

粗关系数据库中的粗函数依赖研究

郭景峰 李 莉 宫继兵

(南京大学计算机新技术国家重点实验室 南京210093) (燕山大学 秦皇岛066004)

摘 要 以粗集理论为研究方法,针对粗关系数据库属性值非原子性的特点,从语义等价的角度改进了已有的粗关系数据库函数依赖定义,提出了其修正定义粗函数依赖,使之更客观地反映粗关系数据库中数据的语义联系,体现现实世界不确定性信息的粗糙性和不完备性。本文还给出了判断粗函数依赖是否成立的算法,并用粗关系实例验证了粗函数依赖的优越性,探讨了基于粗函数依赖的推理规则。

关键词 粗集理论,粗关系数据库,粗函数依赖

The Reseach of Rough Functional Dependency in Rough Relational Databases

GUO Jing-Feng LI Li GONG Ji-Bing

(Key Laboratory of New Computer Technologies Nanjing Univercity, Nanjing 210093) (Yanshan Univercity, Qin Huangdao 066004).

Abstract With Rough Sets Theory as a method, the concept of functional dependency in rough relational databases is modified from the view of semantic equipollence in connection with the characteristic of non-atom in values of rough relational databases. The modificatory concept - rough functional dependency is proposed, so that it can objectively reflect the connections among the data of rough relational database and embody the roughness and incompleteness of the information in the real world. The algorithm on how to judge whether a rough functional dependency holds is also introduced. Additionally, this paper verifies the advantage of rough functional dependency by a rough relational database example. The inference rules based on rough functional dependency are discussed finally.

Keywords Rough sets theory, Rough relational database model, Rough functional dependency

0 引言

传统的关系数据库主要是处理确定知识,对于不确定的数据处理能力较差,为了更好地处理现实世界不断涌现的不确定信息,十分有必要对关系数据库模型进行扩充。T. Beaubouef 等人把粗集理论^[1,2]和传统关系数据库理论结合,提出了粗关系数据库模型^[3](the Rough Relational Database Model,简称RRDM)的概念,使得人们在处理不确定性问题上又向前迈进了一步。RRDM是对关系数据库模型的扩展,它打破了关系数据库模型必须满足第一范式理论的限制,属性值可以是一个由多个原子值组合而成的非原子值。RRDM的这一特性在很大程度上比较符合真实世界的情况。但同样由于失去了原子性,使得关系数据库的许多特性不再适合它。

目前,国内外有关粗关系数据库(the Rough Relational Database,简称RRDB)理论体系的研究还很少,已有的理论主要有:RRDB的粗熵理论^[4]、关系运算理论^[5,7]、粗数据查询理论^[8]、函数依赖理论^[6]等。文[6]最早探讨了粗关系数据库的函数依赖,该文提出了一致度函数的概念,并用此函数度量RRDB数据间函数依赖的程度大小,这为RRDB函数依赖理论的研究奠定了基础,但是,一致度函数理论存在很大的局限性,其判断RRDB两个非原子属性值是否相等的标准很不规范,也无法从语义角度解释函数依赖。

本文的工作以文[6]的RRDB函数依赖理论为基础,运用粗集理论的研究方法,改进其RRDB函数依赖定义,提出其修正定义一粗函数依赖,使之更客观地反映粗关系数据库

中数据的语义联系,体现现实世界不确定性信息的粗糙性和不完备性。基于此,本文将给出判断粗函数依赖是否成立的算法,并用粗关系实例验证粗函数依赖的优越性。

1 基本概念

1.1 RRDM

RRDM同传统的关系模型一样,均是由包含元组的关系构成,元组 t_i 采用 $(d_{i1}, d_{i2}, \dots, d_{im})$ 的形式,其中 d_{ij} 是元组的一个属性值。在传统关系数据库中, $d_{ij} \in D_j$,而在RRDB中, $d_{ij} \subseteq D_j, d_{ij} \neq \Phi$ 。用 $P(D_j)$ 表示 $D_j - \Phi$ 的幂集, $P(D_j)$ 为 d_{ij} 对应属性的真正值域,而不是 D_j 。这样,RRDM与普通的关系数据库的一个不同点就是,其属性值可以由多值(多个属性值)构成。

