连续不确定XML数据管理关键技术

张晓琳　刘立新　编著

科学出版社

连续不确定XML数据管理关键技术

张晓琳　刘立新　编著

科学出版社

北京

内 容 简 介

本书详细介绍作者在连续不确定 XML 数据管理技术领域的最新研究成果，主要内容包括连续不确定 XML 数据模型、连续不确定 XML 数据索引技术、连续不确定 XML 数据编码技术、连续不确定 XML 数据查询、连续不确定 XML Top-k 查询、不确定 XML 关键字查询技术等。

本书可作为计算机专业高年级本科生、研究生的教材，也可供相关科研人员参考使用，有助于读者了解连续不确定 XML 数据管理的相关技术。

图书在版编目 (CIP) 数据

连续不确定 XML 数据管理关键技术/张晓琳，刘立新编著. —北京：科学出版社，2015.5

ISBN 978-7-03-044255-0

Ⅰ. ①连… Ⅱ. ①张…②刘… Ⅲ. ①可扩充语言－程序设计 Ⅳ. ①TP312

中国版本图书馆 CIP 数据核字(2015)第 096481 号

策划编辑：陈 静 / 责任编辑：陈 静 王 苏 / 责任校对：桂伟利

责任印制：张 倩 / 封面设计：迷底书装

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

文林印务有限公司 印刷

科学出版社发行 各地新华书店经销

*

2015 年 5 月第 一 版 开本：720×1 000 1/16

2015 年 5 月第一次印刷 印张：8 3/4

字数：176 400

定价：49.00 元

(如有印装质量问题，我社负责调换)

前　言

传统数据库系统主要针对确定性数据进行管理。随着数据采集和数据处理技术的不断发展，不确定性数据受到广泛关注。不确定性数据包含离散性不确定数据和连续性不确定数据。连续性不确定数据是指数据存在与否未知，而且各个属性存在误差，这种不确定性可以用一个连续分布函数来表示，并普遍存在于军事、电信、经济、物流等领域中。

XML 又称可扩展的标记语言，具有自描述性好、可扩展性高和灵活性好的特点，特别适用于不确定性数据管理。目前，基于 XML 管理不确定数据大多针对离散不确定 XML，然而很多情况下，不确定性数据是连续分布的，如何对连续不确定 XML 数据进行管理具有重要的意义。

连续不确定 XML 数据管理技术研究涉及统计理论、概率理论、数据库技术、网络技术等多个领域，是具有挑战性的研究课题。对其进行理论和方法的研究可以为传感器数据、科学数据和地理学信息等应用提供理论基础和技术支持，具有重要的理论和应用研究价值。

本书内容主要包括：数据模型、数据编码、索引技术、小枝查询、复杂小枝查询、多维数据查询、Top-k 查询、关键字查询等。

本书按照连续不确定 XML 数据管理的内容，共 6 章。

第 1 章为绪论。主要介绍不确定数据的来源及其广泛应用，并论述 XML 表示不确定性数据的优势；概括不确定 XML 数据管理技术的相关研究以及未来发展。

第 2 章为多维连续不确定 XML 数据模型。介绍连续不确定 XML 数据模型支持多维连续随机变量的不确定 XML 数据表示，并能够有效地表示一维及多维连续分布函数，包括二维高斯及均匀分布等标准分布函数，高效地处理不确定 XML 中的多维连续数据。

第 3 章为连续不确定 XML 数据索引。通过重复利用小素数编码并根据码值之间素数因子的包含关系能够快速、精准地判定出 XML 树中任意两个节点的结构关系，并且该编码方式可以支持连续不确定 XML 文档更新；提出连续不确定 XML 数据的节点编码索引技术和结构概要索引技术。

第 4 章为连续不确定 XML 数据查询。在小素数编码的基础上，提出一种非归并不确定 XML 小枝模式查询；包含通配符和复杂谓词的不确定 XML 复杂小枝查询；基于序列的不确定 XML 查询；基于蒙特卡罗思想多维连续不确定 XML 数据查询，以及基于最小二乘法的连续不确定 XML 数据同步多区间查询。

第 5 章为连续不确定 XML 数据 Top-k 查询。扩展 PEDewey 编码支持连续分布类型

节点的编码，定义连续不确定 XML 数据查询结果的概率值计算公式；提出并实现高效的连续不确定 XML 数据 Top-k 查询算法，并设计过滤策略进一步提高算法的效率。

