计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190700164

一种基于模糊集和概率分布的不确定 XML 模型及其代数运算

胡 磊 严 丽
南京航空航天大学计算机科学与技术学院 南京 211100
(1091011194@qq.com)

摘 要 XML 作为一种信息表示和交换的事实标准已被广泛用作不同应用之间的统一数据交换格式,其在实际应用中已经发挥着重要的作用。由于现实中很多信息包含有不确定性,而经典的 XML 不能表示和处理不确定信息,因此有必要对经典 XML 模型进行扩展。考虑到现实世界的复杂性,不确定信息往往同时包含有随机不确定性和模糊不确定,而概率理论和模糊集理论是处理不确定信息的有力工具,因此文中在现有的模糊 XML 和概率 XML 数据模型的基础上,综合利用概率和模糊理论建立一个新的不确定 XML 模型和相关代数,所提出的新的不确定性 XML 模型既能与现有的 XML 模型兼容,又能表达更复杂的不确定信息。

关键词:XML 模型;不确定数据模型;模糊集;概率分布;代数运算
中图法分类号 TP311.131

Uncertain XML Model Based on Fuzzy Sets and Probability Distribution and Its Algebraic Operations

HU Lei and YAN Li
College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211100, China

Abstract As a de-facto standard of information representation and exchange, XML has been widely used as a unified data exchange format between different applications, which has played an important role in real-world applications. However, the real world is filled with uncertain information and classical XML is not able to represent and deal with uncertain data. So it is necessary to extend classical XML model. The real world is complex, which often contains both random and fuzzy uncertainties. Considering that probability theory and fuzzy set theory are powerful tools for dealing with uncertainty, this paper uses both probability theory and fuzzy set theory to build a new uncertain XML model, which is different from the existing fuzzy XML models and probabilistic XML models. The new uncertain XML model is compatible with existing XML models and can represent more complex uncertain information.

Keywords XML model, Uncertain data model, Fuzzy set, Probability distribution, Algebraic operation

1 引言

XML 作为一种灵活的半结构化标签语言,在信息表示和数据交换方面有先天的优势。XML 现在被广泛地应用于各种团体和商界,在很多应用领域,人们已经定义了基于 XML 的应用,如数学领域(MathML)、天文领域(AML)、生物信息领域(BSML)、投资领域(IRML)等。此外,可以作为统一数据交换格式的 XML 在企业信息交换方面也发挥了巨大的作用。XML 已经成为 Web 数据表示与交换的事实上的标准[1]。

传统的 XML 总是假定要表示和处理的数据是完全确定的,然而由于现实世界中处处存在着不确定性,因此在某些情形下很难百分百确定地陈述一些信息。客观上,在科学研究中,实验机器出错、人为的错误或者样本数据的噪声等都可能导致不确定性。主观上,人的思维和认知本身存在着模糊性,这也是不确定性的一个来源。总的来说,传统 XML 模型描述客观世界中事物的困难性主要体现在两个方面,分别是随机不确定性和模糊不确定性。随机不确定性是指即使条件完全相同,事件的结果也不能确切预言。而模糊不确定性是指概念本身没有明确的外延,一个对象是否属于这个概念是难以确定的,因此造成了划分的不确定性。以往将不确定性引入 XML 的工作往往只关注一种不确定性,形成了概率 XML 模型(例如文献[2-3])和模糊 XML 模型(例如文献[4-7])。

正如关系数据库所表明的那样,代数操作对于数据库查

到稿日期:2019-07-23 返修日期:2020-03-31 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:南京航空航天大学研究生创新基地(实验室)开放基金项目(kfjj20181601)
This work was supported by the Open Fund of Graduate Innovation Base (Laboratory) of Nanjing University of Aeronautics and Astronautics (kfjj20181601).
通信作者:严丽(yanli@nuaa.edu.cn)

询优化发挥着非常重要的作用。虽然关于 XML 的代数已被提出[7-11]，但是这些代数操作并不能直接应用到本文提出的不确定 XML 模型。一方面，某些代数提议考虑以元素为代数的基本操作对象，粒度太小，从而导致代数过于复杂。另一方面，某些代数提议虽然以 XML 数据树整体作为基本操作对象，但是其代数只能应用于经典的 XML。基于本文提出的不确定 XML 数据模型，我们提出其形式化的代数。

本文第 2 节介绍了相关工作；第 3 节介绍了相关的基础知识；第 4 节在模型的一个示例的基础上给出不确定 XML 数据树的定义以及 XMLS 结构的定义；第 5 节定义模型的代数操作；最后总结全文。

2 相关工作

利用模糊集理论和概率理论扩展数据库方面的工作由来已久。不确定关系数据库的相关工作有很多[12-15]，其中文献[14]将概率区间引入关系数据库中，并且证明了可以在多项式时间内检查不一致性，并提出了将元组属性的概率区间转化为元组整体概率区间的方法。文献[15]提出了用模糊概率分布来代替概率区间以更好地描述元组的不确定性，并且讨论了相关的代数操作。在面向对象数据库中引入不确定性也被广泛研究[16-21]，其中文献[16]借鉴文献[14]的方法，将概率区间引入面向对象数据库中并且详细讨论了类、属性和方法的继承规则，在此基础上又讨论了相关的代数操作。文献[18]将演绎数据库和面向对象数据库的优点结合起来，并且引入了不确定性。文献[20]将模糊概率分布引入面向对象数据库中代替概率区间，并且详细讨论了在此基础上的代数操作。

XML 作为一种半结构化数据模型，在表示网络数据方面有更大的优势，因此自然引出了如何将这些不确定性数据库中的数据转化为 XML 形式的问题。然而，当前作为数据表示和数据交换的经典 XML 不能处理不确定性信息，因此必然要对不确定性 XML 进行相关研究。

在概率 XML 方面，文献[3]提出了一个概率 XML 模型，可以为元素指定概率，为元素的值指定概率分布，并且给出了在 XML 中引入概率的合理解释以及 DTD 模式的模型定义。然而，文献[3]引入的概率并不是概率区间而是点概率，其次由于 DTD 方式的数据类型过于简单且其语法不基于 XML，因此用 XMLS[22] 结构化 XML 模型会更合适。文献[2]提出了一个概率区间的 XML，并且形式化地提出了弱实例和概率实例等相关概念。此外，文献[2]还提出了模型的两种语义，即全局语义和局部语义。但是其并没有提供模型的 XMLS 方式定义，也没有提供一个形式化的代数。虽然相比文献[3]的点概率而言，在无法精确地给出点概率的情形下，用概率区间表示更加合理。然而，相比概率区间，引入模糊概率分布将会是一种更好的表示方式。首先概率其实也存在着一定的主观因素，某些情况下如果我们很难用一个点概率确定地描述某个事件，那么就可以给概率赋予一个模糊值。其次，概率区间虽然给定义了一个范围，但是我们对范围中的概率了解得很少，而模糊概率分布可以解决这个问题。因此，本文将引入模糊概率分布来代替概率区间。

在模糊 XML 方面，基于 DTD 方式的结构化模糊 XML 模型[6] 和基于 XML Schema 方式的结构化模糊 XML 模型[5] 分别被提出。文献[7]提出了一个模糊 XML 模型，并且讨论了两种模糊性，一种是用模糊值描述的元素的模糊性，另一种是用可能性分布描述元素的值的模糊性。其中元素值的可能性分布又分为两种，一种是表示“排斥或”含义的 disjunctive 分布，另一种是表示“合取”含义的 conjunctive 分布。在此基础上，文献[7]提供了一个形式化的代数，考虑以树为基本的操作单位，来方便查询和优化。

模糊及概率不确定性表示与处理已经引起了国内外研究者的关注。在面向对象数据库的工作[16]中，研究者提出了将类属性赋予一个模糊集，将不确定性类的成员和属性赋予一个概率区间。在关系数据库的工作[15]中，研究者提出了将模糊和概率结合的模糊概率分布。不同于关系数据库和面向对象数据库的研究，模糊及概率不确定性在 XML 框架下的研究工作还很少。

