

多値従属を考慮した関係表現における制約の導出

大阪大学基礎工学部 伊藤 実 岩崎 元昭
 谷口 健一 高 忠雄

1. 予えかき

関係データベース (relational database) における質問 (query) を表現する言語、すなわち、関係言語の一つとして関係代数 (relational algebra) が知られている。R を、 s 個の関係スキーム (relational scheme) R_1, \dots, R_s からなる関係データベーススキーマ (relational database schema) とする。関係代数による質問は、オペランドに関係スキーム名をもち、オペレータとしてその関係代数で定義されている基本演算をもつ一つの式 (以後、関係表現 (relational expression) という) で表される。基本演算として、関係和 (union)、関係差 (set difference)、直積 (cross product)、射影 (projection)、及び、選択 (selection) を考える。(関係積 (intersection) や結合 (join) は、それらの基本演算の組合せで実現できる⁽⁹⁾) 本論文では、選択演算として、(1) 特定の成分 (属性) に指定

された定数をもつ組だけを関係から選ぶものと、(2)特定の2つの成分が同じ値をもつ組だけを関係から選ぶものを考える。

R 中の各関係スキーム R_i は一般にその関係例 (relation instance) が満たすべき制約 (従属性 (data-dependency) 等) が指定されている。 R のデータベース $I = \{r_1, \dots, r_s\}$ に対して、特に各 r_i ($1 \leq i \leq s$) が R_i の関係例ならば、 I を R のデータベース例 (database instance) といい、 R 上の関係表現 E は、 R の各データベース I からそれに対応する一つの関係 $E(I)$ への写像を定義する。 E を利用者による質問とすると、 $E(I)$ はその質問に対する利用者ビュー (user's view) と考えられる。本論文では、 R (すなわち、各 R_i の属性数と制約集合)、及び、 R 上の関係表現 E と、一つの制約 d が任意に与えられたとき、その制約 d が E で成立するかどうか、すなわち、 R 上の任意のデータベース例 I に対して関係 $E(I)$ がその制約 d を常に満たすかどうかの判定問題について考察する。 E で成立する制約を求めることは、質問処理、特に、ビュー上での利用者によるデータの更新処理を行うのに重要である。例えば、 $E(I)$ に対して、利用者が E で成立する制約に反する更新を行おうとすれば、システムはデータベースの統合性 (integrity) を保つためにそのような更新を拒否しなければならないことがある。

上記の問題に対して、文献(5)に次の結果が示されている。

「 R に属する各関係スキーム R_i に対して、関数従属 (functional dependency: FD)⁽²⁾ の集合 F_i が指定されるとする。このとき、

(1) E を R 上の任意の関係表現とし、 d を任意の FD とすると、 d が E で成立するかどうかの判定問題は決定不能である。

(2) E を関係差を含まない R 上の任意の関係表現とし、 d を任意の FD, 又は、value equality (VEQ), 又は、domain equality (DEQ) とすると、 d が E で成立するかどうかの判定問題は決定可能である。但し、VEQ とは、 $A \equiv c$ なる文で、関係に属する各組の A 成分が常に定数 c をもつという制約を表す。また、DEQ とは、 $A = A'$ なる文で、関係に属する各組の A 成分と A' 成分の値が常に等しいという制約を表す。」

本論文では、上記(2)の結果を多値従属 (multivalued dependency: MVD)⁽²⁾ に拡張したときの結果を示す。

データベースを設計する手法として、関係スキーム R を幾つかのサブスキームの集合 $\{R_1, \dots, R_s\}$ に分解する場合、その分解によって元の R で成立しなけければならない制約は保存される必要がある。これを制約保存条件 (constraint preservation condition)⁽⁷⁾ と呼び、 $\{R_1, \dots, R_s\}$ の任意の関係例の集合を $\{r_1, \dots, r_s\}$ とすると、 $r_1 \bowtie \dots \bowtie r_s$ は R の関係例であるか。」と