1.2 RRDB

把RRDB定义为三元组: $r = (U, A, D_i)$,从数据库角度看, U 是所有元组的集合, A 为数据库的属性集, D_i 为某一属性域;对于 $\forall t \in U, A_j \in A$,有 $t(A_j) \subseteq D_{A_j}$ 。按照信息系统的定义,RRDB实际上是特殊的多值信息系统,其中 $U = \{u_1, u_2, \dots, u_{|U|}\}$ 是对象的全体,它是非空有限集,即论域; $A = \{A_1, A_2, \dots, A_{|A|}\}$ 是属性的全体,为非空有限集。

RRDB的属性域 $D_j (1 \leq j \leq |A|)$ 上存在一个等价关系 θ_{A_j} ,构成 D_j 上的等价划分 $P_{A_j}: Z_{A_j}^1, \dots, Z_{A_j}^k$ 。

定义1 一个rough关系 (U, A, D_j) 是 $\Pi_{A_j \in A} P(D_j)$ 的一个子集,且对 $\forall t \in U, 1 \leq i \leq k, |t[A_j] \cap Z_{A_j}^i| \leq 1$ 。

2 RRDB 的函数依赖和推理规则

传统关系数据库的函数依赖理论^[9]已很成熟,而有关 RRDB 函数依赖的研究国内外还很少。文[6]最早探讨了 RRDB 的函数依赖及其推理规则,该文提出了一致度函数的概念,并用此函数度量 RRDB 数据间依赖的程度大小。这里对文[6]的 RRDB 函数依赖理论作一简单描述,并对其不规范处进行修订和补充。

由于 RRDB 属性值的非原子性,不能用简单的相等关系去比较两个非原子数据间的关系。文[6]用等值度来描述 RRDB 非原子数据近似相等的程度。

定义2^[6] 设 (U, A, D_i) 为一 RRDB, $t_p, t_q \in U$ ($1 \leq p, q \leq |U|$), $A_i \in A$ ($1 \leq i \leq |A|$), $t_p[A_i]$ 与 $t_q[A_i]$ 的等值度定义为:

$$EQ(t_p[A_i], t_q[A_i]) = |t_p[A_i] \cap t_q[A_i]| / |t_p[A_i]| \times |t_q[A_i]|$$

若 $X \subseteq A$, 则 $EQ(t_p[X], t_q[X]) = \min_i (|t_p[A_i] \cap t_q[A_i]| / |t_p[A_i]| \times |t_q[A_i]|, 1 \leq i \leq |X|)$ 。

根据定义2, 文[6]定义了 RRDB 的函数依赖。

定义3^[6] 设 (U, A, D_i) 为一 RRDB, $X, Y \subseteq A, \forall t_i, t_j \in U$ ($1 \leq i, j \leq |U|, i \neq j$), 若 $EQ(t_i[Y], t_j[Y]) \geq EQ(t_i[X], t_j[X])$, 则称 X 函数决定 Y , 或 Y 函数依赖于 X , 记为 $X \rightarrow Y$ 。

根据逻辑蕴含理论^[10], 公式 $a \rightarrow b$ 等价于 $\neg a \vee b, \neg a \vee b$ 为真的概率为:

$$P(\neg a \vee b) = \min(1, 1 - P(a) + P(b))$$

把逻辑蕴含理论应用到 RRDB 函数依赖中, 用一致度函数衡量函数依赖的程度大小。

定义4^[6] 设 (U, A, D_i) 为一 RRDB, $X, Y \subseteq A, f: X \rightarrow Y$ 为一函数依赖, $\forall t_i, t_j \in U$ ($1 \leq i, j \leq |U|, i \neq j$), f 的一致度定义为:

$$com(f) = \min_{i \neq j} (1 - X_{ij} + Y_{ij})$$

其中, $X_{ij} = EQ(t_i[X], t_j[X]), Y_{ij} = EQ(t_i[Y], t_j[Y])$ 。

定理1^[6] 在基于粗集理论的扩展关系模型中, 如下推理规则同样成立:

- A1. 自反性: 如果 $Y \subseteq X$, 则 $X \rightarrow Y$ 。
- A2. 增广性: 如果 $X \rightarrow Y$ 且 $Z \cap X \neq \emptyset$, 则 $ZX \rightarrow ZY$ 。
- A3. 传递性: 如果 $X \rightarrow Y$ 且 $Y \rightarrow Z$, 则 $X \rightarrow Z$ 。

3 RRDB 函数依赖定义的不足

RRDB 与传统关系数据库的主要区别就是知识描述的不确定性和属性值的非原子性。因此, 传统关系数据库的很多原理都不能完全解释 RRDB 中的独特之处。比如, RRDB 中同样的知识可能用不同的表达方式来描述, 这就是 RRDB 属性域上存在等价关系的原因。

文[6]提出的函数依赖从集合论的角度判断 RRDB 两个粗数据是否相等, 并以此为基础定义了函数依赖, 并用一致度反映函数依赖的程度。但是, 文[6]的思想没有充分体现 RRDB 的粗特性。在 RRDB 中, 两个语义相同的粗数据, 由于描述方式的差别, 用等值度的概念来衡量, 可能就是等值度很小的数据, 甚至等值度就是零。因此, 基于等值度概念定义的 RRDB 函数依赖并不准确, 不满足语义解释要求。

4 RRDB 函数依赖定义的改进

RRDB 是结合了粗集的一些基本概念如等价类、上近似、

下近似和关系数据库的一些基本原理而形成的, 它是对关系数据库的扩展。本文以粗集理论为基础, 修订文[6]提出的 RRDB 函数依赖定义, 使其既能充分反映 RRDB 的数据粗糙性, 又符合语义解释要求, 体现现实世界不确定知识间的必然联系。

4.1 RRDB 粗糙数据语义等价的思想

RRDB 的每个属性域上都存在一个等价关系, 把该属性域划分成等价类, 每个等价类中的属性值是不可分辨的, 它们描述的是同样的不确定知识, 只不过表达形式不同。等价类中的不同属性值语义上是等价的。基于此, 本文用粗相等(或语义等价)的概念来描述 RRDB 的两个属性值(包括原子属性值和非原子属性值)之间的关系。

定义5 设 (U, A, D_i) 为一 RRDB, $X, Y \subseteq A, \forall t_p, t_q \in U$ ($1 \leq p, q \leq |U|, p \neq q$), $A_i \in A$ ($1 \leq i \leq |A|$), 若 $t_p[A_i] = \{a_1, a_2, \dots, a_k\}, t_q[A_i] = \{b_1, b_2, \dots, b_k\}$, 且 $a_j, b_j \in Z_i^A, Z_i^A$ 为 θ_{A_i} 的一个等价类, 则称 $t_p[A_i]$ 与 $t_q[A_i]$ 粗相等, 记为 $t_p[A_i] \stackrel{R}{=} t_q[A_i]$; 反之, 称 $t_p[A_i]$ 与 $t_q[A_i]$ 不等, 记为 $t_p[A_i] \stackrel{R}{\neq} t_q[A_i]$ 。

