

XML Database Schema Normalization

XML数据库模式规范化

曹丽君 张忠平 著



科学出版社

XML 数据库模式规范化

曹丽君 张忠平 著

科学出版社

北 京

内 容 简 介

本书采用路径表达式和树元组的表示方法分析了 XML 数据库模式的设计和 XML 数据规范化设计,增强了 XML 数据的语义表现力,完善了 XML 数据库规范化理论,从基本定义及符号、XML 函数依赖、XML 范式及文档规范化、XML 键约束、XML 多值依赖、XML 函数依赖和多值依赖的成员籍判定、XML 函数依赖和多值依赖并存下的范式及其规范化、XML 文档相似性度量、基于共享路径处理小枝模式的索引方法、基于 ISP 索引的小枝模式处理方法、基于小枝模式查询的灵活匹配等方面进行了深入研究。

本书具有定位准确、取舍合理、体系得当等特点,可作为高等院校计算机专业本科生、研究生 XML 数据库技术课程的参考教材或课外读物,也适合数据库技术人员和管理人员、科研人员等相关专业学者对 XML 数据库模式规范技术的学习和研究。

图书在版编目(CIP)数据

XML 数据库模式规范化/曹丽君,张忠平著. —北京:科学出版社,2015
ISBN 978-7-03-045316-7

I. ①X… II. ①曹…②张… III. ①可扩展语言-数据库系统-研究
IV. ①TP312

中国版本图书馆 CIP 数据核字(2015)第 181790 号

责任编辑:陈晓萍 / 责任校对:王万红
责任印制:吕春珉 / 封面设计:子时文化

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>

厚 诚 则 铭 印刷

科学出版社发行 各地新华书店经销

*

2015 年 7 月第 一 版 开本: B5 (720×1000)

2015 年 7 月第一次印刷 印张: 11 3/4

字数: 250 000

定价: 56.00 元

(如有印装质量问题,我社负责调换〈厚诚则铭〉)

销售部电话 010-62134988 编辑部电话 010-62138978-2009

版权所有,侵权必究

举报电话: 010-64030229; 010-64034315; 13501151303

前 言

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 数据在大幅度增加,这就对 XML 数据库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。虽然 XML 数据的相关技术研究已有些成果,如 XML 数据的存储与发布技术、XML 数据查询与优化技术等,特别是与关系数据的转换技术已相当成熟。但是,XML 数据已经成为 Internet 上的主流数据,如果仅仅考虑到如何从 XML 文档数据转换到关系数据,且这种转换只是保持了结构信息,而没有从数据库设计的角度来评价 XML 数据库模式,这也必将对以后的 Web 数据处理带来很大的麻烦,势必造成数据的大量冗余和不一致现象。本书从数据库设计的角度出发,对 XML 数据的约束进行深入的研究,直接对 Web 数据进行规范化处理,从而得到良好的 XML 数据库模式,这样不仅完整地保留了 XML 文档数据中的语义和结构信息,满足了数据库设计的要求,一次性地完成了 XML 数据库的设计,避免了现有方法的重复设计,而且减少了数据冗余,保持了 Web 上数据的一致性。因此,XML 数据库模式规范化的研究具有重要的理论意义和实用价值。

本书的主要工作是在已有的 DTD 和 XML-Schema 等规范基础上,采用路径表达式和树元组的表示方法对 XML 数据进行规范化设计与论述。本书共 12 章,分别为绪论、基本定义及符号、XML 函数依赖、XML 范式及文档规范化、XML 键约束、XML 多值依赖、XML 函数依赖和多值依赖的成员籍判定、XML 函数依赖和多值依赖并存下的范式及其规范化、XML 文档相似性度量、基于共享路径处理小枝模式的索引方法、基于 ISP 索引的小枝模式处理方法、基于小枝模式查询的灵活匹配。

第 1 章为绪论。阐述了 XML 数据库模式规范化的背景,分析了国内外 XML 数据库模式规范化的研究现状,论述了 XML 数据库模式规范化的主要内容及理论和实际意义。

