数理解析研究所講究録
458巻 1982年 182-191

分散データベースシステムにおける一般化準結合を用いた質問処理

京都大学 工学部　吉川 正俊
上林 弥彦
矢島 脩三

1. まえがき

分散型データベースにおける質問処理では，全質問処理コストのうちデータ転送コストが支配的であるため，質問処理手順を構築する上では，各地点間のデータ転送量を可能な限り削減することが重要である[1]。異なる地点に存在する関係を結合する場合，データ転送量削減の上で有効な手法として準結合（Semi-join）操作が知られている。

質問処理の最適化を考える上で，等結合質問は重要なクラスである。等結合質問は木型質問と巡回型質問に二分され，このうち木型質問は一般に関係の数をnとすると2(n-1)回の準結合操作を実行することによりすべての関係を解の状態にできることが知られている[2][3]。それに対し巡回型質問は，一般に準結合操作のみを用いて解くことは不可能であり，またたとえ部分的に質問を解く場合でも，必要となる準結合操

作の回数が非常に多くなる[2][3]などの問題があり，現在のところ巡回型質問を効率よく解くための手法は見い出されていない。

このため本稿では準結合操作を一般化準結合操作を一般化した一般化準結合操作を提案する。一般化準結合操作は，2つの関係間の結合属性値に加えて、必要な属性の値をも同時に転送を行なう操作である。質問グラフの全域木を考えることにより，各関係間でどの属性の値の転送が必要かがわかる。またこの全域木に沿って一般化準結合操作を実行することにより，任意の巡回型質問を解くことができることを示す。

2. 諸定義

属性$A_1, A_2, \cdots, A_m$からなる関係Rとは各属性A_i $(i=1,2,\cdots,m)$の定義域の直積の部分集合である。Rの各要素を組と言う。関係Rの属性集合を関係スキーマと言い$R[A_1, A_2, \cdots; A_m]$あるいはRでそれを表わす。関係スキーマの集まりをデータベーススキーマと言い$D[R_1, R_2, \cdots, R_n]$で表わす。データベース状態Dとは，各関係の直積とする。すなわち，

$$D = R_1 \times R_2 \times \cdots \times R_n$$

A_{ik}, A_{jl}をそれぞれR_i, R_jの属性とするとき$R_i.A_{ik} = R_j.A_{jl}$の形をした条件節の論理積結合を等結合条件式と言う。またすべての条件節$R_i.A_{ik} = R_j.A_{jl}$において属性A_{ik}とA_{jl}が等

しいような等結合条件式を半自然結合条件式と言う。さらに，データベーススキーマDにおいて2つの属性A_{ik}とA_{jl}が等しいとき，またそのときに限り条件節$R_i.A_{ik}=R_j.A_{jl}$を持つような等結合条件式を（Dにおける）自然結合条件式と言う。条件式gに対応する質問とはデータベース状態Dをデータベース状態$\{d\in D \mid g(d) \text{ is } true\}$に写す関数であるとする。以後，本稿では質問も条件式と同様にgで表わすものとする。（したがって$\{d\in D \mid g(d) \text{ is } true\}$を単に$g(D)$で表わす。）等結合条件式に対応する質問を等結合質問と言い（半）自然結合条件式に対応する質問を（半）自然結合質問と言う。定義から明らかなように自然結合質問は与えられたデータベーススキーマに対して唯一つ存在する。2つの質問は任意のデータベース状態に対する答が等しいとき等価であると言う。任意の等結合質問は属性名変更することにより自然結合質問に変換することができる[3]。

本稿では質問として自然結合質問を仮定する。

半自然結合質問gに対する質問グラフ$G_g(V_g, E_g, L_g)$はラベル付きの無向グラフである。ここで節点集合V_gはgにおいて参照される関係の集合を表わし，質問中に$R_i.A=R_j.A$なる条件節が存在するときまたそのときに限り節点R_iとR_jの間をラベルAを持つ枝で結ぶ。

