概率XML数据管理技术研究进展

王建卫[1,3] 郝忠孝[1,2]
(哈尔滨理工大学计算机科学与技术学院 哈尔滨 150080)[1]
(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)[2]
(东北林业大学信息与计算机工程学院 哈尔滨 150040)[3]

摘 要 随着网络应用的快速发展,XML数据已大量存在于当前的信息社会,使得XML类型的数据成为当前主流的数据形式,并已经成为Internet中进行数据交换和表示事实上的标准。由于客观世界的复杂性,不确定性是数据常见的内在属性,因此不确定的信息是普遍存在的。通常不确定信息以概率值的形式在XML文件(称为概率XML文件)中表示,因此,研究表示和处理概率XML数据将成为一个新的研究领域。自2001年以来,概率XML数据管理技术取得了一系列研究成果。从概率XML数据模型、PXML代数、查询、原型系统等几个方面综述了概率XML数据管理的研究进展,讨论了目前存在的主要问题和需要进一步研究的方向。

关键词 概率XML数据,数据模型,PXML代数,查询,原型系统
中图法分类号 TP393 **文献标识码** A

Survey of Research on Probabilistic XML Data Management Techniques

WANG Jian-wei[1,3] HAO Zhong-xiao[1,2]
(School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)[1]
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)[2]
(College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China)[3]

Abstract With the rapid development of network application, a large amount of XML data have existed, so the style of XML data becomes the primary data and the standard style of data exchanging and representation on Internet. Because of the complexity of external world, the uncertainty is the common internal attribute of data, the uncertain information universally exists. Usually the uncertain information can be represented as the probability values in XML document (probabilistic XML document), so the research ways of representing and processing the probabilistic XML data will be a new research field. Since 2001, a series of research achievements of the probabilistic XML data management have been obtained. The paper surveyed the research techniques of the probabilistic XML data management including the probabilistic XML data model, the PXML algebra, query and the prototype systems. The existing problems in the current research work and the new research issues were also discussed.

Keywords Probabilistic XML data, Data model, PXML Algebra, Query, Prototype system

1 引言

近年来出现的XML(eXtensible Markup Language,可扩展的标记语言)是一套定义语义标记的规范,是半结构化数据的一种特殊表现形式,正如World Wide Web Consortium(W3C)的XML工作组的定义,"XML是一种通用的标记语言,它能够标记多种不同数据源的信息内容,包括结构化和半结构化文件、关系数据库和对象库等。"随着网络应用的快速发展,符合XML规范的数据(称为XML数据)已大量存在于当前的信息社会,使得XML类型的数据成为主要的数据形式之一,已经成为Internet中进行数据交换和表示事实上的标准[1]。

应用XML文件管理数据面临的问题之一是不得不面对数据的不确定问题。由于客观世界的复杂性,不确定性是数据常见的内在属性,不确定的信息是普遍存在的,通常不确定信息可以用概率值的形式在XML文件中表示。因此,可以把概率XML文件定义为能够以概率值表示不确定信息的普通XML文件。从2001年起,应用概率XML文件管理不确定数据的技术已经取得了一些研究成果[2,3]。本文将从概率XML数据模型、概率XML代数和查询、原型系统等各个方面综述概率XML数据管理技术的研究进展。

2 概率XML数据模型分析

概率XML数据管理的首要问题是如何在普通XML文

到稿日期:2009-01-05 返修日期:2009-03-02 本文受黑龙江省自然科学基金(F200601)资助。
王建卫(1973—),女,博士研究生,主要研究方向为数据库理论及应用,E-mail:jwwang2007@163.com;郝忠孝(1940—),男,教授,博士生导师,主要研究方向为数据库理论及应用。

件中以概率的形式表示不确定信息。通常，概率 XML 文件(可简记为 PXML 文件)指的是普通 XML 文件空间的概率分布，PXML 文件的数据模型对于上层的查询处理和优化有着非常重要的影响。因此，要实现 PXML 文件的管理，必须解决 PXML 文件的数据模型问题。

