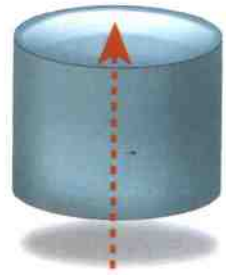
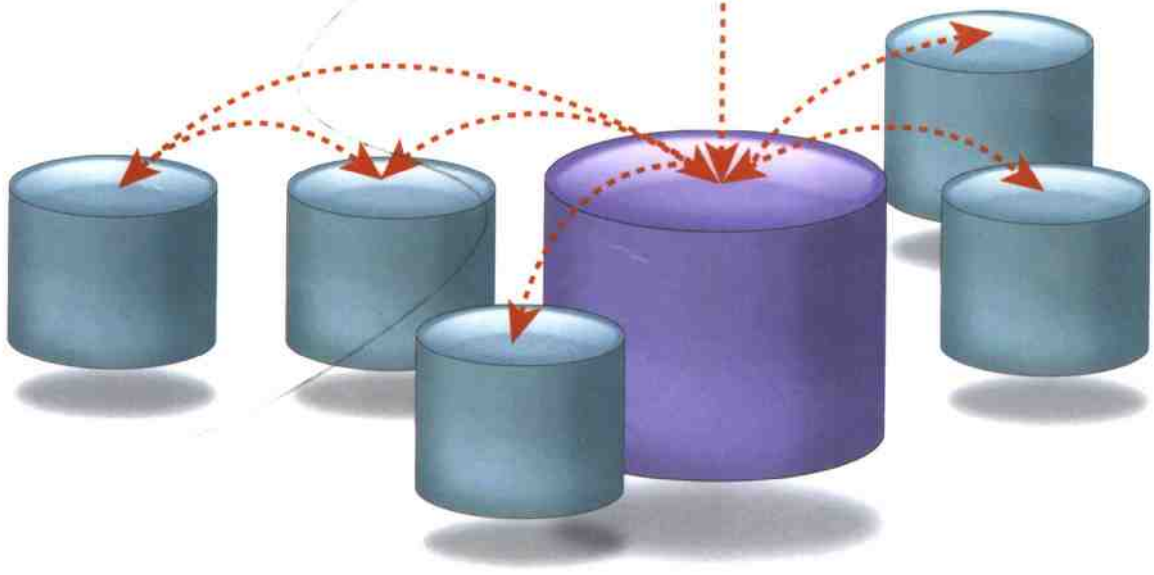
高等学校教材系列

数据库



新理论、方法及技术导论

刘国华　张忠平　岳晓丽　等著



電子工業出版社
Publishing House of Electronics Industry
http://www.phei.com.cn

高等学校教材系列

数据库新理论、方法及技术导论

刘国华　张忠平　岳晓丽　等著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内容简介

本书对近几年数据库领域出现的新理论、方法和技术进行了较全面的阐述。主要内容涉及XML数据的规范化理论、数据库模式匹配方法、对等数据管理系统中数据映射的推导方法、XML动态集成方法、XML访问控制技术、广域传感器数据库中的查询处理技术、数据库视图安全技术、空间数据库中轮廓查询及更新技术、空间网络数据库中最近邻查询技术、数字文档复制检测技术、数据库保序加密技术。

本书内容丰富，知识体系新颖，理论与实践相结合，具有先进性和实用性，可以作为高等学校计算机、信息与计算科学及信息管理与信息系统等专业硕士、博士研究生的教材或参考书，也可供从事信息领域工作的科技人员和工程技术人员以及其他有关人员参考阅读。

图书在版编目（CIP）数据

数据库新理论、方法及技术导论/刘国华等著. -北京：电子工业出版社，2006.12
（高等学校教材系列）
ISBN 978-7-121-03646-0
Ⅰ.数… Ⅱ.刘… Ⅲ.数据库系统 Ⅳ.TP311.13
中国版本图书馆CIP数据核字（2006）第153906号