第 6 章为不确定 XML 关键字查询。首先基于 SLCA 语义，设计了动态 Keyword 数据仓，并基于动态 Keyword 数据仓设计一种求解不确定 XML 关键字查询的算法。然后，提出一种基于最小相关联通子树的 Top-k 语义，根据此语义提出了一种不确定 XML 关键字查询算法，且根据不同的查询条件设计不同的过滤策略来进一步提高查询效率。

在此，首先要感谢作者的导师王国仁教授带作者进入数据库研究领域，他的指导使作者获益良多；其次要感谢本研究团队的老师和研究生，本书凝聚了大家的辛苦付出；最后还要特别感谢实验室的研究生：吕庆、郑春红、霍伟、崔光月、张换香、苏龙超、韩雨童、王鹏、郭丹丹、郝琨。

在本书的撰写和课题的研究过程中，尽管作者投入了大量的时间和精力，但是受知识水平所限，书中难免有不当之处，恳请读者指正并不吝赐教。如果有任何建议和意见，可以发送电子邮件至 zhxl@imust.cn。

张晓琳

2014 年 12 月

此为试读，需要完整PDF请访问：www.ertongbook.com

目　录

第1章 绪 论

1.1 连续不确定 XML 数据

1.1.1 不确定性数据的产生与应用

近年来，传统的确定性数据管理技术已经有了很大的提高，与其有关的数据库产业也有了长足的发展。作为社会基础设施建设的一项重要支撑，数据库系统和数据库相关技术得到了很大的提高。在传统数据库系统中，存储和处理的数据都具有确定性，随着处理数据和采集数据技术的不断发展，不确定性数据受到广泛关注。不确定性数据包含离散的不确定和连续的不确定两种类型。在军事、电信、经济及物流等领域，数据的连续不确定性普遍存在。所谓连续不确定性是指数据的存在性未知，而且各个属性存在误差，这种不确定性可以用一个连续分布的函数来表示。不确定性数据的高效管理不能通过传统的数据库管理技术来实现，这样，就引发了工业界和学术界对新型不确定性数据管理技术研究的高度热情[1]。

产生不确定性数据的原因很多，也很复杂。有可能是最原始的数据本身具有不准确性，也有可能是数据集成过程中造成的不确定性或是采用了不同的数据集合方法造成的不确定性[2]。

（1）从一种数据集合转向另外一种数据集合的过程中会引入不确定性。例如，使用粗粒度数据集合转换到细粒度集合的过程中，假设某地的人口分布数据库是以乡为单位记录的，而在应用中却要求以村为单位进行查询，这样，查询结果就存在不确定性。

（2）产生不确定性数据最直接的原因是原始数据不准确。例如，实验室的物理仪器受到仪器自身精确度的影响导致采集的数据有一定的误差；网络传输过程中，特别是在无线传输过程中，由于受到能量、传输时延和带宽等因素的影响而造成数据的不确定性；在传感器领域[3]和无线射频领域[4-5]，周围环境的变化也会影响到数据的精确性，造成不确定性。

（3）在数据集成的相关应用领域，不同数据源的信息也会产生数据不一致性的问题。在集成的过程中，一些不确定性会被引入。例如，在 Web 应用中存在页面更新的问题，页面的内容不会一成不变，在引用这些原数据时就会产生不确定性。

（4）在处理缺失值的过程中也会出现数据的不确定性。这里缺失值产生的原因很多，包括装备故障、获取信息失败、字段不统一及历史原因等。

（5）数据的不确定性还会出现在一些比较特殊的应用领域。例如，在隐私保护的过程中，要达到保护隐私的目的，对原始数据处理的过程中会造成数据的不确定性。

不确定性数据在以下各个方面有典型的应用。

1）位置服务领域

基于位置服务是目前移动计算通信领域研究中最核心的问题。目前常用的手段为 GPS 技术，通过在电子地图上跟踪移动的物体进行定位从而达到提供空间信息服务的目的。其中的“位置隐私”研究受到了广泛的关注，其目的是保护用户的隐私，给出一个一定范围内的区域，不会精确到一个具体的空间点。这种技术不可避免地存在一定的误差，产生了大量的不确定性数据。