RRDB 属性值粗相等, 那么这两个属性值在语义上等价, 它们是不可分辨的。

定义6 设 (U, A, D_i) 为一 RRDB, $A_i \in A$ ($1 \leq i \leq |A|$), $t_p, t_q \in U$ ($1 \leq p, q \leq |U|, p \neq q$), $t_p[A_i] = \{a_1, a_2, \dots, a_k\}, t_q[A_i] = \{b_1, b_2, \dots, b_k\}$, 若 $t_p[A_i] \stackrel{R}{=} t_q[A_i]$, 则 $t_p[A_i]$ 与 $t_q[A_i]$ 属于同一等价类 $C_T = \{\{a_1, a_2, \dots, a_k\}, \{b_1, b_2, \dots, b_k\}\}$, $T = \{p, q\}$, 称 C_T 为 t_p, t_q 在 A_i 上的等价合类。

等价合类中的元素是语义等价的。

定义7 设 (U, A, D_i) 为一 RRDB, $X \subseteq A, X = \{A_1, A_2, \dots, A_k\}$ ($k \leq |A|$), 若对 $\forall t_p, t_q \in U$ ($1 \leq p, q \leq |U|, p \neq q$), 满足:

$$t_p[A_i] \stackrel{R}{=} t_q[A_i], i = 1, 2, \dots, k$$

则称 $t_1[X]$ 与 $t_2[X]$ 粗相等, 也记为 $t_1[X] \stackrel{R}{=} t_2[X]$ 。

4.2 RRDB 属性子集导出的等价关系

设 $r = (U, A, D_i)$ 为一 RRDB, 对于 $X \subseteq A$, 可以在 r 上导出一个等价关系 θ_X (在不引起混淆的情况下, 简记为 θ), 使得 $u\theta v$ 当且仅当 $u[X] \stackrel{R}{=} v[X], \forall u, v \in r$ 。等价关系 θ 给出了 r 上的一个划分, 记作 r/θ 或 r/X 。

定义8 设 $r = (U, A, D_i)$ 为一 RRDB, $X \subseteq A, \theta$ 为基于 X 的等价关系, 对于 $\forall Z \subseteq r, Z$ 的上、下近似集定义为:

$$\overline{apr}_\theta Z = \bigcup \{Y \in r/\theta \mid Y \cap Z \neq \emptyset\}$$

$$\underline{apr}_\theta Z = \bigcup \{Y \in r/\theta \mid Y \subseteq Z\}$$

定义9 设 $r = (U, A, D_i)$ 为一 RRDB, 令 P 和 Q 为 r 中属性集导出的等价关系, Q 的 P 正域记为 $POS_P(Q)$, 即

$$POS_P(Q) = \bigcup_{X \in r/Q} \underline{apr}_P X$$

4.3 RRDB 函数依赖的修正定义—粗函数依赖

定义10 设 $r = (U, A, D_i)$ 为一 RRDB, $X, Y \subseteq A$, 若对 $\forall t_p, t_q \in r$ ($1 \leq p, q \leq |U|, p \neq q$), 都满足: 当 $t_p[X] \stackrel{R}{=} t_q[X]$ 时, 必有 $t_p[Y] \stackrel{R}{=} t_q[Y]$, 即:

$$t_1[X] \stackrel{R}{=} t_2[X] \Rightarrow t_1[Y] \stackrel{R}{=} t_2[Y]$$

则称属性集 X 粗函数决定 Y 或 Y 粗函数依赖于 X , 记为 $RFD: X \xrightarrow{R} Y$ 。

定理2 设 $r = (U, A, D_i)$ 为一 RRDB, $X, Y \subseteq A$, 如果 X

$\xrightarrow{R} Y$, 且 P, Q 分别为 X, Y 导出的等价关系, 则 $POS_P(Q) = r$, 反之亦然。

证明: 假设 $POS_P(Q) \neq r$, 则 $\exists S \in r/P$, 且对于 $\forall W \in r/Q$, 都有 $S \not\subseteq W$ 。