2.1 概率 XML 数据模型的分类

通常概率数据的存在往往是有约束条件的。PXML 文件的数据模型根据是否考虑满足一定的概率约束条件可分为两大类：无约束条件的概率 XML 数据模型和有约束的概率 XML 数据模型。一般来说，在无约束条件的概率 XML 数据模型的基础上加上具体文件中不同数据之间的概率值的限制条件，就构成了有约束的概率 XML 数据模型[2-8]。因此，无约束条件的概率 XML 数据模型是有约束的概率 XML 数据模型的基础。

2.2 概率属性节点的类型

通常 XML 文件可表示为 XML 文件树，因此，在普通 XML 树中通过添加表示概率属性节点的方法构建概率 XML 树是描述 PXML 文件最常用的方法。在 PXML 树中，根据概率节点对应的概率值的关系，可以把概率属性节点的分布类型分为 4 种：

一是独立类型节点 ind，独立节点在 PXML 树出现的概率是独立的，不受其他节点的影响[2,4]。

二是互斥类型节点 mux，互斥节点在 PXML 树只能出现一个其他节点不出现的节点，或者全都不出现[2,4]。

三是孩子节点组合类型节点 exp[9,10]，exp 节点有多个孩子节点，选择不同的孩子节点组成孩子节点的集合，孩子集合的不同子集 $w_1,\cdots,w_l$ 的概率值分别为 $p^v(w_1),\cdots,p^v(w_l)$，要求 $\sum_{i=1}^{l} p^v(w_i)=1$。

四是外部变量驱动类型节点 cie[11,12]，cie 节点的存在是由独立的外部事件变量 $e_1,\cdots,e_m$ 决定的，对于每个事件 e_i，$(1\leqslant i\leqslant m)$，由已知的 e_i 为真的概率 $p(e_i)$，计算该节点的孩子节点的存在概率。

2.3 PXML 文件的数据模型

文献[2-13]中以上述的分布节点形式表示概率值，提出和应用了以有向图和树为基础的 PXML 模型作为 PXML 文件的数据模型。文献[2]做了概率 XML 数据库研究的第一步，根据不确定数据的特点提出了概率 XML 模型 $PXML^{\{ind,mux\}}$，文献[4]沿用了这一模型，在该模型中，引入了用来说明在 XML 文件中特定的元素不确定性的概率属性节点 Prob(在 XML 文件指定的位置)。Prob 根据(兄弟)节点的概率值之间的关系用节点 Dist 表示，分为互斥型 mux 的和独立型 ind 两种类型，独立节点就是其出现的概率是独立的，不受其他节点的影响，互斥节点就是只能出现一个其他节点不出现的节点，或者全都不出现。该模型能有效地表示概率 XML 文件，缺点是概率 XML 文件的格式与普通 XML 文件有一定的区别，概率 XML 文件中应有节点 Dist 的类型和 Prob 的概率值的说明，因此为适应表示概率数据，源 XML 文件的 DTD 必须作相应的修改，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[5-8]研究了在 XML 文件中不确定信息的数据集成方法，把节点分为概率节点、可能节点和 XML 节点三种，定义概率树 $PT=(T,kind,prob)$，其中 T 为数据树，$kind$ 函数指定节点的类型，$prob$ 函数指定可能节点的存在概率。兄弟可能节点是 mux 类型的分布节点，因此，概率 XML 文件集成的结果是用概率树 $PXML^{\{ind,mux\}}$ 表示概率 XML 文件。这种集成方法为在 XML 文件中表示不确定信息提出了非常实用的方法，能有效地把多个 XML 文件表示为一个概率 XML 树，优点是保持 XML 文件的格式不变，则不必修改 DTD，缺点是集成的结果中可能节点是冗余信息，对于存储和查询处理时间有影响。

