中国计算机学会学术著作丛书

XML数据管理
概念与技术

XML Data Management
Concepts and Techniques

孟小峰 著

清华大学出版社

中国计算机学会学术著作丛书

XML数据管理
概念与技术

XML Data Management
Concepts and Techniques

孟小峰 著

清华大学出版社
北京

内 容 简 介

本书从数据库系统实现的角度，依据作者多年的研究成果全面系统地介绍了 Native XML 数据库系统相关技术。内容涵盖了 XML 数据库存储管理技术(包括存储、编码、索引等方法)；XML 查询处理与优化技术(包括 XML 查询代数、结构查询处理、整体查询处理、近似查询处理、查询优化等)，以及 XML 数据管理新技术(包括 XML/Update 处理、访问控制、关键字查询等)；最后介绍典型 XML 数据库系统和基准测试。

本书主要面向高年级本科生、研究生和研究工作者，它对学习者是很好的教材，对专业人士是很好的参考资料。

图书在版编目(CIP)数据

XML 数据管理：概念与技术/孟小峰著. —北京：清华大学出版社，2009. 10
(中国计算机学会学术著作丛书)
ISBN 978-7-302-20957-7

Ⅰ. X…　Ⅱ. 孟…　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(2009)第 163744 号

责任编辑：薛　慧
责任校对：王淑云
责任印制：王秀菊
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京鑫丰华彩印有限公司
装 订 者：三河市溧源装订厂
经　　销：全国新华书店
开　　本：175×245　**印　张**：21　**字　数**：430 千字
版　　次：2009 年 10 月第 1 版　　**印　　次**：2009 年 10 月第1 次印刷
印　　数：1～3000
定　　价：39.00 元

本书如存在文字不清、漏印、缺页、倒页、脱页等印装质量问题，请与清华大学出版社出版部联系调换。联系电话：(010)62770177 转 3103　　产品编号：034780-01

清华大学出版社计算机学术著作出版基金

评审委员会

中国计算机学会学术著作丛书

名誉主任委员：张效祥
主任委员：唐泽圣
副主任委员：陆汝钤
委　　员：（以姓氏笔画为序）
王　珊　吕　建　李晓明
林惠民　罗军舟　郑纬民
施伯乐　焦金生　谭铁牛

丛书序

Preface

第一台电子计算机诞生于20世纪40年代。到目前为止，计算机的发展已远远超出了其创始者的想象。计算机的处理能力越来越强，应用面越来越广，应用领域也从单纯的科学计算渗透到社会生活的方方面面：从工业、国防、医疗、教育、娱乐直至人们的日常生活，计算机的影响可谓无处不在。

计算机之所以能取得上述地位并成为全球最具活力的产业，原因在于其高速的计算能力、庞大的存储能力以及友好、灵活的用户界面。而这些新技术及其应用有赖于研究人员多年不懈的努力。学术研究是应用研究的基础，也是技术发展的动力。

自1992年起，清华大学出版社与广西科学技术出版社为促进我国计算机科学技术与产业的发展，推动计算机科技著作的出版，设立了“计算机学术著作出版基金”，并将资助出版的著作列为中国计算机学会的学术著作丛书。时至今日，本套丛书已出版学术专著近50种，产生了很好的社会影响，有的专著具有很高的学术水平，有的则奠定了一类学术研究的基础。中国计算机学会一直将学术著作的出版作为学会的一项主要工作。本届理事会将秉承这一传统，继续大

力支持本套丛书的出版，鼓励科技工作者写出更多的优秀学术著作，多出好书，多出精品，为提高我国的知识创新和技术创新能力，促进计算机科学技术的发展和进步作出更大的贡献。

中国计算机学会

2002 年 6 月 14 日

序

Preface

进入新世纪以来，数据库技术面临一场变革，即在原有关系数据库技术成熟之后，新的数据库技术在哪里？一个重要的趋势是具有灵活的半结构化特性的 XML 数据的出现。XML 作为一种数据存储和交换格式，在互联网络环境中扮演着极其重要的角色，它已经成为数据交换事实上的标准，在电子商务、电子政务、金融、出版、科学数据与各种资源的数字化等方面得到越来越广泛和深入的应用。可以想象，在不久的将来，XML 数据的规模将可能达到或者超过各种关系数据库中的数据规模，从而成为继关系数据之后新的主流数据形式。

如何有效管理 XML 数据自然成为寻找突破口的数据库界的热点研究问题。但在 2000 年研究之初，学界和工业界在技术选择上有过一些争论。主流工作认为应当以现有关系数据库为基础，试图建立 XML 数据与关系数据的映射关系，从而可以利用已有的关系数据库系统管理 XML 数据。这显然是受数据库技术历史沿革的影响。因为历史上人们要替代关系数据库系统的多次努力均告失败，尤以演绎数据库、面向对象数据库为代表。最终的结果是被关系数据库所“同化”。但代价是关系数据库这架马车负重累累，越跑越慢了。