質問グラフが木で表わされる半自然結合質問あるいはそれと等価な質問を木型質問と言い，木型質問以外の半自然結合質問を巡回型質問と言う。

[例1] データベーススキーマD[R_1[A,B,C], R_2[A,C,D] R_3[A,B,E]]における以下の3つの質問$Q1$，$Q2$，$Q3$の質問グラフをそれぞれ図1の(a)～(c)に示す。

$Q1$: $(R_1.A=R_2.A)\wedge(R_2.B=R_3.B)\wedge(R_3.C=R_1.C)$

$Q2$: $(R_1.A=R_2.A)\wedge(R_2.B=R_3.B)\wedge(R_3.A=R_1.A)$

$Q3$: $(R_1.A=R_2.A)\wedge(R_2.A=R_3.A)\wedge(R_2.B=R_3.B)$

質問$Q2$はその質問グラフに閉路を持つが，質問$Q3$と等価であるため木型質問である。それに対し質問$Q1$は巡回型質問である。□

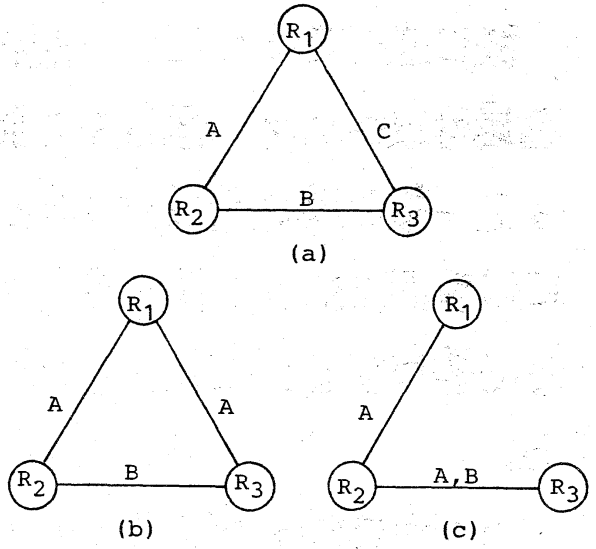

図1. 質問グラフ

R_1とR_2の自然結合は$R_1 \bowtie R_2$で表わす。また$R_1, R_2, \cdots, R_n$のn個の関係の自然結合は簡単に$\bowtie_{i=1}^{n} R_i$で表わす。

R_2によるR_1の準結合は$R_1 \ltimes R_2$で表わされ、次式で定義される： $R_1 \ltimes R_2 \triangleq (R_1 \bowtie R_2)[R_1]$

Xを$R_1 \cap R_2 \subseteq X \subseteq R_2$を満足する属性集合とするとき$R_1 \bowtie (R_2[X]$をXにおける$R_2$による$R_1$の一般化準結合と言い，それを$R_1 \overset{X}{\ltimes} R_2$

で表わす。$R_1 \cap R_2 = X$のときは$R_1 \overset{X}{\times} R_2 = R_1 \ltimes R_2$が成立する。一般化準結合を実行することにより関係スキーマ自体が動的に変化する。

本稿では分散型データベースを仮定し，各関係は互いに異なる一つの地点に存在するものとする。

3. 一般化準結合を用いた質問処理

3.1 質問全域木

本節では，任意の巡回型質問を一般化準結合を用いて解く方法について述べる。

既に述べたように，一般化準結合操作は2つの関係間で結合属性以外の他の属性をと転送を行なう操作である。したがって一般化準結合操作を用いた質問処理手順を構築するためには質問に応じて関係間でどの属性値を転送するかを決定する必要がある。次に，まずそのための写像としてattrを与える。

巡回型質問qが与えられたとき質問グラフG_qの根付き全域木を質問全域木と言い，それをT_qで表わす。T_qは木であるため，節点間の親子関係が定義できる。T_qにおいて節点R_iとその親の節点を結ぶ枝をa_iで表わし，節点R_iとその子孫から成る節点の集合をC_iで表わすものとする。一般性を失うことなくT_qの根をR_1とする。T_qの枝の集合をE_qとするとき，T_qの節