第 2 章为基本定义及符号。介绍了形式化的 DTD、XML 文档和路径表

达式等定义, 定义了 XML 树元组、结点值相等、树包含等相关符号概念, 并结合实例对这些定义和符号进行了说明。

第 3 章为 XML 函数依赖。本章基于路径表达式和树元组给出了 XML 函数依赖的形式化定义, 并在此基础上给出了 XML 函数依赖的推理规则和 XML 路径集闭包的概念, 并对 XML 函数依赖推理规则的有效性和完备性进行了证明。接着, 在 XML 函数依赖和推理规则的基础上引入了 XML 函数依赖的逻辑蕴涵与覆盖的概念, 包括无冗余覆盖、规范覆盖以及最小覆盖等, 给出了相关的覆盖的求解算法, 并对算法的正确性、可终止性和时间复杂性进行了分析和证明。最后, 给出 FD_{XML} 集的最优覆盖概念。

第 4 章为 XML 范式及文档规范化。基于 XML 函数依赖形式化定义, 给出 XML 范式的定义。提出 XML 文档规范化规则——元素提升规则和元素创建规则, 并在此基础上给出 XML 文档规范化算法。对算法的正确性、可终止性和时间复杂性进行证明分析, 并实验证明了规范化后的文档在查询时间和存储空间效率上都有明显的改善。

第 5 章为 XML 键约束。本章在分析已有半结构化数据键定义的优劣基础上, 结合 FD_{XML} 定义, 首先给出了基于路径表达式的 XML 键定义。该定义支持多个元素和属性, 表达绝对和相对的现实语义约束, 不受限于 XML 文档的任何类型的规范。其次, 给出了绝对键和相对键的推理规则, 并对其有效性和完备性进行了证明。最后, 基于 XML 键及其相关概念和推理规则, 给出多项式时间求解 FD_{XML} 集的一个候选键的算法, 并对算法的正确性、可终止性和时间复杂性进行了证明和分析。

第 6 章为 XML 多值依赖。本章针对 XML 文档中的多值依赖问题, 分析了 XML 文档中由多值依赖而引起的数据冗余和各种操作异常现象, 包括插入异常、删除异常、更新异常等。给出了有关 XML 多值依赖的一些基本概念, 包括 MVD_{XML} 的逻辑蕴涵、等价与覆盖, 路径集闭包, 无冗余覆盖, 简化的 MVD_{XML} 等。基于基本的概念, 给出了 XML 多值依赖有效和完备的推理规则, 并对有效性和完备性进行了证明。同时, 给出了相应问题的求解算法, 包括 MVD_{XML} 成员籍判定算法, 左、右部简化的 MVD_{XML} 集算法, 无冗余 MVD_{XML} 集检验算法以及求解 MVD_{XML} 无冗余覆盖算法。上述算法从正确性、可终止性和时间复杂性上给出了分析证明。XML 文档中的多值依赖问题的研究更好地表达了现实世界中实体的一对多语义约束关系。

第7章为XML函数依赖和多值依赖的成员籍判定。本章针对XML函数依赖和多值依赖并存情况下的成员籍问题。基于成员籍问题的描述和讨论，给出了求解路径依赖基、成员籍和最小依赖集的算法，并对这三个算法的可终止性、正确性和时间复杂度进行分析和证明。

第8章为XML函数依赖和多值依赖并存下的范式及其规范化。对XML中的冗余和键进行了描述，给出了3XNF和4XNF定义，在此基础上给出了规范化规则和无冗余判定定理，进一步提出了XML文档规范化算法，并对算法的可终止性、正确性及时间复杂度进行分析和证明，最后通过实验证明该算法的有效性。

第9章为XML文档相似性度量。本章着重讨论了XML文档基于路径集合和代价的相似性度量，这对XML文档相似性度量，聚类XML文档树编辑距离度量提供了更丰富的度量方法，给出了规范化XML文档相似性度量方法——集合度量方法、线性度量方法和代价度量方法，并提出基于权重代价的机器学习的相似性度量算法。实验证明该算法扩展了XML文档查询范围，提高了文档的查全率和查准率。

第10章为基于共享路径处理小枝模式的索引方法。论述了小枝模式查询处理和索引技术的相关研究现状，指出了现有小枝模式查询处理方法中存在的 key 问题，结合已经存在的索引思想，给出了一种基于共享路径的索引ISP，利用该索引技术可以高效地处理小枝模式查询。

第11章为基于ISP索引的小枝模式处理方法。对基于ISP索引的小枝模式处理方法进行了比较深入的分析。首先分析了小枝模式处理的过程，针对其处理过程为每一步骤提出了详细的处理方法或算法，并对基于ISP索引的小枝模式处理算法进行了实验分析。

