XML 弱函数依赖及单依赖集合研究

刘先锋[1],李高仕[1],黄海燕[2]
LIU Xian-feng[1],LI Gao-shi[1],HUANG Hai-yan[2]

1.湖南师范大学 数学与计算机科学学院,长沙 410081
2.湖南科技学院 教务处,湖南 永州 425100
1.College of Mathematics and Computer Science,Hunan Normal University,Changsha 410081,China
2.Office of Teaching Affairs,Hunan University of Science and Engineering,Yongzhou,Hunan 425100,China

LIU Xian-feng,LI Gao-shi,HUANG Hai-yan.Research on weak functional dependencies and its mono-dependent set for XML.Computer Engineering and Applications,2009,45(3):155-157.

Abstract: Weak functional dependency is a sort of functional dependency about XML database existed null value.In this paper, based on concepts of null value,incomplete tree tuple,the definitions of weak functional dependency,mono-dependent set are given,and the theorem of determine mono-dependent set and the terminate-able theorem of determine mono-dependent are proved.
Key words: weak functional dependency;mono-dependent;null value

摘 要: XML 弱函数依赖是在 XML 数据库中引入空值理论后的函数依赖。在空值、不完全树元组等概念的基础上,定义了弱函数依赖、单依赖集合,证明了单依赖集合判定定理和单依赖集合判定可终止定理。
关键词: 弱函数依赖;单依赖;空值
DOI: 10.3778/j.issn.1002-8331.2009.03.046 **文章编号:** 1002-8331(2009)03-0155-03 **文献标识码:** A **中图分类号:** TP311

在关系数据库中,不完全关系是指包含空值[1]的关系,即某些值到目前为止还不知道。同样,对于 XML 数据,它的树型结构特点决定它存在空值的可能性,即对于一棵给定的 XML 树[2-3],它的某些路径是不完全的,用空值表示那些不完全的路径值,一棵 XML 不完全树中的空值都用允许取的实际值取代后就成为一棵完全树。设树 T 是一棵不完全树,T 中的空值用各种可能值代替后所形成的完全树的集合称为这棵树的可能世界,并将其记做 $Poss(T)$。众所周知,函数依赖在 XML 数据库设计及应用中起着重要作用。那么,函数依赖在不完全树中怎样体现呢? 这里有以下定义:给定一个 DTD D,设树 T 是 D 上的一棵不完全树,$S_1,S_2\in paths(D)$,$S_1\rightarrow S_2$ 是一个函数依赖。如果存在一个完全树 $T'\in Poss(T)$,而 T' 适合(满足)函数依赖 $S_1\rightarrow S_2$,则称 T 适合函数依赖 $S_1\rightarrow S_2$。那么,什么是“单依赖”集合? 即满足“交特性”和“分特性”的函数依赖集合。

1 XML 弱函数依赖

关于 DTD、XML 树路径以及树元组的定义请参考文献[2-4]。空值关系模型的定义参考文献[5]。偏序、嵌入、扩展 XML 树、构造完全树的定义请参考文献[6]。

定义 1[6](相容) 用记号 $\perp$ 表示未知空值,即到目前尚不知道其值的那一种值;用记号 $\perp_{inc}$ 表示不相容空值,即由于数据不相容引起的空值。设给定一个 DTD D,对于定义在 D 上的树元组 t,及任一定义在这个树元组上的路径 $p\in paths(D)$,$t.p\neq\perp_{inc}$,$t.p\neq\perp$,则称 t 是 D 的完全树元组,否则 t 是不完全树元组;如果存在 $t.p=\perp_{inc}$,则称 t 是不相容树元组,否则称 t 是相容的。设给定树 T 与 D 相容,记 $T<D$,对于所有 $p\in paths(D)$,任意 $t\in tuples_D(T)$ 都是完全的,称 T 是完全的,否则称 T 是不完全的;如果 t 是不相容的,则称 T 是不相容的,否则称 T 是相容的。

定义 2[6](弱函数依赖) 一个 DTD D 上的基于树元组的弱函数依赖是形为:$s_1\rightarrow s_2$ 的一个命题,其中,$s_1,s_2\subseteq paths(D)$,它的含义是,对于一棵树 T,有 $T<D$,当存在一棵树 $T'\in Poss(T)$,$s_1,s_2\subseteq paths(T')$满足以下条件:$\forall t_1,t_2\in tuples_D(T')$,若 $t_1.s_1=t_2.s_1$,则必有 $t_1.s_2=t_2.s_2$,称 T 弱满足(也称适合)函数依赖 $s_1\rightarrow s_2$,记做 $T|\approx s_1\rightarrow s_2$ 。