显然，上述方法沿用了这一惯性思维，但其致命弱点是不言而喻的，即在将 XML 数据映射为关系数据的同时，XML 数据中某些属性值的缺失和重复，将导致大量关系表的产

生，进而影响 XML 数据的存储和重构的效率，导致该方法的效果和效率均不理想。

与之截然相反的方式是另起炉灶，针对 XML 数据的树形结构特点，开发纯(Native)XML 数据库管理系统，以便高效管理 XML 数据。学术界对这一技术进行了广泛深入的研究，取得若干重要成果。2006 年 IBM 率先将纯 XML 数据库技术引入其 DB2 的最新扩展版本，并开始应用的推广示范。纯 XML 数据库技术至此成为新世纪数据库变革的重要成果之一。迄今为止，国内外尚没有一本从系统构建的角度系统阐述纯 XML 数据库技术的书籍，而本书的出版填补了这一空白。

《XML 数据管理：概念与技术》一书从构建纯 XML 数据库系统的角度出发，对现有的关键技术分 14 章逐一进行详细深入的介绍，内容涵盖了纯 XML 数据库存储、编码、索引、查询处理与优化(包括查询代数、结构连接和整体匹配、查询简化、代价分析、查询分解)、XML 数据更新、XML 访问控制以及最新的 XML 技术(包括近似查询处理技术、XML 关键字查询技术等)，最后详细介绍了典型的纯 XML 数据库管理系统及基准测试。全书内容系统全面，结构清晰，深入浅出，既反映了 XML 数据管理领域的最新研究成果，又具有良好的可读性。

贯穿全书，作者对基本问题给出了深入细致的讨论，对行之有效的技术给出了详细具体的描述。对概念方面的内容，辅以一些仔细设计的图表加以说明。对技术方面的内容，则给出实现算法和实例，有效提高了读者对技术问题的了解与认识。

本书作者孟小峰教授长期从事数据库系统和理论的研究工作，20 世纪 90 年代参与我国第一个大型数据库系统 COBASE 的研制、基于曙光并行机的并行数据库系统 PBase 的开发，主持开发了我国第一个基于掌上电脑的嵌入式移动数据库系统——“小金灵”，从而系统深入地掌握了数据库系统的内核技术，奠定了扎实的基础。2000 年年初，孟小峰教授即较早在国内开展了纯 XML 数据库系统的研究，在 XML 数据的存储、索引、编码、查询优化、查询处理、数据更新、访问控制、关键字查询以及近似查询等方面进行了系统的工作，发表了多篇高水平论文，积累了大量有价值的研究成果。2002 年率先在国内开发了纯 XML 数据库系统 OrientX，历经 8 年，先后发布 6 个版本，功能日渐完善，成为本领域一个有代表性的成果。本书是作者多年来对 XML 数据管理技术研究的结晶，它所体现出的坚持长线研究，并将理论研究和系统实现有机结合的研究理念和风格，应为当下学术界所推崇。

相信本书对广大的科研工作者和研究生具有重要的学术参考价值，也可作为高等院校相关专业的研究生教材。

何新贵

2009 年 7 月 6 日

前言

Foreword

随着信息技术的迅猛发展，人们可以通过互联网从世界各地接收和发送信息，而信息交换过程中的一个突出问题就是数据格式的异构性，这将极大地阻碍对信息进行有效的使用。XML 正是针对这一问题而提出的解决方案。随着计算机和网络技术的不断发展，XML 技术的应用也将不断扩展。该技术不仅可以用于银行之间进行数据交换、证券公司对其上市公司相关的数据进行统计、图书馆对其馆藏书目进行查询检索、企事业单位对其文件档案进行管理，还可用于电子商务、出版行业、医疗管理、科学数据管理等领域。

XML 又称可扩展标记语言(eXtensible Markup Language)，是由 W3C 组织于 1998 年 2 月发布的一种数据标准。作为 SGML 的一个简化子集，它集成了 SGML 功能丰富与 HTML 易用性的特点，以一种开放、自描述的方式定义数据结构。XML 可以同时描述数据内容和结构特性，通过这些结构特性，可以了解数据之间的语义关系。与 HTML 相比，HTML 是写给人看的，而 XML 则是写给机器看的。与 SGML 相比，XML 显得更为简单，同时也可用于设计文档描述语言。

XML 技术在当前的互联网络和 IT 环境中扮演着越来越重要的角色，它事实上已经成为数据交换的标准、SOA 架构的基石。据 Gartner 预测，XML 文件的使用率在 2007 年达到 40%，在 2008 年占据支配地位。IDC(国际数据公司)最近发布的一份报告显示，在 500 家受访企业的 IT 部门中，有