不妨设 $X = \{A_1, A_2\}, Y = \{A_3, A_4\}, t_1, t_2 \in S$, 且 $t_1 \in W_1, t_2 \in W_2 (W_1, W_2 \in r/Q, W_1 \cap W_2 = \emptyset, S \not\subseteq W_1, S \not\subseteq W_2)$, 且有

$$\forall t_i \in S, t_i[A_i] \in C_S, i=1, 2$$

$$\forall t_{w1} \in W_1, t_{w1}[A_j] \in C_{W_1}, j=3, 4$$

$$\forall t_{w2} \in W_2, t_{w2}[A_j] \in C_{W_2}, j=3, 4$$

由假设条件可知: $C_{W_1} \neq C_{W_2}$ 。

由于 $t_1, t_2 \in S$, 且 $t_1 \in W_1, t_2 \in W_2$, 根据属性集导出等价关系的基本原理, 可得:

$$t_1[A_i], t_2[A_i] \in C_S (i=1, 2),$$

$$t_1[A_j] \in C_{W_1}, t_2[A_j] \in C_{W_2}$$

很显然, $t_1[X] \stackrel{R}{=} t_2[X], t_1[Y] \neq t_2[Y]$ 。

又根据已知条件 $X \xrightarrow{R} Y$, 可以得到:

$$t_1[X] \stackrel{R}{=} t_2[X] \Rightarrow t_1[Y] \stackrel{R}{=} t_2[Y]$$

这与假设的 $t_1[Y] \neq t_2[Y]$ 矛盾, 所以假设不成立, 即有: $POS_P(Q) = r$ 。

反过来, 如果 $POS_P(Q) = r$, 根据 Q 的 P 正域和粗函数依赖定义易知 $X \xrightarrow{R} Y$ 。

定理3 一个粗函数依赖 $X \xrightarrow{R} Y$ 与下列命题等价:

$$(1) \theta_X \subseteq \theta_Y;$$

$$(2) \bigcap_{a \in X} \theta_a \subseteq \bigcap_{a \in Y} \theta_a.$$

证明: 由定义10可知, $X \xrightarrow{R} Y$ 成立当且仅当以下条件成立:

$\forall t_1, t_2 \in r$, 若 $t_1[a] \stackrel{R}{=} t_2[a] (a \in X)$, 则有 $t_1[a] \stackrel{R}{=} t_2[a] (a \in Y)$ 成立; 即

$$t_1 \theta_X t_2 \mid = t_1 \theta_Y t_2.$$

基于此, 可得:

$$\forall t_1 (\bigcap_{a \in X} \theta_a) t_2 \mid = t_1 (\bigcap_{a \in Y} \theta_a) t_2;$$

$$\forall (\bigcap_{a \in X} \theta_a) \subseteq (\bigcap_{a \in Y} \theta_a).$$

5 判断粗函数依赖的算法描述

根据定理3, 本文给出判断粗函数依赖 $X \xrightarrow{R} Y$ 是否成立的算法。

设 $r = \{U, A, D_i\}$ 为一 RRDB, $X, Y \subseteq A$, 且有:

$$U/\theta_X = \{X_{11}, X_{12}, \dots, X_{1i}, \dots, X_{1|\pi_1|}\}, |\pi_1| \leq |U|;$$

$$U/\theta_Y = \{X_{21}, X_{22}, \dots, X_{2j}, \dots, X_{2|\pi_2|}\}, |\pi_2| \leq |U|;$$

算法如下:

(O1)[Initialize]. Set $i=1, j=1$.

(O2)[$X_{1i} \subseteq X_{2j}$?]. If $X_{1i} \subseteq X_{2j}$, then (for this i we have already found a j such that $X_{1i} \subseteq X_{2j}$, so) go to (O3) (to check next i).

If $X_{1i} \not\subseteq X_{2j}$, then go to (O5) (to check next j).

(O3)[$i = |\pi_1|$?]. If $i = |\pi_1|$ then the algorithm is completed with the answer: $X \xrightarrow{R} Y$.

If $i < |\pi_1|$ then go to (O4).

(O4)[Increase i]. Set $i \leftarrow i+1$, go to (O2).