いう問題として表すことができる。但し、 \bowtie は自然結合 (natural join)⁽⁹⁾ を表す。これは、 R で指定される各制約が、 $R_1 \bowtie \dots \bowtie R_s$ で成立するかどうかの判定問題に帰着できる。 R 及び R_1, \dots, R_s で指定される制約が FD 及び MVD の場合、この判定問題の決定可能性は、文献(7)に未解決の問題の一つとして挙げられていたが、本論文の結果を用いれば、任意に与えられた FD 又は MVD が $R_1 \bowtie \dots \bowtie R_s$ で成立するかどうかの判定問題が多項式時間で解けることが示せる。

2. 諸定義

値の可算無限集合を $\Gamma = \{c_1, c_2, \dots\}$ とする。次数 (degree) m の関係 (relation) r とは、 Γ に属する値を成分としてもつような m 字組 (m -tuple) の有限集合である。本論文では、関係の直積や和集合をとる演算を考えるので、関係に属する組の各成分を、通常の属性 (attribute) を用いて識別するよりも、成分番号 (domain number) を用いて何番目の成分であるかという形式で識別する方が便利である。

$U = \{1, \dots, m\}$ とおく。関係 r に属する組 t 、及び、 U に属する番号 A に対して、 t の A 成分を $t[A]$ と表す。更に、 U の部分集合 $X = \{A_1, \dots, A_k\}$ に対して、 $\langle t[A_1], \dots, t[A_k] \rangle$ を $t[X]$ と表す。特に各成分の順番が問題になる場合 (例え

ば、組の成分の置換)には、 $\sigma(X) = A_{i_1} \cdots A_{i_k}$ なる系列化関数 σ を定義し、組 $\langle t[A_{i_1}], \dots, t[A_{i_k}] \rangle$ を $t[\sigma(X)]$ と表す。但し、 (i_1, \dots, i_k) は $(1, \dots, k)$ のある置換 (permutation) を表す。以下では、番号の集合 X, Y に対して、その和集合 $X \cup Y$ を XY と略記することがある。次に、本論文で使用する関係に対する 5 種類の演算を定義する。

$Y \in U$ 上の関係とし、 $X \in U$ の部分集合とする。このとき、関係 Y の $\sigma(X)$ 上への射影 (projection) とは、 $\{t[\sigma(X)] \mid t \in Y\}$ なる関係で、 $Y[\sigma(X)]$ と表す。但し、 σ は X に対する系列化関数である。 $\sigma \in U$ に対する系列化関数とすれば、射影 $Y[\sigma(U)]$ によって Y の各列の置換が行える。

Y_1 を次数 m の関係とし、 Y_2 を次数 n の関係とする。このとき、 Y_1 と Y_2 の直積とは、 $\{t_1 \cdot t_2 \mid t_1 \in Y_1, \text{ 且つ } t_2 \in Y_2\}$ なる次数 $m+n$ の関係であり、 $Y_1 \times Y_2$ と表す。但し、 $t_1 \cdot t_2$ は t_1 と t_2 の連接、すなわち、 $t_1 = \langle a_1, \dots, a_m \rangle$, $t_2 = \langle b_1, \dots, b_n \rangle$ とおけば、 $t_1 \cdot t_2 = \langle a_1, \dots, a_m, b_1, \dots, b_n \rangle$ である。同様に、次数 m_1 の関係 Y_1, \dots , 次数 m_k の関係 Y_k (Y_1, \dots, Y_k はすべて異なる関係である必要はない) に対して、 $Y_1 \times \cdots \times Y_k$ は $\{t_1 \cdots t_k \mid t_1 \in Y_1, \dots, t_k \in Y_k\}$ なる次数 $\sum_{i=1}^k m_i$ の関係である。このとき、 Y_i に属する組 t_i の j 番目の成分は、 $Y_1 \times \cdots \times Y_k$ では $(\sum_{l=1}^{i-1} m_l + j)$ 番目の成分であるが、簡単のため、 $\sum_{l=1}^{i-1} m_l + j$ を