责任编辑：李秦华
印　　刷：北京市天竺颖华印刷厂
装　　订：三河市金马印装有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路173信箱　邮编：100036
开　　本：787×980　1/16　　印张：22.25　　字数：570千字
印　　次：2006年12月第1次印刷
定　　价：35.00元

凡所购买电子工业出版社的图书有缺损问题，请向购买书店调换；若书店售缺，请与本社发行部联系。联系电话：（010）68279077。邮购电话：（010）88254888。
质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。
服务热线：（010）88258888。

序　言

数据库是一个极富挑战性的研究领域，也是一个十分活跃的研究领域。多年来，在国内外学者的共同努力下，数据库领域生机勃勃。通过与多学科的有机结合产生了一系列新型的数据库，如面向对象数据库、分布式数据库、并行数据库、演绎数据库、知识库、多媒体数据库、移动数据库、工程数据库、统计数据库、科学数据库、空间数据库、地理数据库、实时数据库、Web 数据库等。同时又出现了很多新的研究方向，如数据仓库、数据挖掘、XML 数据管理、数据集成、对等数据管理、传感器数据管理、数据库安全与加密、空间数据管理、流数据管理、数据网格等，形成了很多新理论、新方法和新技术。本书的内容是这些理论、方法和技术的一个真子集，主要涉及 XML 数据的规范化理论、数据库模式匹配方法、对等数据管理系统中数据映射的推导方法、XML 动态集成方法、XML 访问控制技术、广域传感器数据库中的查询处理技术、数据库视图安全技术、空间数据库中轮廓查询及更新技术、空间网络数据库中最近邻查询技术、数字文档复制检测技术、数据库保序加密技术。他们是本书作者所在的燕山大学数据库课题组近几年研究工作的一个阶段性总结，是课题组全体成员集体智慧的结晶。

随着我国数据库事业的发展，从事数据库研究工作的人员越来越多，如何及时了解国内外的研究动态，掌握新知识是人们面临的一大难题。为了帮助大家克服这个困难，近几年来，作者所在的数据库课题组一直紧跟国际前沿，对 VLDB，SIGMOD 和 PODS 等著名国际会议上的论文进行认真归类、筛选，然后结合自身优势进行重点研究，取得了一定的研究成果，这些成果被分配到本书的每一章中。为了便于读者阅读和对问题进行更深入的研究，本书的每一章为一个主题，主要内容包括研究现状分析、本章所需的基础知识及最新的研究成果，这样安排可以使读者用最短的时间掌握新知识，找到问题研究的切入点。

本书的形成得到了教育部科学技术研究重点项目（No.205014）、燕山大学博士基金项目（No.B125, B144,B81）及燕山大学研究生课程建设项目（No.XX0402）的支持，在此一并表示感谢！

本书由刘国华、张忠平和岳晓丽主持和统稿，参加本书撰写工作的还有刘欣、荣凌燕、刘通、钱颖、郜时红、李旭、韩梅、马君、张淑芝、张坤、李静、候士江、于醒兵等。限于我们的水平，不妥之处恳请读者指正。

前　言

数据库的研究和开发经历了四代的演变，取得了辉煌的成就，成为世界各国信息基础设施的核心技术和重要基础。

数据库是一个极富挑战性的研究领域，也是一个十分活跃的研究领域。从 20 世纪 60 年代开始至今，大量的研究成果不断涌现，每年都有数千篇学术论文在重要学术会议和学术刊物上发表，每隔几年就会出现一大批新的挑战性问题，随之又会出现大量解决这些问题的研究成果和新产品。近年来，随着计算机软硬件系统、数据库应用、数据容量和类型的迅速变化，数据库领域的活跃程度和变化速度与日俱增。

针对数据库技术的进展和我国数据库应用水平的提高，结合作者多年来在数据库领域的研究成果和实践，并查阅了国内外大量数据库研究成果和文献，把数据库领域的新理论、新方法和新技术纳入《数据库新理论、方法及技术导论》一书。