2）传感器网络领域

在工业、军事等领域，传感器网络作为新兴的数据收集与传输技术，有着广泛的应用。在自组的网络中，传感器节点以无线网络分布方式处理采集到的数据。有时节点使用较低成本的电子元器件就会直接导致精度的降低。外界环境包括带宽、传输时延、消耗的能量等，加之条件多变，均会使数据产生不确定性。对于一些温度或湿度的采集数据往往具有连续不确定性。这在传感器网络中被广泛使用。

3）电信服务领域

电信服务领域的数据量庞大，文件的传输数据、日志数据、用户的各类通话数据等都是海量的。这些数据具有一些共同的特点：质量要求高、规模庞大、产生速度快。对这些数据进行分析处理的过程中，难免会产生一些不确定性的因素，产生不确定性的数据。

4）互联网领域

当前互联网技术得到空前发展，信息迅速膨胀，大量信息分布在互联网上，成为了历史上最大规模的数据库。中国是网民数量最大的国家，截至 2008 年年底，中国的站点总数已经达到 287.8 万个，网页数更是超过 160 多亿，网页中的总字数达到了 470237487269KB。互联网作为一个管理系统，具有分散性，其中的数据质量较差，各站点信息的发布和数据维护复杂，互联网海量数据抽取结果的准确性也较低。

5）金融服务领域

金融领域的很多数据包含不确定信息。这些数据包括金融机构自身的数据、企业自身的数据、交易数据、监管和审计数据。很多时候是人为引入的，这些人为引入的不确定性信息使得金融诈骗犯罪案件层出不穷，造成恶劣的社会影响。因此在检测和分析金融数据时，必须考虑信息的不确定性和其中的虚假成分。

6）数据挖掘领域

从大量的原始数据、分散数据中分析对比、挖掘出有用的信息是数据挖掘的基本定义。挖掘任务的成功与否取决于原始数据的质量。当数据信息量大且客观准确时，从中获取的有用信息就越多；反之，获取的有用信息就越少，获得的知识越无价值。物理设备故障、信息无法获取、数据不一致等都会导致原始数据质量的下降，造成不确定性数据。数据预处理往往会改变数据自身的一些特性，也会产生不确定性数据。

如前所述，在诸多应用中，不确定性数据广泛存在。在各应用中，数据的描述方式各不相同。在传统的关系型数据库中，不确定数据通过关系表存储，其不确定性通过概率值存储在二维关系表中标识。但是，在关系型数据库中的数据具有结构化的特点，不适于一些嵌套结构和复杂对象的表达，而对于不确定性数据，大多在多种粒度基础上存在，因此需要考虑用新的结构来对不确定性数据进行存储。其他重要的数据形式还包括半结构化数据（XML）、流数据、多维数据和空间数据等[3]。

1.1.2 XML 文档表示不确定性数据的优势

可扩展的标记语言（extensible markup language，XML）是具有一定结构的标记语言，可以用于标记电子文件，在文档中可以定义数据的类型也可以用来存储数据，允许用户自己对标记语言进行定义[6]。XML 也是通用的标记语言（SGML）的子集，适合 Web 的传输应用。XML 中提供统一的方法来描述和交换结构化的数据，作为互联网中数据交换和表示的事实标准[7]，其主要由 3 个要素组成：文档类型定义[8]（DTD）与 XML 模式[9]（XML schema）、可扩展样式语言[10]（XSL）、可扩展链接语言[11]（XLL）。DTD 与 XML schema 共同作用，规定了 XML 文档的逻辑结构，提供语法分析程序，检查 XML 文件的规范性；XSL 规定了文档的样式语言；XLL 提供文档间的链接。

XML 本身有很多优点，这使它在表达概率数据模型方面有很大的优势。首先，可扩展性是 XML 的一个主要特点，其实质是指 XML 文档可以根据自身的需要定义数据关系，形成自己独特的标准。目前，各个行业均在建立行业化的标准规范。以电子商务为例，在网络上处理交易时，可以通过前台的 Web 站点把后台的数据信息呈现出来。一个 XML 文档可以看成一个小型的数据库，定义了各种数据关系，存储了各种属性的关系化数据，实现了信息的传输、上下文的检索和数据的交换等。XML 具有可扩展性好的特点，这使其在数据集成、挖掘、信息抽取领域也有广泛的应用。