$j^{(i)}$ と略記する。同様に、 $U_i = \{1, \dots, m_i\}$ とすると、 U_i の任意の部分集合 $X = \{A_1, \dots, A_m\}$ に対して、 $X^{(i)} = \{A_1^{(i)}, \dots, A_m^{(i)}\}$ とする。

Y_1, Y_2 をそれぞれ次数 m の関係とする。このとき、 Y_1 と Y_2 の次数が等しいので、 Y_1 と Y_2 の関係和が定義できる。 Y_1 と Y_2 の関係和とは、 $\{t \mid t \in Y_1, \text{ 又は } t \in Y_2\}$ なる次数 m の関係で $Y_1 \cup Y_2$ と表す。もし、 Y_1 と Y_2 の次数が異なれば、 $Y_1 \cup Y_2$ は定義されない。長個の関係 Y_1, \dots, Y_k がすべて同じ次数をもつならば、それらの関係和を $Y_1 \cup \dots \cup Y_k$ と表す。

選択には 2 種類の演算 value equality (VEQ) と domain equality (DEQ) がある。 Y を次数 m の関係とし、 $U = \{1, \dots, m\}$ とする。VEQ とは $A \equiv c$ なる文である。但し、 $A \in U$ 、且つ、 $c \in \Gamma$ である。このとき、 $Y[A \equiv c]$ を $\{t \mid t \in Y \text{ 且つ } t[A] = c\}$ なる関係とする。VEQ の集合 $\{A_1 \equiv c_1, \dots, A_n \equiv c_n\}$ をまとめて、 $A_1 \dots A_n \equiv c_1 \dots c_n$ と書くことがある。DEQ とは $A = A'$ なる文である。但し、 $A, A' \in U$ である。このとき、 $Y[A = A']$ を $\{t \mid t \in Y \text{ 且つ } t[A] = t[A']\}$ なる関係とする。DEQ の集合 $\{A_{i_1} = A_{j_1}, \dots, A_{i_n} = A_{j_n}\}$ をまとめて、 $A_{i_1} \dots A_{i_n} = A_{j_1} \dots A_{j_n}$ と書くことがある。

関数従属 (FD) とは、 $X \rightarrow A$ なる文である。但し、 $X \subseteq U$ 且つ $A \in U$ である。次数 m の関係 Y に属する任意の組 t_1, t_2

に対して、もし $t_1[X] = t_2[X]$ ならば、 $t_1[A] = t_2[A]$ であるとき、関係 r は $X \rightarrow A$ を満たすという。FD の集合 $\{X \rightarrow A_1, \dots, X \rightarrow A_m\}$ をまとめて $X \rightarrow A_1 \dots A_m$ と表すことがある。

多値従属 (MVD) とは、 $X \twoheadrightarrow Y$ なる文である。但し、 $X, Y \subseteq U$ である。 $Z = U - XY$ とおく。次数 m の関係 r に属する任意の組 t_1, t_2 に対して、もし $t_1[X] = t_2[X]$ ならば、 $t_1[XY] = t_2[XY]$ 且つ $t_1[XZ] = t_2[XZ]$ なる組 t_3 が r に属するとき、関係 r は $X \twoheadrightarrow Y$ を満たすという。

ある組 t に対して、 $t[A] = c$ であるとき、 t は $V \in Q$ $A \equiv c$ を満たすという。更に、関係 r に属する任意の組 t が $A \equiv c$ を満たすとき、関係 r は $A \equiv c$ を満たすという。同様に、ある組 t に対して、 $t[A] = t[A']$ であるとき、 t は $D \in Q$ $A = A'$ を満たすという。関係 r に属する任意の組 t が $A = A'$ を満たすとき、関係 r は $A = A'$ を満たすという。

r をある関係とし、 $M \in MVD$ の集合とする。このとき、 r に対する M のもとでの chase とは、 r (に属する組) を含み、且つ、 M に属するすべての MVD を満たす最小の関係で、 r^* と表す。 r^* を実際に求める手続は文献(6)に書かれている。 r に含まれる組の個数を n とすれば、一般に r^* を記述するのは $O(n \cdot \deg(r))$ の領域が必要である。但し、 $\deg(r)$ は関係 r の次数を表す。

