数理解析研究所講究録
458巻 1982年 152-161

関係データベースにおける Representative Instance に関する一結果

阪大 基礎工 岩崎元昭
伊藤実
嵩忠雄

1. まえがき

最近、universal instanceを一般化した概念として representative instance が提案された[4][6][8]。representative instanceは、データベースに属する各関係が universal instance の射影でない場合においてもデータベース全体の内容を表現できるひとつの関係として提案され、この関係がデータベーススキームで指定されたすべての制約を満たさなければならないと仮定されている。本論文では、制約として関数従属だけを考え、次の2つの多項式時間手続きを示した。

(1) 与えられたデータベーススキームが整合性を保つかどうか、すなわち、与えられたデータベーススキームの任意のデータベースに対して、そのrepresentative instanceが常にすべての関数従属を満たすかどうかを判定する。

(2) データベーススキームが整合性を保つとき、represen-

tative instance の全射影[6]を効率よく計算するための関係表現を構成する。

これらの結果は、文献(7)の結果をさらに一般化したものである。

2. 諸定義

$R=\{A_1,\cdots,A_v\}$ を属性の集合とし、各属性 A_i の定義域を $dom(A_i)$ とする。直積 $dom(A_1)\times\cdots\times dom(A_v)$ の有限な部分集合 r を R 上の関係という。t を $dom(A_1)\times\cdots\times dom(A_v)$ の元（組という）とし、A を R に属する属性とすると、t の A 成分を $t[A]$ と表す。同様に、R の部分集合 $X=\{A_i',\cdots,A_i'\}$ に対して、$(t[A_i'],\cdots,t[A_i'])$ を $t[X]$ と表す。以下では、属性を表すのに A, B, C, … 等の文字を用いて、属性の集合を表すのに X, Y, …等の文字を用いる。属性集合 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。また、ひとつの属性だけからなる集合 $\{A\}$ を単に A と書くことがある。

R 上で定義される関数従属（FD）とは、$X\to A$ なる文である。但し、$X\subseteq R$, $A\in R$ である。R 上の関係 r に属する任意の二つの組 t, s に対して、もし $t[X]=s[X]$ ならば $t[A]=s[A]$ であるとき、r は $X\to A$ を満たすという。FD の集合

$\{X \to A_1, \cdots, X \to A_i\}$ をまとめて $X \to A_1 \cdots A_i$ と書くことがある。FDのある集合Fに対して、Fに属するすべてのFDを満たす任意の関係において常に成立する（F以外の）FDが存在しうる。そのようなすべてのFDの集合を（Fも含めて）F^+ と表す。Rの部分集合Xに対して、Fに関する閉包(closure)を $\text{closure}(X, F) = \{A \mid X \to A \in F^+\}$ と定義する。$\text{closure}(X, F)$ を求める $O(\|F\|)$ 時間の手続きが知られている(1)。

本論文では、一般に変数を含む関係を考える。すなわち、関係に属するある組のある成分が定数ではなく変数であるものも許す。2つの組 t_1, t_2 に対して、$t_1[A] = t_2[A]$ であるのは、t_1 と t_2 がA成分に同じ定数または同じ変数をもつ場合である。Yを一般に変数を含むR上の関係とし、FをR上で定義されるFDの集合とする。Yに対するFのもとでの chase(5) とは、Fに属する各FDに対する次に定義するFD規則を適用してYに現れる変数を定数または変数に置き換える操作である。

［FD規則］ $X \to A$ をFに属するFDとし、t_1, t_2 を $t_1[X] = t_2[X]$ かつ $t_1[A] \neq t_2[A]$ なるYに属する組とする。t_1 はA成分に変数ひをもつとする。$X \to A$ に関するFD規則によって、YのA成分に現れる変数ひはすべて $t_2[A]$ に置き換えられる。このとき、$t_1[A]$, $t_2[A]$ とも異なる定数をも

つ場合には、FD規則は適用されず、Yに対するFのもとでのchaseはFを満たさないという。

ある属性集合U上のデータベーススキームとは、2字組の集合$\mathbb{R}=\{\langle R_1, F_1\rangle, \cdots, \langle R_n, F_n\rangle\}$である。但し、$U=R_1\cup\cdots\cup R_n$かつ各$F_i$は$R_i$上で定義されるFDの集合である。各$R_i$を関係スキームという。関係の集合$I=\{r_1, \cdots, r_n\}$に対して、もし各$r_i$が$R_i$上の関係であり、かつ、$F_i$を満たすならば、Iを$\mathbb{R}$のデータベースという。すなわち、Iはデータベーススキーム$\mathbb{R}$のある時点での値を表す。Iに属する関係には変数は現れないものとする。以下では、データベーススキーム$\mathbb{R}=\{\langle R_1, F_1\rangle, \cdots, \langle R_n, F_n\rangle\}$が与えられたとき、$U=R_1\cup\cdots\cup R_n$、$F=F_1\cup\cdots\cup F_n$と表し、次の2つの仮定をおく。

