

XML 概率数据库中空值处理方法研究

周智增 王新军

(山东大学计算机科学与技术学院 济南 250101)

摘 要 不确定性和不完全性是现实世界对数据库的挑战,空值、XML 和概率数据库三者的结合可以更好地处理数据,但同时也增加了数据库的复杂性。阐释了空值在 XML 概率数据库中的两种意义,其中一种解释会产生概率区间,然后采用一种折中的方法解决了引入空值的基于 XML 的概率数据库产生的概率区间问题,并证明其正确性。最后提出一种新的 N 次矩运算,用来对数据库数据进行全面分析。

关键词 空值,XML,概率数据库, N 次矩运算

中图法分类号 TP311 **文献标识码** A

Research on Null Values in Probabilistic XML Database

ZHOU Zhi-zeng WANG Xin-jun

(College of Computer Science and Technology, Shandong University, Jinan 250101, China)

Abstract Uncertainty and Incompleteness are a real-world challenges of the database. The combination of NULL value and probabilistic XML database can better deal with data, but also increase the complexity of the database. We proposed two interpretation what means the null value in the XML probabilistic database, one of them can generate probabilistic interval, so we presented a compromise solution to the probabilistic interval problem in probabilistic XML database with NULL value. Finally, we proposed an N^{th} moment method for anglicizing probabilistic database comprehensively.

Keywords Null value, XML, Probabilistic database, N^{th} moment operation

1 引言

通常认为,我们已经步入了信息社会,信息量与日俱增。而与此相矛盾的是,在某一方面,信息量又显得非常匮乏,所掌握的信息也存在不确定性和不完全性。信息的不确定性和不完全性主要来源于 3 个方面:

(1) 认知水平还没有达到足够的高度,主要表现在科学研究中;

(2) 收集信息的手段还没有达到足够先进,比如在科学预测中;

(3) 掌控信息的一方由于各种原因而采取了严格的保密措施,使信息只能在小范围流通,像在军事上和商业中。

对于不确定信息,在 1988 年, Klir G J 和 Folger T A 在文献[2]中就将不确定性数据分为模糊型和概率型,此后对这两方面的研究就逐步深入扩展,但直到 1996 年, Dev 和 sarkala 在文献[3]中才提出一种概率关系数据模型及代数,并在研究中用系统的观点阐述了概率数据库的有关理论,奠定了在数据库中应用概率论来处理不确定信息的基础。而早在 1984 年, C. Zaniolo 在文献[1]中就提出了将空值理论应用于关系数据库中来处理不完全信息。

Green T J 和 Tannen V 在文献[6]中提出一种新的模型: probabilistic c-tables 模型,用来描述不完全和不确定信息,该模型采用完整和封闭的关系代数,使一般概率数据库模型的查询语言具有完整性和封闭性。但是该模型对不完全信息和概率信息分别拥有不同的描述系统,而没有很好地结合起来;并且,该模型主要针对关系型数据建模,而对半结构化数据无能为力。

NIERMAN. A 和 JAGADISH. H. V 在文献[4]中提出了一种基于 XML 的概率数据模型 ProTDB,解决了某些应用所涉及的数据无法用关系模式描述的问题,文献[5,7]也从不同方向提出不同的 XML 概率数据模型,但是这些模型无法描述不完全信息。本文在文献[4]的基础上引入空值来解决这个问题。而空值的引入增加了数据库的复杂性,会带来一系列的计算问题,本文将在第 3 节提出一种针对概率区间的折中处理办法,在第 4 节提出一种全新的运算: N 次矩运算。

2 表示方法

本文在文献[4]中提出的 ProTDB 模型的基础上引入空值,在这里只对这类数据库加以描述和分析,而暂时不会提出完整的模型理论。下面首先给出对空值的表示方法的定义。

来稿日期:2009-10-30 返修日期:2009-12-30 本文受自然科学基金(90818001),山东省自然科学基金(Y2007G38),山东省优秀中青年科学家科研奖励基金(2005BS01002),山东省科技攻关计划(2008GG30001005),山东省科学技术发展计划(2007GG1QX01036),山东省博士后创新项目专项资金(200703084)资助。

周智增(1979-),男,硕士生,主要研究方向为数据库,E-mail:zhouzhizeng@sdu.edu.cn;王新军(1968-),男,教授,博士生导师,主要研究方向为数据库和工作流。

