲贤和张琳硕士论文基础上整理而成，第八章语义桌面部分参考了周毅的硕士论文。在引用国内外专家的研究成果和技术背景知识时，笔者尽量在参考文献中列出引用成果的作者及出处，如有遗漏之处还请读者见谅。陈和平教授的学生张琳、杨玲贤老师和周静参与了 OBSA 原型系统的设计工作。学生周毅、胡博和冯琳协助整理了大量参考资料，周毅负责设计了语义桌面的原型系统 OntoBook。陈和平教授审阅了全稿，并提出了许多具体的修改意见。

本书所进行的研究得到了中国博士后科研基金（20060400275）、江苏省博士后科研基金

（0601009B）、湖北省教育厅科学研究项目（Q200711004）、湖北省自然科学基金(2007ABA296)和软件工程国家重点实验室（武汉大学）开放基金（SKLSE05-03）的资助。本书的顺利出版也得到了武汉科技大学计算机科学与技术学院学科建设资金和武汉科技大学青年人才科研启动资金的资助。作者感谢武汉科技大学计算机科学与技术学院和武汉大学计算机学院领导的关心与支持，感谢武汉大学出版社计算机事业部黄金文副编审的支持与帮助，也感谢两位作者家人的支持。作者之一顾进广要特别感谢何炎祥教授、徐宝文教授、陈建勋教授、张晓龙教授、陈和平教授、许先斌教授、方康玲教授、黄传河教授的支持与帮助。

作　者

2007 年 8 月

目 录

第一章 引 言

1.1 概述

计算机产业的迅速发展使得以计算机存储设备为载体的电子信息愈来愈多，根据信息的格式可以将其划分为结构化信息和非结构化信息两大类。结构化信息能够用统一的结构加以表示，有着非常良好的数据结构，如关系数据库、面向对象数据库中的数据或符号等；非结构化信息往往由自然语言表示，一般没有统一的结构。非结构化信息所涵盖的内容十分广泛，以企业信息化领域为例，非结构化信息主要可分为：

- 营运内容：如合约、票据、工作流及交易记录等；
- 部门内容：如各类文档、电子表格、电子邮件及日程安排等；
- Web 内容：如 HTML 网页及 XML 格式的信息等；
- 多媒体内容：如音频文件、视频文件、图像文件等。

信息时代给人类带来了迅速膨胀的信息量，一直从事数据方面研究的加州大学伯克利分校的统计结果表明：全球每年产生的信息多达 20 亿千兆字节，人均约 250 兆字节，而结构化信息只占到其中的 10%，其余 90%都是非结构化信息，并且这种增长势头还在持续。非结构化信息无疑在人类生活中扮演着越来越重要的角色，往往一个备忘录、一封邮件等这些“死角”里都会隐藏着非常重要的信息资源，用户对非结构化信息处理的要求也随之从简单的存储逐步上升为识别、检索和深度加工。

纵观非结构化信息的应用现状，主要存在着下列几大问题：

（1）信息的大量膨胀。大多数企业所处理的信息量平均每 12~18 个月就会扩展一倍，这种几何级数的增长速度的确使得相关人员感到无所适从。

（2）大量信息孤岛的存在。随着计算机系统的普及，企业先后使用各种相互独立的网络系统、应用系统(邮件管理，人事管理，销售管理等)，在部分提高效率的同时，这些系统的相互独立性也为企业的整体管理设置了障碍，它们缺乏一个统一的界面，没有相互连接的信息渠道，数据通常都被封存在企业的不同数据库、主机、文件服务器上，只有少数有特许访问权的用户能看到这些数据。为了查找一个问题，一般会要在各个系统中不停地切换，才能找到自己想要的信息。孤岛的存在越来越给企业的整体信息化带来了屏障。

（3）缺少个性化的信息。董事会、企业领导、企业员工、客户、供应商、合作伙伴等，这些都是企业信息的提供者和需求者，而他们所切入的角度和关注重点是不一样的。这种“个性化”体现在各个方面，如内容(一般的知晓/情报)、频率(例外/定期/持续)、结构(同类文件/各种来源的文件)、安全(加密/公开)、存取(个人/团队/公司)、集成(内部/集成/外部)等。而现在的企业信息化系统往往是“千人一面”，只实现了“如果你肯找，最终反正能找到”的这样一种被动式、大众化的信息提供方式，而没有实现个性化的信息存取。