易看出,在不完全 XML 树中强函数依赖需要在任意 $T'\in Poss(T)$上成立,而弱函数依赖只要存在一个 $T'\in Poss(T)$满足此函数依赖即可。

2 XML 单依赖集合

用 Σ^+表示从 Σ 出发用文献[6]中的公理系统推出的函数依赖集合。

基金项目: 国家自然科学基金(the National Natural Science Foundation of China under Grant No.10571052);湖南省高校青年骨干教师资金;湖南省教育厅科研资金资助。
作者简介: 刘先锋(1964-),男,教授,研究方向:数据挖掘、电子商务技术;李高仕(1974-),男,硕士研究生,研究方向:XML 半结构化数据、数据库。
收稿日期: 2008-01-03 **修回日期:** 2008-04-08

定义 3(覆盖) 给定 DTD D,设 Σ_1 与 Σ_2 是 D 上的两个函数依赖集合,若 $\Sigma_1^+=\Sigma_2^+$,则称 Σ_1 是 Σ_2 的一个覆盖(根据对称 Σ_2 也是 Σ_1 的覆盖),也称 Σ_1 与 Σ_2 等价。

定义 4 (冗余路径) 设 Σ 是给定 DTD 上的一个函数依赖集,$S_1\rightarrow S_2\in\Sigma$,$S_1\rightarrow S_2$ 中相对于 Σ 的冗余路径(extraneous path)p 定义为如下:(1)如果 $p\in S_1$,且 Σ 逻辑蕴涵$(\Sigma-\{S_1\rightarrow S_2\})\cup\{(S_1-p)\rightarrow S_2\}$,则称路径 p 是 $S_1\rightarrow S_2$ 左部的冗余路径;(2)如果 $p\in S_2$,且逻辑蕴涵$(\Sigma-\{S_1\rightarrow S_2\})\cup\{S_1\rightarrow(S_2-p)\}$,则称路径 p 是 $S_1\rightarrow S_2$ 右部的冗余路径。

定义 5(既约) 设 Σ 是给定 DTD 上的一个函数依赖集,$s_1\rightarrow s_2\in\Sigma$。如果 Σ 的任何真子集都不与等价,则称 Σ 是无冗余的,否则是冗余的。如果 $s_1\rightarrow s_2$ 不包含左部冗余路径,称 $s_1\rightarrow s_2$ 是左部既约的。如果 $s_1\rightarrow s_2$ 不包含右部冗余路径,称 $s_1\rightarrow s_2$ 是右部既约的。如果 $s_1\rightarrow s_2$ 同时是左部既约和右部既约的,称 $s_1\rightarrow s_2$ 是既约的函数依赖。如果 Σ 中的每一个函数依赖都是左部既约的(右部既约的,既约的),称 Σ 是左部既约的(右部既约的,既约的)。

例 1 $\Sigma_1=\{AB\rightarrow C,A\rightarrow B,B\rightarrow C\}$,$\Sigma_2=\{A\rightarrow B,B\rightarrow C\}$,任取 Σ_1 中的函数依赖,总能在 Σ_2 中通过文献[6]中的公理系统推出,反之亦成立,可知,Σ_1 和 Σ_2 等价,但 $\Sigma_2\subset\Sigma_1$,故 Σ_1 冗余。$\Sigma_3=\{A\rightarrow B\}$,$\Sigma_4=\{B\rightarrow C\}$,由于 $\Sigma_3^+\neq\Sigma_2^+$,$\Sigma_4^+\neq\Sigma_2^+$,可知 Σ_2 是无冗余的。

例 2 $\Sigma_1=\{A\rightarrow BC,AB\rightarrow D,B\rightarrow C\}$既不是左既约的,也不是右既约的,$\Sigma_2=\{A\rightarrow BC,B\rightarrow C,A\rightarrow D\}$ 是左既约的,$\Sigma_3=\{A\rightarrow B,B\rightarrow C,AB\rightarrow D\}$是右既约的,$\Sigma_4=\{A\rightarrow B,B\rightarrow C,A\rightarrow D\}$既是右既约的,又是左既约的,故是既约的。

定义 6(正则覆盖) 给定 DTD D,设 Σ 是 D 上的一个函数依赖的集合,如果 Σ 中的函数依赖右边都只有一个路径,即都是形如 $X\rightarrow A$ 的函数依赖,而且 Σ 是无冗余的,是左既约的,则称 Σ 是正则的。又如果 Σ 是 Σ' 的覆盖而且是正则的, 则称 Σ 是 Σ' 的正则覆盖。