PXML 文件除了可以表示为以树为基础的数据模型，还可以表示为以有向图为基础的数据模型。文献[9]提出一种以有向图为基础的概率半结构数据(Probabilistic Semistrutured Data，简记为 PSD)模型 $PXML^{\{exp\}}$，分布节点的类型为 exp。该模型中把概率实例定义为半结构实例及其对应的概率值，所有的半结构实例构成一个可能世界(possible world，简记为 PW)集合。文献[10]提出了第一个以概率区间值表示 PXML 的正式的数据模型。这种方法的优点是提供了一种新的以有向图为基础的 PXML 文件建模方式。由于通常把 XML 文件描述为文件树，因此要应用有向图描述 PXML 文件的存储、查询方法，必须要做相应的修改。因此，数据模型的主要工作还是以 PXML 树为主。

文献[11]提出了一种简单的概率树(Simple Probabilistic tree，简记为 sp_tree)模型 $PXML^{\{ind\}}$(树中分布节点类型为 ind)和概率树(Probabilistic tree，简记为 prob_tree)模型 $PXML^{\{cie\}}$(树中分布节点类型为 cie)。文献[12]中应用了前一种模型，并分析了查询的复杂度。

上述数学模型共同的特点是只考虑了在 XML 中如何以概率的形式表示、记录不确定信息，而没有考虑概率数据之间的限制条件。文献[13]提出了 PXML 文件的数据模型 PXDB (probabilistic XML database，概率 XML 数据库)，PXDB 定义为 $\tilde{D}=(\tilde{P},C)$，其中 d 为 PXML 文件 $\tilde{P}$ 的可能世界集合，C 为约束条件的集合，P 是概率空间 $\tilde{P}$ 的随机文件，而且 $d\vdash C$，$\Pr(P\vdash d)>0$。由 PXDB 的定义可知，PXDB 为包含必须满足的一组约束条件的 PXML 文件实例的概率子空间。与能表达概率依赖的已知模型相比，PXDB 的优点在于约束的重要性首先在于保持数据的完整性，其次约束也表达了概率数据之间的依赖关系。

3 PXML 代数

按照关系代数的理论，XML 查询代数是对遵循一定数据模型的 XML 文件集合的操作集，那么 PXML 代数的实现方式主要有两种思路：一是扩展 XML 代数(extended XML Algebra，简记为 e_ XML Algebra)的方法，在已建立的 XML 代数的基础上增加概率数据的操作；二是在关系代数的基础上实现概率 XML 的操作，这种方法适用于以使能 XML 数据库为基础实现概率 XML 数据的管理。目前，关于 PXML 代数的主要工作集中在第二个方面。

3.1 e_ XML Algebra

扩展 XML 代数的方法在文献[2]中首先得到了应用，代

数系统以 TIMBER 的实现为基础，在查询解析器和查询执行器中增加概率 XML 数据管理的函数。这种方法直接应用了 TAX 代数系统，实现比较简单，但修改和扩展 XML 代数的方式不灵活。

3.2 SP_algebra

文献[14-19]针对不同的 PXML 数据模型说明了基于关系代数的 PXML 代数 SP_algebra，操作对象为半结构概率对象 SPO，SP_algebra 包括标准的集合操作（如并、交和差等），还扩展了关系代数运算（如选择、投影、笛卡儿积和连接等）以便管理概率数据，并定义了一个新的与概率计算有关的操作符号边缘操作（conditionalization）。这种方法的操作过程为首先 SPO 转换为关系表，然后对关系表应用 PXML 代数 SP_algebra，所作的工作没有涉及定义聚集函数和 PXML 代数的完备性证明。

文献[9]定义了针对数据模型 PSD 的代数运算，具体操作有投影（如祖先投影）、选择（如根据对象的选择、根据值的选择）和笛卡尔积。其中投影操作符包括祖先投影、后裔投影和单一节点投影；选择操作符根据对象选择和值选择两种类型的选择条件可分为对象选择运算和值选择运算。

