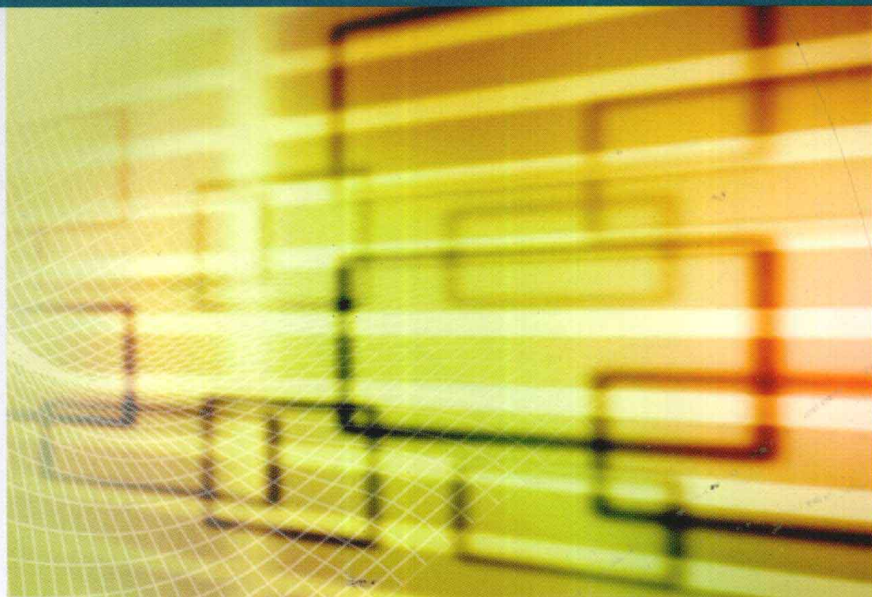


数据库应用理论系列图书

「不完全信息下 XML数据库基础」

郝忠孝 著



科学出版社

数据库应用理论系列图书

不完全信息下XML数据库基础

郝忠孝 著

科学出版社

北 京

内 容 简 介

本书系统论述和分析了不完全信息 XML 数据库和概率 XML 数据库等若干新的技术和理论。

本书共 8 章。主要内容包括:有关 XML 数据库的一些基本概念、基于编码的 XML 数据库存储方法;以语义型不完全信息下 XML、逻辑型不完全信息-概率 XML 数据为主线,讨论了不完全信息下 XML 的强函数依赖推理规则、不完全信息下 XML 数据依赖规范化、存在 XSFD 的 XML Schema 规范化、概率数据模型分析和数据的转换、概率 XML 查询代数系统和 EXQuery、概率 XML 树的结点概率查询算法等。

本书可作为计算机科学与技术学科、数据库领域、Agent 技术、软件设计元素的交换、网络服务领域、EDI、电子商务等领域相关专业的高年级本科生或硕士生选修课教材,也可供从事上述领域研究的博士生、科研人员及工程技术人员等参考。

图书在版编目(CIP)数据

不完全信息下 XML 数据库基础/郝忠孝著. —北京:科学出版社,2011
(数据库应用理论系列图书)
ISBN 978-7-03-031564-9

I. ①不… II. ①郝… III. ①可扩展语言, XML-程序设计
IV. ①TP312

中国版本图书馆 CIP 数据核字(2011)第 113198 号

责任编辑:耿建业 / 责任校对:陈玉凤

责任印制:赵 博 / 封面设计:耕者设计工作室

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>

骏杰印刷厂印刷

科学出版社发行 各地新华书店经销

*

2011 年 6 月第 一 版 开本:B5(720×1000)

2011 年 6 月第一次印刷 印张:13

印数:1—2 500 字数:246 000

定价:55.00 元

(如有印装质量问题,我社负责调换)

作者简介



郝忠孝,教授,山东蓬莱人,1940年12月生,中共党员,曾任原东北重型机械学院副校长,齐齐哈尔大学副校长,哈尔滨理工大学学术委员会主席。现任哈尔滨工业大学博士生导师(兼)、哈尔滨理工大学博士生导师。原机械电子工业部有突出贡献专家、享受国务院政府特殊津贴、全国优秀教师、黑龙江省共享人才专家、黑龙江省级学科带