其次，XML 还具有自描述性的特点。XML 是 SGML 定义的具有描述能力的文档。在存储信息时，不是采用应用软件独有的数据格式，而是采用易懂、易记的自描述标记形式来表现，因此，XML 最适合作为数据交换的标准。自描述性也使其广泛应用于电子商务等领域，作为数据交换的标准。XML 所具有的数据描述性的特点使其能够应用到各个领域。这也是 XML 被人们关注的原因。

另外，XML 还具有灵活性的特点。在传统的数据库当中，对需要进行具体描述的数据，可以根据已定义的模型来呈现，而在 Web 上的数据却不一样，这些数据具有动态性、可变性、自描述性等特点，数据形式特别复杂，模型描述形式不确定，且站点上的数据都是各自独立化设计的。在 Web 上的数据是具有一定的结构，但是因为自述层次的存在，它也是一种具有不完全结构的数据即半结构化数据。很多研究中已经把 XML 看做一种半结构化的数据模型，用到的不确定性信息一般涉及多种粒度、多种层次。XML 所具有的灵活性能够很自然地表示多粒度的不确定性信息，甚至可以是嵌套的不确定性信息。

此为试读，需要完整PDF请访问：www.ertongbook.com

综前所述，XML 具有的优点包括：好的自描述性、高的可扩展性和灵活性。因此，在很多产生不确定性信息的应用中特别适用，如信息抽取、数据集成等领域。

1.2 连续不确定 XML 数据管理技术发展

1.2.1 数据模型

在普通 XML 文件中，可以通过附加概率的形式来表示信息的不确定性。作为不确定 XML 数据管理的关键问题之一，通过获取普通 XML 文件的空间概率分布将普通 XML 文件转化为概率 XML 文件（可简记为 PrXML 文件）是一项重要的研究内容。因此，通过采用附加概率值在 XML 文件中表示不确定信息的方式，所构成的概率 XML 数据作为一种新的不确定数据的表示方法，已成为目前不确定 XML 数据的主要表现形式。对于不确定 XML 数据管理领域，数据模型的建立和操作与建设一座大楼的地基一样重要，模型选择与建立的好坏直接影响后期查询处理、代数优化等诸多概率 XML 文件操作的效果。那么，高效管理概率 XML 数据文件并实现相关的复杂数据操作首先要解决概率 XML 文件的数据模型问题。

选择或建立与待处理问题相适应的数据模型是不确定性数据管理操作中的重要问题。可能世界模型[2,12]在传统的不确定数据管理领域中是一种常用的数据模型。通过该模型逐渐引申出了许多其他数据模型，因此，可能世界模型语义是目前不确定 XML 数据模型的基础。可能世界模型由确定性的可能世界实例构成，通过数据实例将不确定数据库确定化并量化表示。文献[13]提出了一种基于关系数据库管理技术的不确定 XML 数据的管理方法，其中对可能世界模型进行了较详细的分析。

一般情况下，使用 XML 文件树[14]来表示 XML 文件。XML 文件树中包含了若干类型的概率属性节点。依据不同的概率对应关系，概率 XML 文件树中的概率属性节点有以下 6 种类型[14]。

（1）ind 节点[15]，即独立类型节点。该类型节点出现的概率不受其他节点的影响。

（2）det 节点[16]，即确定类型节点。该类型节点在概率 XML 树中出现的概率为 1。

（3）mux 节点[17]，即互斥类型节点。互斥类型的节点在概率 XML 树中不能同时出现。

（4）exp 节点[18-19]，即孩子节点组合类型节点。被选择的孩子节点组成孩子节点集进而构成 exp 节点。

（5）cie 节点[20-21]，即外部变量驱动类型节点。独立的外部事件变量 $e_1,\cdots,e_m$ 决定了 cie 节点的存在性。

（6）cont 节点[22]，即连续分布类型节点。具有 cont(D)的形式，其中 D 描述了实数范围内随机变量的概率分布。与上述分布类型节点不同，cont 节点只能出现在叶子节点上。