4 PXML 查询方案分析

由于 PXML 是特殊的 XML 文件，PXML 查询既有 XML 数据库查询的特点，也有概率关系数据库查询的特点。

4.1 查询语义

根据查询结果的性质，PXML 的查询语义可分为两种，一是面向对象的语义，查询结果为概率 XML 树的节点对象和节点属性；二是面向数值的语义，查询结果为节点的标签。在第二种语义中，查询结果为概率数值时，概率计算的准确性要依赖于分布节点类型的判断[13]。

4.2 查询方案

目前研究者们已经提出了很多 XML 数据库的查询方法，其中 twig 模式查询是 XML 数据库查询的有效方法之一，文献[20,21]在 PXML 中沿袭了 twig 模式查询思路。twig 模式查询可以在 PXML 上进行，因此，PXML 的查询方案的研究可以归纳为以下两个方向。

一是在 PXML 文件对应的所有随机文件上执行查询。一般地，一个 PXML 文件对应于实际中的很多随机文件，根据 PXML 文件的集成方法可以列出在 PXML 文件对应的所有随机 XML 文件及其概率，然后在所有的随机 XML 文件上执行查询[22]。显然，这是一种比较简单的方法，但由于需要列出 PXML 文件对应的所有随机 XML 文件，查询效率显然不高。

二是直接在概率 XML 文件上执行查询[2,23]。这是可行的查询方案，查询的关键问题是如何选择一组需要的概率节点。文献[15]和[16]中采用的实现方式是在较成熟的 XML 原型系统上增加关于概率操作的函数。文献[21]从查询语义的角度研究了 PXML 数据库中 twig 匹配的查询方式，根据查询执行方式的不同或者返回结果的情况将查询的语义分为 3 种：一是布尔查询语义，就是检查可满足性的查询，即对于给定的文件和查询，确定文件中是否存在满足条件的解；二是完全语义查询语义，就是对于给定的文件和查询，找出所有完全匹配查询条件的解；三是不完全语义查询，就是对于给定的文件和查询，找出不完全满足查询的要求的解。在这 3 种语义查询中，求解匹配的最大子树是查询过程的核心问题。不足之处在于文中对子树的粒度问题没有深入研究，不能保证返回结果的语义完整性，返回的结果中可能含有无关信息和对用户意义不大的无效信息。

从以上的分析可以看出，由于概率 XML 数据库的查询与基于 XML 的概率数据的表示方式有关，随着新的基于 XML 的概率数据的表示方法的提出，PXML 的查询方案不应仅局限于所有的随机文件查询和概率 XML 文件查询两种方案。

5 实现的原型系统分析

自 2001 年以来研究者们开始设计概率 XML 数据库的原型系统，并在原型系统上作了一系列的查询工作。这里介绍 3 种较成熟的概率 XML 数据库系统（ProTDB，PEPX 和 SPDBMS）。

5.1 ProTDB 系统结构分析

文献[2]中实现的 ProTDB 系统是第一个概率 XML 数据库系统，该系统构建在 Timber 的基础上，能够实现单一节点查询和连接查询。ProTDB 系统的结构如图 1 所示。

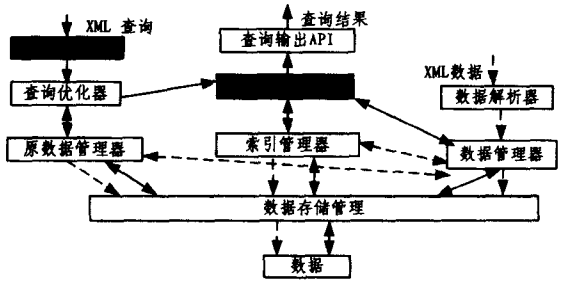

图 1 ProTDB 系统的基本结构

在图 1 中，——→表示程序流程；……→表示数据装入；■表示增加的 ProTDB 函数。该系统与 Timber[24]相比，在查询解析器和查询执行器中增加了概率 XML 文件的管理函数。