関係スキームを $R \langle m, D \rangle$ と表す。但し、 R はスキーム名、 m は次数、 D は FD 及び MVD の集合とする。次数 m の関係 r が D に属するすべての FD 及び MVD を満たすならば、その関係 r を R の関係例（又は R の制約を満足する関係）という。関係スキームの集合 $R = \{R_1 \langle m_1, D_1 \rangle, \dots, R_s \langle m_s, D_s \rangle\}$ を関係データベーススキーマという。各関係 r_i が次数 m_i であるような関係の集合 $I = \{r_1, \dots, r_s\}$ を R のデータベースという。更に、各 r_i が R_i の関係例であれば、 I を R のデータベース例（又は R の制約を満足するデータベース）という。

R 上の関係表現とは、オペランドに R に属する関係スキーム名をもち、オペレータとして射影、直積、関係和、 $\vee \in Q$ 及び $D \in Q$ をもつ一つの式である。従って、 $R_1, R_2[\sigma_1(x)], (R_1[\sigma_2(y)]) \times R_2 \times R_3, R_1 \cup (R_2[\sigma_3(z)]), R_1[A \equiv c], R_2[A = A']$ 等はすべて R 上の関係表現である。 E を R 上の関係表現とし、 $I = \{r_1, \dots, r_s\}$ を R のデータベースとする。このとき、 E は I から一つの関係 $E(I)$ への写像を定義する。その値は、 E のオペランドに現れる各関係スキーム名 R_i に関係 r_i を代入することによって得られる。以下では、射影、 $\vee \in Q$ 及び $D \in Q$ は直積よりも強い演算で、且つ、直積は関係和よりも強い演算とみなして、誤解のない限り関係表現に現れる括弧を省略する。 R 上の関係表現 E に対して、 E の次数 ($\text{deg}(E)$) と表

す) を、関係 $E(I)$ の次数と定義する。

R に属する関係スキーム $R_i \langle M_i, D_i \rangle$ 及び一つの制約 (すなわち、FD, MVD, VEQ 又は DEQ) d に対して、もし R_i の任意の関係例が d を常に満たすならば、 d は R_i で成立する (valid) という。同様に、 R 上の関係表現 E が与えられたとき、もし R の制約を満たすデータベース I に対して、 $E(I)$ が制約 d を常に満たすならば、 d は E で成立するという。

E を関係和を含む R 上の関係表現とする。このとき、 E は $(R_{k_1} \times \dots \times R_{k_m}) [Z \equiv V] [P = Q] [\sigma(W)]$ という標準形に変換できることが知られている。⁽⁵⁾ この変換は $O(|E|^2 + \text{deg}(E))$ 時間で容易に行える。但し、 $|E|$ は E の記述の長さを表す。 E の標準形から射影を除いた $(R_{k_1} \times \dots \times R_{k_m}) [Z \equiv V] [P = Q]$ を E のフレーム (frame) という。

3. 結果

[定理1] $R = \{R_1 \langle M_1, D_1 \rangle, \dots, R_s \langle M_s, D_s \rangle\}$ を関係データベーススキーマとする。但し、各 D_i は任意のFD及びMVDの集合とする。このとき、 R 上の任意の関係表現 E 及び任意の制約 d が与えられたとき、 d が E で成立するかどうかの判定問題は決定可能である。 \square