例 3 易知 $\Sigma_1=\{A\rightarrow B,A\rightarrow D,BI\rightarrow J,A\rightarrow C,A\rightarrow E\}$ 是正则的。易证 Σ_1 是 $\Sigma_2=\{A\rightarrow BCE,AB\rightarrow DE,BI\rightarrow J\}$的正则覆盖。

请注意,正则函数依赖集合中的函数依赖右部都只有一个路径,所以这些函数依赖也都是右既约的,于是也都是既约的。这样,正则函数依赖集合一定也是既约函数依赖集合。

定义 7(单依赖集合) 设 Σ 是给定 DTD D 上的函数依赖集合,如果 Σ 满足以下条件,则称 Σ 为“单依赖”(mono-dependent)集合:对 D 上的所有路径 A:(1)只要 $X\rightarrow A$,$Y\rightarrow A\in\Sigma^+$是不可比较的,则必有 $X\cap Y\rightarrow A\in\Sigma^+$。这里,$X\rightarrow A$,$Y\rightarrow A$ 是不可比较的是指 $X\not\subseteq Y$ 而且 $Y\not\subseteq X$。(2)若有 $XB\rightarrow A\in\Sigma^+$,$YA\rightarrow B\in\Sigma^+$,则或者有 $Y\rightarrow B\in\Sigma^+$,或者有$(X\cap Y)A\rightarrow A\rightarrow B\in\Sigma^+$。称性质(1)为“交特性”;(2)为“分特性”,并称形如 $XB\rightarrow A$,$YA\rightarrow B$ 的两个函数依赖为“循环函数依赖”。

例 4 $\Sigma_1=\{A\rightarrow B,C\rightarrow D\}$平凡的是单依赖集合,$\Sigma_2=\{BC\rightarrow A,AC\rightarrow B\}$是单依赖集合,因为 $X=C$,$Y=C$,$(X\cap Y)A\rightarrow B$ 是 $CA\rightarrow B$。$\Sigma_3=\{A\rightarrow B,B\rightarrow C\}$不是单依赖集合,因为 $A\rightarrow C\in\Sigma_3^+$,但 $A\cap B\rightarrow C$ 是 $\Phi\rightarrow C$ 不属于 Σ_3^+。

定理 1 (单依赖集合判定定理) 设 Σ 是给定 DTD 上的函数依赖集合,Σ'是它的正则覆盖,当且仅当对 D 中的每一个路径 A, 下列两个条件均为真时,Σ 是单依赖集合:(1)Σ'中最多只能有一个形如 $X\rightarrow A$ 的函数依赖,而且如果 Σ'中对这个路径 A 确实有 $X\rightarrow A$ 的函数依赖,则对 X 中的每个路径 B 都有 $A\notin(D-AB)_\Sigma^+$。(2)若 $X\rightarrow A$,$Y\rightarrow B\in\Sigma'$,则或者 $A\notin Y$,或者 $Y\subseteq(X\cap Y)A$。

证明 证明分为(当)和(仅当)两部分。(当)设上述两条件为真而且 Σ'是 Σ 的正则覆盖。

针对单依赖集合定义的第(1)部分,需证:若 $X\rightarrow A$,$Y\rightarrow A\in\Sigma^+$,则 $X\cap Y\rightarrow A\in\Sigma^+$。

设 $W\rightarrow A$ 是 Σ'中惟一的决定 A 的函数依赖。如果 $W\not\subseteq X$,则 $W\neq\phi$,而且存在 $B\in W$ 但 $B\notin X$,这样,由 $X\subseteq D-AB$,可推出$(D-AB)\rightarrow A$,于是 $A\in(D-AB)_\Sigma^+$,这与 $A\notin(D-AB)_\Sigma^+$矛盾,所以 $W\notin X$ 不正确,故 $W\subseteq X$。类似可知:$W\subseteq Y$,这样,$W\subseteq X\cap Y$,于是从 $W\rightarrow A$ 可推出 $X\cap Y\rightarrow A$。

对于单依赖集合的定义的第(2)部分,设有循环 $XB\rightarrow A$,$YA\rightarrow B$,需证或者有 $Y\rightarrow B\in\Sigma^+$,或者有$(X\cap Y)A\rightarrow B\in\Sigma^+$。