本书的内容大致安排如下：

第 1 章介绍了 XML 数据的规范化理论，包括 XML 文档定义、XML 函数依赖定义、XML 函数依赖推理规则集、XML 范式及文档规范化、XML 多值依赖定义、XML 多值依赖规则集以及 XML 多值依赖下的范式等。

第 2 章介绍了数据库模式匹配方法，包括模式匹配的相关基础知识、模式匹配技术的多种分类标准、基于副本的完整模式匹配方法、复杂模式匹配方法和基于全集的复杂模式匹配方法等。

第 3 章介绍了对等数据管理系统中数据映射的推导方法，包括数据映射表、映射关系模型——语义图、改进的映射推导算法、数据映射推导系统以及对等数据管理原型系统等。

第 4 章介绍了面向模式的 XML 动态集成方法，包括 XML 数据库模式、属性集成语法、面向模式的 XML 动态集成框架等。

第 5 章介绍了 XML 递归模式的访问控制技术，包括基于安全视图的访问控制模型、XML 递归安全视图的查询重写算法以及展开递归结点算法等。

第 6 章介绍了广域传感器数据库中的查询处理技术，包括查询等价分解方法、传感器网络中的多查询以及多查询处理体系结构等。

第 7 章介绍了数据库视图安全技术，包括敏感信息泄漏类型、基于关键元组的消除信息泄漏方法、基于先验知识的消除信息泄漏方法、基于关系覆盖的 k-匿名保护法以及基于熵的消除信息泄漏方法等。

第 8 章介绍了空间数据库中轮廓查询及更新技术，包括基于动态窗口查询的轮廓查询技术、轮廓更新技术、轮廓体更新技术以及数据流环境中的轮廓体查询技术等。

第 9 章介绍了空间网络数据库中最近邻查询技术，包括静态的 k-NN 查询算法、增量 k-NN 查询算法、动态的 k-NN 查询监视、增量 k-NN 监视算法以及群组 k-NN 监视算法等。

第 10 章介绍了数字文档复制检测技术，包括文档复制检测的定义、发展、通用系统结构、文档特征提取方式、文本块的选择和基于串匹配方法的文档复制检测技术等。

第 11 章介绍了关系数据库中的字符数据加密技术，包括数据加密的算法、技术、加密系统的功能要求和基于数值型数据的保序加密的性质和原理等。

我们在本书的写作过程中，查阅了国内外大量数据库领域的研究成果和文献，将作者所在的燕山大学数据库课题组近几年研究工作的阶段性成果的新理论、新方法和新技术纳入本书，使之能够反映数据库领域的最新进展。但是，由于才疏学浅，时间紧迫，不足之处在所难免，如蒙读者指正，不胜感谢。

作者

2006 年 10 月 30 日于燕山大学

目　录

第 1 章　XML 数据的规范化理论

随着 XML 成为 Web 上的数据表示和数据交换的标准，需要通过 Web 交换和处理的 XML 数据大幅度增加，这就对 XML 数据的模式提出了更高的要求。同关系数据库类似，如果 XML 数据模式设计得不好，同样会引起插入、删除和更新等异常。由于 Web 的开放性，XML 数据异常的危害性要远远大于关系数据异常的危害性。在关系数据库领域中，规范化是模式设计的基础，它是评价一个模式好坏的主要依据。对于 XML 领域，讨论相应的问题也具有极其重要的意义。本章从数据库设计的角度出发，在 XML 数据中引入函数依赖和多值依赖这两种重要的数据约束，在此基础上，对 XML 数据的规范化理论进行讨论。

1.1　引言

1998 年 2 月，万维网联盟（W3C）推出了可扩展的置标语言 XML（eXtensible Markup Language）作为 Web 上进行半结构化数据传输与交换的标准。随着 XML 的出现，XML 数据相关技术研究成为热点。例如，XML 数据的存储技术与发布技术的研究；XML 数据查询与优化技术等。这些方面的研究都是基于现有的 XML 数据进行直接地存储、转换、查询与优化，等等。现有的 XML 存储方法，都是一个从 Web 世界到机器世界直接的转换过程，它们仅仅考虑到如何完整地保留 XML 文档中的结构信息，而没有从数据库设计角度来评价所得到的关系数据库模式，这样必将对后来的数据处理带来很大的影响，因为这样的数据库可能存在着插入、删除和修改异常。为了避免这些更新异常的出现，就必须对得到的关系数据库模式按照数据库设计中的要求进行改进。也就是说，现有的方法看起来简单，实际上要想真正得到一个优秀的数据库却是非常复杂的。