第12章为基于小枝模式查询的灵活匹配。分析了小枝模式匹配的国内外研究现状和现有匹配方法的不足，阐述了小枝模式匹配中的基本概念，提出了基于小枝模式匹配的灵活匹配方法，并定义了相关匹配条件，并通过具体的实例对所提出的方法进行了性能分析和说明。

本书由河北科技师范学院学术著作出版基金资助，同时也是数学与信息科技学院网络工程专业改革试点阶段性成果之一，由河北科技师范学院曹丽君统稿并编写第1~6章和第10~12章（共20万字），燕山大学张忠平统稿并编写第7~9章（共5万字），本书在编写过程中也得到了河北科技师范学

院王海明、马国光、赵立强、李玉香、李密生的帮助与支持，在此一并表示深深的感谢与敬意。

对于本书的编写，作者尽量体现科研思维和成果，虽竭尽全力，但限于能力和水平，书中难免存在疏漏和错误之处，希望广大读者批评指正。

曹丽君

于河北科技师范学院

张忠平

于燕山大学

目 录

前言

第 1 章 绪论	1
1.1 背景	1
1.2 国内外现状	8
1.3 内容及意义	13
1.3.1 XML 规范化的内容	13
1.3.2 理论和实际意义	15
第 2 章 基本定义及符号	17
2.1 XML 简介	17
2.1.1 XML 与标签	17
2.1.2 XML 特性	19
2.1.3 XML 数据库	20
2.1.4 XML 约束	21
2.2 DTD	27
2.3 XML 树	28
2.4 结点值相等	30
2.5 其他定义与符号	30
2.6 小结	31
第 3 章 XML 函数依赖	32
3.1 XML 函数依赖定义	32
3.2 XML 函数依赖蕴涵问题	36
3.3 XML 函数依赖推理规则	37
3.3.1 推理规则正确性	38
3.3.2 推理规则完备性	39
3.3.3 推理规则的应用	41
3.4 XML 函数依赖集的覆盖问题	43
3.4.1 等价与覆盖	43

3.4.2	XML 函数依赖集的非冗余覆盖	44
3.4.3	左部路径冗余与规范覆盖集	46
3.4.4	XML 函数依赖集的最小覆盖	48
3.5	小结	49
第 4 章	XML 范式及文档规范化	50
4.1	XML 范式	50
4.1.1	XML 范式定义	51
4.1.2	XML 范式级别	51
4.1.3	模式分解	51
4.2	规范化规则	52
4.2.1	元素提升规则	52
4.2.2	元素创建规则	53
4.3	规范化算法	55
4.3.1	无损连接算法	55
4.3.2	算法和实验分析	56
4.4	小结	59
第 5 章	XML 键约束	60
5.1	XML 键的定义	60
5.2	XML 键的推理规则	64
5.2.1	XML 绝对键的推理规则	64
5.2.2	XML 相对键的推理规则	66
5.3	XML 候选键求解算法	67
5.4	小结	71
第 6 章	XML 多值依赖	72
6.1	XML 多值依赖定义	72
6.2	XML 多值依赖推理规则	76
6.2.1	推理规则的有效性	78
6.2.2	推理规则的完备性	81
6.3	XML 多值依赖的简化	83
6.4	XML 多值依赖的蕴涵和覆盖	84
6.4.1	XML 多值依赖的蕴涵	84

6.4.2 XML 多值依赖的覆盖	90
6.5 小结	93
第 7 章 XML 函数依赖和多值依赖的成员籍判定	94
7.1 成员籍	94
7.2 求解路径依赖基算法	96
7.2.1 算法描述	96
7.2.2 算法分析	97
7.3 成员籍判定	100
7.3.1 算法描述	101
7.3.2 算法分析	101
7.4 最小依赖集	102
7.4.1 算法描述	102
7.4.2 算法分析	103
7.5 小结	104
第 8 章 XML 函数依赖和多值依赖并存下的范式及其规范化	105
8.1 引言	105
8.2 XML 函数依赖下的范式	105
8.2.1 有效变化和冗余	105
8.2.2 XML 第三范式	109
8.2.3 XML 函数依赖规范化设计和算法	110
8.3 XML 函数依赖和多值依赖并存下的范式	114
8.3.1 第四范式	114
8.3.2 XML 多值依赖规范化设计和算法	118
8.4 实验分析	122
8.4.1 实验设置	122
8.4.2 实验结果及性能分析	122
8.5 小结	125
第 9 章 XML 文档相似性度量	126
9.1 相关工作	126
9.2 基本定义	127
9.3 XML 相似性度量方法	128