29%的企业宣称正在大量使用 XML 数据库。XML 的广泛应用使得高效的 XML 数据管理成为一种迫切的需求。

XML 数据具有树形结构的特点，这和关系数据模型是不同的，因此，如何有效地管理 XML 数据成为数据库领域新的挑战问题。

XML 数据管理经历两个阶段：

一是基于关系的 XML 数据管理。其特点是在关系型数据库内核层的基础上，将 XML 的树形结构数据拆散、重组转换成关系型数据存入关系数据库，在提取 XML 数据时，利用 SQL 语言的优化将表格型数据取出并还原成 XML 树结构数据。实践表明，这种方法在效率和效果上均不理想。

另一种称为纯 XML 数据库系统(Native XML Database Systems)，其特点是以自然的方式处理 XML 数据。纯 XML 数据库系统能够保持 XML 数据的树形结构，可以将结点或者子树作为存储单元，针对 XML 数据存储和查询特点专门设计适用的数据模型和方法，从数据库核心层直至其查询语言都采用与 XML 直接配套的技术。因此，纯 XML 数据库系统得到了研究者的广泛关注。

本书作者自 2000 年起即对纯 XML 数据管理进行了连续多年的系统性研究。本书基于作者多年在 XML 数据管理系统方面的研究积累，从数据库系统实现的角度，全面系统地介绍了纯 XML 数据库系统相关技术。内容涵盖了 XML 数据库存储管理技术(包括存储、编码、索引等方法)；XML 查询处理与优化技术(包括 XML 查询代数、结构查询处理、整体查询处理、查询优化等)，以及 XML 数据管理新技术(包括 XML/Update 处理、访问控制、关键字查询、近似查询处理等)；最后介绍典型的纯 XML 数据库系统和基准测试。

本书的内容和组织结构

本书按 XML 数据管理系统自底向上的逻辑层次，共分 14 章。

第 1 章是本书的总述，对 XML 数据管理的基本概念及技术进行了概括，并对未来 XML 数据管理技术发展趋势做了展望。

第 2 章介绍了 XML 基础知识。着重介绍如何使用 XML DTD 和 XML Schema 来验证 XML 文档的正确性和有效性，如何使用 XPath 和 XQuery 等查询语言来对 XML 文档进行查询，如何使用 XQuery/Update 来实现 XML 文档的更新，以及如何使用 SAX 和 DOM 来进行 XML 文档的访问。

第 3 章介绍了 XML 数据库存储技术。分析了 XML 存储需要考虑的存储粒度和存储策略，并介绍了 4 种主要的存储策略，最后着重介绍了支持 XML 更新的存储方法。

第 4 章介绍了 XML 数据编码技术。介绍了 4 种常见的编码方法，即：区域编码、前缀编码、k 分树编码以及支持动态更新的编码方法。

第 5 章介绍了 XML 数据索引技术。讨论了 XML 数据管理中的索引问题，并介绍了两种类型的索引方法：路径索引与序列索引。

第 6 章介绍了 XQuery 的代数处理策略，包括 TAX，Xtasy，TTX，OreintXA 这几种已有的 XML 查询代数。其中着重介绍了 XML 代数 OreintXA，以及扩展支持 XQuery/Update 的代数。

第 7 章介绍了 XML 查询处理技术中的基本处理方法，包括二元结构连接算法、以目标结点为导向的 XML 路径查询处理算法和基于区域划分的结构连接算法。

第 8 章介绍了 XML 查询处理技术中的整体处理方法，包括整体的 Twig 查询处理、基于序列匹配的 Twig 查询处理以及复杂 Twig 查询处理。

第 9 章讨论了 XML 查询优化技术。主要介绍了 XML 查询优化中查询树简化、代价分析和查询分解三个主要问题，并给出对应的解决方法。

第 10 章阐述 XQuery/Update 中 Transform 查询的处理以及优化技术。介绍了基于代数处理的 XQuery/Update 查询的优化策略以及策略选择的方法。