头人、黑龙江省计算机学会副理事长。

主要研究领域:①空值数据库理论。在国内外首次提出了空值数据库数据模型,完成一系列相关研究,形成了比较完整的理论体系,著有国内外第一部该领域的论著《空值环境下数据库导论》。②数据库 NP-完全问题的求解问题。首次基本解决了求全部候选关键字、主属性,基数为 M 的候选关键字,最小候选关键字等问题,著有《关系数据库数据理论新进展》一书。③数据库数据组织的无环性理论研究。在无 α 环、无 β 环、无 γ 环的分解条件与规范化理论研究方面有了突破性进展,著有《数据库数据组织无环性理论》。④时态数据库理论研究。系统提出并完成了时态数据库中基于全序、偏序、多粒度环境下的各种时态理论问题研究,著有《时态数据库设计理论》的论著。⑤主动数据库理论研究。著有国内外第一部该方面的论著《主动数据库系统理论基础》。⑥空间、时空数据库理论研究。首次解决了空间数据库线段最近邻查询的问题,著有《时空数据库查询与推理》。⑦不完全信息下 XML、概率 XML 数据库理论研究。首次解决了不完全信息下 XML 数据库部分理论研究问题,著有《不完全信息下 XML 数据库基础》。

发表学术论文 230 余篇,其中,国家一级论文 160 余篇、在《计算机研究与发展》上发表个人学术论文专辑两部,被 SCI、EI 等检索 140 余篇。1991 年发表学术论文数居中国科技界第 5 位(并列)。著书 7 部。

前 言

数据库技术和计算机网络技术已成为当今世界计算机应用中两个最重要的基础领域。可扩展标识语言 XML, 目前已经成为 Internet 上信息表示和数据交换的一个重要标准。XML 具有良好的格式以及严格的验证机制, 数据编码为 XML 格式的应用将使信息能够以一种简单可用的方式加以表达, 这些信息的提供者也能非常简便地提供互操作性; 将内容与其表示相分离使得 XML 更能专注于表示数据的语义信息; 用户自定义标签和无限嵌套的能力赋予 XML 无限的可扩展性, 这使得 XML 能够清晰自然地表示各行业的数据。XML 解析器和访问接口实现简单, 各种平台都很容易获取稳定的 XML 解析器和访问接口, 这使得开发基于 XML 的应用变得简单高效。

XML 数据库系统是 Web 技术和数据库技术相结合的产物。XML 有利于信息的表达和结构化组织, 从而使数据搜索更有效。XML 能够使 Web 的维护更方便, 也使 Web 的应用更稳定。

建立数据库系统的目的是模拟现实客观世界, 在现实世界中常常存在大量不完全信息, 所以数据库必须考虑不完全信息的表示和处理, 而且能够表示和处理不完全信息的数据库系统更具有现实意义, 也更具有应用价值。现有的 XML 数据库都不具备这种能力。

概率 XML 数据作为一种特殊的 XML 数据, XML 数据管理理论的建立为概率 XML 数据的管理理论奠定了基础。概率 XML 数据管理理论的研究开始于 2002 年, 研究内容集中在概率 XML 数据中不确定信息的表示方法、概率 XML 代数、概率 XML 数据库查询优化方法、原型系统的设计等几个方面。概率 XML 数据是一种特殊的半结构数据, 先有数据后有模式, 信息的不确定性含义与关系型数据的含义是不同的, 概率关系数据库理论不完全适用于概率 XML 数据的管理。

作者出版此书的目的之一是因为目前国内、外不完全信息环境下 XML 数据库规范化理论有些问题刚刚起步, 远未达到解决实际问题的程度; 概率 XML 数据库理论的研究不系统, 还未达到满意程度; 尚未有有关不完全信息下 XML 数据库理论研究的著作问世, 使作者有抛砖引玉的想法。

另一个目的是作者从事数据库理论研究工作三十余年, 作为一种责任想把研究的一些结果留给年轻同志, 使有兴趣的同行和读者深入到这一领域。

本书共 8 章。为了使读者更好地读懂本书, 作者在第 1 章中概略介绍有关 XML 数据库的一些基本概念; 第 2 章讨论基于编码的 XML 数据库存储方法; 第 3