文献[14]～文献[23]使用上述几种分布节点表示概率值，提出并使用了基于有向图和树的概率 XML 文件的数据模型。文献[15]根据不确定数据的特点提出了使用 ind 和 mux 类型节点表示的概率 XML 模型，是第一个关于概率 XML 数据库方面的研究。文献[18]继续使用上述模型，通过在 XML 文件中指定的位置引入概率属性节点 prob 来指出 XML 文件中特定元素的不确定性。prob 使用 dist 节点表示（兄弟）节点概率值之间的关系，分为 mux 和 ind 两种类型。这种模型虽然能够有效地表示概率 XML 文件，但由于文件格式不同于普通 XML 文件，在概率 XML 文件中应该指出 prob 的概率值和 dist 节点的类型，因此，为了表示概率数据，需要适当修改源 XML 文件的文档定义类型（DTD），初始的 dist 和 val 定义如下：

```
<! ELEMENT dist (val+)>
<! ATTLIST dist type (independent|multually-exclusive) "independent">
<! ELEMENT val (#PCDATA)>
<! ATTLIST val prob CDATA"1">
```

文献[24]～文献[27]研究了不确定信息在 XML 文件中数据集成的方法。将节点分为概率节点、可能节点和普通 XML 节点 3 种。此外，概率 XML 文件还可以以有向图为基础建立数据模型。文献[18]提出了一种基于有向图的半结构概率数据模型 $\text{PrXML}^{\{\text{exp}\}}$（分布节点类型为 exp），模型中所有的半结构实例通过与相应的概率值对应，构成一个可能世界（possible world，PW）集合。文献[19]是第一个通过概率区间值表示概率 XML 的数据模型，基于有向图提出了一种新的建模概率 XML 文件的方法。文献[20]提出了一种简单的概率树（simple probabilistic tree，sp_tree）模型 $\text{PrXML}^{\{\text{ind}\}}$（分布节点类型为 ind）和概率树（probabilistic tree，prob_tree）模型 $\text{PrXML}^{\{\text{cie}\}}$（分布节点类型为 cie）。文献[21]使用了 sp_tree 模型，分析了查询复杂度。通常情况下，XML 文件以文件树的形式表示。那么使用有向图描述概率 XML 文件需要进行必要的修改。因此，数据模型的主要研究还是基于概率 XML 树的形式。文献[28]第一个将概率树 PT = (T, kind, prob)扩展，用于建模连续不确定数据，其中，可能节点保存概率密度函数和概率块。因此，每个可能节点保存了无数种可能性。这种集成方法特别实用，在不需要修改 DTD 的同时可以有效地通过一个概率 XML 树表示多个 XML 文件，但是可能节点作为集成结果中的冗余信息，增加了存储和查询处理的时间。

上述研究提出与使用的数据模型均采用概率的形式表示 XML 文件中的不确定数据，但都没有对概率数据之间的相互关系进行相应的研究。文献[23]提出了概率 XML 数据库（probabilistic XML database，PXDB）数据模型。PXDB 通过增加外部约束条件不仅能够在 XML 中以概率的形式表示并记录不确定信息，而且可以表达不确定概率数据间的限制条件，因此，该模型在表达概率数据之间的依赖关系方面体现出其优势。

不同类型分布节点扩展和组合表示的数据模型构成了概率 XML 文件的家族 PrXML^{C}，其中，$C \subseteq \{\text{mux, ind, exp, cie}\}$，它们因使用的分布节点的类型而不同。例如，$\text{PrXML}^{\{\text{cie,exp}\}}$ 使用了 cie 和 exp 类型的分布节点。文献[24]～文献[27]研究了 p-documents 家族中不同节