(O5)[$j = |\pi_2|$?]. If $j = |\pi_2|$ (now we have an i such that $X_{1i} \not\subseteq X_{2j}$ for $j = 1, 2, \dots, |\pi_2|$) then the algorithm

is completed with the answer: $X \xrightarrow{R} Y$.

If $j < |\pi_2|$ then go to (O6)

(O6)[Increase j]. Set $j \leftarrow j+1$, go to (O2).

该算法的复杂性主要由 (O2) 步骤中需进行的比较引起, 次数约为 $|\pi_1| \times |\pi_2| \leq |U|^2$, 所以该算法的时间复杂度为 $O(|U|^2)$ 。

6 粗关系实例分析

为了更好地理解粗函数依赖, 下面给出一个粗关系实例。

土壤分析员往往根据土壤的一些特征来对土壤样品分类。假设要基于以下属性分析: 土壤颜色、土壤颗粒大小和土壤质量, 并对三个属性的值域划分如下:

Color = {[black, ebony], [brown, tan, sienna], [white], [gray], [orange]}

Particle-size = {[big, large], [huge, enormous], [medium], [small, little, tiny]}

Quality = {[good], [fertile, productive], [normal], [poor, bad]}

土壤样品的信息存入以下粗关系数据库模式中: SOIL (No, Color, particle-size, quality), No 为土壤样品容器编号, 是 SOIL 模式的关键字。粗关系实例见表1。

根据表1, 我们可以得到:

$$r/color = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}\}$$

$$r/particle-size = \{\{1, 6\}, \{2, 5, 7\}, \{3\}, \{4\}, \{8\}\}$$

$$r/quality = \{\{1, 6\}, \{2, 5, 7\}, \{3, 4\}, \{8\}\}$$

在以下公式中, color, particle-size, quality 导出的等价关系分别简记为 c, p, q 。

$$POS_c(P) = \{1, 2, 3, 4, 5, 6, 7, 8\}$$

$$POS_p(q) = \{1, 2, 3, 4, 5, 6, 7, 8\}$$

因此, 根据定理1, 有:

$$color \xrightarrow{R} particle-size,$$

$$particle \xrightarrow{R} sizequality.$$

根据 $particle \xrightarrow{R} sizequality$ 和 RRDB 描述不确定知识的特性可以得出以下蕴涵的语义信息: 土壤颗粒大小影响土壤质量, 颗粒越大, 土壤质量越好; 颗粒越小, 质量也越差。

用文[6]的函数依赖定义分析粗关系实例表, 根据定义3不难发现, 该表的属性之间不存在函数依赖, 数据相对独立。

因此, 文[6]的函数依赖定义并不能充分反映出 RRDB 粗糙数据之间所蕴含的依赖关系。

表1 土壤信息粗关系实例

No	Color	Particle-size	Quality
1	Brown	Medium	Normal
2	Black, tan	Large	Normal, fertile
3	Gray	Medium, small	Poor
4	Black	Tiny	Poor
5	Gray, brown	Large	Normal, productive
6	Gray, white	Medium	Normal
7	Gray, black	Big	Normal, fertile
8	White, gray, tan	Enormous, big	Good, productive, normal

7 RFD 的推理规则

给定 RRDB $r = \{U, A, D_i\}$, 对于本文的 RRDB 粗函数依赖修正定义, 同样满足定理1关于扩展粗关系数据库模型的三条推理规则。

(下转第95页)

系统状态不会继续发生变化,达到静止状态。定理1正确。

3)综上所述,定理2正确。

2. 算法终止性。假设循环不可终止。存在 ' i ' ($i \leq k$) 使得 $\Phi' = \text{false}$, 也就是规则没有被触发, 或规则的条件为假, 或者规则的动作对系统状态没有影响。这就说明了没有规则可能被触发, 循环状态不可能无限地持续下去, 肯定是可终止的。如果 $\Phi' = \text{true}$ ($i > k$), 规则集可能不可终止, 但程序会因为 $i > k$ 而终止。