章讨论不完全信息下 XML 强函数依赖推理规则;第 4 章讨论不完全信息下 XML 数据依赖规范化;第 5 章讨论存在 XSFD 的 XML Schema 规范化;第 6 章讨论概率数据模型分析和数据的转换;第 7 章讨论概率 XML XQuery 和 XML EXQuery;第 8 章讨论概率 XML 树的结点概率查询算法。

本书以语义型不完全信息下 XML、逻辑型不完全信息-概率 XML 数据为主线,力求用通俗易懂的语言全面、系统地进行讨论,做到条理清晰、逻辑性强、易于理解。

本书可作为计算机科学与技术学科、数据库领域、Agent 技术、软件设计元素的交换、网络服务领域、EDI、电子商务等领域相关专业的高年级本科生或硕士生选修课教材,也可供从事上述领域研究的博士生、科研人员及工程技术人员等参考。

本书的出版得到了哈尔滨理工大学领导的鼓励和支持,科学出版社林鹏和耿建业同志的大力支持与帮助,在此表示衷心感谢。

我的学生殷丽凤、王建卫、路燕、张丽平等博士做了大量富有成效的工作。本书所有图表的绘制分别由张丽平、王建卫、郝晓红、李松等完成。在此书出版之际,对他们给予的帮助表示诚挚的感谢。

由于作者水平有限,书中难免会有各种纰漏和疏忽,敬请读者批评指正。

作 者

2011 年 4 月于哈尔滨

目 录

前言

第 1 章 XML 概述	1
1.1 XML 和 DTD 简介	1
1.2 XML 的优势与特点	3
1.2.1 XML 的优势	3
1.2.2 XML 的特点	4
1.2.3 XML 应用概述	5
1.3 基于编码的 XML 数据库存储方法	6
1.4 不完全信息下 XML	7
1.4.1 不完全信息下关系数据库数值理论	7
1.4.2 不完全信息下 XML 数值理论	7
1.5 概率 XML 数据	9
1.6 本章小结	10
第 2 章 基于编码的 XML 数据库存储方法	12
2.1 DTD 完全一致性判定的相关定理和算法	12
2.2 基于编码的 XML 关系数据库存储方法	26
2.2.1 编码方法	28
2.2.2 模式映射	30
2.2.3 路径查询的实现	31
2.3 XML 文档的重组与更新	35
2.3.1 XML 文档的重组	35
2.3.2 XML 文档的更新	38
2.3.3 DTD 的更新	39
2.3.4 递归模式的处理	40
2.4 本章小结	41
第 3 章 不完全信息下 XML 强函数依赖推理规则	42
3.1 基础知识	42
3.1.1 不完全信息的语义分类	42
3.1.2 不完全信息之间的关系	44
3.1.3 不完全信息函数依赖的保持条件	46

3.2	不完全信息下 XML 文档树相关基本概念	46
3.3	不完全信息下 XML 强函数依赖的性质	49
3.4	不完全信息下 XML 强函数依赖推理规则集	52
3.4.1	XSFD 推理规则集的正确性	52
3.4.2	XSFD 推理规则集的完备性	54
3.5	本章小结	55
第 4 章	不完全信息下 XML 数据依赖规范化	56
4.1	不完全信息下 XML 强闭包依赖	56
4.1.1	不完全信息下 XML 强闭包依赖的定义	56
4.1.2	不完全信息下 XML 强闭包依赖的性质	59
4.2	不完全信息下强闭包依赖的推理规则集	60
4.2.1	不完全信息下强闭包依赖推理规则集的正确性	60
4.2.2	不完全信息下强闭包依赖推理规则集的完备性	61
4.3	XSIND 范式	63
4.3.1	XSFD 和 XSIND 之间的关系	63
4.3.2	XSIND 产生数据冗余的原因	64
4.3.3	XSFD 和非循环 XSIND 互不影响的判定条件	64
4.3.4	XSIND 范式的判定条件	69
4.3.5	规范 S 为 XSIND 范式的算法	72
4.4	不完全信息下 XML 强多值依赖	76
4.5	不完全信息 XML 文档树满足 XSMVD 范式的条件	77
4.6	规范不完全信息 XML 文档树为 XSMVD 范式	82
4.7	本章小结	87
第 5 章	存在 XSFD 的 XML Schema 规范化	88
5.1	基本定义	88
5.2	XSFD 的推理规则集	93
5.3	XSFD 的成员籍问题	96
5.4	不完全信息下规范 XML Schema 为 XSFD 范式	99
5.4.1	XML 双类结点强函数依赖范式	99
5.4.2	XML 双类结点强函数依赖规范规则	100
5.4.3	XML Schema 规范化算法	102
5.5	存在 XSMVD 的性质	103
5.6	XSMVD 的推理规则集	106
5.6.1	XSMVD 推理规则集的正确性	106
5.6.2	XSMVD 推理规则集的完备性	109