设 $W\rightarrow A$ 及 $Z\rightarrow B$ 分别是 Σ'中决定 A 及决定 B 的函数依赖。如果 $W\not\subseteq XB$,则 $W\neq\phi$,而且存在 $C\in W$,但 $C\notin XB$,于是 $XB\subseteq D-AC$,这样,由 $XB\rightarrow A$ 可以推出$(D-AC)\rightarrow A$,于是,$A\in(D-AC)_\Sigma^+$,与本定理的第(1)个条件相矛盾,所以 $W\not\subseteq XB$ 是不对的,故 $W\subseteq XB$。由于 $A\neq B$,类似可证 $Z\subseteq YA$。这样,如果 $A\notin Z$,则 $Z\subseteq Y$,而 $Z\rightarrow B\in\Sigma'$,所以 $Y\rightarrow B\in\Sigma^+$。如果 $A\in Z$,则有 $Z\subseteq(W\cap Z)A$,这样,由 $Z\rightarrow B\in\Sigma'$可知,$(W\cap Z)A\rightarrow B\in\Sigma^+$,而且它不是平凡依赖。再由 $W\subseteq XB$ 以及 $Z\subseteq YA$ 可知:$(W\cap Z)A\subseteq(XB\cap YA)A$。然而 $A\notin X$,$B\notin Y$,所以$(XB\cap YA)=(X\cap Y)$,这样,$(W\cap Z)A\subseteq(X\cap Y)A$,于是$(X\cap Y)A\rightarrow B\in\Sigma^+$。

(仅当)设 Σ 是单依赖集合。

对于条件(1),由于 Σ'是正则的,所以最多只有一个 $X\rightarrow A\in\Sigma'$。如果对于这个 A 确实有一个 $X\rightarrow A\in\Sigma'$,那么需证明对所有 $B\in X$ 都有 $A\notin(D-AB)_\Sigma^+$。分两种情况:

(1)$X=D-A$ 或 $X=\Phi$。

这时平凡地对每一个 $B\in X$,显然都必有 $A\in(D-AB)_\Sigma^+$。

(2)$\Phi\subset X\subset D-A$。

用反证法。设有一个 $B\in X$,使 $A\in(D-AB)_\Sigma^+$,为了书写方便,记$(D-AB)=W$,于是 $A\in w_\Sigma^+$,即 $W\rightarrow A\in\Sigma^+$。由于 $B\in X$,但 $B\notin W(=D-AB)$,所以 $X\cap W\neq X$。这样,由于 Σ'正则,而 $X\rightarrow A\in\Sigma'$,故 $X\cap W\rightarrow A\notin\Sigma'$,从而 $X\cap W\rightarrow A\notin\Sigma^+$。然而,$W\rightarrow A\in\Sigma^+$,$X\rightarrow A\in\Sigma^+$,$X\cap W\rightarrow A\notin\Sigma^+$将与 Σ 是单依赖集合矛盾, 所以

对所有 $B \in X$，都必有 $A \notin (D-AB)^+_\Sigma$。

对于上述条件（2），设 $X \to A$，$Y \to B \in \Sigma'$，证明或者有 $A \notin Y$，或者有 $Y \subseteq (X \cap Y)A$。

用反证法，设 $A \in Y$，而且 $Y \nsubseteq (X \cap Y)A$，分两种情况：

情况 1　$B \notin X$。

因为 $X \to A$，所以可以推出 $X(Y-A) \to A(Y-A)$，由于 $Y \subseteq (Y-A)A$，所以由 $Y \to B$ 可以推出 $(Y-A)A \to B$。于是可推出 $X(Y-A) \to B$，即 $X(Y-A) \to B \in \Sigma^+$。这样，首先 $X(Y-A)$ 不会是 Y 的真子集，因为如果是的话，将有 $X(Y-A) \neq Y$，而 $Y \to B$ 与 $X(Y-A) \to B$ 同属 Σ^+，从而同属 $(\Sigma')^+$。这与 Σ' 是正则的相矛盾。另外，由于 $A \in Y$，而 $A \notin X(Y-A)$，所以 $X(Y-A) \neq Y$，所以 $X(Y-A) \nsubseteq Y$。其次，$A \in Y$，而 $A \notin X(Y-A)$ 也说明 $Y \nsubseteq X(Y-A)$，所以 Y 与 $X(Y-A)$ 是不可比较的。这样，由于 Σ 是单依赖集合，所以 $Y \cap (X(Y-A)) \to B$ 应属于 Σ^+，但 $Y \cap (X(Y-A)) \subset Y$，所以这与 Σ' 是正则的相矛盾（Σ 是 Σ' 的覆盖）。这样，即得 $A \in Y$ 或者 $Y \subseteq (X \cap Y)A$。