点组合表示的数据模型表达现实世界的能力及查询评估效率，研究证明了不同概率 XML 文件家族之间的转换，并指出了不同模型上的查询复杂度。研究表明，$\text{PrXML}^{\{cie\}}$由于允许数据项之间复杂的概率依赖关系，所以是全局的概率数据模型。$\text{PrXML}^{\{mux,ind\}}$是局部的概率数据模型，更精确地，它只允许兄弟间的依赖，并且 $\text{PrXML}^{\{mux,ind\}}$概率文档可以转化为一个等价的更加简洁的 $\text{PrXML}^{\{mux,det\}}$概率文档。所以，ind 类型和 det 类型的分布节点可以转化。$\text{PrXML}^{\{cie\}}$和 $\text{PrXML}^{\{mux,det\}}$都能表示任何离散的 px-space。但是，它们在简洁性和查询效率的权衡方面依旧是不相同的。文献[16]研究证明，$\text{PrXML}^{\{cie\}}$是比 $\text{PrXML}^{\{mux,det\}}$呈指数倍精简的表示模型，$\text{PrXML}^{\{mux,det\}}$与 $\text{PrXML}^{\{mux,ind\}}$一样精简的表示模型。但是，在查询处理效率方面，Kimelfeld 等在文献[29]和文献[30]中证明，在 $\text{PrXML}^{\{mux,det\}}$表示的数据模型中，计算所有树模式查询的概率都存在一个可解的多项式算法，而所有有用的查询在 $\text{PrXML}^{\{cie\}}$上都是 #p-complete 的。综合上述研究结果，$\text{PrXML}^{\{ind,mux\}}$符合简洁性、查询复杂度及效率之间的权衡，因此，被目前大多数研究所采用。

文献[24]增加了一种附加类型的分布节点，即 cont 连续类型节点，扩展了不确定 XML 数据文件的语法，能够由前述数据模型只支持叶子节点上为离散分布的随机变量扩展到具有连续分布类型的随机变量。需要指出的是，cont 节点只会出现在叶子节点上。对于离散分布类型的随机变量，一个概率 XML 文档定义了一个指定了相应概率的树的有限集。对于连续分布类型的随机变量，一个概率文档定义了一个包含为树的集合指定了概率取值的连续分布的树的无限集，因此，cont 类型节点的存在可以使概率 XML 文件的表达对象由离散随机变量扩展到支持连续分布类型的随机变量。

1.2.2　编码方法

随着 XML 技术的快速发展，对 XML 数据的高效索引技术和查询技术的需求也越来越迫切。在查询过程中，为了避免反复读取源 XML 文档并能快速判断两个节点之间的关系，人们提出了 XML 文档的编码方案。编码之前需要了解文档采用的数据模型，在模型的基础上再进行编码的操作，在编码的基础上完成查询算法的设计。有关普通 XML 编码技术的研究已经取得了一定的成果，但对于不确定 XML 编码技术的研究还较少。

1. 普通 XML 编码技术的研究现状

为了达到既能提高查询效率，又能减少对源 XML 文档的读取，并能快速判断出两个节点的结构关系的目的，近年来人们提出了很多编码方法，并且在编码的基础上设计了各种各样的 XML 数据索引和查询算法。编码技术在整个 XML 数据管理领域所起的作用越来越重要。

目前最常见的编码方法有 3 种，分别为区域编码、前缀编码和素数编码。

1）区域编码

区域编码方案的编码思想是：为 XML 文档树中的每一个节点分配一对数字，这对数字表示此节点所覆盖的区域范围。给出任意的两个区域 r_1 和 r_2，如果 r_1 包含 r_2，则 r_1 节

点是 r_2 节点的祖先。基于这样的思想，人们又提出了多种区域编码方法。文献[31]最早提出了起止性区间编码的概念。该编码由一个三元组组成，三元组表示为(start, end, level)，其中，start 表示节点在文档的开始位置，end 表示节点在文档的结束位置，level 表示节点在整个文档树中所处的层数。这里所说的位置可以是物理概念上的位置，也可以是逻辑概念上的位置。物理位置是指节点在文档中的开始位移量和结束位移量；逻辑位置是指深度优先遍历节点所在的 XML 树时，每一个节点被遍历访问两次，第一次访问该节点时为 start 赋值，第二次为 end 赋值。利用这种编码可以快速判断两节点的嵌套关系，但当文档发生变化时，更新的代价很高。文献[32]提出了前序-后序的编码方案，节点的编码是一个二元组(preorder, postorder)，其中，preorder 是前序遍历 XML 文档树时节点对应的前序遍历序列值，postorder 是后序遍历 XML 文档树时节点对应的后序遍历序列值。给定两个节点 u 和 v，当 u.preorder < v.preorder 并且 u.postorder > v.postorder 时，u 是 v 的祖先节点。以上提到的两种编码能够判断节点间的结构关系，但是文档发生变化时更新的代价很高。据此，文献[33]提出了一种扩展的区间编码。该编码由二元组(order, size)组成，其中，order 代表遍历序列，size 代表节点的大小，size 可以任意调整，从而支持更新。这种编码方法在保持原有高效判断结构关系的基础上，极大地降低了更新的代价。编码时可以随意调整 size 的值，使其适当大于该节点的实际大小。在插入新的节点时就可以利用这些预留的空间，从而减少更新的代价，也不会影响结构关系的判定。