5.7	XSMVD 弱范式	112
5.8	规范 XML Schema 为 XSMVD 弱范式	115
5.8.1	规范 XML Schema 为 XSMVD 弱范式规范规则	115
5.8.2	规范 XML Schema 为 XSMVD 弱范式算法	118
5.9	本章小结	119
第 6 章	概率数据模型分析和数据的转换	121
6.1	概率 XML 数据管理简述	121
6.1.1	概率 XML 数据模型	121
6.1.2	概率 XML 代数	124
6.1.3	概率 XML 数据查询	125
6.1.4	概率 XML 数据库系统	126
6.2	概率数据模型分析	126
6.2.1	概率关系数据模型	127
6.2.2	概率 XML 数据模型	128
6.2.3	概率 XML 树的路径表达式的类型	129
6.2.4	概率 XML 树的路径表达式的关系	133
6.2.5	概率 XML 树的路径表达式中结点的关系	135
6.3	概率关系数据与概率 XML 数据的转换	135
6.3.1	概率关系模式转换为概率 XML 模式	136
6.3.2	概率 XML 模式转换为概率关系模式的算法	141
6.3.3	概率关系数据转换为概率 XML 数据树的算法	143
6.3.4	概率关系数据转换为概率 XML 数据的算法	146
6.4	本章小结	147
第 7 章	概率 XML XQuery 和 XML EXQuery	149
7.1	以概率 XML 元素单元为操作对象的概率运算	149
7.1.1	叶子结点概率运算	151
7.1.2	概率阈截取运算	152
7.1.3	平凡化运算	152
7.2	概率 XML 树的基本运算	153
7.2.1	并运算	153
7.2.2	交运算	154
7.2.3	差运算	156
7.2.4	连接运算	157
7.3	概率 XML 代数系统的相关特性	160
7.3.1	概率 XML 代数系统的封闭性	160

7.3.2	查询操作有效性和完备性	161
7.4	EXQuery 函数	161
7.4.1	与路径表达式有关的函数	162
7.4.2	与路径表达式转换有关的函数	164
7.4.3	概率计算有关的函数	165
7.4.4	路径表达式的划分有关的函数	166
7.4.5	路径表达式关系判断的有关的函数	168
7.4.6	结点关系判断有关的函数	172
7.4.7	结点概率的计算函数	174
7.4.8	与树类型的判断有关的函数	174
7.5	本章小结	177
第 8 章	概率 XML 树的结点概率查询算法	178
8.1	子树个数的估计算法	178
8.2	基于可能世界原理的结点概率查询算法	182
8.3	基于路径表达式划分的结点概率查询算法	186
8.4	本章小结	192
参考文献		193

第 1 章 XML 概述

1.1 XML 和 DTD 简介

XML(Extensible Markup Language)是标准通用标记语言(SGML)的一种变体,它采用了优化浏览器显示文档视图的严格的规则集合。从 SGML 中经过精心修剪而来的 XML,既保持了 SGML 的功能,又减少了 SGML 的复杂性。与 HTML 不同,XML 允许文档开发人员创建描述数据的标记,并使开发人员可以创建被称为文档类型定义(Document Type Description,DTD)的规则集合。任何标准的 XML 语法分析器都可以读取、解码和检验这种基于文本的自描述文档,并以独立于平台的方式提取数据元素,因此,使应用程序可以通过另一种名为文档对象模型(DOM)的标准访问数据对象。