定义 1 在 XML 概率数据库中,所存储的不完全信息用空值来代替,记作*,表示该信息暂时为空值,终究会被实际值取代而成为完全信息。

有了上面的定义,就可以在数据库中描述不完全信息了,具体实例如下。

例 1 我军经过侦察,得知敌方 A 团的基本情况(部分),但对最近出现在该团的一部分人员暂时存在不明确信息,该部分人员不妨称为 E 部,它可能隶属于该团的某个营,也可能是上级派来考查或者学习的,一段时间后就会离开。我方经过多方面的分析,得出如图 1 所示的数据。我方需要根据不同情况,制定相应的作战计划。从图 1 中可以清楚地看到空值的表示方法。

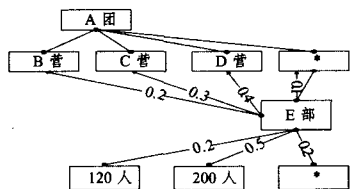


图 1 敌方某团部分信息

在包含空值的概率关系中,含空值的节点的概率属性在整体概率中所起的作用有两种解释。第一种认为,这些概率不包含已经在兄弟节点的概率属性,而已经存在的兄弟节点的概率属性也不会发生变化。第二种解释认为,空缺的概率包含已经存在的兄弟节点的概率属性,那么在这种解释下,结点的概率属性变成了一个区间,只有当空值变成确定值,对相应的节点概率属性进行更新,概率属性才能变成单个值。

如上面的例子所示,用第一种解释可以得出,E 部隶属于 B 营、C 营、D 营的概率分别为 0.2、0.3、0.4,而剩下的 0.1 表示 E 部不属于这 3 个营中的任何一个的概率;E 部有 120 人和 200 人的概率分别是 0.2 和 0.5,而还有 0.2 的概率表示 E 部人数为其他状况。使用第二种解释时,E 部隶属于这 3 个营的概率分别是 $[0.2, 0.3]$ 、 $[0.3, 0.4]$ 和 $[0.4, 0.5]$,而还有一种可能是以概率为 0.1 不属于这 3 个营;而 E 部人数为 120 人和 200 人的概率分别是 $[0.2, 0.4]$ 和 $[0.5, 0.7]$,还可能以 0.2 的概率存在其他情况。

大多数情况下,需要在概率数据库中对各个节点的概率属性为单个值时进行运算,而对概率区间的操作是困难的。下面提出一种折中的算法,在需要时使用单个概率值来代替概率区间。

3 解决方案

在这部分先给出一个预备定义,然后引出一个区间概率函数,并且证明存在一个概率分布与之相适应,最后对这个区间概率函数进行严格化。

定义 2 对于包含空值的基于 XML 的概率数据库的一个实例 T , N 是它的一个非根节点,如果节点 n 是 N 的子女,记作 $n \in N$ 。

给定一个区间 $I=[x, y]$,通常用符号 $I.lb$ 来代表 x ,而用 $I.ub$ 来代表 y 。区间函数 ι 与一个非根节点 N 关联,对于每个子节点 $n_i \in N$,一个闭合子区间 $[lb(s), ub(s)] \subseteq [0, 1]$ 。如果 $\sum_{n_i \in N} lb(n_i) \leq 1$ 且 $\sum_{n_i \in N} ub(n_i) \geq 1$, ι 被称为区间概率函数。用区间概率函数 ι 表示的非根节点 N 的概率分布是一个

映射 $P: N \rightarrow [0, 1]$, 当

$$(1) \forall n_i \in N, lb(n_i) \leq p(n_i) \leq ub(n_i), \text{ 且}$$

$$(2) \sum_{n_i \in N} p(n_i) = 1$$

引理 1 对于任何一个非根节点 N 和它的区间概率函数 ι , 都存在一个概率分布 $P(N)$ 与之相适应。

证明:对于任何一个区间概率函数 ι , 存在多种可能的分布与之相适应,一种解决方案是让分布尽可能地接近区间的“中间”。让 $\sum_{n_i \in N} lb(n_i) = L$ 且 $\sum_{n_i \in N} ub(n_i) = U$ 。由区间概率函数的定义知道, $L \leq 1$ 且 $U \geq 1$ 。与区间约束相一致的的概率函数是: $p(n_i) = lb(n_i) + (ub(n_i) - lb(n_i)) * \frac{1-L}{U-L}$ 。