以往的不确定 XML 模型的相关工作存在一定的局限。首先，研究集中在概率不确定性[2-3] 或者模糊不确定性[4-7]。这就形成了一种对立：概率不确定性 XML 模型无法处理模糊信息，而模糊 XML 模型又无法处理概率信息。但现实世界是复杂的，往往同时存在着两种不确定性，只引入一种不确定性的数据模型对现实事物的解释能力会存在局限性。其次，以往工作中扩展 XML 模型表示概率不确定性的方式通常是点概率或者概率区间，本文采用的方式是引入模糊概率分布。模糊概率分布相比概率区间而言，显得更加直观而且可以附带更多的不确定信息。统计界的贝叶斯学派认为：一个事件的概率是人们根据经验对该事件发生的可能性给出的个人信念。相比确定的概率区间而言，模糊概率分布可以涵盖更多的信息。例如我们可以为某个事件发生的概率指定一个模糊概率分布[0.3/0.7,0.5/0.8,0.6/0.9]，并且解释为事件发生概率为 0.7 的可能性为 0.3，事件发生概率为 0.8 的可能性为 0.5。模糊概率分布事实上是概率与模糊两种主观信念相结合的产物。最后，以往 XML 模型相关工作中的代数要么粒度太小导致代数过于复杂而难以在实际中使用，要么只能应用于经典的 XML 模型。考虑到上述问题，本文将综合利用概率理论和模糊集理论，将模糊概率分布和模糊可能性分布引入经典 XML 中，使得新的不确定性 XML 模型能更完善地表示现实中的概念。此外，考虑到数据查询和优化的需要，本文将为提出的不确定 XML 模型建立一个完整的代数，能方便地应用于本文提出的 XML 模型，且不会过于复杂。

3 基础知识

在不确定性关系数据库模型中，概率和模糊值可以与属性关联，也可以与元组整体关联。与属性关联的概率值和模糊值可以转化为与元组整体关联的相应值。而元组是关系数据库中的核心概念，与元组相连的不确定属性可以自然地解释为在给定关系中存在的概率或者可能性。在 XML 模型中引入不确定性，自然也需要解释这种不确定性是相对于什么而言的不确定性。本文将在 XML 中的元素中引入模糊概率

分布属性，将模糊概念相关元素的属性值引入可能性分布。

3.1　概率分布与概率 XML

由于本文引入的概率解释为条件概率，因此在此简单介绍条件概率的定义，以及后面分析会用到的公式。

定义 1（概率的公理化定义）　设 Ω 为一个样本空间，Γ 为 Ω 的某些子集组成的一个事件域。如果对任一事件 $A\in\Gamma$，定义在 Γ 上的一个实值函数 $P(A)$ 满足：

1）非负性公理，若 $A\in\Gamma$，则 $P(A)\geqslant 0$；

2）正则性公理，$P(\Omega)=1$；

3）可列可加性公理，若 $A_1,A_2,\cdots,A_n,\cdots$ 互不相容，则：

$$P(\bigcup_{i=1}^{\infty}A_i)=\sum_{i=1}^{\infty}P(A_i)$$

则称 $P(A)$ 为事件 A 的概率，称三元素 (Ω,Γ,P) 为概率空间。

定义 2（条件概率定义）　设 A 与 B 是样本空间 Ω 中的两事件，若 $P(B)>0$，则称

$$P(A|B)=\frac{P(AB)}{P(B)}$$

为“在 B 发生下 A 的条件概率”，简称为条件概率。

性质 1　条件概率是概率（只要证明满足概率的公理化定义即可）

性质 2（全概率公式）　设 $B_1,B_2,\cdots,B_n$ 为样本空间 Ω 的一个分割，即 $B_1,B_2,\cdots,B_n$ 互不相容，且 $\bigcup_{i=1}^{n}B_i=\Omega$，如果 $P(B_i)>0,i=1,2,\cdots,n$，则对任一事件 A 有：

$$P(A)=\sum_{i=1}^{n}P(B_i)P(A|B_i)$$

在关系数据库中，我们假定元素的概率属性是在给定关系的集合中存在的概率，但是在 XML 中并不存在与关系这类集合对应的概念。这种情况下，可以将 XML 中的元素和以它为根的子树看作一个整体，并且定义 XML 中元素的概率为这个元素和以它为根的子树在现实世界中存在的概率。当然，在 XML 中元素的概率都是假定其父结点存在时，当前结点存在的条件概率（当父结点不存在时当前结点的概率为 0）。这种情况下，我们可以根据给定的条件概率，利用全概率公式计算出当前元素的存在概率。考虑一条结点链，$A\rightarrow B\rightarrow C$，$A$ 是 B 的父结点，B 是 C 的父结点，假定我们已知 A 的概率和 B 的条件概率，可以利用全概率公式得到 B 的概率。假定 A 的概率是 $prob(A)$，B 的条件概率是 $prob(B|A)$，利用全概率公式得 $prob(B)=prob(A)prob(B|A)+prob(\overline{A})prob(B|\overline{A})$，因 A 不存在时，B 的概率区间为 0，所以有 $prob(B|\overline{A})=0$。因此得到 $prob(B)=prob(A)prob(B|A)$，这样就得到了 B 的概率。同样地，可以得到 $prob(C)=prob(C|B)prob(B|A)prob(A)$。上述讨论说明，虽然在模型中给定的是条件概率，但是概率也隐含在模型中，如果需要查询给定元素的概率，可以利用上述方法计算出来。

3.2　模糊集与可能性分布

现有的处理模糊信息的方法大多依据模糊集理论以及可能性分布理论。模糊信息作为不确定信息的一种表现形式，在现实世界中广泛存在，如“年轻人”“高个子”等。这些模糊信息都需要用模糊集或者延伸的可能性分布来表述。

设在论域 U 上给定一个映射：

$$A:U\rightarrow[0,1]$$

$$u\mapsto A(u)$$

则称 A 为 U 上的模糊集，$A(u)$ 称为 A 的隶属度函数（或称为 u 对 A 的隶属度）。

有限模糊集可表示为：

$$A=\left\{\frac{A(u_1)}{u_1},\frac{A(u_2)}{u_2},\cdots,\frac{A(u_n)}{u_n}\right\}$$

可能性是人们经常用到的概念，例如“这次比赛可能取胜”“这个机器可能有问题”等。

F 约束定义为：以 U 为论域，X 是在 U 上取值的一个变量。设 F 是 U 上的模糊集，若 F 对 X 的取值起一种可伸缩的约束作用，则称 F 为 X 的 F 约束。具体可表示为：

$$X=u:F(u)$$

其中，$F(u)$ 可以解释成 F 对 X 取值为 u 时的约束程度。

可能性分布定义为：设 F 是论域 U 上的模糊集，而 $F(u)$ 解释为 u 与标为 F 的概念的相容度。设 X 是在 U 上取值的变量，而 F 起着与 X 相关联的 F 约束 $R(X)$ 的作用，则命题“X 是 F”可以表示为 $R(X)=F$，由此，与 X 有关的可能性分布记为 Π_X，假定它等于 $R(X)$，即 $\Pi_X=R(X)$，相应地，与 X 有关的可能性分布函数（或 Π_X 的可能性分布函数）用 π_X 表示，并且在数值上定义等于 F 的隶属度，即：

$$\forall u\in U,\pi_X(u)=F(u)$$

也就是说，$X=u$ 的可能性 $\pi_X=F(u)$。

前面已经简单介绍过模糊集的相关理论，本文引入模糊属性值分布，主要是作为对模糊概念的解释。例如，如果 XML 文档中含有模糊概念“年轻人”，那么我们可以很自然地引入一个可能性分布作为“年轻人”这个模糊概念的进一步解释。但是，如果仅仅局限在这一步，模型会有很大的局限性。现在我们考虑理论上进一步扩展可能性分布，为与模糊概念相关的精确概念引入可能性分布。

首先以一个例子来解释传统的可能性分布理论。设论域 $U=\{1,2,3,4\cdots\}$，且 F 为小整数 F 集，定义为：

$$\text{小整数}=\left\{\frac{1}{1},\frac{1}{2},\frac{0.8}{3},\frac{0.6}{4},\frac{0.4}{5},\frac{0.2}{6}\right\}$$

如果 X 的取值为 4，则有 $X=4:0.6$，这个式子表示当 X 取值为 4 时，F 约束为 0.6。那么由 $\pi_X(u)=F(u)$，即可能性等于相容度，可以说，当 $X=4$，X 为小整数的可能性为 0.6，也就是说变量 X 伴随的可能性分布为：

$$\Pi_X=\left\{\frac{1}{1},\frac{1}{2},\frac{0.8}{3},\frac{0.6}{4},\frac{0.4}{5},\frac{0.2}{6}\right\}$$