随着非结构化信息应用范围的日趋扩大，如何有效地对它们加以利用已经成为进一步提高信息化水平的主要障碍。传统的数据库虽然在处理结构化的数据、文字和数值信息方面拥有非常成熟的技术，在金融、电信等领域的数值计算和实时事务处理上也得到了广泛应用，但由于自身底层结构的缘故，它们在管理非结构化数据方面显得有些先天不足，特别是对这些海量非结构化数据进行查询时速度较慢。非结构化信息能够表达的内容丰富多样，针对不同类型的非结构化信息编写专门的应用程序虽然可以达到信息访问的目的，然而不同场合不同时间同一种信息所需的处理方式、侧重点及力度可能都不一样，这就导致了应用程序需求的多变性和复杂性，不仅开发成本高，而且对非结构化信息的处理效果也不理想。

非结构化信息的类型多样，通常无法抽象成单一的信息模型，并且还具有异构性、分布性、增长性和变化性等显著特征：

（1）异构数据源。表现在各系统采用不同的软硬件平台、不同的数据模型以及不同的数据库来表示和存储数据。

（2）分布自治性。各系统都独立设计、实现并自治运行，具有各自完整的功能，相互之间的关联很弱。具有相同语义内容的数据往往表现方式完全不同。

（3）变化频繁、增长速度快。在各系统尤其是像 Web 站点这样的系统中，数据一直处于变化之中，不仅数量增长快，而且数据类型、数据格式以及表现数据的方式也在不断变化。

诚然，我们无法找到一种统一的方式来处理完全非结构化的信息，但是可以采取某种方式（如数据挖掘和机器学习的方法、基于规则的方法等）构建这些非结构化信息之间的关键信息，并采用合适的方式进行描述或者标记，这就是半结构化的数据表示方法。1998 年初，W3C(万维网联盟)完成了 XML 的初步设计。1999 年，W3C 在原有基础上制定了一系列标准，完善了 XML。在最近几年，围绕 XML 为基础的信息表示及处理标准不断丰富。XML 开始显示了可以承担描述半结构化数据重任的特征。

以 XML 及相关技术为基础的半结构化信息表示技术正影响着信息技术领域和计算机技术领域发生着重大的变化，这些变化表现为以下方面：

（1）资源共享的方式发生着重大的变化。早期的网络环境下，人们只能通过 Telnet、Gopher 等方式共享文本的信息资源，而且由于网络的限制，共享的范围也仅限于科学研究者或有限的大学师生。随着世界上第一个支持 HTTP 协议的互联网浏览器（基于 HTML 语言的浏览器）的诞生，互联网获得了蓬勃发展，开创了资源共享的新时代，普通的计算机用户也可以通过个人计算机浏览丰富多彩的互联网资源。同时企业计算模式也从传统的 C / S（客户 / 服务器）模式逐渐转化向互联网模式过渡，开创了电子商务和电子政务的新时代。但是，目前的互联网仅仅是方便人与人之间的资源共享，计算机与计算机之间依然无法采用一种统一的标准进行资源的共享。未来的互联网将进入语义网[1,2,3]（Semantic Web，如图 1-1 所示）时代，计算机与计算机之间能够通过共同的信息交换标准 XML 相互理解对方所表示的资源，Semantic Web、Semantic P2P、Semantic Grid 以及 Knowledge Grid[3,4]是这一时代主要技术发展趋势，资源共享的级别也将从传统的半结构化信息升级到知识共享领域。

（2）分布式计算技术标准化。XML 技术不仅促进了半结构化信息资源的共享，也促进了计算资源共享的标准化。早期的分布式计算技术构建于请求—服务的模式基础上，虽然中间件技术（如 DTP 中间件、数据访问中间件如 ODBC、JDBC 等）简化了这种分布式计算的复杂度，但计算资源（如接口、数据通信方式等）并没有统一的标准。组件技术（如 COM+、CORBA 和 J2EE）发展了这一需求，它通过接口标准化和通信方式标准化为分布式计算提供

了一个可以普遍遵循的标准，并通过引入标准的软件设计体系结构和设计模式（Design Patterns）简化系统构建过程，提高软件开发的效率。Web Service 和 Grid Service 技术更进一步地将这种标准化的趋势引入互联网环境下的分布式计算领域，它采用基于 XML 的语言描述计算机资源提供可提供的计算服务及接口标准（WSDL），计算需求方通过统一的服务中介机构获取相应的服务资源。服务的提供者与需求者之间的通信也构建于标准的基于 XML 的通信协议 SOAP 之上。