一个 XML 文档包含了从根元素开始的嵌套元素结构。每个元素有一个标签(Tag)与之相对应,除了字符串数据,它还可以包含属性值对和嵌套子元素。同时包含字符串数据和嵌套子元素的元素被称为具有混合内容。XML 中的元素标签必须良好嵌套,这样的 XML 文档是良构的。

DTD 规定了 XML 文档的结构,可视为 XML 文档的模式。DTD 通过描述子元素和属性的名字及类型来定义 XML 元素的结构。嵌套子元素的出现次数用类似正则表达式的如下符号规定:* (包含零或多个元素的集合),+ (至少包含一个元素的集合),? (包含零或一个元素)和没有修饰(包含且仅包含一个元素)。同时 DTD 也严格规定了 XML 文档中元素的子元素的出现顺序,即按照元素定义的正则表达式顺次展开。元素允许包含至多一个类型为 ID 的属性,由该属性在文档域内唯一标识该元素,同时其他元素也可以通过类型为 IDREF 的属性对其加以引用。IDREF 属性并没有指明该属性所指向的元素类型,在这一意义上说,IDREF 属性是没有类型的。属性定义中可以用 #REQUIRED 和 #IMPLIED 等修饰符类似描述其在出现次数上的约束,同时属性的取值范围也可在 DTD 中定义。一个良构且满足 DTD 的定义的 XML 文档被视为是有效的。

很多学者也提出了各种 XML 存储模型,根据存储方式的不同,处理 XML 数据的基本方式可分为三类:

(1) 文件系统。HTML 的存储管理都是基于这种平面的文件系统。但 XML 和 HTML 有着很大的区别,XML 是层次结构的,而 HTML 不是。主要方法是

XML 数据以 BLOB 的形式进行存储,然后在每一次使用时将它进行解析。这存在着很大缺陷,为了提高查询效率可以建立索引,索引的维护是需要解决的关键问题。

(2) 数据库管理系统。这是一种基于层次数据库的存储管理技术。由于 XML 本身是层次结构的,所以可以将 XML 数据存储于层次数据库中。在层次数据库中,可以利用:①pseudo query 的语法来编写查询语句对数据进行查询。但层次数据库技术非常的不成熟,而且操作非常复杂,因此不是一种很有效的方法。②基于关系数据库的存储管理技术。关系数据库是目前最成熟的数据库技术,可以将 XML 的模式映射到关系数据库中。例如,将 DTD(文档类型定义)映射到关系数据库中,这种基于关系数据库的存储管理技术被应用的非常广泛。③基于面向对象数据库的存储管理技术。面向对象数据库用自身的方法、关系和语义来管理分层 XML 树,同时提供了强大的导航和链接的功能。虽然面向对象数据库的体系结构非常适合存储 XML 数据,但它本身技术还不成熟,从而限制了它的应用。

(3) XML 数据库。XML 数据库是一个 XML 文档的集合,这些文档是持久的并且是可以操作的。XML 文档趋向于面向文档处理或趋向于面向数据处理。面向文档处理的文档是使用 XML 来获取自然语言的那些文档,侧重于给用户信息的最终表示。面向数据处理的文档是 XML 主要用于数据传送的文档,侧重于它被应用程序的使用和交换。

XML 数据是一个面向数据处理的文档的集合。XML 在数据库与应用程序之间,以及在多对应用程序之间进行数据交换是很有用的。在数据交换中,被交换的数据可能需要被保留在一个存档文件中。一个可搜索的 XML 数据库可以作为数据交换的中心部件,并且它也可以与其他使用 XML 的应用程序进行交互,或者获取 Web 服务的摘要信息。

如果数据库把数据保存为 XML,那么在存储粒度上存在几个选择。XML 数据可以被保存为一个文档、被分割为更小的部分并被保存为片段,或者被分解为单独的元素。即使数据不保存为 XML,它也可以作为 XML 被输出,提供给其他应用程序或用户。通过解析文档,并把数据保存在数据库结构(如关系)中。数据可以作为 XML 被添加到数据库中。

XML 数据库比传统数据库更具有表现力,能够对半结构化数据进行有效的存储和管理,提供了对标签名称的操作,并且包括了对路径的操作,显示丰富多样。但 XML 数据库检索速度和修改效率较低,解析手段有缺陷(基于文件解析的 SAX 方式速度慢,而基于内存解析的 DOM 资源消耗较大),且安全性和并发操作机制也是需要重点解决的问题之一。