以其中任何一项为例，例如 $\frac{0.2}{6}$，表示已知“X 是 6”，则“X 是小整数”的可能性为 0.2。

当 XML 中存在模糊概念“小整数”时，我们引入形式化的元素和属性，以便用对应的可能性分布来解释模糊概念。同样地，我们也可以使得精确的概念伴随一个可能性分布。现在考虑另外一个例子，问李四是否是一个年轻人，我们取年龄作为论域 U。显然“年轻人”是定义在 U 上的模糊集。假定模糊集“年轻人”的定义如下：

$$\text{年轻人}=\left\{\frac{0.7}{25},\frac{0.85}{26},\frac{0.9}{27},\frac{0.8}{28},\frac{0.75}{29},\frac{0.7}{30}\right\}$$

那么按照可能性分布理论，我们可以给“年轻人”定义一个可

能性分布。如果知道了确切年龄，如张三的年龄是 26 岁，我们可以推测张三是年轻人的可能性是 0.85。

前面的方式都是由精确推理模糊，现在可以反过来考虑问题，由模糊推理精确。如果知道了张三是年轻人(模糊概念)，则据此来推测张三的年龄。那么就可以给年龄这个精确的概念定义一个同样的可能性分布：

$$年龄=\left\{\frac{0.7}{25},\frac{0.85}{26},\frac{0.9}{27},\frac{0.8}{28},\frac{0.75}{29},\frac{0.7}{30}\right\}$$

并且可以做出如下直观解释，年龄是 25 岁的可能性是 0.7，年龄是 29 的可能性是 0.75。这样就为与模糊概念相关的精确概念也定义了一个可能性分布，因此无论是精确概念还是模糊概念，都可以合理地引入可能性分布。

4 不确定 XML 模型的表示

第 3 节解释了在不确定 XML 模型中引入两种不确定的合理性。不确定性可以存在于 XML 文档的元素或者属性值上，对于元素的不确定性，我们引入将概率和模糊性结合的模糊概率分布来表述；对于元素属性值的不确定性，我们引入概率分布或者可能性分布来表述。

首先给出一个不确定 XML 示例，如图 1 所示。下面以图 1 为例详细阐述不确定 XML 引入的新属性和新元素。很明显，首先应该介绍与概率相关的属性，FP 表示一个模糊概率分布。图 1 中的第 2 行表示对应元素存在的概率为 0.4 的可能性为 0.55，概率为 0.5 的可能性为 0.7，以此类推。与模糊值相关的属性是 poss，它的取值范围为[0,1]，如图 1 中的第 29 行表示刘峰年龄是 23 的可能性是 0.4。为了便于利用 XMLS 定义模型的结构，我们引入了两种类型的概率值元素，即 PNleafval 和 PLeafval。PNleafval 表示对应此模糊概率分布的元素不是叶子结点，而 PLeafval 表示对应此模糊概率分布的元素是叶子结点。PLeafval 主要用于概率分布，作为概率分布其中的一项。同样地，我们也引入了两种类型的模糊值元素，即 FNleafval 和 FLeafval，其中 FLeafval 主要用于进一步解释模糊概念的可能性分布。

在不确定 XML 模型中，子元素之间的概率可能存在某种关系，要表达这种关系就不得不形式化地引入概率分布。但是由于现实世界的关系是错综复杂的，如果考虑每一种情形，那么模型将变得异常复杂。但是有两种情形是很常见的，一种是“排斥”，另一种是“相容”，“排斥”表示几种情形只有一种会发生，而“相容”表示几种情形可以同时发生。PDisjunctive 元素表示概率上的排斥。如图 1 中的第 6－17 行所示，我们引入 PDisjunctive 表示第 7－11 行和第 12－16 行这两种情况只有其中一种会发生，我们观察一下对应的 XML 文档也可以看到，第 7－11 行的性别是男而第 12－16 行的性别是女，一个人的性别不可能同时是男和女，因此满足排斥的情形。PConjunctive 元素表示概率上的相容。如图 1 中的第 38－41 行表示这个人有可能同时有两个邮箱，或者只有两个邮箱其中的一个(模糊概率分布已经给出)。同样地，我们也引入两种可能性分布，一种用 FDisjunctive 表示，另一种用 FConjunctive 表示。如图 1 中的第 28－34 行所示，因为一个人的真实年龄只有一个值，所以我们用 FDidjunctive 表达这种含义。而第 22－25 行则表达了一个相容的可能性分布。

```
1.<provinces>
2. <PNleafval FP="0.55/0.4 0.7/0.5 0.8/0.6 0.85/0.7">
3.  <province PName="shanxi">
4.   <city CName="xian">
5.    <employee ID="67214398">
6.     <PDisjunctive>
7.      <PNleafval FP="0.4/0.2 0.9/0.3 0.8/0.4">
8.        <name>LiMing</name>
9.        <position>engineer</position>
10.       <sex>male</sex>
11.     </PNleafval>
12.     <PNleafval FP="0.8/0.5 0.9/0.6 0.75/0.7">
13.       <name>LiMing</name>
14.       <position>senior employee</position>
15.       <sex>female</sex>
16.     </PNleafval>
17.    </PDisjunctive>
18.   </employee>
19.   <student SID="47582139">
20.    <name>LiuFeng</name>
21.    <honor>
22.     <FConjunctive>
23.       <FLeafval poss="0.7">Swimming Champion
              </FLeafval>
24.       <FLeafval poss="0.6">Racing Champion
              </FLeafval>
25.     </FConjunctive>
26.    </honor>
27.    <age>
28.     <FDisjunctive>
29.       <FLeafval poss="0.4">23</FLeafval>
30.       <FLeafval poss="0.6">25</FLeafval>
31.       <FLeafval poss="0.8">27</FLeafval>
32.       <FLeafval poss= "0.3">28</FLeafval>
33.       <FLeafval poss="0.2">29</FLeafval>
34.     </FDisjunctive>
35.    </age>
36.    <sex>male</sex>
37.    <email>
38.     <PConjunctive>
39.       <PLeafval FP="0.6/0.5 0.75/0.6    0.3/0.7">
            LiuFeng@yahoo.com</PLeafval>
40.       <PLeafval FP="0.9/0.2 0.85/0.3 0.75/0.4" >
            LFeng@163.com</PLeafval>
41.     </PConjunctive>
42.    </email>
43.   </student>
44.  </city>
45. </province>
46. </PNleafval>
47. <province PName="shandong"/>
48.</provinces>
```

图 1 不确定 XML 文档的示例

Fig. 1 Example diagram of uncertain XML document

4.1 不确定 XML 数据模型

本节进一步给出不确定 XML 数据模型的正式定义。不确定性 XML 数据模型的基本数据结构是根据有向图定义的，因此先介绍一些简单的概念。设 V 是一顶点的集合，$E\subset V\times V$ 是边的集合，并且 $L:V\rightarrow\Sigma$ 是一个从结点到标签集合 Σ 的映射。那么四元组 $G=(V,E,\Sigma,L)$ 是一个带标签的有向图。根据有向图的定义，我们给出定义 3。

定义 3(XML 数据树) XML 数据树是一个带标签的有向图，并且满足如下条件：

1)有且仅有一个根结点；

2)图中无环；

3)除了根结点外，每个结点仅有一个父结点；

4)每个结点都有一个标签；

5)元素的次序是有意义的。

由于在 XML 中元素的次序是有意义的，因此必须引入二元关系，在二元关系的基础之上引入偏序关系。

定义 4(二元关系) 对于给定的集合 V，$V\times V$ 的任何子集 R 称为集合 V 上的二元关系。

定义 5（偏序关系）　设 R 为非空集合 V 上的二元关系，如果 R 是自反的、反对称的和传递的，则称 R 为 V 上的偏序关系，简称偏序，记为 $\prec$。