2）前缀编码

前缀编码被称为基于路径的编码。在一棵采用前缀编码的文档树中，父亲节点的编码是其孩子节点编码的前缀，因此，在判定节点间的结构关系时，如果一个节点是另一个节点的前缀，那么该节点是另一个节点的祖先节点。在文档树中插入一个新的节点时，只会影响插入节点的父亲节点和它所有的子孙节点。这样，编码更新就限制在了父亲节点的范围内。文献[34]中最早提到一种前缀编码（Dewey 编码），它也是最经典的前缀编码，父亲节点的编码是孩子节点编码的前缀，通过判断编码间的前缀关系来判断节点之间的结构关系。但前缀编码方案都需要分隔符，不仅浪费了存储空间，还给数据的表示带来了不便。文献[35]中提出一种基于前缀的 BitPath 编码方法，它与 Dewey 编码类似，不同的是，它用 0、1 二进制数来代替整数，并且不需要分隔符，节省了存储空间并且有更高的判断效率。文献[36]中提到的 CSSU 编码也是基于前缀编码的，这种编码不再用纯数字编码，而是采用英文字母编码。在进行更新操作时，不必对已经编码的节点重新编码。文献[37]中提出的 OrdPath 编码是一种前缀编码方法，首先用奇数编码，预留偶数为将来新插入的节点编码。文献[38]中提出的 PPSCU 编码也是一种前缀编码，利用数字和字母的混合方式进行编码。这种方法能快速判断结构的关系，当文档发生改变时，二次编码率为零。

3）素数编码

素数编码的基本思想是：用素数为 XML 文档树中的节点编码。根据素数的唯一性及

素数之间的整除关系判断结构关系。首先对根节点进行编码，为 Prime_lable 分配素数 1，然后自顶向下为每一个节点指定一个唯一的素数值，即 self_lable，每个节点的 Prime_lable 是其父亲节点的 Prime_lable 与自己本身的 self_lable 的乘积，通过 Prime_lable 来判定节点间的结构关系。文献[39]中第一次提到了素数编码方式，用素数为 XML 文档中的节点编码，在更新时不需要重新编码已经存在的节点，但是当文档树深度过大时会造成大素数的溢出。文献[40]中提到的 Pri-Order 编码方式是一种基于素数的编码方式，采用三元组的形式编码，三元组的三部分相互结合保持了文档的树形结构，能够在文档动态更新和存储恢复间达到平衡，但算法的实现较复杂。文献[41]提出的 PSB 编码方案是一种新的基于素数整除关系和序列化的文档编码机制，不仅可以快速判断结构关系，还支持对编码的更新操作，这种方案同样存在算法实现较复杂的问题。

2. 不确定 XML 编码技术的研究现状

目前，对 XML 编码技术的研究主要是对普通 XML 的研究，对不确定 XML 编码技术的研究也只是针对离散的不确定 XML，而对连续不确定的研究几乎没有。文献[42]第一次将区间编码应用于离散的不确定 XML 中，但仍然没有克服区间编码的缺点。文献[43]研究了离散的不确定 XML 的 Top-k 查询，用到了前缀编码，但是它的工作主要还是处理离散不确定性数据。当文档结构深度过大时，在 p-文档型基础上进行前缀编码会浪费存储空间。

1.2.3　索引技术

根据索引面向的数据模型和查询模式及实现机制的不同，研究人员提出了种类繁多的索引方法，大体上可分为三类：节点信息记录类索引、结构摘要类索引和 XML 数据与全文数据的联合索引。下面将对此三类索引进行概括与总结。

1. 节点信息记录类索引

节点信息记录类索引的主要思想是：建立一个数据库，把 XML 数据分解为多个记录单元，构成一个集合，集合中不仅有记录单元还有记录单元在 XML 文档中的位置信息。将这些信息统一存储在数据库中，当用户查询时，首先在数据库中找到数据的位置信息，然后进行简单的处理，得到节点之间的结构关系。具体可采用两种方法：第一种是为节点编码，第二种是为节点记录路径信息。所以研究人员将节点信息记录类索引总结为三类：通过为节点编码后形成的索引、通过记录节点的路径信息形成的索引、通过前两种相结合形成的索引。