9.3.1	集合度量方法	128
9.3.2	线性度量方法	130
9.3.3	代价度量方法	131
9.4	XML 文档相似性度量算法	134
9.4.1	基于权重代价的度量算法	134
9.4.2	算法分析	136
9.5	小结	137
第 10 章	基于共享路径处理小枝模式的索引方法	138
10.1	引言	138
10.2	基本概念	140
10.2.1	共享路径	140
10.2.2	模式匹配	140
10.2.3	编码方式	141
10.2.4	XML 模型	142
10.2.5	索引简介	142
10.3	关键问题	146
10.4	ISP 索引	147
10.4.1	ISP 构建	147
10.4.2	ISP 结构	148
10.4.3	ISP 构造算法	149
10.5	小结	150
第 11 章	基于 ISP 索引的小枝模式处理方法	151
11.1	小枝模式查询处理过程	151
11.2	小枝模式查询预处理	152
11.2.1	小枝模式预处理思想	152
11.2.2	小枝模式预处理算法	153
11.3	模式树预匹配	155
11.3.1	匹配思想	155
11.3.2	匹配算法	155
11.4	简化模式树匹配	157
11.4.1	简化模式树思想	157

11.4.2 简化模式树匹配思想	158
11.4.3 多合并匹配算法	159
11.5 实验	161
11.5.1 实验环境设置	161
11.5.2 实验数据设置	161
11.5.3 实验结果及性能分析	163
11.6 小结	166
第 12 章 基于小枝模式查询的灵活匹配	167
12.1 引言	167
12.2 相关工作	167
12.3 基本概念	168
12.4 灵活匹配方法	170
12.4.1 灵活匹配方法的主要思想	170
12.4.2 灵活匹配方法的实例分析	171
12.4.3 性能分析	172
12.5 小结	173
参考文献	174

第 1 章 绪 论

1.1 背 景

互联网（Internet）已经成为新经济时代的标志，它极大地影响了人类的生活方式、商业模式，并将继续对人类社会的进步起着巨大的推动作用。2015 年 2 月 3 日，中国互联网络信息中心（CNNIC）发布了《第 35 次中国互联网络发展状况统计报告》，报告显示：

截至 2014 年 12 月，我国域名总数为 2060 万个，年增长 11.7%。其中“.CN”域名总数为 1109 万，年增长 2.4%，占中国域名总数比例为 53.8%；“.COM”域名数量为 795 万，占比为 38.6%；“.中国”域名总数达到 28.5 万。中国网站数量为 335 万个，年增长 4.6%。中国网页数量为 1899 亿个，年增长 26.6%。其中，静态网页数量为 1127 亿，占网页总数的 59.36%；动态网页数量为 772 亿，占网页总量的 40.64%。我国 IPv4 地址数量为 3.32 亿，拥有 IPv6 地址 18797 块/32，年增长 12.8%。国际出口带宽为 4118663Mb/s，年增长 20.9%。

2014 年，台式计算机、笔记本电脑等传统上网设备的使用率保持平稳，移动上网设备的使用率进一步增长，新兴家庭娱乐终端网络电视的使用率达到一定比例。通过台式计算机和笔记本电脑接入互联网的比例分别为 70.8%、43.2%，与 2013 年底基本持平；通过手机接入互联网的比例继续增高，较 2013 年底提高 4.8%；平板电脑的娱乐性和便捷性特点使其成为网民的重要娱乐设备，2014 年底使用率达到 34.8%，并在高学历（本科及以上学历网民使用率 51.0%）、高收入人群（月收入 5000 元以上网民使用率 43.0%）中拥有更高使用率；随着网络技术和宽带技术的发展，网络电视融传统电视和网络为一身，其共享性、智能性和可控性迎合现代家庭娱乐需求，逐渐成为一种新兴的家庭娱乐模式，截至 2014 年 12 月，网络电视使用率已达到 15.6%。

作为互联网最主要应用的 Web 正成为整个世界的窗口，它实现了全球用户和机构信息的共享。Web 已经成为人类社会的主要信息源、媒体和商务的门户，互联网已经成为信息传播的介质和商务运作的基础平台。