5.2 PEPX 系统结构分析

文献[23]中提供了另外一种设计概率 XML 数据库系统的思路，PEPX 是基于概率编码的概率 XML 数据库系统，由解析模块、解释器、查询解析器等组成，其中解析模块的作用是概率 XML 文件解析为关系数据库，模式为{name, value, probability, me, start, end, level, pid}，主键为{start}，以{name, start}创建索引，属性{start, end, level}用来检测两个 XML 节点的祖先-后裔关系；解释器是用户接口，接收来自用户的查询或更新请求，调用查询解析器用来解析查询或更新语句；查询解析器产生输入到 SQL 转换器的查询树，SQL 语句通过关系查询引擎执行。实现的 PEPX 系统要优于 ProTDB 系统，不同之处是以概率链的形式记录概率属性，而在 ProTDB 中记录的是条件概率。PEPX 的结构如图 2 所示。

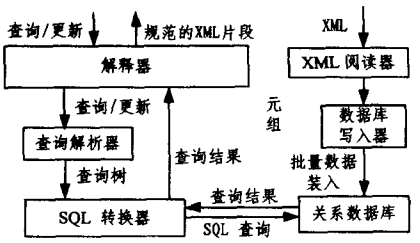

图 2 PEPX 系统的基本结构

5.3 SPDBMS 系统结构分析

文献[14,22]在关系数据库之上设计半结构概率数据库原型系统的框架，并实现了原型系统 SPDBMS，该系统结构的核心部分是 SPDBMS 应用服务器，用于处理来自不同客户要求的查询结果。SPDBMS 的结构如图 3 所示。

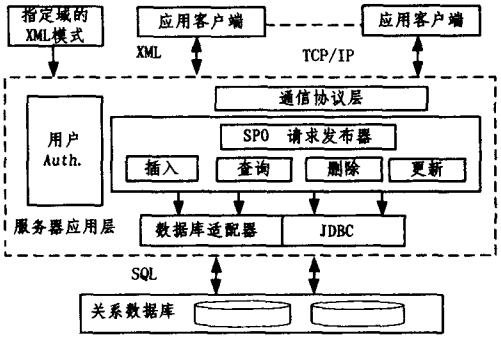

图 3 SPDBMS 系统的基本结构

目前，SPDBMS 版本是基于关系数据库实现的，首先把 SPO 表示为关系表，把 SP-Algebra 操作符转换为 SQL 段的序列，下一步应研究在原生 XML 数据库系统上如何将其转换为 SP-Algebra，以支持 XQuery 语句。

综上，文献中实现的原型系统分别依赖于关系数据库系统、原生 XML 数据库系统，因此开发设计真正的概率 XML 数据库系统仍然是艰巨的研究任务。

结束语 本文从 PXML 文件的数据模型、代数系统、查询和实现的原型系统等几个方面综述了 PXML 数据库的研究进展，后续的研究工作为下述几个方面：

一是确定在普通 XML 文件中记录概率数据的标准，这是解析文件的基础；

二是分布节点的 4 种类型是否能够表示 PXML 文件中节点之间的所有关系，是否需要增加分布节点新的类型；

三是现有的 PXML 代数系统依赖于关系代数或 XML 代数，定义哪些基本的代数操作使 PXML 代数完备化，这也是查询优化工作的前提和基础；

四是在 PXML 中查询结果必然与概率有关，而且列出所有的结果会增加查询算法的复杂度，因此查询结果的取舍问题也是研究内容之一。

参 考 文 献

[1] http://www.w3.org/TR/

[2] Nierman, Jagadish H V. ProTDB: Probabilistic data in XML[C]//Proceedings of the 28th VLDB Conference. Hong Kong, China, 2002