定义 6（偏序）　集合 V 与集合 V 上的偏序关系 R 一起称为偏序集，记为 $\langle V,R\rangle$。

尽管在 XML 中元素的次序是有意义的，但属性的次序确实是无关紧要的。比如，下面的两个元素是等价的：〈person lastname＝“Woo” firstname＝“Jason”/〉，〈person firstname＝“Jason” lastname＝“Woo”/〉。因此，在 XML 模型中我们需要一种更为细致的树形概念，本文在 XML 中引入类型映射 $t:V\to T$，其中 T 是类型的集合，用来区分不同类型的结点（如元素结点、属性结点等）。

根据上述的定义和讨论，我们正式地引出不确定 XML 数据树。不确定 XML 数据树是一个 XML 数据树，并且附加 t 属性、ρ 属性和 FP 属性。其形式化定义如定义 7 所示。

定义 7（不确定 XML 数据树）　不确定 XML 数据树是一个七元组，$U=(V,E,L,t,\prec,\rho,FP)$，其中：

1）$L:V\to\Sigma$，这里 Σ 是一个标签集。对于每一个对象 $v\in V$，指定一个标签 $\xi\in\Sigma$。

2）$t:V\to T$，这里 T 是类型的集合，类型集合用于区分不同结点，如果是元素结点，那么结点将含有序的信息，如果是属性结点，那么将不含有序的信息。

3）$\prec$ 是定义在 V 上元素结点的偏序信息。对于元素结点它保留不确定 XML 数据树中的次序信息，对于属性结点对应信息为空。

4）ρ 是一个从任意结点 $v\in V$ 到本地可能性的映射。它定义当某个对象的父对象存在时该对象的孩子集的可能性分布。

5）FP 是一个从任给结点 $v\in V$ 到本地模糊概率分布的映射。它定义当某个对象的父对象存在时该对象的条件模糊概率分布。

定义 8（不确定 XML 子树）　假定 $U=(V,E,L,t,\prec,\rho,FP)$ 和 $u'=(V',E',L',t',\prec',\rho',FP')$ 是两个不确定数据树。我们称 u' 是 U 的子树，记为 $u'\angle U$，如果满足下述条件：

1）$V'\subseteq V,E'\subseteq E,\prec'\in\prec$；

2）对于任意 $v\in V'$，有 $L(v)=L'(v),t(v)=t'(v)$；

3）如果 $i\in V'$ 且 $(j,i)\in E$，那么 $j\in V'$，也就是说子树要自上而下地保持原不确定 XML 数据树中的结构；

4）对于任意的 $v\in V',\rho'\leqslant\rho$ 且 $FP'\leqslant FP$。

定义 9　一个不确定 XML 数据森林是不确定 XML 数据树的集合。

定理 1　给定一个不确定 XML 数据树的集合 $U=\{u_1,u_2,\cdots,u_n\}$，存在一个不确定 XML 数据森林与 U 匹配。

定义 10（不确定 XML 数据树同构）　设 $u_1=(V_1,E_1,L_1,t_1,\prec_1,\rho_1,FP_1)$ 和 $u_2=(V_2,E_2,L_2,t_2,\prec_2,\rho_2,FP_2)$ 是 $U=(V,E,L,t,\prec,\rho,FP)$ 的子树。u_1 和 u_2 是同构的，记为 $u_1\cong u_2$，如果存在一个双射函数 $h:V_1\to V_2$，则满足：

1）对于任意 $v\in V_1,h(v)\in V_2,L_1(v)=L_2(h(v))$ 且 $t_1(v)=t_2(h(v))$；

2）对于任意 $(u,v)\in E_1,(h(u),h(v))\in E_2$。

定理 2　不确定 XML 数据树的同构是一种等价关系。

证明：

1）自反性：考虑映射 $h:V\to V$ 满足 $\forall v\in V,h(v)=v$。h 是一个双射函数，满足 $\forall v\in V,L(v)=L(h(v)),t(v)=t(h(v))$ 且 $\forall(u,v)\in E,(h(u),h(v))\in E$。因此 h 是不确定数据树到自身的同构映射，故同构满足自反性。

2）对称性：给定两个不确定 XML 数据树 u_1 和 u_2。设 $h:V_1\to V_2$ 是从 u_1 到 u_2 的同构映射，那么 h 是一个双射函数 $h(v_1)=v_2,v_1\in V_1$，满足 $L_1(v_1)=L_2(h(v_1)),t_1(v_1)=t_2(h(v_1))$，$\forall v_1\in V_1$ 且 $(h(u),h(v))\in E_2$，$\forall(u,v)\in E_1$。因为 h 是双射，且 $h(v_1)=v_2,\forall v_1\in V_1$，所以 $h^{-1}(v_2)=v_1$，$\forall v_2\in V_2$，由 $(h(u),h(v))\in E_2,\forall(u,v)\in E_1$ 得 $(h^{-1}(u),h^{-1}(v))\in E_1,\forall(u_2,v_2)\in E_2$。

由 $L_1(v_1)=L_2(h(v_1)),t_1(v_1)=t_2(h(v_1)),\forall v_1\in V_1$ 得 $L_1(h^{-1}(v_2))=L_2(v_2),t_1(h^{-1}(v_2))=t_2(v_2),\forall v_2\in V_2$。所以可以得到从 u_2 到 u_1 的同构映射 $h^{-1}:V_2\to V_1$。也就是说 $u_1\cong u_2\Rightarrow u_2\cong u_1$。

3）传递性：给定 3 个不确定 XML 数据树 u_1,u_2 和 u_3。假设 $h_1:V_1\to V_2$ 和 $h_2:V_2\to V_3$ 分别是 u_1 到 u_2 和 u_2 到 u_3 的同构映射。

由于 $L_1(v_1)=L_2(h(v_1)),t_1(v_1)=t_2(h(v_1)),\forall v_1\in V_1$，$L_2(v_2)=L_3(h(v_2)),t_2(v_2)=t_3(h(v_2)),\forall v_2\in V_2$，则有 $L_1(v_1)=L_3(h_2(h_1(v_1))),t_1(v_1)=t_3(h_2(h_1(v_1))),\forall v_1\in V_1$ 成立。

同理，有 $(h_2(h_1(u)),h_2(h_1(v)))\in E_3,\forall(u,v)\in E_1$ 成立。因此 $h_2\circ h_1$ 是 u_1 到 u_3 的同构映射，传递性成立。

综上所述，不确定 XML 数据树的同构是一种等价关系。

4.2　Schema 定义

如图 1 所示，经典 XML 文档必须被扩展以便能够容纳不确定数据，因此我们引入了相应的结构来表示不确定数据。为了容纳这些结构，必须修改相应的 XML Schema。本节将详细介绍 XML Schema 的修改。首先给出“FP”类型的定义如下：

```
〈xs:simpleType name=“FPType”〉
〈xs:list〉
〈xs:simpleType〉
〈xs:union〉
〈xs:simpleType name=“Prob”〉
〈xs:restriction base=“Float”〉
〈xs:minInclusive value=“0”/〉
〈xs:maxInclusive value=“1”/〉
〈/xs:restriction〉
〈/xs:simpleType〉
〈xs:simpleType name=“fuzzy”〉
〈xs:restriction base=“Float”〉
〈xs:minInclusive value=“0”/〉
〈xs:maxInclusive value=“1”/〉
〈/xs:restriction〉
〈/xs:simpleType〉
〈/xs:union〉
〈/xs:simpleType〉
```

```
</xs:list>
</xs:simpleType>
```

PLeafval 元素的定义如下：

```
<xs:element name="PLeafval" type="PLeafvalType"/>
<xs:complexType name="PLeafvalType">
<xs:extension base="built-in data types">
<xs:attribute name="FP" type="FPType"
use="default" value="1"/>
</xs:extension>
</xs:complexType>
```

PNleafval 元素的定义如下：

```
<xs:element name="PNleafval" type="PNleafvalType"/>
<xs:complexType name="PNleafvalType">
<xs:sequence>
<xs:element name="original-definition"
type="originalType" minOccurs="0" maxOccurs="unbounded"/>
</xs:sequence>
<xs:attribute name="FP" type="FPType"
use="default" value="1"/>
</xs:complexType>
```

Conjunctive 元素的定义如下：

```
<xs:element name="PConjunctive" type="PConjunctiveType"/>
<xs:complexType name="PConjunctiveType">
<xs:choice>
<xs:element name="PNleafval" type="PNleafvalType"
minOccurs="1" maxOccurs="unbounded"/>
<xs:element name="PLeafval" type="PLeafvalType"
minOccurs="1" maxOccurs="unbounded"/>
</xs:choice>
</xs:complexType>
```