本章的思想不是从 Web 世界到机器世界直接转换，而是着眼于数据库设计，直接对 Web 世界中的 XML 模式进行处理，对其数据约束进行规范化研究，从而得到规范化的 XML 数据模式，不仅完整地保留了 XML 文档中的语义和结构信息，而且一次性地完成了一个良构数据库的设计。在规范化的 XML 数据上进行存储、集成、发布和传输交换数据，保证了数据在互联网上的一致性，提高数据质量，在存储效率和查询优化上具有重要的实用价值。目前需要通过 Internet 交换和处理的 XML 数据大大增加，对于 XML 数据规范化理论的深入研究将有力地促进企业的信息化、电子商务以及电子政务的发展，具有巨大的应用前景和经济效益。

1.1.1　半结构化数据及 XML 模式

XML 作为一种半结构化数据的表示模型，从提出到现在只不过几年的时间，但它作为一种跨

产品、跨界面、跨平台的互联网的标准语言，已经显现出其强大的应用前景，并受到了政府、企业和各大软件厂商的广泛关注。各个行业，如金融机构、海关、媒体产业正制订各自行业的 XML 文档类型定义（DTD，Document Type Definition），以利于数据以公认的格式进行交换与集成。随着 XML 数据的增多，相关行业标准 DTD 的制订，人们也开始越来越多地希望以对待数据库的方式来处置和管理 XML 文档。人们关注的首要问题是，用 XML 表示的数据之间有什么联系？有什么约束？截止到 2006 年 6 月，大约就有多达 12 种以上的 XML 模式语言被提出。比如，XML DTD，XML-Schema，XDR，SOX，Schematron 以及 DSD 等。它们之间的特性比较如表 1.1 所示。

表 1.1 特性比较概要

特性	DTD1.0	XML-Schema1.0	XDR 1.0	SOX 2.0	Schematron 1.4	DSD 1.0
模式						
XML 中的语法	否	是	是	是	是	是
命名空间	否	是	是	是	是	否
包含	否	是	否	是	否	是
输入	否	是	否	是	否	否
数据类型						
内置类型	10	37	33	17	0	0
用户定义类型	否	是	否	是	否	是
域约束	否	是	否	部分	是	是
显示空值	否	是	否	是	否	否
属性						
默认值	是	是	是	是	否	是
选择性	否	否	否	否	是	是
可选与必需	是	是	是	是	是	是
域约束	部分	是	部分	部分	是	是
条件定义	否	否	否	否	是	是
元素						
默认值	否	部分	否	否	否	是
内容模型	是	是	是	部分	是	是
有序序列	是	是	是	是	是	是
无序序列	否	是	是	否	是	是
选择性	是	是	是	是	是	是
最小和最大次数	部分	是	是	是	是	部分
开放模型	否	否	是	否	是	否
条件定义	否	否	否	否	是	是
继承						
扩展简单类型	否	否	否	否	否	否
约束简单类型	否	是	否	是	否	否
扩展复杂类型	否	是	否	是	否	否
约束简单类型	否	是	否	否	否	否

（续表）

特性	DTD1.0	XML-Schema1.0	XDR 1.0	SOX 2.0	Schematron 1.4	DSD 1.0
唯一性或键						
属性唯一性	是	是	是	是	是	是
非属性唯一性	否	是	部分	否	是	否
属性键	否	是	否	否	是	否
非属性键	否	是	否	否	是	否
属性外键	部分	是	部分	部分	是	是
非属性外键	否	是	否	否	否	是
其他						
动态约束	否	否	否	否	是	否
版本	否	否	否	否	否	是
文件/记录	否	是	否	是	是	是
嵌入HTML	否	是	否	是	部分	是
自描述性	否	部分	否	否	部分	是