[3] Zhao Wenzhong, Dekhtyar A, Goldsmith J. Representing Probabilistic Information in XML[M]. Lexington: University of Kentucky Department of Computer Science, 2003

[4] Kimelfeld B, Sagiv Y. Matching twigs in probabilistic XML[C]//VLDB'07. Vienna, Austria, 2007

[5] van Keulen M, de Keijzer A, Alink W. A probabilistic XML approach to data integration[C]//International Conference on Data Engineering, Proceedings-21st International Conference on Data Engineering(ICDE 2005). 2005: 459-470

[6] de Keijzer A. Probabilistic XML in Information Integration[C]//Proceedings of the VLDB 2006. Ph. D. Workshop Seoul. Rep of Korea, 2006

[7] de Keijzer A, van Keulen M. User Feedback in Probabilistic Integration[C]//DEXA'07. 18th International Workshop on Database and Expert Systems Applications. 2007: 377-381

[8] Bos W. Probabilistic XML Integration [C]// the sequel. 7 th Twente Student Conference on IT. Enschede, 2007

[9] Hung E, Getoor L, Subrahmanian V S. PXML: A Probabilistic Semistructured Data Model and Algebra[C]//Procedings of the 19th International Confernce on Data Engineering(ICDE'03). 2003: 467-478

[10] Hung E, Getoor L, Subrahmanian V S. Probabilistic interval XML[C]//Proceedings of the 9th International Conference on Database Theory. Lecture Notes In Computer Science. 2572. 2003: 361-377

[11] Abiteboul S, Senellart P. Querying and updating probabilistic information in XML[C]//EDBT 2006. Volume 3896. 2006: 1059-1068

[12] Senellart P, Abiteboul S. On the complexity of managing probabilistic XML data[C]//Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. Beijing, China, 2007: 283-292

[13] Cohen S, Kimelfeld B, Sagiv Y. Incorporating Constraints in Probabilistic XML [C] // Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems. Vancouver, Canada, 2008: 109-118

[14] Zhao Wenzhong, Dekhtyar A, Goldsmith J. A Framework for Management of Semistructured Probabilistic Data[J]. Journal of Intelligent Information Systems, 2005, 25(3): 293-332

[15] Zhao Wenzhong, Dekhtyar A, Goldsmith J. Databases for Interval Probabilities[J]. International Journal of Intelligent Systems, 2004, 19(9): 789-815

[16] Magnani M, Montesi D. Management of interval probabilistic data[J]. Acta Informatica, 2008 (45): 93-130

[17] Dekhtyar A, Mathias K K, Gutti P. Structured Queries for Semistructured Probabilistic Data[C]//TDM'2006

[18] Dekhtyar A, Goldsmith J, Hawkes S R. Semistructured Probabilistic Databases[C]//Proc. Statistical and Scienti Database Management Systems. 2001

[19] Hung E. Managing uncertainty and ontologies in databases[D]. University of Maryland at College Park College Park, MD, USA, 2005

[20] Kimelfeld B, Kosharovsky Y, Sagiv Y. Query Efficiency in Probabilistic XML Models[C]//Proceedings of the 2008 ACM SIGMOD international conference on management of data. Vancouver, Canada, 2008

[21] Kimelfeld B, Sagiv Y. Matching Twigs in Probabilistic XML[C]//VLDB '07. Vienna, Austria, 2007

[22] Hung E, Subrahmanian V S. Managing uncertainty and ontologies in databases[D]. University of Maryland at College Park, 2005

[23] Li Te, Shao Qihong, Chen Yi. PEPX: A Query-Friendly Probabilistic XML Database[C]//Proceedings of the 15th ACM International Conference on Information and Knowledge Management. CIKM '06. ACM Press, 2006: 848-849

[24] Paparizos S, Jagadish H V. The Importance of Algebra for XML Query Processing[C]//EDBT 2006 Workshops. LNCS 4254. 2006: 126-135