容易验证 $lb(n_i) \leq p(n_i) \leq ub(n_i)$ 且 $\sum_i p(n_i) = 1$ 。

关于概率分布,Goldman 和 Rivest^[8]曾经提出过最大熵理论。关于 N 的区间概率函数 ι 是严格的,当且仅当对任何一个关于 N 的区间概率函数 ι' , 每个关于 ι 和 ι' 的概率分布 P 相等, $\iota(n_i).lb \geq \iota'(n_i).lb$ 且 $\iota(n_i).ub \leq \iota'(n_i).ub$ 当 $n_i \in N$ 。如果每个关于 ι' 的概率分布 P 与关于 ι 的概率分布相等,那么 ι 与严格 ι' 相等。严格化操作 *tight*, 是从区间概率函数到区间概率函数的映射,所以 *tight*(ι) 结果与 ι 严格相等。下面的结论(文献[9]定理 2)说明总能严格化一个区间概率函数。

定理 1 假设 ι 和 ι' 是关于节点 N 的区间概率函数且 $tight(\iota') = \iota$ 。使 $n_i \in N$ 。那么:

$$\iota(n_i) = [\max(\iota'(n_i).lb, 1 - \sum_{n_i' \in N \wedge n_i' \neq n_i} \iota'(n_i').ub), \min(\iota'(n_i).ub, 1 - \sum_{n_i' \in N \wedge n_i' \neq n_i} \iota'(n_i').lb)]$$

根据上面的引理以及定理,可以用与区间约束相一致的的概率函数对这样的概率区间进行折中处理。

例 2 在例 1 中,对于 E 部的隶属问题, $L=0.9, U=1.2$ 符合引理 1 中概率区间函数的要求,则可以得出 E 部隶属于 A、B、C 营的概率分布为 0.233、0.333、0.433;而在 E 部人数问题中, $L=0.7, U=1.1$, 人数为 120 人和 200 人的概率分布为 0.35 和 0.65。可以验证 $lb(n_i) \leq p(n_i) \leq ub(n_i)$ 且 $\sum_i p(n_i) = 1$ 。

可以看到,计算过程中,空值暂时不作考虑,这和下面的 N 次矩运算有些类似,而当空值被实际值所取代时,概率区间也就消失了。上面的折中方案,有充分的理论依据,解决此类问题比较简便。

4 N 次矩运算

在这部分提出一个新的运算: N 次矩运算,这个运算能得到基于原始分布的、不同属性集合上的聚合性质。在定义这个运算之前,先按照统计学的观点,给出 N 次矩的定义,并说明它的用途。

定义 3 设 ψ 是域 Ψ 上的一个随机变量,它的概率密度函数是 $f_\psi(x), x \in \Psi$, 则它的 N 次矩记作 $\mu_N(\psi)$, 定义如下:

$$\mu_N(\psi) = \int_{x \in \Psi} x^N f_\psi(x) dx$$

分布的矩可用来得到分布的综合性性质,诸如平均值、标准差、失真度、突出度等。例如,随机变量 ψ 的标准差就可以很容易地从 1 次及 2 次矩得到:

$$stdv(\psi) = \sqrt{\mu_2(\psi) - \mu_1(\psi)^2}$$

这些综合性性质不仅可用来了解分布的整体面貌,而且可用来对两个不同的分布进行比较。这样,矩在统计分析中就

非常重要。 N 次矩运算可全面地完成数据的各种统计分析并进而帮助我们形成对客观世界中对象性质的整体看法。

根据上面 N 次矩的定义,可以对 N 次矩运算进行定义。设对于包含空值的基于 XML 的概率数据库的一个实例 T , n 是 T 上的非根节点,用 pS 表示节点上表示概率的属性值,则 T 上的 n 的 N 次矩记作 $\mu_{n,N}(T)$,定义如下:

$$\mu_{n,N}(T) = \{x(T) \mid (\exists y \text{ 是 } n \text{ 的子女})(y(T) \downarrow \wedge x(T) = y(T)) \wedge (\forall A \text{ 是 } T \text{ 上除 } n \text{ 以外的其它非根节点})(x(A) = m_{n,T,N}(y, A))\}$$

其中

$$m_{n,T,N}(y, A) =$$

$$\begin{cases} \frac{\sum_{\substack{y \in n \\ y(A) \downarrow \\ y(T)=x(T)}} (y(A))^N y(pS)}{\sum_{\substack{y \in n \\ y(A) \downarrow \\ y(T)=x(T)}} y(pS)} & \text{当 } pS \text{ 不为 } 0 \text{ 且 } A \text{ 是数值型} \\ \Omega & \text{其他} \end{cases}$$

关于 N 次矩运算有以下几点要说明:

(1)实际上这是一族运算,因为每给一个正整数 N ,就有一个运算。

(2)用 N 次矩运算可以定义其他一些运算,例如,标准差、失真度及突出度等。

(3)由定义可以看出,空值 $*$ 在计算矩时不会被考虑,所以只有分布中显示给出值的部分才参与矩运算。

(4) N 次矩计算对于非数值型节点在运算时会产生一种特殊的空值“ Ω ”。

例3 把该运算应用到例1,当使用1次矩运算后,得出的E部人数期望值为177,我方可以再通过使用标准差、失真度、突出度等参数,得出相应的结论,从而制定有针对性的作战计划。该运算同样可以用于科学研究和科学预测。

结束语 基于XML的概率数据库引入空值,使数据库对不确定信息和不完全信息的描述进行结合,可以更好地表示现实世界。该方法增加了对数据库操作的难度,虽然本文用了折中的方法解决了区间概率的运算问题,但是该方法在大多数情况下,还得具体情况具体分析,而且,该方法只适用

于子节点概率值上限之和大于等于1的情况,而对小于1的情况无能为力;针对空值概率的两种解释,对于建立数学模型描述也是很不利,需要在语义上加以解决,使其对这两种解释都适用。下一步要设计推广的代数运算,使问题可以在代数层面加以解决。

参考文献

- [1] Zaniolo C. Database relations with null values [J]. Journal of Computer and System Sciences, 1984, 28(1): 142-166
- [2] Klir G J, Folger T A. Fuzzy sets, uncertainty and information [M]. New Jersey: Prentice-Hall, 1988
- [3] Dey, Sarkar S. A Probabilistic Relational Model and Algebra [J]. ACM Transactions on Database Systems, 1996, 21(3): 339-369
- [4] Nierman A, Jagadish H V. ProTDB: Probabilistic data in XML [C]//Proceeding of the 28th International Conference on Very Large Data Bases. Hong Kong: Morgan Kaufmann Publishers, 2002: 646-657
- [5] Hung E, Getoor L, Subrahmanian V S. PXML: a probabilistic semistructured data model and algebra [C]//Proceedings of the 19th International Conference on Data Engineering. Bangalore, India: IEEE Computer Society Press, 2003: 467-482
- [6] Green T J, Tannen V. Models for incomplete and probabilistic information [J]. IEEE Date Engineering Bulletin, 2006, 29(1)
- [7] 张群, 王新军, 吴欣. 一种扩展的基于XML的概率数据模型 [J]. 山东大学学报: 理学版, 2007, 42(9)
- [8] Goldman S, Rivest R. A non-iterative maximum entropy algorithm [C]//Proceedings of the 2nd Annual Conference on Uncertainty in Artificial Intelligence (UAI'86). Elsevier Science Publishing Company, Inc., New York, NY, 1986: 133-148
- [9] Dekhtyar A, Goldsmith J, Hawkes S. Semistructured probabilistic databases [C]//Proceedings of the Conference on Statistical and Scientific Database Management (SSDBM). Fairfax, VA, USA, 2001: 36-45

(上接第130页)

- [7] Bradshaw R W, Holt J E, Seamons K E. Concealing Complex Policies with Hidden Credentials [C]//Proceedings of the 11th ACM Conference on Computer and Communications Security. Washington, DC: ACM Press, October 2004: 146-157
- [8] Li Jiangtao, Li Ninghui. OACerts: Oblivious Attribute Certificates [C]//Proceedings of the 3rd International Conference on Applied Cryptography and Network Security (ACNS 2005), Volume 3531 of Lecture Notes in Computer Science. New York, USA: Springer, 2005: 301-316
- [9] Castelluccia C, Jarecki S, Tsudik G. Secret Handshakes from Camouflage Encryption [C]//Proceedings of the 10th International Conference on the Theory and Application of Cryptology and Information Security, Volume 3329 of Lecture Notes in Computer Science. Springer, 2004: 293-307
- [10] Bertino E, Ferrari E, Squicciarini A. Privacy-Preserving Trust Negotiation [R]. CERIAS-TR-2004-75. Center for Education and Research in Information Assurance and Security, Purdue University, 2005
- [11] 廖俊国, 洪帆, 李俊, 等. 在信任协商中保密证书的敏感属性 [J]. 通信学报, 2008, 29(6): 20-25
- [12] Ajayi O, Sinnott R, Stell A. Formalising Dynamic Trust Negotiations in Decentralised Collaborative e-Health Systems [C]//Proceedings of the 2nd International Conference on Availability, Reliability and Security. 2008: 1-8