及び枝から属性集合への写像attrを図2に示す手続きによって定義する。ただしlabel(e)は枝e(∈E_g)に付いている属性ラベルを表わすものとする。Sを質問中に現われるすべての結合属性の集合とすると、図2の手続きは$O(|E_g||V_g||S|)$の時間ですべての枝及び節のattrの値を計算する。attrの定義より次の命題が成立する。

[命題1] 節点R_iの親をR_jとするとき

$$attr(a_i) = attr(R_i) \cap attr(R_j)$$

```
do for each  e∈E'q;
     attr(e) = {label(e)};
end;
do for each  Ri∈Vq;
     attr(Ri) = Ri;
end;
do for each  e = (Rj,Rk)∈Eq - E'q;
     do for each  e'∈Tqにおけるパス(Rj,Rk)に含まれる枝の集合;
          attr(e') = attr(e') ∪ {label(e)};
     end;
     do for each  Ri∈Tqにおけるパス(Rj,Rk)に含まれる節の集合;
          attr(Ri) = attr(Ri) ∪ {label(e)};
     end;
end;
```

図2　attrを計算する手続き

［例2］　図3に質問全域木の例を示す。各節及び枝に写像attrを適用した結果は以下のようになる。

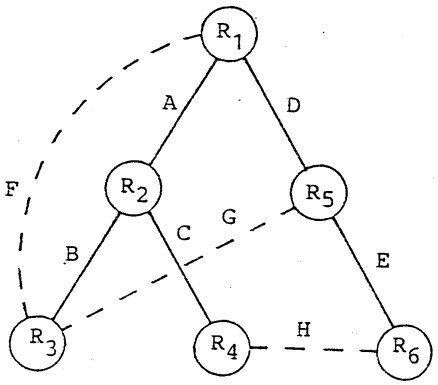

図3　質問全域木

$attr(R_1)=\{A, D, F, G, H\}$

$attr(R_2)=\{A, B, C, F, G, H\}$

$attr(R_3)=\{B, F, G\}$

$attr(R_4)=\{C, H\}$　　$attr(R_5)=\{D, E, G, H\}$

$attr(R_6)=\{E, H\}$

$attr(a_2)=\{A, F, G, H\}$　　$attr(a_3)=\{B, F, G\}$

$attr(a_4)=\{C, H\}$　　$attr(a_5)=\{D, G, H\}$　　$attr(a_6)=\{E, H\}$ □

$(\bowtie_{j=1}^{n} R_j)[attr(R_i)]$ をR_iの既約という。既約な関係は与えられた自然結合条件式を真にするような最小の状態になっている。次に一般化準結合を用いて各関係を既約にするための手続きを示す。

3.2　質問処理手続きUP

T_Qにおいて節R_iがm個($m \geq 0$)の子$R_{i1}, \cdots, R_{im}$を持つとする。このとき，R_iに対する手続きUP(Q, R_i)を以下のように再帰的に定義する。

(i) $m=0$のとき　$UP(Q, R_i)=\langle\ \rangle$

(ii) $m \neq 0$のとき

$$UP(Q, R_i)=\langle UP(Q, R_{i1}), \cdots, UP(Q, R_{im}), R_i \overset{}{\underset{attr(a_{i1})}{\triangleright}} R_{i1}, \cdots, R_i \overset{}{\underset{attr(a_{im})}{\triangleright}} R_{im} \rangle$$

ただし，< >内の手続きは左から順に処理を行なうものとする。直感的には，UP($\mathfrak{L}$, R_i) は$T_{\mathfrak{L}}$の葉から節R_iに向けて順に一般化準結合操作を繰り返す手続きである。手続きUP($\mathfrak{L}$, R_i)を単にUPで表わす。

$\hat{R}_i$に属するすべての関係の自然結合を$\bowtie\hat{R}_i$で表わすものとする。このとき次の補題が成立する。

[補題1] 手続きUP($\mathfrak{L}$, R_i)の結果，関係R_iは$\bowtie\hat{R}_i[attr(R_i)]$となる。

（証明） 略 □

補題1よりただちに次の定理が成立する。

[定理1] 手続きUPの結果，関係R_iは既約となる。□

3.3 全域木の根以外の関係の既約化

前項で述べた手続きUPにより，全域木の根に相当する関係R_1は既約となることがわかったが，R_1以外の関係は一般にまだ既約とは言えない。本稿ではR_1以外の関係を既約にするための手続きについて考察する。