从表1.1中可以看出：

从“使用容易”角度来看，DTD是最容易学习的模式语言。Schematron模式语言描述是相对简单的，但要求用户仍需学习另一种语言XPath，才能展示更多的功能。DSD模式比XML-Schema和Schematron模式倾向于更加详细一些，由于XML-Schema和DSD支持广泛的性质集，相对而言是比较难学的，但是从DTD很容易移植到其他模式语言中。

从“语言”角度来看，XML模式语言可以从多个方面来划分，比如基于语法与基于模式，面向定义与面向有效性，面向结构与面向约束，等等。DTD，XML-Schema，XDR和SOX属于基于语法分组，而Schematron属于基于模式分组。DSD介于两者之间，同时支持两种性质。基于语法分组在XML查询中有优势，已知模式结构和定义可以帮助用户书写更多的优化查询。另一方面，基于模式语言划分可以在表达中能更好地描述约束。

从“数据库”角度来看，没有一种语言彻底地满足需要。SQL DDL（Data Definition Language，数据定义语言）描述不仅仅是关系和属性集的规范，还包括每个属性、完整性约束、每个关系安全性等索引相关的值域信息。XML模式支持各种固定的域类型，并不能表达关系模式所描述的全部内容。尽管Schematron或DSD能表达完整性约束，但它们不支持物理索引描述功能。

上述几种典型的模式语言，都是从语言设计角度定义的，缺少对XML模式约束的描述。因为约束是数据语义的重要组成部分。然而XML文档作为半结构化数据的特例，虽然它很容易表达来自不同数据源的数据，但是，由于DTD与XML-Schema这些模式定义方法对于约束的描述都是有限的，其所能表示的语义信息也是相对有限的。

XML是半结构化数据的特例。半结构化数据是界于严格结构化的数据（如关系数据库中的数据）和完全无结构的数据（如声音、图像文件）之间的数据形式，它具有如下一些特点：

(1)**隐含的模式信息**：半结构化数据是具有一定的结构，但其结构与数据混在一起，没有显式的模式定义，如 HTML 文件。

(2)**不规则的结构**：一个数据集合可能有异构的元素组成。例如，学生集合中某些学生有电子邮件地址，而另一些学生则没有。同样的信息可能由不同类型的数据表示。例如，某些姓名是字符串，而另一些则是由 first name 和 last name 组成的复杂结构。

(3)**没有严格的类型约束**：由于没有一个预先定义的模式，以及数据在结构上的不规则性，所以缺乏对数据的严格类型约束。

目前，国内外关于半结构化数据的研究主要集中在新的数据模型、查询模式、存储技术以及优化技术等方面。在众多的研究课题中，对半结构化数据结构的研究是一个非常重要的方向。半结构化数据存在一定的结构，但这些结构或者没有被清晰地描述，或者是经常动态变化的，或者过于复杂而不能被传统的模式定义来表现。

没有强制性的模式约束，使得半结构化数据具有很大的灵活性，能够满足网络这种复杂分布环境的需要，但是也给数据的处理带来了很大的困难，使得数据处理的效率低下，很难具有实用性。半结构化数据模式在实际的数据处理中有着很广泛的用途。主要有：

(1)**用户界面**：由于半结构化数据没有明确的模式，给用户查询带来了很大的困难。模式信息有助于用户了解数据的结构，从而提出更精确和有效的查询。

(2)**查询优化处理**：模式信息有助于查询处理器对查询计划进行优化，大大缩减查询的搜索空间。

(3)**改进数据存储**：了解模式信息，可以更好地设计数据的物理存储结构和索引结构，从而提高查询执行的效率。

(4)**异构数据源的集成**：了解不同数据源的模式信息，有助于选择适当的集成模式和定义转换规则。

由于认识到半结构化数据模式的重要性，近年来学者们已经在这方面做了很多研究工作，有许多工作目前仍在进行当中。

对于半结构化数据的模式，目前已经提出了多种描述形式，比较有代表性的有基于逻辑的形式和基于图的形式。无论是哪种描述形式，其讨论的基础都是采用带标记的有向图作为半结构化数据模型，最典型的就是 OEM 模型，概括来说目前有两类描述方法：