1. (universal relation scheme assumption) もし$X\to A\in F^+$かつ$XA\subseteq R_i$ならば$X\to A\in F_i^+$である。この仮定は、各属性はただ一つの役割しかもたない、すなわち、任意の属性集合Xに対して、Xの属性間の関係(relationship)は高々一つしか存在しないことを表す。これは関係データベースの設計理論に関する多くの論文で仮定されている。

2. (FDの集合の最適化) 各F_iに属する任意のFD $X\to A$に対して、

(a) $X\to A\notin (F_i-\{X\to A\})^+$

(b) 任意の $X' \subseteq X$ に対して、$X' \to A \in F_i^+$

が成立する。与えられた F_i が(a)、(b)を満たさない場合、$O(|F_i||F_i|)$ 時間でそれらの性質を満たすように変換できる[2]。

本論文では、関係に対する演算として射影、結合、及び、関係和を考える。R 上の関係 r 及び R の部分集合 X に対して、r の X 上への射影を、$r[X] = \{t[X] \mid t \in r\}$ と定義する。R_1 上の関係 r_1 及び R_2 上の関係 r_2 に対して、r_1 と r_2 の結合を、$r_1 \bowtie r_2 = \{t \mid t[R_1] \in r_1$ かつ $t[R_2] \in r_2\}$ と定義する。R 上の 2 つの関係 r_1, r_2 に対して、r_1 と r_2 の関係和を、$r_1 \cup r_2 = \{t \mid t \in r_1$ または $t \in r_2\}$ と定義する。R 上の関係表現とは、R に属する関係スキーム $R_1, \cdots, R_n$ を変数とし、射影、結合、関係和を基本演算とする式で、次のように定義される。

(1) R_i はそれ自身で関係表現である。

(2) E_1, E_2 を関係表現とすると、$E_1[X]$, $E_1 \bowtie E_2$, $E_1 \cup E_2$ は関係表表である。

$I = \{r_1, \cdots, r_n\}$ を R のデータベースとすると、E は I から一つの関係 E(I) への写像を定義する。E(I) の値は、E 中に現れる各変数 R_i に関係 r_i を値として代入し、基本演算の定義に従って計算することにより求まる。R_1 上の関係 r_1 及び R_2 上の関係 r_2 に対して、r_2 が $R_1 \cap R_2 \to R_1$ なる FD を満たすなら

ば、$r_1 \bowtie r_2$を拡張結合(extension join)という[3]。

$I = \{r_1, \cdots, r_n\}$をRのデータベースとする。r_iに属する各組のU-R_i成分に相異なる変数を追加することにより、r_iはU上の関係とみなせる。r_iのU上への関係への拡張を、

$aug_U(r_i) = \{t \mid t[R_i] \in r_i$ かつ U-R_iに属する各属性Aに対して、tのA成分は他のどこにも現れない変数$\}$

と定義する。更に、$aug_U(I) = \bigcup_{i=1}^{n} aug_U(r_i)$と定義する。このとき、Iの表現例(representative instance)とは、$aug_U(I)$に対するFのもとでのchaseによって最終的に得られる関係のことであり、rep(I)と書く。

Rの任意のデータベース$I = \{r_1, \cdots, r_n\}$に対して、その表現例rep(I)が常にFを満たすならば、Rは整合性を保つという。従って、Rが整合性を満たせば、Rの任意のデータベースに対して、$aug_U(I)$に対するFのもとでのchaseは必ずFを満たす。

3. 整合性判定手続き

本節では、データベーススキームRが任意に与えられたときに、Rが整合性を保つかどうかを判定する手続きを示す。

また、手続きの時間複雑度の評価を行う。但し、手続きの正当性の証明は省略する。

158

[整合性判定手続き]

入力：データベーススキーム $\mathbb{R} = \{<R_1, F_1>, \cdots, <R_n, F_n>\}$

方法：ある関係スキーム R_i に対して次の手続き EXAM(R_i) が "no" を返せば $\mathbb{R}$ は整合性を保たない。すべての R_i に対して EXAM(R_i) が "yes" を返せば $\mathbb{R}$ は整合性を保つ。

procedure EXAM(R_i)

begin

a. i 以外の各 j ($1 \leqq j \leqq n$) に対して、$G_j \leftarrow F_j$, $H_j \leftarrow \phi$ とする。(G_j は F_j に属するFDでまだ調べていないもの、H_j は既に調べられたものの集合をそれぞれ表す。)

b. $S \leftarrow R_i$ (以下では、S に対して $R_i \rightarrow S \in F^+$ が成立する。)

c. ある G_j ($1 \leqq j \leqq n$, $j \neq i$) に $X \subseteq S$ なるFD $X \rightarrow A$ が存在する限り、次の1～2を実行する。

1. (G_j に $X \subseteq S$ なるFD $X \rightarrow A$ が存在するとして)
G_j から

(i) $X_1 \subseteq S$, $\cdots$, $X_p \subseteq S$

(ii) closure(X_1, F_j) = $\cdots$ = closure(X_p, F_j)