[定理2] $T_{\mathfrak{L}}$において関係R_iの状態が$\bowtie\hat{R}_i[attr(R_i)]$であり、$attr(a_i) \subseteq attr(R_j)$を満足する既約な関係$R_j$が存在するならば、一般化準結合操作$R_i \overset{}{\underset{attr(a_i)}{\ltimes}} R_j$により，$R_i$もまた既約になる。

（証明） 略 □

$T_{\underline{g}}$においてR_jがR_iの親の場合は、命題1より$attr(a_i) \subseteq attr(R_j)$が成立する。したがって定理2より，手続きUPを適用後以下で定義される手続きDOWNを実行することにより，すべての関係を既約にできることがわかる。手続きDOWNとUPと同様に以下に再帰的な定義を与える。

(i) $m=0$のとき　$DOWN(\underline{g}, R_i) = \langle \quad \rangle$

(ii) $m \neq 0$のとき　$DOWN(\underline{g}, R_i) = \langle R_{i1} \overset{}{\underset{attr(a_{i1})}{\boxtimes}} R_i, \cdots, R_{im} \overset{}{\underset{attr(a_{im})}{\boxtimes}} R_i,$

$DOWN(\underline{g}, R_{i1}), \cdots, DOWN(\underline{g}, R_{im}) \rangle$

UPと同様に手続き$DOWN(\underline{g}, R_i)$を$DOWN$で表わす。

［例3］　図4に質問グラフと質問全域木を示す。この例では$attr(R_1) = \{A, B, D\}$，$attr(a_4) = \{A, B\}$である。したがって定理2より，手続きUPが終了した時点で一般化準結合操作$R_4 \overset{}{\underset{AB}{\boxtimes}} R_1$を実行することにより，関係$R_4$は既約となる。すなわち，この例の場合，手続きUP終了後，各関係を既約にするため一般化準結合操作の系列としては

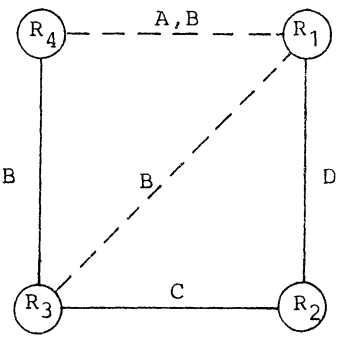

図4. 関係の既約化

$R_2 \overset{}{\underset{ABD}{\boxtimes}} R_1 \Rightarrow R_3 \overset{}{\underset{ABC}{\boxtimes}} R_2 \Rightarrow R_4 \overset{}{\underset{AB}{\boxtimes}} R_3$

及び

$R_4 \overset{}{\underset{AB}{\boxtimes}} R_1$，$R_2 \overset{}{\underset{ABD}{\boxtimes}} R_1 \Rightarrow R_3 \overset{}{\underset{ABC}{\boxtimes}} R_2$

の2通りが考えられる。□

5. あとがき

本稿では、省略したが，通常の準結合に比べて余分に転送が必要な属性値をデータ圧縮することにより一般化準結合を用いた巡回型質問の処理は通常の準結合操作と同等の転送データ量で実行が可能であると考える。

謝辞　御討論頂いた矢島研諸氏に深謝する。なお、本研究は一部文部省科学研究費による。

参考文献

(1) Rothnie,J.B.,Jr. and Goodman,N. "A Survey of Reserch and Development in Distributed Database Management," Proceedings of International Conference on VLDB, pp.48-62, Oct. 1977.
(2) Bernstein,P.A. and Chiu,D.M. "Using Semi-Joins to Solve Relational Queries," JACM, Vol.28, No.1, pp.25-40,Jan. 1981.
(3) Bernstein,P.A. and Goodman,N. "The Theory of Semi-Joins," CCA Rep., No.CCA-79-27, Nov.15, 1979.