情况 2　$B \in X$。

首先，由于 $A \in Y$，而 Σ' 正则，所以 $(Y-A) \to B \notin \Sigma^+$。其次，因 $A \in Y$，所以 $(X \cap Y)A \subseteq Y$。而由前提 $Y \nsubseteq (X \cap Y)A$ 可知，有 $(X \cap Y)A \subset Y$。这样，由于 Σ' 是正则，得 $(X \cap Y)A \to B$ 也不属于 Σ^+。于是 $(Y-A) \to B \notin \Sigma^+$ 和 $(X \cap Y)A \to B \notin \Sigma^+$ 同时成立，与 Σ 是单依赖集合相矛盾。故必有 $A \notin Y$ 或者 $Y \subseteq (X \cap Y)A$。

定理 1 证毕。

定理 2（单依赖集合判定可终止定理）给定 DTD 上的函数依赖集合 Σ 是否是单依赖的检验可在 $|\Sigma|$ 的多项式次数内完成。

证明由定理 1 可知，这个测试工作可以通过求出 Σ 的正则覆盖 Σ' 后，对 D 上的每个路径 A 做定理 1 中的两项检查而实现。而这两项检查及求 Σ 的正则覆盖都可在 $|\Sigma|$ 的多项式次数内完成，所以整个工作可在 $|\Sigma|$ 的多项式次数内完成。

3　结论

XML 函数依赖在 XML 数据库研究中占有重要地位。将空值理论引入 XML 数据中，给出了 XML 弱函数依赖的概念，并给出单依赖集合的定义；证明了单依赖集合判定定理和单依赖集合判定可终止定理。下一步的工作是在本文的基础上对弱函数依赖的可加性，以及可加性和单依赖集合之间的重要联系进行研究，进一步完善 XML 数据库理论。

参考文献：

[1] Levene M，Loizou G.The additivity problem for functional dependencies in incomplete relations[J].Acta Inf，1997，34（2）：135-149.

[2] Arenas M，Libkin L.A normal form for XML documents[J].ACM Transactions on Database Systems，2004，29（1）：195-232.

[3] Vincent M W，Liu Ji-xue，Liu Cheng-fei.Strong functional dependencies and their application to normal forms in XML[J].ACM Trams Database Syst，2004，29（3）：445-462.

[4] 胡小明.XML 函数依赖的推理规则与蕴涵问题研究[D].秦皇岛：燕山大学，2004.

[5] 马垣.非经典关系数据库理论[M].北京：清华大学出版社，2005.

[6] 苏召，刘国华.XML 弱函数依赖及其推理规则[J].计算机应用，2007，27（5）：1228-1231.

（上接 154 页）

由算法时间复杂度分析可知，对于复杂路径上对等点的数据映射推导，在映射表数目较少时，一般对等数据映射、长路径数据映射和最有路径数据映射三种对等点数据映射的时间复杂度在一个数量级上，对映射推导的时间影响相差不大，但在映射表数目较多时，最优路径的对等数据映射推导更有利于用户在短时间内推导出新的对等点数据映射表。

4　结束语

给出了一种基于最优路径的对等点数据映射推导算法，用于解决复杂路径上任一对等点的数据映射推导，使映射推导更具有普遍性。

由时间复杂度分析可知，基于最优路径的对等点的数据映射推导可以在不增加复杂度的前提下，使用户在短时间内寻求最合适的路径进行数据映射推导，解决了复杂路径上任一对等点的数据映射推导问题。

参考文献：

[1] Kementsietsidis A，Arenas M，Miller R J.Mapping data in peer-to-peer systems：semantics and algorithmic issues[C]//Proc of the ACM SIGMOD International Conference on Management of Data，2003：325-336.

[2] Bernstein P，Giunchiglia F，Kementsietsidis A，et al.Data management for peer-to-peer computing：a vision[C]//Proceedings of the 5th International Workshop on the Web and Databases，2002：89-94.

[3] Kementsietsidis A，Arenas M.Data sharing through query translation in autonomous systems[C]//Proc of the International Conference on Very Large Data Bases，2004：468-479.

[4] Rahman M A，Kiringa I，Saddik A E.Generalization of an algorithm for checking consistency of mapping tables in a peer-to-peer system[C]//Proc of the International Conference on Computer and Information Technology（ICCIT），2004.

[5] 刘国华，张忠平，岳晓丽，等.数据库新理论、方法及技术导论[M].北京：电子工业出版社，2006.