PDisjunctive 元素的定义如下：

```
<xs:element name="PDisjunctive" type="PDisjunctiveType"/>
<xs:complexType name="PDisjunctiveType">
<xs:choice>
<xs:element name="PNleafval" type="PNleafvalType"
minOccurs="1" maxOccurs="unbounded"/>
<xs:element name="PLeafval" type="PLeafvalType"
minOccurs="1" maxOccurs="unbounded"/>
</xs:choice>
</xs:complexType>
```

fuzzy 类型的定义如下：

```
<xs:simpleType name="fuzzy">
<xs:restriction base="Float">
<xs:minInclusive value="0"/>
<xs:maxInclusive value="1"/>
</xs:restriction>
</xs:simpleType>
```

FLeafval 元素的定义如下：

```
<xs:element name="FLeafval" type="FLeafvalType"/>
<xs:complexType name="FLeafvalType">
<xs:extension base="built-in data types">
<xs:attribute name="poss" type="fuzzy"
use="default" value="1"/>
</xs:extension>
</xs:complexType>
```

FNleafval 元素的定义如下：

```
<xs:element name="FNleafval"
type="FNleafvalType"/>
<xs:complexType name="FNleafvalType">
<xs:sequence>
<xs:element name="original-definition"
type="originalType" minOccurs="0" maxOccurs="unbounded"/>
</xs:sequence>
<xs:attribute name="poss" type="fuzzy"
use="default" value="1"/>
</xs:complexType>
```

FConjunctive 元素的定义如下：

```
<xs:element name="FConjunctive" type="FConjunctiveType"/>
<xs:complexType name="FConjunctiveType">
<xs:choice>
<xs:element name="FNleafval" type="FNleafvalType"
minOccurs="1" maxOccurs="unbounded"/>
<xs:element name="FLeafval" type="FLeafvalType"
minOccurs="1" maxOccurs="unbounded"/>
</xs:choice>
</xs:complexType>
```

FDisjunctive 元素的定义如下：

```
<xs:element name="FDisjunctive" type="FDisjunctiveType"/>
<xs:complexType name="FDisjunctiveType">
<xs:choice>
<xs:elementname="FNleafval"type="FNleafvalType"
minOccurs="1" maxOccurs="unbounded"/>
<xs:element name="FLeafval" type="FLeafvalType"
minOccurs="1" maxOccurs="unbounded"/>
</xs:choice>
</xs:complexType>
```

有了以上定义，我们就可以修改经典 XML 中元素的定义，使得每个元素都可以含有模糊概率分布。对于含有不带概率和模糊信息的叶子结点的元素，定义如下：

```
<xs:elementname="original-definition"type="originalType"minOccurs="0" maxOccurs="unbounded"/>
```

对于含有单个模糊概率分布或者模糊值的叶子结点的元素，定义如下：

```
<xs:element name="leafelement" type="leafelementType"/>
<xs:complexType name="leafelementType">
<xs:choice>
<xs:element name="FLeafval" type="FLeafvalType"/>
<xs:element name="PLeafval" type="PLeafvalType"/>
</xs:choice>
</xs:complexType>
```

对于含有概率分布或者可能性分布的叶子结点集合的元素，定义如下：

```
<xs:element name="leafelement" type="leafelementType"/>
<xs:complexType name="leafelementType">
<xs:all>
<xs:choice>
```

```
〈xs:element name="PConjunctive" type="PConjunctiveType"/〉
〈xs:element name="PDisjunctive" type="PDisjunctiveType"/〉
〈/xs:choice〉
〈xs:choice〉
〈xs:element name="FConjunctive" type="FConjunctiveType"/〉
〈xs:element name="FDisjunctive" type="FDisjunctiveType"/〉
〈/xs:choice〉
〈/xs:all〉
〈/xs:complexType〉
```

对于不含有叶子结点但是含有单个模糊概率分布或者模糊元素的元素,定义如下:

```
〈xs:element name="nonleafelement" type="nonleafelementType"/〉
〈xs:complexType name="nonleafelementType"〉
〈xs:choice〉
〈xs:element name="FNleafval" type="FNleafvalType"/〉
〈xs:element name="PNleafval" type="PNleafvalType"/〉
〈/xs:choice〉
〈/xs:complexType〉
```

对于不含有叶子结点但是含有概率分布或者可能性分布的元素,定义如下:

```
〈xs:element name="nonleafelement" type="nonleafelementType"/〉
〈xs:complexType name="nonleafelementType"〉
〈xs:all〉
〈xs:choice〉
〈xs:element name="PConjunctive" type="PConjunctiveType"/〉
〈xs:element name="PDisjunctive" type="PDisjunctiveType"/〉
〈/xs:choice〉
〈xs:choice〉
〈xs:element name="FConjunctive" type="FConjunctiveType"/〉
〈xs:element name="FDisjunctive" type="FDisjunctiveType"/〉
〈/xs:choice〉
〈/xs:all〉
〈/xs:complexType〉
```

以上方式为定义本文提出的不确定 XML 模型提供了一个标准的框架,根据模型需要,既可以引入模糊概率分布又可以引入可能性分布。

5　代数操作

代数查询语言是数据库的一种基于关系代数构造的重要查询语言,代数方法是数据库中进行查询处理的有效方法。本节介绍如何从概念上设计不确定 XML 代数的问题。如果考虑的粒度太小,如考虑 XML 中的元素作为操作对象,那么会使得代数过于复杂,难以处理,因此本文所有的操作符的粒度考虑为单个的不确定 XML 数据树或者不确定 XML 数据树的集合。

5.1　模糊概率分布和相应操作

定义 11　定义在不确定 XML 数据树 u 上的模糊概率分布为 $u(FP)=\{\pi_u(p_1)/p_1,\pi_u(p_2)/p_2,\cdots,\pi_u(p_n)/p_n\}$,其中每个概率度 p_i 伴随一个可能性度 $\pi_u(p_i)$ 并且满足 $\pi_u(p_i)\in[0,1]$,$p_i\in[0,1]$。

给定两个不确定 XML 数据树 u_1 和 u_2,其中 $u_1(FP)=\{\pi_{u_1}(p_1)/p_1,\pi_{u_1}(p_2)/p_2,\cdots,\pi_{u_1}(p_m)/p_m\}$ 且 $u_2(FP)=\{\pi_{u_2}(q_1)/q_1,\pi_{u_2}(q_2)/q_2,\cdots,\pi_{u_2}(q_n)/q_n\}$。

本文定义 4 种类型的合并操作,这些合并操作将用于不确定 XML 数据模型的相应代数操作。当然这些合并操作都定义在同构的不确定 XML 数据树上。

1)合并加操作。不确定 XML 数据树 u_1 和 u_2 的合并加操作表示为⊕,定义为 $u=u_1\oplus u_2$,其中 u,u_1,u_2 是同构的。并且有:

$$u[FP]=\{\max(\min_{i,j}(\pi_{u_1}(p_i),\pi_{u_2}(q_j))/\min_{i,j}(1,p_i+q_j))\mid 1\leqslant i\leqslant m,1\leqslant j\leqslant n\}\quad(1)$$

2)合并减操作。不确定 XML 数据树 u_1 和 u_2 的合并减操作表示为 Θ,定义为 $u=u_1\Theta u_2$,其中 u,u_1,u_2 是同构的。并且有:

$$u[FP]=\{\max(\min_{i,j}(\pi_{u_1}(p_i),\pi_{u_2}(q_j))/\max_{i,j}(p_i-q_j,0))\mid 1\leqslant i\leqslant m,1\leqslant j\leqslant n\}\quad(2)$$

3)合并乘操作。不确定 XML 数据树 u_1 和 u_2 的合并乘操作表示为⊗,定义为 $u=u_1\otimes u_2$,其中 u,u_1,u_2 是同构的。并且有:

$$u[FP]=\{\max(\min_{i,j}(\pi_{u_1}(p_i),\pi_{u_2}(q_j))/(p_i\times q_j))\mid 1\leqslant i\leqslant m,1\leqslant j\leqslant n\}\quad(3)$$