其中,经过了 40 多年发展的关系数据库,具有管理方便、存储容量小、检索速

度快、修改效率高、安全性好等特点,技术已十分成熟,利用其管理 XML 数据能够保障 XML 数据的可移植性和完整性;另一方面,目前大量的 Web 数据主要存放在关系数据库中,所以,不少的研究人员都致力于 XML 的关系数据库存储方法的研究。

基于关系数据库的 XML 存储方法的研究重点是 XML 文档的模式到关系模式的映射。近几年所提出的模式映射方法,主要分为两类:

(1) 映射产生的关系模式依赖于 XML 模式。它要求预先确定 XML 文档的结构,对 DTD(或 Schema)进行简化、分解等预处理,再将 DTD 中的元素、属性映射成关系数据库中的表或表的属性,即根据 DTD(或 Schema)产生存放 XML 文档数据的一个或几个关系表。这种方法在存储特定的 XML 文档的情况下,具有较高的存储和查询效率,但是,如果需要存放的是海量的、具有不同结构的 XML 数据,如 Internet 上得到的各种数据,则具有很大的局限性:①不同的 DTD(或 Schema)会产生大量的不同模式的关系表;②如果 XML 文档的结构发生了变化,那么相应的关系表必须适应这种变化,修改模式定义,重新对 XML 数据进行存储,更新粒度较大。

(2) 映射方法产生的关系模式独立于 XML 模式。这类方法通常是将 XML 文档解析成树或图结构,然后提供相应的关系模式来储存这些图结构,关系数据库中不但保存了 XML 数据,且同时保存了 XML 文档的结构。能够有效地解决第一类映射方法中出现的部分问题,但对于 XML 路径查询,它通常需要遍历整个数据库,进行多次连接操作才能确定路径的正确性完成查询工作,因而效率相对第一类方法要差些。

1.2 XML 的优势与特点

1.2.1 XML 的优势

可扩展标识语言 XML 由万维网联盟 W3C 于 1998 年 2 月发布为 W3C 的 1.0 版本标准,是一种用于网络信息描述、组织和显示的语言,是 Internet 的“世界语”,目前它已经成为 Internet 上信息表示和数据交换的一个重要标准。它在当今世界的网络信息表示和数据交换上如此重要是因为它具有如下优势:

(1) XML 具有良好的格式以及严格的验证机制,数据编码为 XML 格式的应用将使信息能够以一种简单可用的方式加以表达,这些信息的提供者也能非常简便的提供互操作性。

(2) 将内容与其表示相分离使得 XML 更能专注于表示数据的语义信息。

(3) 用户自定义标签和无限嵌套的能力赋予 XML 无限的可扩展性,这使得

XML 能够清晰自然的表示各行业的数据。

(4) XML 解析器和访问接口实现简单,各种平台都很容易获取稳定的 XML 解析器和访问接口,这使得开发基于 XML 的应用变得简单高效。

鉴于 XML 的上述优势,XML 具有广阔的应用前景。基于 XML 的行业标记语言目前较多,如数学领域的 MathML、化学领域的 CML 以及音乐领域的 MusicML 等;在电子商务等领域中逐渐代换传统的电子数据交换(EDI)技术,成为数据交换的新标准;业界领先的软件提供商提供面向 XML 的工具和产品;众多企业和研究者以 XML 格式发布它们的数据。种种现象表明,XML 数据及面向 XML 数据的应用将持续快速发展,使 XML 迅速成为数据交换的唯一公共语言。

1.2.2 XML 的特点

XML 有以下几个特点。

(1) 开放的标准。

XML 是基于 W3C 定制的开放的标准,从而使得基于 XML 的应用越来越广泛。

(2) 具有良好的格式。

XML 标记大都是成双成对的,极少数不在此列的情况也需要遵守特殊格式定义起始。

(3) XML 支持对文档内容进行验证,或称文档有效性保证。

目前,DTD 和 XML Schema 已经成为 XML 两种主流模式。DTD 是一个专门的文件,用来定义和检验 XML 文档中的标记。XML Schema 用 XML 语法描述,它比 DTD 优越,多个 Schema 可以满足使用 XML 名称空间,可以详细定义元素的内容和属性值的数据类型。