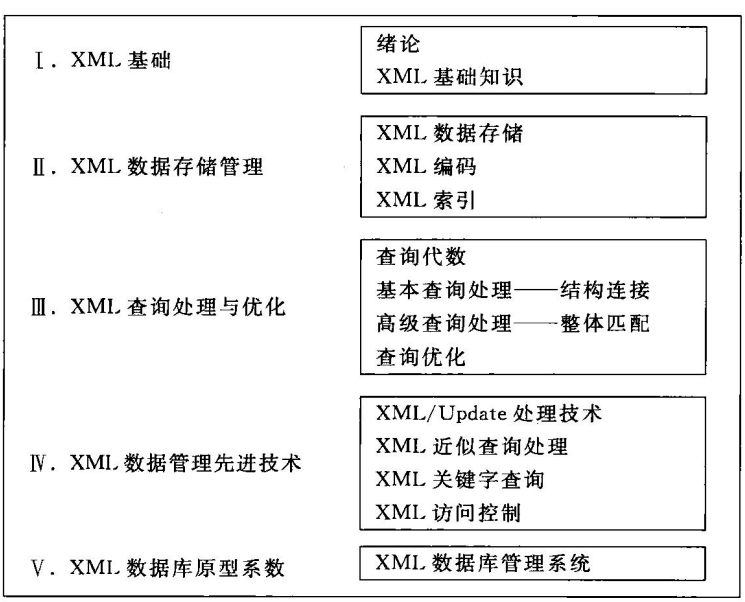
第 11 章介绍了 XML 近似查询处理技术。重点讨论了如何有效查询复杂结构 XML 文档的问题。

第 12 章介绍了 XML 关键字查询技术，包括当前经典的 XML 关键字查询语义和查询算法。

第 13 章讨论了如何扩展基于角色的访问控制技术来进行 XML 访问控制的方法。

第 14 章详细地介绍了典型的纯 XML 数据库管理系统及基准测试。

全书每章均附有习题和参考文献，书末附有总参考文献和词汇索引。



章节结构图

本书的对象和使用方法

本书从数据库系统的实现角度出发，深入详细地介绍了 XML 数据库相关技术，包括 XML 数据存储、编码、索引、查询代数、基本查询处理（结构连接）、高级查询处理（整体匹配）、查询优化、原型系统和基准测试等，内容系统全面；本书同时强调内容的先进性，将 XML 数据管理的最新成果涵盖进来，包括 XML/Update 处理、XML 近似查询、关键字查询、访问控制等。本书在阐述上力求简洁明了，对概念方面的内容，辅以一些图表加以说明，对技术方面的内容，则给出精炼的实现算法。本书在编写中考虑到教学的需要，给出了大量的实例。每章的内容安排大致相当，均附有习题、相关文献以及文献导读。书后附有总文献和词汇索引。

本书主要面向系统开发人员，高年级本科生、研究生和科研工作者，它对学习者是很好的教材，对专业人士是很好的参考资料。基于本书可开设的课程有如下建议：

XML 数据管理技术概述：面向高年级本科生选修课程，内容包括第Ⅰ，Ⅱ，Ⅲ，Ⅴ；

XML 数据库实现技术：面向研究生的课程，包括第Ⅰ～Ⅴ；

XML 数据存储管理专题：与本科数据库管理系统课程结合，补充内容Ⅱ；

XML 数据查询处理专题：与本科数据库管理系统课程结合，补充内容Ⅲ；

XML 数据管理先进技术专题：与相关研究生课程结合，补充内容Ⅳ。

致谢

作者从 2000 年起即开始从事 XML 数据管理的研究，2002 年自主开发的纯 XML 数据库原型系统 OrientX 在 W3C 网站上发布。本书是作者在多年来形成的研究成果的基础上，经过总结和整理而成。

这里首先要感谢我的导师王珊教授，她带我进入数据库研究这一殿堂，并有机会参与研发了我国第一个大型国产数据库系统 COBASE（国家“八五”、“九五”科技攻关计划，1991—1998）、基于曙光并行机的并行数据库系统 PBase（国家“863”计划，1996—1998）、中文自然语言查询系统 NChiql（国家自然基金重点项目，1996—2000）、嵌入式移动数据库系统“小金灵”，以上经历使我对数据库系统的内核技术有了系统深入的了解，从而奠定了扎实的基础。

其次要感谢国家自然基金委多年长期的资助和国家“863”计划的一贯支持。说来奇怪，本书的研究成果一直是在没有明确的课题资助的情况下凭借作者的执著和坚持而取得的，可见研究贵在坚持。

本书的形成凝聚了中国人民大学网络与移动数据管理实验室（http://idke.ruc.edu.cn）的集体智慧。特别感谢实验室的博士研究生和硕士研究生周军锋、王宇、王

静、罗道峰、安靖、陆世潮、蒋瑜、陈妍、欧建波、王小锋、朱金清、王伟、徐俊劲、富丽贞等。特别是周军锋、朱金清、王伟、徐俊劲、富丽贞在资料收集和文献整理方面做了大量工作。

本书涉及面广，内容丰富，涉及的文献众多。值得指出的是，在全书的撰写和课题的研究中，尽管作者投入了大量的精力、付出了艰苦的努力，然而受知识水平所限，书中不当之处在所难免，诚恳读者批评指正并不吝赐教。如果有任何建议或意见，可发电子邮件至 xfmeng2006@gmail.com。

孟小峰

2009 年 3 月于北京

目 录

Contents