1）通过为节点编码后形成的索引

节点编码类索引依赖某种遍历序列中的序号来确定节点的位置信息。所以首先要对 XML 文档树设计一种遍历节点的策略，根据确定的遍历策略遍历节点形成一个由节点组成的序列。序列中的序号和节点的标签一一对应，根据序号确定标签的次序关

系，进一步表明节点间的次序关系，反映节点间的结构关系，所以这类索引可以依据序号信息确定节点间的子孙后代关系。

这类索引的实质是根据树的特点确定一种遍历方式，根据此遍历方式进行遍历，得到节点的遍历序列，这个序列对应的就是相应节点的编码；对于 XML 文档树中的任意两个节点，对它们的编码构建某种运算，节点之间的结构关系就依赖这种运算的结果。根据遍历方式的不同，编码可归结为如下三类：根据区间编码和根据 *B-E-L* 模式编码、根据路径编码。

（1）根据区间编码。该方式的基本思路是采用前序或后序遍历节点形成的序列中，前后两个序号之间的间隔就是节点的区间编码。区间编码可表示为(tabL, pre, post)，其中 tabL 为节点的标签。虽然这种编码思想简单、容易确定，但是也存在一定的缺点（如有时无法确定父子关系），所以，研究人员在此基础上又引入一个参数，就是节点所在的层数，这种编码可表示为(tabL, pre, post, *l*)，其中 *l* 表示层数。这种编码称为扩展的区间编码。Dietz 编码、Li-Moon 编码和 Zhang 编码都是其中典型的代表。Dietz 编码是将一个节点的前序遍历序号和后序遍历序号组合成一个二元组为(pre, post)，然后根据节点的二元组信息来判断祖孙关系和父子关系。Li-Moon 编码是将节点的扩展前序遍历序号、子孙节点的范围和节点的深度构成一个三元组，表示为<order, size, depth>（其中，order 是为了给节点预留一定的空间而给出的不连续的前序遍历序号，size 是节点的子孙的范围，depth 是节点深度），然后根据节点的三元组信息来判断祖孙关系和父子关系。Zhang 编码是将 XML 文档树中的每个节点在前序遍历时分别被访问两次，产生两个遍历序号。这两次分别是访问子孙节点之前和访问子孙节点之后，这样把 XML 文档树中的每个节点都赋予一个三元组<*b*, *e*, *l*>（其中 *b* 表示访问子孙节点之前节点的遍历序号，*e* 表示访问子孙节点之后的遍历序号，*l* 为节点的层数），然后根据节点的三元组信息来判断祖孙关系和父子关系。

（2）根据 *B-E-L* 模式编码。*B-E-L* 模式中 *B* 表示节点起始标签在 XML 字符流中的序号，*E* 表示节点终止标签在 XML 字符流中的序号，*L* 表示节点的层数。这类编码主要是由 *B*、*E* 和 *L* 构成的数字对作为节点的编码。由此看出，这类编码和区间编码有相似之处。

（3）根据路径编码。这类编码是根据文档树具有树形结构的特点，把从根节点到后代节点之间的每条路径上的每一个节点都进行编码，Dewey 编码、PBiTree 编码、k-ary 编码和 eXist 编码都是其中典型的代表。Dewey 编码的思想是将每一个节点的父节点的编码作为本节点的编码前缀，根据这个特点来确定祖孙关系和父子关系。k-ary 编码是将 XML 文档树中的节点都映射到一棵完全 *k* 叉树中，根据 *k* 叉树的特点来确定祖孙和父子关系；k-ary 编码不利于底层节点的编码，因为底层节点数量太大，空间浪费较严重。所以，又引入了 eXist 编码，不将 *k* 值固定，*k* 值根据每层节点的子节点数来确定。PBiTree 编码是将 XML 文档树中的节点映射到一棵完全二叉树 PBiTree 上。根据二叉树的结构关系来确定祖孙关系和父子关系。

区间编码往往是将查询语句分解，形成一系列的二元结构关系，然后进行结构关

此为试读，需要完整PDF请访问：www.ertongbook.com