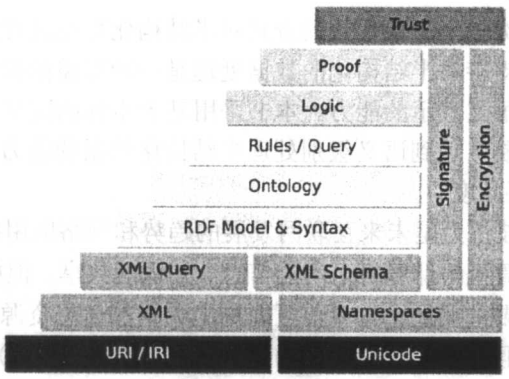

图 1-1 Semantic Web 的体系结构

尽管如此，由于半结构化信息在结构和语义上的异构性，要想达到在分布式环境下共享半结构化信息资源依然存在一定的困难。结构上的异构性表明在一个分布式的环境下，不同的节点可能会采用不同的结构表示其所包含的信息；语义上的异构性则表明不同的节点包含的信息所实际表示的含义可能存在的差异性。有关这个方面的研究，表现出以下的趋势：

（1）XML 具有数据模式的表示方法，它有着丰富的内容和关系、语法和语义的分离、内容和表现的分离等特性。因此采用基于 XML 的语言描述半结构化信息是目前普遍认可的方案。尽管它具有这些优势，但需要指出的是 XML 毕竟只是一种定义文档结构的描述性语言，并且具有语法的多样性。它无法消除半结构化信息在语法和语义上的异构性。

（2）利用本体来描述隐藏于非结构化信息之中的语义信息及知识，并通过这些语义信息来克服不同节点之间的语义的异构性是目前比较普遍的解决方案。本体(Ontology)[4,5]为特定领域内应用系统的设计提供共享的概念体系，能够减少或消除概念及术语上的混乱，使计算机对特定领域的知识处理更为精确、更为便捷。如果将本体与 XML 文档结构相关联，势必可以提高 XML 表达非结构化信息的语义丰富性和 XML 文档中各元素联系的准确性。

（3）分布式信息集成是信息共享的一种主要方式，目前日益受到重视的网格（Grid）技术使得信息集成实现机制标准化。如何在一个信息集成环境下提供一个一致的全局语义环境是信息集成技术目前急需要解决的问题，也是一个研究的热门话题。

（4）在分布式环境下，基于 XPath 的查询机制对半结构化信息进行处理是一个普遍接受的方案，特别是 XQuery 语言，它兼有结构化查询语言和过程化编程语言的特点，更成为半结构化信息查询事实上的工业标准。结合基于本体的语义信息提高查询和处理的准确度、优化查询过程、减少查询过程中的冗余，制定更为合理的查询规划方案也是目前半结构化信息处理研究的重点问题之一。

本书的主要目的就是充分利用各节点所隐含的语义信息或知识，消除不同节点的半结构化信息之间的语义差异性，在语义级别上共享、查询和处理不同节点上的半结构化信息。

1.2 研究意义

从以下几个角度讨论本书研究的意义：

1. 从技术发展的角度

由于目前无法找到类似于关系代数的方式对半结构化数据进行形式化的描述，扩展现有数据与信息表示机制来支持对半结构化的数据处理是一种有益的探索。XML 作为一种信息框架描述语言，不具备语义表示的能力。本书采用基于本体的语义表示机制扩展 XML 表示的信息及其查询处理机制，达到语义级别处理半结构化数据的能力。

2. 从互联网的发展角度

语义网、语义网格被认为是未来互联网发展的趋势和网络应用的基础平台，其核心基础是实现人与人之间、计算机与计算机之间的信息、知识的共享。但如本书所述，要想达到这一个目标还存在两个主要的问题：①目前互联网上大量的信息资源均以 HTML 或其他非结构化的形式来表示，而且在很长一段时间内这种状况不会改变。②由于网络环境的分布性，不同节点之间表示机制上存在表示结构和语义上的差异性。

充分利用基于本体的语义资源来描述各节点的语义信息，并以此为基础进行数据的查询及其他操作将是实现语义网或语义网格平台的基础。它将提高整个互联网资源的利用率，并从根本上改善现有互联网上一些应用的性能和质量，包括互联网资源的搜索、电子商务和电子政务等。

3. 从潜在应用前景的角度

（1）从企业信息集成的角度

如前所述，消除企业内部的“信息孤岛”，实现全面的企业信息化管理是企业信息集成目前的主要研究与应用方向。然而，由于企业内部信息系统形式多样，不同系统之间并没有统一的表示信息的标准（例如统一制定数据库的字段、结构，采用统一的语言描述企业的某一项业务等），不同系统之间同样存在语义和结构上的差别。另外，企业集成系统的另外一个特点是基于工作流和知识流基础上的工作协同。本书研究的目标将为企业内部或企业间信息集成提供一种通用的解决方案。它将体现以下显著的特征：