(1)**基于逻辑的描述形式**：在已经提出的半结构化数据模式的描述形式中，基于逻辑的描述形式是重要的一类，如一阶逻辑（first-order logic）、描述逻辑（description logic）以及 Datalog 等。它们非常相似，但在表达能力等方面有所差别，这方面比较典型的是基于 Datalog 的模式描述形式。

(2)**基于图的描述形式**：半结构化数据模式的另一种重要形式是基于图的形式。由于半结构化数据一般采用带标记的有向图来表示，所以这种描述形式的一个显著优点是模式和数据采用同一种数据模型（图模型），给处理带来了很大的方便。模式图通常是一个有根，边上带

标记的有向图，其边上的标记可以与数据图相同，也可以加以扩充，如允许类似于“nameladdress”的形式，或采用特定形式的规则（如一元谓词）等。模式图中的结点可以加一定的注释，表明其代表的语义或其他特定的含义。有许多学者还提出了其他形式的模式图，但本质上基本相同，这里就不再一一介绍。

为了更有效地进行XML数据的处理，学者们提出了许多关于其模式描述的方案，最主要的是文档类型定义（DTD）。与半结构化数据的模式相比，DTD作为模式的优点是它的正则语法支持定义半结构数据。

但XML文档置标语言的烙印使DTD无法符合数据库的观点为数据提供非常适合的模式。因此，很多研究采用XML1.0标准提出了更适合表达XML数据模式的方法，如XML-Data，XML-Schema，DCD等。它们的共同点是：(1)要与XML DTD兼容并提供更丰富的描述能力；(2)使用XML 1.0的语法规范定义自身。但是这些规范的研究还不成熟，在很多方面还未达成共识。相比之下，基于DTD规范已被广泛接受。

本章讨论的规范化模式问题就是基于DTD约束的。

1.1.2 研究现状

有关XML数据规范化理论的研究到目前为止已经取得了一些初步成果，但还没有形成统一的规范和完整的理论体系。

规范化理论源于数据库领域。在关系数据库里，键和外键是数据库概念设计的基础，它们提供了如何唯一识别一个元组和如何引用另一个元组的方法。函数依赖理论是关系模式设计的核心部分，是范式研究的基础。基于范式的关系模式，可以消除插入、修改和删除等异常现象，保证数据的完整性和一致性。在关系数据库领域中，对于函数依赖有深入的研究，还包括一些复杂的扩充模型，如嵌套函数依赖和嵌套关系上的范式。此外，在有限的面向对象数据模型上也有关于函数依赖的讨论。它们的研究局限于传统的数据库范畴，XML所特有的树状结构和路径导航的文档模型显然超出了它们的表述能力。而XML数据约束对于研究XML数据的查询优化、数据集成以及XML数据与其他形式数据库的转换都极为重要。

在半结构化数据领域中，由于半结构化数据的树状结构，大多数研究比较集中在路径约束上。由于半结构化模式语言表达能力的缺陷，造成对键的约束表述能力的缺乏。一些更广泛的约束是基于Web的管理进行研究的，显然，这样的研究并不是建立在统一模型上的。

DTD, XML Schema和XML Data提供了XML上基本的约束定义能力。在DTD中，通过为属性标注ID和IDREF提供了定义键和外键的方法。但是这种机制的表述能力相对有限，并存在缺陷。首先，被标注了ID的属性必须在整个文档内唯一，而不是仅仅在某个特定类型的元素内。这样的要求显然局限性太强。在现实中，很多情况下键值都是采用自然数序列来生成的。如果是这样，那么文档中就不能存在两个这样的键了。其次，ID或IDREF只能被标注在单个属性上，这样在关系数据库中很常见的复合键就没有办法在XML中表示了。最后，在DTD中IDREF既没有类