4)合并最大值操作。不确定 XML 数据树 u_1 和 u_2 的合并最大值操作表示为©,定义为 $u=u_1\copyright u_2$,其中 u,u_1,u_2 是同构的。并且有:

$$u[FP]=\{\max(\min_{i,j}(\pi_{u_1}(p_i),\pi_{u_2}(q_j))/\max_{i,j}(p_i,q_j))\mid 1\leqslant i\leqslant m,1\leqslant j\leqslant n\}\quad(4)$$

5)合并最小值操作。不确定 XML 数据树 u_1 和 u_2 的合并最小值操作表示为⊙,定义为 $u=u_1\odot u_2$,其中 u,u_1,u_2 是同构的。并且有:

$$u[FP]=\{\max(\min_{i,j}(\pi_{u_1}(p_i),\pi_{u_2}(q_j))/\min_{i,j}(p_i,q_j))\mid 1\leqslant i\leqslant m,1\leqslant j\leqslant n\}\quad(5)$$

例 1　设 u_1 和 u_2 是同构的不确定 XML 数据树,并且 $u_1[FP]=\{0.4/0.3,0.6/0.4\}$,$u_2[FP]=\{0.5/0.2,0.6/0.3\}$,应用上述定义的操作,可以得到:

1)$u[FP]=\{\max(\min(0.4,0.5)/\min(1,0.3+0.2),\min(0.4,0.6)/\min(1,0.3+0.3),\min(0.6,0.5)/\min(1,0.4+0.2),\min(0.6,0.6)/\min(1,0.4+0.3))\}=\{\max(0.4/0.5,0.4/0.6,0.5/0.6,0.6/0.7)\}=\{0.4/0.5,0.5/0.6,0.6/0.7\}$。

2)$u[FP]=\{\max(\min(0.4,0.5)/\max(0.3-0.2,0),\min(0.4,0.6)/\max(0.3-0.3,0),\min(0.6,0.5)/\max(0.4-0.2,0),\min(0.6,0.6)/\max(0.4-0.3,0))\}=\max(0.4/0.1,0.4/0.0,0.5/0.2,0.6/0.1)\}=\{0.6/0.1,0.5/0.2\}$。

3)$u[FP]=\{\max(\min(0.4,0.5)/(0.3\times 0.2),\min(0.4,0.6)/(0.3\times 0.3),\min(0.6,0.5)/(0.4\times 0.2),\min(0.6,0.6)/(0.4,0.3))\}=\{\max(0.4/0.06,0.4/0.09,0.5/0.08,0.6/0.12)\}=\{0.4/0.06,0.5/0.08,0.4/0.09,0.6/0.12\}$。

4)$u[FP]=\{\max(\min(0.4,0.5)/\max(0.3,0.2),\min(0.4,0.6)/\max(0.3,0.3),\min(0.6,0.5)/\max(0.4,0.2),\min(0.6,0.6)/\max(0.4,0.3))\}=\{0.4/0.3,0.4/0.3,0.5/0.4,0.6/0.4\}=\{0.4/0.3,0.6/0.4\}$。

5) $u[FP]=\{\max(\min(0.4,0.5)/\min(0.3,0.2),\min(0.4,0.6)/\min(0.3,0.3),\min(0.6,0.5)/\min(0.4,0.2),\min(0.6,0.6)/\min(0.4,0.3))\}=\{0.4/0.2,0.4/0.3,0.5/0.2,0.6/0.3\}=\{0.5/0.2,0.6/0.3\}$。

5.2 集合操作

本文提出4种标准的集合操作:并、交、差、笛卡尔积。集合操作将分为两种不同的类型,一种是定义在单个不确定XML数据树上的模型集合操作,另一种是定义在不确定XML数据森林上的实例集合操作。模型集合操作将改变原不确定XML数据树的某些结构,而实例集合操作不改变单个不确定XML数据树的结构,只改变不确定XML数据森林的结构。

定义12(模型并操作) 设 $u_1=(V_1,E_1,L_1,t_1,\prec_1,\rho_1,FP_1)$ 和 $u_2=(V_2,E_2,L_2,t_2,\prec_2,\rho_2,FP_2)$ 是不确定XML数据树。模型并操作的定义如下:

$u_1\bigcup u_2=(V_r,E_r,L_r,t_r,\prec_r,\rho_r,FP_r)$

其中,$V_r=V_1\bigcup V_2$,$E_r=E_1\bigcup E_2$,$L_r=L_1\bigcup L_2$,$t_r=t_1\bigcup t_2$,$\pi_r=\prec_1\bigcup\prec_2$。

根据概率理论和模糊集理论,有:

$\rho_r=\max(\rho_1,\rho_2)$

$FP_r=\{\max(\min_{x,y}(\pi_{u_1}(p_x),\pi_{u_2}(q_y))/\max_{x,y}(p_x,q_y))\mid 1\leqslant x\leqslant m,1\leqslant y\leqslant n\}$

定义13(实例并操作) 设 $U_1=\{u_{11},u_{12},\cdots,u_{1m}\}$ 和 $U_2=\{u_{21},u_{22},\cdots,u_{2n}\}$ 是不确定XML数据森林。实例并操作的定义如下:

$U_1\bigcup U_2=\{u\mid u\in U_1 \text{ or } u\in U_2\}$

图2和图3给出了执行模型并操作的示例,图中FP代表模糊概率分布,操作结果由公式 $FP_r=\{\max(\min_{x,y}(\pi_{u_1}(p_x),\pi_{u_2}(q_y))/\max_{x,y}(p_x,q_y))\mid 1\leqslant x\leqslant m,1\leqslant y\leqslant n\}$ 计算得出。图3中FDisjunctive代表可能性分布,操作的结果由公式 $\rho_r=\max(\rho_1,\rho_2)$ 计算得出。

```
〈PNleafval FP="0.4/0.2 0.6/0.4"〉
  〈name〉LiMing〈/name〉
  〈position〉engineer〈/position〉
  〈sex〉male〈/sex〉
〈/PNleafval〉
```
R_1

```
〈PNleafval FP="0.7/0.3 0.6/0.4"〉
  〈name〉LiMing〈/name〉
  〈position〉engineer〈/position〉
  〈sex〉male〈/sex〉
〈/PNleafval〉
```
S_1

```
〈PNleafval FP="0.8/0.3 0.7/0.2"〉
  〈name〉LiMing〈/name〉
  〈position〉engineer〈/position〉
〈/PNleafval〉
```
C_1

```
〈PNleafval FP="0.4/0.3 0.6/0.4"〉
  〈name〉LiMing〈/name〉
  〈position〉engineer〈/position〉
  〈sex〉male〈/sex〉
〈/PNleafval〉
```
$R_1\bigcup S_1$

图2 基于模糊概率分布的并操作

Fig. 2 Union based on fuzzy probability distribution

```
〈age〉
〈FDisjunctive〉
  〈FLeafval poss="0.4"〉23〈/FLeafval〉
  〈FLeafval poss="0.6"〉25〈/FLeafval〉
  〈FLeafval poss="0.8"〉27〈/FLeafval〉
  〈FLeafval poss="0.3"〉28〈/FLeafval〉
  〈FLeafval poss="0.2"〉29〈/FLeafval〉
〈/FDisjunctive〉
〈/age〉
```
R_2

```
〈age〉
〈FDisjunctive〉
  〈FLeafval poss="0.3"〉23〈/FLeafval〉
  〈FLeafval poss="0.7"〉25〈/FLeafval〉
  〈FLeafval poss="0.8"〉27〈/FLeafval〉
  〈FLeafval poss="0.4"〉28〈/FLeafval〉
  〈FLeafval poss="0.3"〉29〈/FLeafval〉
〈/FDisjunctive〉
〈/age〉
```
S_2

```
〈age〉
〈FDisjunctive〉
  〈FLeafval poss="0.2"〉23〈/FLeafval〉
  〈FLeafval poss="0.6"〉25〈/FLeafval〉
  〈FLeafval poss="0.7"〉27〈/FLeafval〉
  〈/FDisjunctive〉
〈/age〉
```
C_2

```
〈age〉
〈FDisjunctive〉
  〈FLeafval poss="0.4"〉23〈/FLeafval〉
  〈FLeafval poss="0.7"〉25〈/FLeafval〉
  〈FLeafval poss="0.8"〉27〈/FLeafval〉
  〈FLeafval poss="0.4"〉28〈/FLeafval〉
  〈FLeafval poss="0.3"〉29〈/FLeafval〉
〈/FDisjunctive〉
〈/age〉
```
$R_2\bigcup S_2$