[定理 2] $R = \{R_1 \langle m_1, F_1 \rangle, \dots, R_s \langle m_s, F_s \rangle\}$ を任意の関係データベーススキーマとする。但し、各 F_i は FD の集合とする。 E を R 上の任意の関係表現とし、 d を任意の制約 (FD, $\forall E Q$ 又は $\exists E Q$) とする。このとき、 d が E で成立しないうかがどうかの判定問題は NP 完全である。 \square

$R = \{R_1 \langle m_1, D_1 \rangle, \dots, R_s \langle m_s, D_s \rangle\}$ を関係データベーススキーマとする。但し、各 R_i に対して、 $D_i = F_i \cup M_i$ とし、 $U_i = \{1, \dots, m_i\}$ とおく。まず、2, 3 の定義を与える。 E を関係組を含む R 上の任意の関係表現とし、 $U = \{1, \dots, \text{deg}(E)\}$ とおく。

E で成立するすべての $\exists E Q$ によって U は同値類 $\mathcal{L}(E)$ に分割される。すなわち、 U の分割 $\mathcal{L}(E) = \{L_1, \dots, L_p\}$ は、「任意の $A, A' \in U$ に対して、 $A = A'$ が E で成立するとき、かつそのとき限り、 A と A' は $\mathcal{L}(E)$ の同じブロックに属する (すなわち、ある i が存在して、 $A \in L_i$ 且つ $A' \in L_i$) 」という性質をもつ。 $\forall E Q$ $A \equiv C$ が E で成立するための必要十分条件は、ある $A' \equiv C$ なる $\forall E Q$ が E に現れており、且つ、 A と A' が $\mathcal{L}(E)$ の同じブロックに属することである。従って、分割 $\mathcal{L}(E)$ が求まれば、任意の $\exists E Q$ 又は $\forall E Q$ が E で成立するかどうかの判定が行える。 $\mathcal{L}(E)$ の同じブロックに属す

る A, A' に対して、 $A \equiv C$ 及び $A' \equiv C'$ が E に現れており、且つ、 $C \neq C'$ ならば、 E は、 R の任意のデータベース I に対して $E(I) = \emptyset$ なる関係表現である。又 $f(E)$ が求まれば、この判定は可能である。

X を U の部分集合とする。左辺が X で、且つ、 E で成立する FD の右辺全体の集合を X の閉包 (closure) といひ、 $f(X, E)$ と表す。すなわち、 $f(X, E) = \{A \mid X \rightarrow A \text{ は } E \text{ で成立}\}$ と定義する。また、 $X \twoheadrightarrow B$ は E で成立し、且つ、任意の B' ($\emptyset \neq B' \subsetneq B$) に対して $X \twoheadrightarrow B'$ は E で成立しないような右辺 B の集合を X の従属基 (dependency basis) といひ、 $M(X, E)$ と表す。 $M(X, E)$ は U の分割で、 $X \twoheadrightarrow Y$ が E で成立するための必要十分条件は、 Y が $M(X, E)$ に属する幾つかのブロックの和集合で表されることである。従って、 $f(X, E)$ (又は、 $M(X, E)$) が求まれば、左辺が X の任意の FD (又は MVD) が成立するかどうかの判定が行える。関係スキーム $R \langle m, D \rangle$ に対して、もし D が FD 及び MVD の集合ならば、任意の $X \subseteq \{1, \dots, m\}$ に対して $f(X, R)$ 及び $M(X, R)$ を求める $O(m \cdot \|D\|)$ 時間の手続きが知られている⁽⁸⁾。但し、 $\|D\|$ は D の記述の大きさを表す。

関係和を含まない R 上の任意の関係表現は、多項式時間で $(R_{k_1} \times \dots \times R_{k_m}) [Z \equiv V] [P = Q] [W]$ なる標準形に変形できる。

$E = (R_{R_1} \times \cdots \times R_{R_m}) [\exists \equiv V] [P = Q]$, 且つ, $U = \{1, \dots, \text{deg}(E)\}$ とおく。