(4) 面向对象的特性。

XML 文件是树状结构,也有属性,这非常符合面向对象的编程,而且也体现出对象方式的存储,XML 是信息的对象化语言。

(5) 选择性更新。

通过 XML,数据可以在选择的局部小范围内更新,提高服务器的性能。

(6) 支持高级搜索。

在 Internet 上若 Web 页是 XML 格式的,不仅可以搜索数据,而且还可以在搜索中加入与数据相关的上下文信息,形成更精确的搜索机制。

(7) XML 是电子数据交换的格式。

XML 是为互联网的数据交换而设计的,能够应用于电子商务、电子政务等各个领域的数据交换。

(8) XML 具有一套完整的方案。

XML 是一个技术大家族,有一系列相关技术,包括文件数据验证、显示输出、文件转换、文档对象和链接等。XML 文档是对象实例,DTD 和 Schema 是界面或类,XSL 是方法和实现,XML 的资源描述框架是信息导航、浏览、搜索的用户接口标准。

1.2.3 XML 应用概述

XML 是一个网络信息描述、组织和显示的语言,它的开放性、严谨性、结构性和灵活性备受网络开发者的青睐。在以下领域,XML 有广阔的应用前景。

1) 数据库领域

关系型数据库行业的三大巨头——Oracle 公司、IBM 公司和 Microsoft 公司都在它们的数据库产品中提供了对 XML 的支持。从本质来看,XML 文档就是数据库,是数据的集合,每个文件都含有某种类型的数据。在许多方面它和其他文件没什么区别,但作为一种“数据库”格式,XML 有一些优点,例如,它是可交换的、自描述的,能够以树状或图形结构描述数据。

2) Agent

若把 XML 结构化的数据送到 Agent(智能体),Agent 就能很容易地理解这些数据的含义,很容易理解这些数据与已有知识的关系。基于 XML 的数据交换对于解决 Agent 的交互性问题起着重要的作用,XML 及相关技术必将促进 Agent 技术的发展。

3) 软件设计元素的交换

XML 能够描述软件设计中有关的设计元素,如对象模型,甚至能够描述最终设计出来的软件。这些基于 XML 的设计元素可以借助 Web 在开发组内进行交换,在不同的开发工具之间交换。

4) 网络服务领域

XML 有利于信息的表达和结构化组织,从而使数据搜索更有效。XML 能够使用 URL 别名使 Web 的维护更方便,也使 Web 的应用更稳定;XML 还能够使用数字签名使 Web 的应用更广阔。信息发布在企业的竞争发展中起着重要的作用,服务器只需发出一份 XML 文档,客户就可根据自己的需求选择和制作不同的应用程序以处理数据。加上 XSL 的帮助,能够实现广泛的、通用的分布式计算。

5) EDI

传统的 EDI(电子数据交换)标准缺乏可扩展性和灵活性。使用 XML,程序能够理解在交换数据中所表示的商务数据及概念,根据明确的商务规则来进行数据处理。

6) 电子商务领域

XML 的丰富置标能够描述不同类型的单据,如保险单、信用证、索赔单以及

各种发票等。结构化的 XML 文档发送至 Web 的数据可以被加密,并且很容易附加上数字签名。

1.3 基于编码的 XML 数据库存储方法

由于逐渐广泛的应用,XML 正逐步将 Web 转化为一个巨大的数据库,但将数据表示为 XML 只是对 Web 这一巨大数据库进行有效处理的第一步。如何对大量的 XML 数据进行有效的管理已经成为数据库研究领域的一个重要的研究问题。同时,对传统数据库管理技术提出了巨大的挑战,也为数据库研究提供了良好的机遇。

目前大多数商业化的关系数据库支持 XML,关系数据库和对象关系数据库系统也提供了各种 XML 数据的扩展和插件。同时,为 XML 数据管理量身打造的 XML 原生数据库系统也不断涌现,也出现了 XML 数据管理的标准。XML 数据库系统是 Web 技术和数据库技术相结合的产物。很多学者对 XML 数据库管理技术进行了系统深入的研究,取得了很多有意义的理论研究成果,XML 数据库的规范化理论是其中研究成果之一。