图3 基于可能性分布的并操作

Fig. 3 Union based on possibility distribution

定义14(模型交操作) 设 $u_1=(V_1,E_1,L_1,t_1,\prec_1,\rho_1,FP_1)$ 和 $u_2=(V_2,E_2,L_2,t_2,\prec_2,\rho_2,FP_2)$ 是不确定XML数据树。模型交操作的定义如下:

$u_1\bigcap u_2=(V_r,E_r,L_r,t_r,\prec_r,\rho_r,FP_r)$

其中,$V_r=V_1\bigcap V_2$,$E_r=E_1\bigcap E_2$,$L_r=L_1\bigcap L_2$,$t_r=t_1\bigcap t_2$,$\prec_r=\prec_1\bigcap\prec_2$。

根据概率理论和模糊集理论,有:

$\rho_r=\min(\rho_1,\rho_2)$

$FP_r=\{\max(\min_{x,y}(\pi_{u_1}(p_x),\pi_{u_2}(q_y))/\min_{x,y}(p_x,q_y))\mid 1\leqslant x\leqslant m,1\leqslant y\leqslant n\}$

定义15(实例交操作) 设 $U_1=\{u_{11},u_{12},\cdots,u_{1m}\}$ 和 $U_2=\{u_{21},u_{22},\cdots,u_{2n}\}$ 是不确定XML数据森林。实例交操作可以定义为:

$U_1\bigcap U_2=\{u\mid u\in U_1 \text{ and } u\in U_2\}$

图4给出了执行模型交操作的一个示例,图中FP代表模糊概率分布,操作的结果由公式 $FP_r=\{\max(\min_{x,y}(\pi_{u_1}(p_x),\pi_{u_2}(q_y))/\min_{x,y}(p_x,q_y))\mid 1\leqslant x\leqslant m,1\leqslant y\leqslant n\}$ 计算得出。FDisjunctive代表可能性分布,操作的结构由公式 $\rho_r=\min(\rho_1,\rho_2)$ 计算得出。

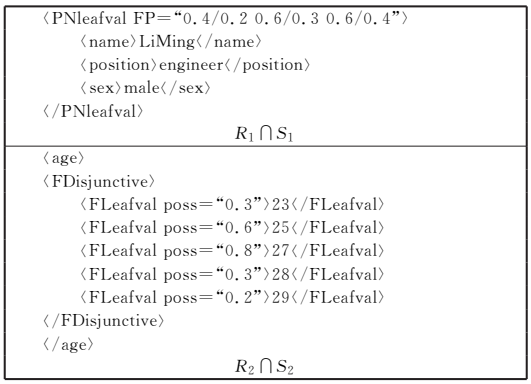

图 4　不确定交操作

Fig. 4　Uncertain intersection

定义 16(模型差操作)　设 $u_1=(V_1,E_1,L_1,t_1,\prec_1,\rho_1,FP_1)$和 $u_2=(V_2,E_2,L_2,t_2,\prec_2,\rho_2,FP_2)$是不确定 XML 数据树。模型差操作的定义如下:

$$u_1-u_2=(V_r,E_r,L_r,t_r,\prec_r,\rho_r,FP_r)$$

其中,$V_r=V_1-V_2$,$E_r=E_1-E_2$,$L_r=L_1-L_2$,$t_r=t_1-t_2$,$\prec_r=\prec_1-\prec_2$。

根据概率理论和模糊集理论,有:

$$\rho_r=\min(\rho_1,\rho_2')\text{,}\rho_2'=1-\rho_2$$

$$FP_r=\{\max(\min_{x,y}(\pi_{u_1}(p_x),\pi_{u_2}(q_y))/\max_{x,y}(p_x-q_y,0))\mid 1\leqslant x\leqslant m,1\leqslant y\leqslant n\}$$

定义 17(实例差操作)　设 $U_1=\{u_{11},u_{12},\cdots,u_{1m}\}$和 $U_2=\{u_{21},u_{22},\cdots,u_{2n}\}$是不确定 XML 数据森林。实例差操作的定义如下:

$$U_1-U_2=\{u\mid u\in U_1 \text{ and } u\notin U_2\}$$

图 5 给出了模型差操作的一个示例,图中 FP 代表模糊概率分布,操作的结果由公式 $FP_r=\{\max(\min_{x,y}(\pi_{u_1}(p_x),\pi_{u_2}(q_y))/\max_{x,y}(p_x-q_y,0))\mid 1\leqslant x\leqslant m,1\leqslant y\leqslant n\}$计算得出。FDisjunctive 代表可能性分布,操作的结果由公式 $\rho_r=\min(\rho_1,\rho_2')\text{,}\rho_2'=1-\rho_2$计算得出。

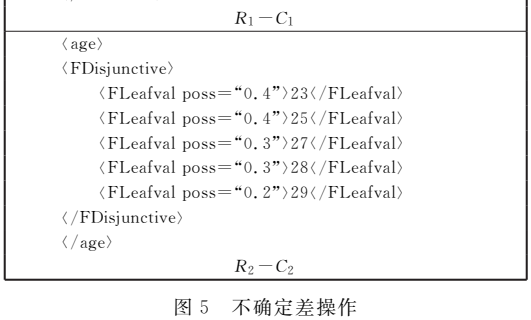

图 5　不确定差操作

Fig. 5　Uncertain difference

定义 18(模型笛卡尔积)　设 $u_1=(V_1,E_1,L_1,t_1,\prec_1,\rho_1,FP_1)$和 $u_2=(V_2,E_2,L_2,t_2,\prec_2,\rho_2,FP_2)$是不确定 XML 数据树。模型笛卡尔积的定义如下:

$$u_1\infty u_2=(V_r,E_r,L_r,t_r,\prec_r,\rho_r,FP_r)$$

其中,$V_r=V_1\infty V_2$,$E_r=E_1\infty E_2$,$L_r=L_1\infty L_2$,$t_r=t_1\infty t_2$,$\prec_r=\prec_1\infty\prec_2$。根据概率理论和模糊集理论,有:

$$\rho_r=\min(\rho_1,\rho_2)$$

$$FP_r=\{\max(\min_{x,y}(\pi_{u_1}(p_x),\pi_{u_2}(q_y))/(p_x\times q_y))\mid 1\leqslant x\leqslant m,1\leqslant y\leqslant n\}$$

定义 19(实例笛卡尔积)　设 $U_1=\{u_{11},u_{12},\cdots,u_{1m}\}$和 $U_2=\{u_{21},u_{22},\cdots,u_{2n}\}$是不确定 XML 数据森林。实例笛卡尔积的定义如下:

$$U_1\infty U_2\in\{(u_1,u_2)\mid u_1\in U_1 \text{ and } u_2\in U_2\}$$

定理 3　假定 u_i 和 u_j 是不确定 XML 数据树,则有:

1)$u_i\cup u_j$ 是不确定 XML 数据树;

2)$u_i\cap u_j$ 是不确定 XML 数据树;

3)u_i-u_j 是不确定 XML 数据树;

4)$u_i\infty u_j$ 是不确定 XML 数据树。

定义 20　设 $u_1=(V_1,E_1,L_1,t_1,\prec_1,\rho_1,FP_1)$,$u_2=(V_2,E_2,L_2,t_2,\prec_2,\rho_2,FP_2)$,$\cdots$,$u_n=(V_n,E_n,L_n,t_n,\prec_n,\rho_n,FP_n)$是不确定 XML 数据树。本文中反复利用定理 3 得到的不确定 XML 数据树 $U=(V,E,L,t,\prec,\rho,FP)$是 $u_1,u_2,\cdots,u_n$ 的重构。

5.3　选择操作

在不确定 XML 模型中,选择操作可以根据给定的不确定 XML 树模式来筛选不确定 XML 树。选择操作接受不确定 XML 数据树作为输入,输出与给定的模式相匹配的不确定 XML 数据树的子树集合。