① 提供统一的基于语义的工作流或知识流平台，实现各节点间基于语义的协同。

② 在企业内部或企业之间构建统一的语义门户，采用同样的标准和操作方式对信息进行管理和操作。

③ 在语义级别实现企业信息系统之间的融合。

（2）从个人信息管理角度

本书所讨论的内容虽然基于分布式环境，但对于个人业务平台或者个人平台之间的协作同样具有参考价值。目前，随着个人计算机的处理能力不断增强，个人计算机的信息管理表现为：

① 基于文件目录的个人信息管理。按照文件的目录方式管理用户的文件（如 Word 文档、影音文件、电子邮件等）。

② 不同应用程序所描述的信息，即使在语义上表现的内容是相同的，也不得不按照不

同的方式来进行管理，例如用 Microsoft Word 编写的一篇论文和用 Wordperfect 编写的同样内容的论文，在语义上应该是相同的，但按目前的个人计算机的管理方式来说，二者没有任何相同之处。

本书所讨论的内容可以帮助个人计算机用户按照语义来管理个人计算机或者网络，构建个人计算机或者局域网络的语义门户。如果更进一步，可以协助用户在一个 P2P 环境下实现基于语义的信息共享。实现这一目标也是我们今后的工作。

1.3 本书的主要研究内容

本书主要探讨了在分布式环境下基于语义的 XML 信息集成技术。重点探讨三个方面的问题：

（1）分布式环境下基于语义的半结构化信息互操作机制，重点探讨基于本体的集成模式下的半结构化信息的互操作机制。对于半结构化互操作机制，主要考虑了设计分布式环境下针对半结构化数据处理的全局视图和局部视图的形式，定义基于本体的语义映射机制。对于集成机制，主要考虑了 Meidator-Wrapper 模型下基于语义相似度的复杂映射与集成机制。

（2）基于语义的半结构化信息查询机制，重点探讨如何利用语义信息扩展 XML 代数进行半结构化信息的查询，以及在一个集成环境下如何利用语义信息重写（Rewrite）查询请求，制定分布式环境下各节点的查询规划。

（3）在主流应用环境下的实现问题。

针对上述三个研究问题，重点研究了两种环境下的实现机制：

① 在一个分布式的基于 Mediator-Wrapper 模式的信息集成机制环境下的实现机制，包括基于语义的半结构化信息的表示以及扩展 XML 查询语言以支持基于语义的查询处理等。

② 数据网格是大规模基于数据处理系统的一个发展趋势，目前对于数据网格的研究主要在高效的存储、复制、安全等方面，基于语义的数据网格的处理目前研究的相对较少，本书探讨了在数据网格环境下如何查询和处理基于语义的半结构化信息的问题。

另外，如何利用信息集成机制来管理个人信息资源将是未来的一个研究热门课题，这里所描述的个人信息资源将不仅局限于本地个人计算机所存储和处理的信息资源，也包括互联网上某个人感兴趣的信息资源。本书在这个方面也作了一些探讨，重点探讨了语义桌面及其原型系统。

1.4 本书的组织形式

本书的第二章和第三章主要介绍了后续章节所需要的基础知识。其中第二章主要介绍了 XML 信息处理的基本机制，包括 XML 查询语言和 XML 查询代数，讨论了在分布式环境下 XML 的处理机制及目前所遇到的主要问题。第三章重点对本体进行了综述性讨论，并重点介绍了基于 F-Logic 的本体表示机制、RDF/RDFs 语言及 OWL 语言等。这两章所介绍的基本概念和相关的定义是后续章节的基础。第四章则介绍了信息集成的基本机制和模型。

第五章至第八章则重点讨论了三种环境下基于 XML 的信息集成机制。其中第五章重点讨论了基于本体的半结构化信息集成机制及基于本体的查询机制，其内容主要源于 OBSA 原型系统，主要讨论了 OBSA 集成环境中基于语义的半结构化信息表示和查询机制的具体实

现。第六章在第五章的基础上重点探讨了 Mediator-Wrapper 环境下基于语义相似度的信息集成机制。第七章探讨了如何在开放的数据网格平台环境实现对基于语义的数据处理。第八章则讨论了语义桌面及其原型系统，并探讨了扩展至语义 P2P 环境下基于语义的 XML 信息集成机制。

由于本书是围绕基于语义的 XML 信息处理的多个研究子项目的总结，因此每一章所讨论的命题具有一定的独立性，为保证每章的这种独立性，本书在每一章的开始给出了本章所涉及内容的综述，介绍目前需要解决的问题，然后介绍本书的解决方案。

1.5 本章小结

本章首先介绍了本书研究背景，指出基于半结构化的数据处理已经成为社会经济生活中一个非常重要的组成部分，但由于语义和结构上的异构性，使得半结构化数据在互联网应用领域、企业应用领域及个人信息管理领域等并没有发挥出应有的作用，同时本章介绍了本书研究的主要内容及其意义、全文的组织结构等。