[定理3] 次の条件1又は条件2が成立するとき, $\mathcal{L}(E)$ が (R, E) の記述長に関して) 多項式時間で求まる。すなわち, 任意の $V \in Q$ 又は $D \in D$ が E で成立するかどうかの判定は多項式時間で行える。

(条件1) 各 M_i ($1 \leq i \leq s$) が空集合, すなわち, 各関係スキームで MVD が制約として与えられていない。

(条件2) E 中には各関係スキーム名が高々2回しか現れない。 \square

[定理4] 次の条件3が成立するとき, 任意の $X \subseteq U$ に対して $\mathcal{F}(X, E)$ 及び $\mathcal{M}(X, E)$ が多項式時間で求まる。すなわち, 任意の FD 又は MVD が E で成立するかどうかの判定は多項式時間で行える。

(条件3) E 中には DEQ が現れない。すなわち, $E = (R_{R_1} \times \cdots \times R_{R_m}) [\exists \equiv V]$ という形をしている。 \square

[定理5] $\mathcal{L}(E)$ が既知ならば, 任意の $X \subseteq U$ に対して, $\mathcal{F}(X, E)$ 及び $\mathcal{M}(X, E)$ は $R, E, X, \mathcal{L}(E)$ の記述長に関して多項式時間で求まる。従って, 定理3の結果より, 条件1

又は条件2が成立するとき、任意のFD又はMVDがEで成立するかどうかの判定は多項式時間で行える。□

[定理6] 上記のすべての定理は、Eが射影を含む場合にも容易に拡張できる。この結果の応用として、任意のFD又はMVDが $R_1 \times \dots \times R_s$ で成立するかどうかの判定は多項式時間で行える。□

6. あとがき

本論文で示した手法を用いて、次のような拡張は容易に行える。

(1) R に属する各関係スキーム R_i の制約の集合 D_i がFD及び結合従属(join dependency: JD)⁽⁶⁾からなる集合とする。JDはMVDの一つの一般化であり、特別な場合としてMVDを含む。このとき、 R 上の任意の関係表現 E 、及び、任意のFD、JD、VE θ 又はDE θ d が与えられたとき、 d がEで成立するかどうかの判定問題は決定可能である。

(2) 上記(1)の結果に加え、関係表現Eに現れる演算として、関係積を許しても、 d がEで成立するかどうかの判定問題は決定可能である。

残された問題の一つは、 R 上の任意の関係を含む関係表現 E に対して、同値類 $\mathcal{E}(E)$ を求める能率のよい手続きを（もし存在するならば）見つけることである。

参考文献

- (1) Beeri, C. On the membership problem for functional and multivalued dependencies in relational databases. *ACM Trans. Database Syst.* 5, 3 (Sept. 1980), 241-259.
- (2) Beeri, C., Fagin, R. and Howard, J. H. A complete axiomatization for functional and multivalued dependencies in database relations. *Proc. 3rd ACM SIGMOD Int. Conf. Management of Data, Toronto, 1977*, pp. 47-61.
- (3) Hagiwara, K., Ito, M., Taniguchi, K. and Kasami, T. Decision problems for multivalued dependencies in relational databases. *SIAM J. Comput.* 8, 2 (May 1979), 247-264.
- (5) Klug, A. Calculating constraints on relational expressions. *ACM Trans. Database Syst.* 5, 3 (Sept. 1980), 260-290.

- (6) Maier, D., Mendelzon, A. O. and Sagiv, Y. Testing implications of data dependencies. ACM Trans. Database Syst. 4,4 (Dec. 1979) 455-469.
- (7) Maier, D., Mendelzon, A., Sadri, F. and Ullman, J. Adequacy of decompositions of relational database.
- (8) Sagiv, Y. An algorithm for inferring multivalued dependencies that works also for a subset of propositional logic. J. ACM 27.2 (Apr. 1980), 250-262.
- (9) Ullman, J. D. Principles of Database Systems. Computer Science Press, 1980.