定义 21(不确定 XML 树模式)　不确定 XML 树模式是一个四元组 $P=(V_P,E_P,F_V,R_V)$,其中

1)V_P 是顶点的有限集合。

2)E_P 是有向边的有限集合。

3)F_V 是定义在 V_P 上的函数,对于给定的顶点 $v\in V_P$,$F_V(v)$指定了一个谓词,该谓词是原子谓词的布尔组合。

4)$R_V:V_P\rightarrow re(R)$是定义在 V_P 上的函数。对于每一个 $v\in V_P$,$re(R)$是顶点的路径表达式。其中 R 可以通过表达式 $R::=e\mid R_1\cdot R_2\mid R_1\mid R_2\mid R^+$递归地构造。

定义 22(选择)　设 $U=(V,E,L,t,\prec,\rho,FP)$是不确定 XML 数据树。对于一个给定的不确定 XML 树模式 $P=(V_P,E_P,F_V,R_V)$,有如下定义:

$$\sigma_P(U)=\{u\mid u=\kappa(P,U)\}$$

其中,u 是 U 的子树,函数 $\kappa(P,U)$可以根据给定的模式 P 匹配不确定 XML 树 U。所有与不确定 XML 树 U 匹配的结果构成了一个不确定 XML 森林。

定理 4　$\sigma_P(U)$是一个不确定 XML 森林。

证明:根据定义 22,选择操作将返回一个集合,集合中的每个元素都是一个与给定的不确定 XML 树匹配的不确定 XML 子树,根据定义 9 知 $\sigma_P(U)$是不确定 XML 森林。

5.4　投影操作

在关系数据库中,投影和选择是一对正交的操作,选择从行的角度出发,而投影从列的角度出发。然而,在不确定 XML 模型中,这两个操作的区分不是很明显。不确定 XML 模型中的选择操作生成一个不确定 XML 森林,选择操作并没有保留原树中的层次结构,而投影操作将保留原不确定 XML 树中的层次结构。因此,投影操作可以视为不确定 XML 数据树中结点的消除而不是结点的指定,并且在结点消除产生的子结构中保留原不确定 XML 数据树中的层次关系。

定义 23(投影)　假定 $U=(V,E,L,t,\prec,\rho,FP)$是不确

定 XML 数据树，Φ 是一个投影函数，$P=(V_P, E_P, F_V, R_V)$ 是给定的树模式，PL 是一个投影列表。那么投影操作可以定义为：

$$\pi_{P,PL}(U)=\{t \mid t=\Phi(P,PL,U)\}$$

5.5 连接操作

连接操作基于树模式连接不确定 XML 数据树，它是笛卡尔积和选择操作的混合操作。选择条件是比较第一棵不确定数据树和其他不确定数据树的属性。

定义 24(连接) 假定 $U_1=(V_1, E_1, L_1, t_1, \prec_1, \rho_1, FP_1)$ 和 $U_2=(V_2, E_2, L_2, t_2, \prec_2, \rho_2, FP_2)$ 是不确定 XML 数据树。$P=(V_P, E_P, F_V, R_V)$ 是给定的树模式。那么连接操作可以定义为：

$$U_1 \otimes_P U_2=\{t \mid t=\sigma_P(U_1 \infty U_2)\}$$

其中，P 是匹配 $(U_1 \infty U_2)$ 的不确定 XML 树模式，满足 F_V 中至少有一个谓词是 $L_1(v_1)=L_2(v_2)$，这里 v_1 是 U_1 中的点，v_2 是 U_2 中的点。因此，$L_1(v_1)$ 代表 U_1 中的结点标签，$L_2(v_2)$ 代表 U_2 中的结点标签。

结束语 本文提出了一个不确定 XML 模型来管理 XML 中的两种不确定信息(概率不确定和模糊不确定)。我们在没有改变当前 XML 标准的情况下，扩展了经典 XML 处理不确定信息的能力。基于概率理论和模糊集理论，我们为元素赋予了一个模糊概率分布，用来表示在父元素存在的情况下，当前元素存在的条件概率。此外，本文可以为元素的值赋予可能性分布，用来描述元素值之间的特定关系。为了能够使经典 XML 含有表达不确定的相关元素和属性，我们用 XMLS 定义了相关元素和属性的结构。本文定义的模型既能与经典 XML 相容，又能表达完整的不确定信息。

此外，本文提出了一个形式化的不确定 XML 模型的代数框架，它接受不确定 XML 树集合作为输入，以不确定树集合为输出。本文提出的代数能支持不确定数据的查询和优化。

参 考 文 献

[1] Extensible Markup Language (XML)[OL]. https://www.w3.org/TR/2008/REC-xml-20081126/.

[2] HUNG E, GETOOR L, SUBRAHMANIAN V S. Probabilistic interval XML [C]// International Conference on Database Theory. 2003: 361-377.

[3] NIERMAN A, JAGADISH H V. ProTDB: Probabilistic data in XML[C]// Proceedings of the 28th International Conference on Very Large Databases(VLDB'02). 2002: 646-657.

[4] MA Z, YAN L. Modeling fuzzy data with XML: A survey [J]. Fuzzy Sets and Systems, 2016, 301: 146-159.

[5] YAN L, MA Z M, LIU J. Fuzzy data modeling based on XML schema[C]// Proceedings of the 2009 ACM symposium on Applied Computing. 2009: 1563-1567.

[6] MA Z M, YAN L. Fuzzy XML data modeling with the UML and relational data models [J]. Data & Knowledge Engineering, 2007, 63(3): 972-996.

[7] MA Z M, LIU J, YAN L. Fuzzy data modeling and algebraic operations in XML[J]. International Journal of Intelligent Systems, 2010, 25(9): 925-947.

[8] GETTA J R. An XML algebra for online processing of XML documents [C]// Proceedings of International Conference on Information Integration and Web-based Applications & Services. ACM, 2013: 503.

[9] JAGADISH H V, LAKSHMANAN L V, SRIVASTAVA D, et al. TAX: A tree algebra for XML[C]// International Workshop on Database Programming Languages. 2001: 149-164.

[10] BURATTI G, MONTESI D. A data model and an algebra for querying XML documents[C]// 17th International Workshop on Database and Expert Systems Applications. 2006: 482-486.

[11] CHE D, SOJITRAWALA R M. DUMAX: a dual mode algebra for XML queries[C]// Proceedings of the 2nd International Conference on Scalable Information Systems. 2007: 52.

[12] MA Z M, ZHANG F, YAN L. Fuzzy information modeling in UML class diagram and relational database models [J]. Applied Soft Computing, 2011, 11(6): 4236-4245.

[13] MA Z M, ZHANG F, YAN L, et al. Extracting knowledge from fuzzy relational databases with description logic[J]. Integrated Computer-Aided Engineering, 2011, 18(2): 181-200.

[14] LAKSHMANAN L V, LEONE N, ROSS R, et al. Probview: A flexible probabilistic database system[J]. ACM Transactions on Database Systems (TODS), 1997, 22(3): 419-469.

[15] YAN L, MA Z M. A fuzzy probabilistic relational database model and algebra [J]. International Journal of Fuzzy Systems, 2013, 15(2): 244-253.

[16] EITER T, LU J J, LUKASIEWICZ T, et al. Probabilistic object bases[J]. ACM Transactions on Database Systems (TODS), 2001, 26(3): 264-312.

[17] YAN L, MA Z M. Conceptual design of object-oriented databases for fuzzy engineering information modeling [J]. Integrated Computer-Aided Engineering, 2013, 20(2): 183-197.

[18] CAO T H, ROSSITER J M. A deductive probabilistic and fuzzy object-oriented database language[J]. Fuzzy Sets and Systems, 2003, 140(1): 129-150.

[19] CAO T H, NGUYEN H. Uncertain and fuzzy object bases: a data model and algebraic operations[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2011, 19(2): 275-305.

[20] YAN L, MA Z. A Probabilistic Object-Oriented Database Model with Fuzzy Measures[M]// Advances in Probabilistic Databases for Uncertain Information Management. 2013: 23-38.

[21] YAN L, MA Z M. Comparison of entity with fuzzy data types in fuzzy object-oriented databases[J]. Integrated Computer-Aided Engineering, 2012, 19(2): 199-212.

[22] XML Schema[OL]. W3C Recommendation. https://www.w3.org/TR/2004/REC-xmlschema-1-20041028/.

HU Lei, born in 1993, postgraduate. His main research interests include data and knowledge engineering.

YAN Li, born in 1964, Ph.D, professor, is a member of China Computer Federation. Her main research interests include data and knowledge engineering.