随着互联网时代的到来，数据越来越多地开始以网络在线的方式进行发布、交换和集成。1998 年 2 月，万维网协会 W3C(World Wide Web Consortium)推出了可扩展的标记语言 XML(eXtensible Markup Language)，将它作为一种互联网进行数据表示和交换的标准。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、媒体产业正制定各自行业的 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 的制定，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？表 1-1 中列出了常见的几种 XML 模式语言。

表1-1 常见的几种XML模式语言特性比较

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

续表

特性	DTD1.0	XML-Schema1.0	XDR 1.0	SOX 2.0	Schematron 1.4	DSD 1.0
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
元素						
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
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”角度看，XML 模式语言可以从多个方面来划分，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。DTD, XML-Schema, XDR 和 SOX 属于基于语法分组, 而 Schematron 属于基于模式的分组。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更好地描述约束。

从“数据库”角度看, 没有一种语言彻底地满足需要。数据定义语言 (Data Definition Language, SQL DDL) 描述的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。

由于 XML 作为 Internet 上数据标准的出现, XML 数据的相关研究成为热点。如图 1-1 所示, 从现实世界中客观事物到其在计算机中的具体表示, 实际上经历了三个领域——现实世界、信息世界和机器世界。存在于人们头脑之外的客观世界, 称为现实世界。信息世界是现实世界在人脑中的反映, 人们把它用文字和符号记载下来。信息世界的信息以某种数据形式存储在机器世界里。本书认为 Web 世界应当与机器世界是同一层次, 即它们都位于信息世界下。信息世界的信息被分别存储在这两个世界中。关系模型和面向对象模型都是机器世界中的数据模型。与之相似, XML 则应当被看成 Web 世界中的数据模型, DTD 则应当被看成 XML 数据的模式。

现有的 XML 存储方法, 如图 1-1 所示, 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麻烦, 因为这样的数据库可能存在着插入、删除和更新异常。为了避免这些异常的出现, 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

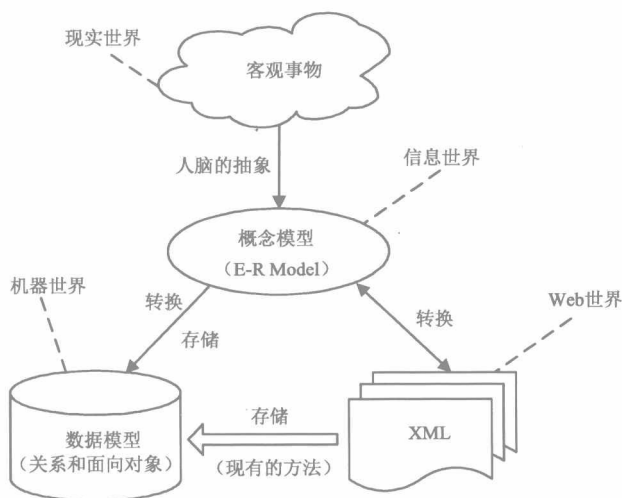


图 1-1 信息的转化过程

到一个优秀的数据库却是非常复杂的。本书的思想不是从 Web 世界到机器世界直接的转换，而是着眼于数据库设计，对 Web 世界中的模式（即 DTD）数据约束进行规范化研究，直接进行 XML 数据库模式的设计，这样不仅完整地保留了 XML 文档中的语义和结构信息，而且满足了数据库设计的要求，一次性地完成了一个优秀数据库的设计，避免了现有方法的数据库重复设计。

随着 XML 应用范围的拓广，为了提高数据质量，作为 XML 文档的主要模式定义方法，一个设计良好的 DTD 是必须的。那么，如何设计良好的 DTD？进一步地，如何将设计不好的 DTD 转化为相应的设计良好的形式就变得很重要。可以从不同的角度来判断 DTD 是否设计良好，一个重要的考虑因素——数据冗余。这是因为，和关系数据库一样，数据冗余往往是导致操作异常产生的根源。数据被重复地多次存储，必然使相应的插入、删除和更新操作变得困难。如果 XML 的模式设计得不好，在 XML 文档中同样也会造成数据冗余和操作异常现象。这必将造成 XML 文档中数据存储空间的浪费，查询优化也会受到影响。对于 XML 文档操作产生的各种异常，直接影响 XML 作为不同应用程序之间数据表示和交换标准的应用。由于 Internet 的开放性，XML 数据操作异常的危害性要远远大于关系数据库操作异常的危害性，这就对 XML 数据库模式提出了更高的要求。

XML 是半结构化数据的特例。半结构化数据是介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