(iii) closure(X_1, F_j) は G_j に属するFDの中で極小 ($X \subseteq S$ なる任意のFD $X \rightarrow A \in G_j$ に対して closure(X, F_j) $\neq$ closure(X_1, F_j)) を満たすFDの集合 $X_1 \rightarrow A_1$, $\cdots$, $X_p \rightarrow A_p$ を任意に選ぶ。

2. 上記1.で選ばれた各FD $X_g \rightarrow A_g$ に対して次の条件 (不

整合条件という）を満たすかどうかを調べる。

［不整合条件］ $S \cap closure(X_g, F_j) - closure(X_g, H_j) \neq \emptyset$

もし、不整合条件を満たすものが存在すれば"no"を返す。さもなければ、

$S \leftarrow S \cup closure(X_i, F_j)$

$H_j \leftarrow H_j \cup \{ X \rightarrow A \mid X \rightarrow A \in G_j$ かつ $X \subseteq closure(X_i, F_j) \}$

$G_j \leftarrow G_j - H_j$

とする。

d.（Cで"no"が返されずに終了した場合）"yes"を返す。

end　EXAM　□

［定理1］ データベーススキーム $\mathbb{R} = \{\langle R_1, F_1 \rangle, \cdots, \langle R_n, F_n \rangle\}$ が整合性を保つかどうかは、$O(n|F||\mathbb{F}|)$ 時間で判定できる。但し、$F = F_1 \cup \cdots \cup F_n$ とする。□

4. 表現例の全射影の計算

$\mathbb{R} = \{\langle R_1, F_1 \rangle, \cdots, \langle R_n, F_n \rangle\}$ を整合性を保つデータベーススキームとし、$I = \{r_1, \cdots, r_n\}$ を $\mathbb{R}$ のデータベースとする。利用者によるデータベースに対する質問が属性集合 Z を参照するとすると、その答を求めるためには表現例の Z 上への射影を求めることが必要である。利用者は Z 成分に変数を含まない組だけに興味があると仮定する。一般に変数を含む関係

r の Z 上への全射影を $r[Z\text{-total}] = \{t[Z] \mid t \in r$ かつ t は Z 成分に変数を含まない$\}$ と定義する。$rep(I)[Z\text{-total}]$ を求めるのに、$rep(I)$ を求めた後 Z 上への射影をとることは多くの時間を必要とし効率的ではない。本節では、$R_1 \cup \cdots \cup R_n$ の任意の部分集合 Z が与えられたとき、R の任意のデータベース I に対して $E_Z(I) = rep(I)[Z\text{-total}]$ を満たす関係表現 E を多項式時間で求める方法を示す。E_Z は $E_1 \cup \cdots \cup E_n$ の形をしており、各 E_i は拡張結合の系列になっている。従って、$E_Z(I)$ は効率よく求めることができる。

$E_Z = E_1 \cup \cdots \cup E_n$ の求め方を簡単に説明すると、各 E_i は $EXAM(R_i)$ において、$Z \subseteq S$ となるような最短の FD の選び方を、$X_1 \to A_1 \in F_{j1}, \cdots, X_m \to A_m \in F_{jm}$ とする。このとき、$E_i = R_i \cup R_{j1}[closure(X_1, F_{j1})] \cup \cdots \cup R_{jm}[closure(X_m, F_{jm})]$ とする。但し、$closure(X_1, F_{j1}) \subseteq R_{j1}$ である。もし、Z を $closure(R_i, F_1 \cup \cdots \cup F_n)$ ならば E_i は定義されない。

[定理2] R を整合性を保つデータベーススキームとする。属性集合 Z が任意に与えられたとき、R の任意のデータベース I に対して $rep(I)[Z\text{-total}] = E_Z(I)$ となる関係表現 E_Z は $O(n|F||F|)$ 時間で求まる。□

参考文献

(1) Beeri. C. and P. A. Bernstein, "Computational problems related

to the design of normal form relational schemes," ACM Trans. on Database Systems, Vol.4, No.1 (March 1979), pp. 30-59.

(2) Bernstein, P.A., "Synthesizing third normal form relations from functional dependencies," ACM Tran. on Database Systems, Vol.1, No.4 (Dec. 1976), pp. 277-298.

(3) Honeyman, P., "Extension joins," Proc. Int. Conf. on VLDB, 1980, pp. 239-244.

(4) Honeyman, P., "Testing satisfaction of functional dependencies," Proc. XPI Conf., Stonybrook, N.Y., June, 1980.

(5) Maier, D., A.O. Mendelzon and Y. sagiv, "Testing Implications of dependencies," ACM Tran. on Database Systems, Vol.4, No.4 (Dec. 1979), pp. 455-469.

(6) Sagiv, Y., "Can we use the universal instance assumption without using nulls?" Proc. ACM-SIGMOD Int. Conf. on Management of Data, Ann Arbor, April 1981, pp. 108-120.

(7) Sagiv, Y., "A characterization of globally consistent databases and their correct access paths."

(8) Vassiliou, Y., "A formal treatment of imperfect information in database management," Technical Report CSRG-123, University of Tronto, Nov., 1980.