XML 作为一种数据交换的标准和半结构化数据的表示模型,人们关注的问题是:用 XML 表示的数据之间有什么联系以及有什么约束? 因此,研究 XML 文档中普遍存在的关键字约束问题就非常重要,随着 XML 成为 Web 上一种通用的数据格式,将关键字(简称为键)的概念引入到 XML 领域。键也在不变的数据上构成了一个重要的约束类,特别的,键与地址或者内部对象标识符相比,它在现实世界中的一个对象与它在数据库所表示的事物之间提供了一个不变连接。这种连接在更改数据库作为世界模型的改变中是非常关键的。随着 XML 越来越多地应用在数据管理中,很自然就需要一个基于值的方法来确定一个 XML 文档中元素的位置。XML 数据通常也是来自数据库,而在数据库中关键字通常被用来传达数据的语义的最本质的部分,为了使用 XML 来表示当前驻留在数据库中的数据,我们必须能够表示数据的原始语义,特别是键。键对索引、XML 数据的存档以及设计关系存储都是不可或缺的。为了获得更加有效的应用,键所给出的约束和结构经常在这样的系统中得到开发。XML 关键字在 XML Query 的数据模型中也起到了非常重要的作用。因为索引通常建立在键之上,键为数据的结构提供了信息,所以,XML 键在研究队列优化的过程中也是非常重要的。这也是进行其他 XML 相关研究的基础,如 XML 模式的规范化、查询优化、结构优化、路径约束等。

目前常见的 XML 文档模式定义规范主要包括 DTD 和 XML Schema。这些模式定义方法对数据约束提供了有限支持,其所能表述的内容主要集中在键和外

键上,其他方面的数据约束没有给予清晰完整的表达。

DTD 定义了 XML 文档的语法和整体结构,具有简洁的文档规则、参数实体和拥有众多的可用工具等优点。

遵守 XML 语法规则并遵守相应 DTD 文件规范的 XML 文档,称为有效的 XML 文档。一个语法正确的 DTD 是一致的,当且仅当至少存在一个有效的 XML 文档满足该 DTD;如果不存在任何有效的 XML 文档满足该 DTD,则该 DTD 是不一致的。不一致的 DTD 没有实际用途,应当尽可能地避免。DTD 的一致性已经成为 XML 研究中的一个重要问题。

1.4 不完全信息下 XML

1.4.1 不完全信息下关系数据库数值理论

在客观世界中,常存在:某些学生因故缺考而暂时无成绩;历史档案中有时出现“生日不详”;会议发言人有时会宣布议事日程“有待公布”;警方记录常出现“嫌疑人下落不明”等信息。这些不确定的信息在经典数据库中是无法描述和处理的,这些信息称为不完全信息。

客观世界中的不完全信息有很多语义,为了对不完全信息的描述更接近现实世界,本书作者在文献[1]中,根据不完全信息本身所固有的重要语义信息(即不完全信息是否可用实值来取代该不完全信息的完全信息)把不完全信息分成了三大类:不存在型不完全信息、存在型不完全信息以及占位型不完全信息。不存在型不完全信息表示在数据库中不存在任何实值,但应该给出某种表征指示信息不存在。存在型不完全信息表示信息的真正缺失,体现在取值当前未知,只是取值范围已知。它与不存在型不完全信息之间的根本区别在于它将来可被一个已知实值替代,且该实值一定是语义范围内的常规值。占位型不完全信息是最不确定的一类,可能是不存在型不完全信息,也可能是存在型不完全信息,要随着时间的推移才能知道,是最不确定的一类。

由于存在型不完全信息在客观世界中存在的比例最大,把该类不完全信息引入关系模型中,对关系模型进行了扩展,研究不完全信息下关系数据库的规范化理论。

1.4.2 不完全信息下 XML 数值理论

建立数据库系统的目的是模拟客观世界,又由于客观世界存在大量不完全信息,所以数据库必须考虑不完全信息的表示和处理,而且能够表示和处理不完全信息的数据库系统更具有现实意义,也更具有应用价值。XML 语言本身能够更好