3. 算法复杂性。相对于以前所有节点的两两比较, 程序的每次循环只需要计算 n 个节点, 知道存在一个 ' i ' 使得 $\Phi' = \text{false}$, 或者 $i > k$ 退出循环, 其复杂性为 $O(ni)$, 比以前的 $O(n^2)$ 大大降低。证毕。

显然, 循环次数越多, 可满足性判定的代价就越高。如果两次以内都是可满足的, 我们可以首先检验其单调性。如果循环中每个被修改的划分都是单调的, 那么这个规则集是可终止的。我们还可以先利用有界性对比较复杂的触发环进行分析。如果 V 是一个节点的表达式。假设其子表达式 E 近似于有界区间 $[a, b]$ ($a \in E \in b$) 包含于表达式 V' 。如果 V' 是不可满足的, 那么 V 也是不可满足的。

到目前为止, 我们都仅谈到了简单触发环。实际上, 有很多触发图包含复杂触发环。对于这样的情况, 我们的程序需要用到嵌套的循环语句。那些循环可以分别检测, 然后进行组合。

结束语 我们提出了一种静态的分析 ECA 规则集可终止性的方法, 把主动规则的执行翻译成一段循环程序。通过这

种方法, 规则集的可终止性变得比较清楚。我们就可以检查循环条件的可满足性及嵌套循环条件的可满足性来判定规则集行为的可终止性。对于复杂触发环我们可以首先进行单调性分析及有界性分析来协助可终止性分析。本文提出的算法可以获得更高的精度及更小的复杂度, 相对于以前的节点对的比较, 对于简单触发环, 其复杂性由 $O(n^2)$ 降到了 $O(n)$ 。我们以后的工作将集中在两个方面:

1. 通过增量分析、施加应用限制来改进可满足性分析程序;

2. 对嵌套循环(复杂触发环)条件的可满足性作进一步的研究。

参考文献

- 1 Baralis E, Bianco A. Performance Evaluation of Rule Semantics in Active Databases. In: 13th ICDE Conf. Proc. April 1997. 365~374
- 2 Baralis E, Ceri S, Paraboschi S. Compile-Time and Runtime Analysis of Active Behaviors. IEEE Transactions on Knowledge and Data Engineering, 1998, 10(3): 353~370
- 3 Bonifati A, Ceri S, Paraboschi S. Active Rules for XML: A New Paradigm for E-Services. In: 1st Workshop on E-Services (co-held with the 26th Very Large Data Bases Conference) Proc. Sep. 2000
- 4 Chakravarthy S, Le R, Dasari R. ECA Rule Processing in Distributed and Heterogeneous Environments. In: Intl. Symposium on Distributed Objects and Applications, 1999
- 5 Bonifati A, Ceri S, Paraboschi S. Pushing Reactive Services to XML Repositories using Active Rules. www 2001. 633~641
- 6 Bailey J, Poulouvasilis A. Termination Analysis of Active Database Rules Techniques for Increased Precision. [Technical Report TR-98-06]. Department of Computer Science, King's College London, April 1998

(上接第92页)

I 自反性: 若 $Y \subseteq X$, 则 $X \xrightarrow{R} Y$ 在 r 上成立。

II 增广性: 若 $X \xrightarrow{R} Y$ 在 r 上成立, 且 $Z \subseteq U$, 则 $XZ \xrightarrow{R} YZ$ 。

III 传递性: 若 $X \xrightarrow{R} Y$ 和 $Y \xrightarrow{R} Z$ 在 r 上成立, 则 $X \xrightarrow{R} Z$ 在 r 上成立。

证明: (1) 自反性: $\forall t_1, t_2 \in r$, 因为 $Y \subseteq X \subseteq A$, 所以 $\theta_x \subseteq \theta_y$, 即有 $\bigcap_{a \in X} \theta_a \subseteq \bigcap_{a \in Y} \theta_a$, 根据定理3可知 $X \xrightarrow{R} Y$ 在 r 上成立。

(2) 增广性: 用反证法。假设 $\exists t_1, t_2 \in r$ 违反 $XZ \xrightarrow{R} YZ$, 即 $t_1[XZ] = t_2[XZ]$, 但 $t_1[YZ] \neq t_2[YZ]$ 。

从 $t_1[XZ] = t_2[XZ]$ 可知, $t_1[X] = t_2[X]$, $t_1[Z] = t_2[Z]$ 。

从 $t_1[YZ] \neq t_2[YZ]$ 可知, $t_1[Y] \neq t_2[Y]$ 或 $t_1[Z] \neq t_2[Z]$ 。

如果 $t_1[Y] \neq t_2[Y]$, 由于 $t_1[X] = t_2[X]$, 则与已知的 $X \xrightarrow{R} Y$ 矛盾; 如果 $t_1[Z] \neq t_2[Z]$, 则与推导出的 $t_1[Z] = t_2[Z]$ 矛盾, 因此假设不成立, 所以 II 是正确的。

(3) 传递性: 用反证法。假设 $\exists t_1, t_2 \in r$ 违反 $X \xrightarrow{R} Z$, 即 $t_1[X] = t_2[X]$, 但 $t_1[Z] \neq t_2[Z]$ 。

如果 $t_1[Y] \neq t_2[Y]$, 则与已知的 $X \xrightarrow{R} Y$ 矛盾; 如果 $t_1[Y] = t_2[Y]$, 则与已知的 $Y \xrightarrow{R} Z$ 矛盾, 因此假设不成立, III 是正确的。

结论 本文中运用粗集理论的研究方法, 改进了 RRDB 中的函数依赖定义, 并给出了判断粗函数依赖是否成立的算

法, 使之更客观地反映 RRDB 数据的语义联系, 体现现实世界不确定性信息的粗糙性和不完备性。本文用粗关系实例验证了粗函数依赖定义的优越性。在此基础上, 探讨了粗函数依赖的推理规则, 并得出了一些结论, 这对 RRDB 理论的进一步研究奠定了基础。

当然, RRDB 数据依赖理论待研究的问题还有很多, 比如: 基于 Rough 关系数据库的数据约束和数据完整性、函数依赖与键的性质等, 这些问题的研究是建立 RRDB 完善的理论体系的基础。

参考文献

- 1 Pawlak Z. Rough sets[J]. International of Information and Computer Science, 1982, 11(5): 341~356
- 2 Pawlak Z. Rough sets-theoretical aspects of reasoning about data [M]. Dordrecht: Kluwer Academic Publishers, 1991. 68~162
- 3 Beaubouef T, Petry F, Buckles B. Extension of the relational database and its algebra with rough set techniques. Computational Intelligence, 1995, 11: 233~245
- 4 Beaubouef T, Petry F, Arora G. Information theoretic measures of uncertainty for rough sets and rough relational databases. Information science, 1998, 109: 185~195
- 5 Beaubouef T, Petry F. Fuzzy Set Quantification of Roughness in a Rough Relational Database Model. In: IEEE Intl. Conf. on Fuzzy Systems. 1994. 172~177
- 6 Nakata M, Murai T. Data Dependencies over Rough Relational Expressions. In: IEEE Intl. Fuzzy Systems Conf. 2001. 1543~1546
- 7 安秋生, 徐久成, 沈钧毅, 等. Rough 关系数据库模型及其关系操作. 计算机科学, 2002, 29(7): 72~89
- 8 安秋生, 徐久成, 沈钧毅, 等. 基于粗糙关系数据库的粗糙数据查询. 西安大学学报, 2002, 36(8): 859~862
- 9 施伯乐, 丁宝康, 周傲英等 编著. 数据库系统教程. 高等教育出版社, 1999, 12: 120~148
- 10 Hailperin T. Probability Logic. Notre Dame Journal of Formal Logic, 1984, 25(3): 198~212