

条件式グラフの変形による分散データベースの 質問処理

京都大学・工学部 吉川 正俊 (Masatoshi YOSHIKAWA)

九州大学・工学部 上林 弥彦 (Yahiko KAMBAYASHI)

1. まえがき

準結合は分散データベース等における質問処理コストを削減する上で有用であるが、その能力は限られている。関係間の不等号結合で表現される自然不等号質問の場合、一般に単数属性準結合のみで処理できない質問は以下のいずれかの条件を満たす[1][2]。(1)一対の関係間に複数の結合条件が定義されている。(2)条件式グラフ内に、収束対を1個だけ含むか収束対を含まない閉路が存在する。本稿では、(1)の質問を扱うために複数属性準結合を導入する。また、文献[3]で導入した一般化準結合を不等号の場合へ拡張し、(2)の質問を含む巡回型質問の処理方法を示す。さらに、これらの処理方法を利用した分散データベースの質問処理について考察する。

2. 基本的事項

関係、関係スキーマをそれぞれ R, \underline{R} で表すものとする。条件式 q に対し、 $(\sigma_q(R_1 \times R_2 \times \dots \times R_n))[\underline{R}_i]$ ($1 \leq i \leq n$)を、 R_i の部分解と呼ぶ。本稿では条件式と質問を同一視する。以下、3,4節では、 q に対しすべての関係の部分解を求

める手法を与え、5節ではそれらをもとにより一般的な場合について考察する。A,Bをそれぞれ $R_i, R_j (i \neq j)$ の属性とするとき、 $(R_i.A \theta R_j.B)$ の形の結合節の論理積で表現される質問を不等号結合質問と呼ぶ。ただし、 θ は $=, <, >, \leq, \geq$ のいずれかである。不等号結合質問のうち、制約演算による前処理と属性名変更によって、結合節をすべて $(R_i.A \theta R_j.A)$ の形に変換できるものを自然不等号質問と呼ぶ。以下、自然不等号質問を単に質問と呼ぶ。条件式グラフは質問を表現するラベル付き無向グラフであり、各節点は関係に対応する。また、質問中に R_i と R_j の間で定義される結合節が存在するとき(それらを $c_{ij}^1, c_{ij}^2, \dots, c_{ij}^k$ とする)またそのときに限り、 R_i と R_j に対応する節点間を枝で結ぶ。各結合節に対応してラベル $l_{ij}^1, l_{ij}^2, \dots, l_{ij}^k$ が枝 $\langle R_i, R_j \rangle$ に付加される。 $c_{ij}^h (1 \leq h \leq k)$ が $(R_i.A \theta R_j.B)$ のとき、 l_{ij}^h は $((A,B), \theta)$ とする。また、 $l_{ji}^h = (l_{ij}^h)^{-1} = ((B,A), \theta^{-1})$ と定義する。ただし、 θ^{-1} は θ の逆である("＜", "≧", "="は、それぞれ"＞", "≦", "="の逆とする)。さらに、 $at(c_{ij}^h, R_i) = A, at(c_{ij}^h, R_j) = B, C_{ij} \equiv c_{ij}^1 \wedge c_{ij}^2 \wedge \dots \wedge c_{ij}^k, at(C_{ij}, R_i) = \{at(c_{ij}^1, R_i), at(c_{ij}^2, R_i), \dots, at(c_{ij}^k, R_i)\}, L_{ij} = \{l_{ij}^1, l_{ij}^2, \dots, l_{ij}^k\}$ とし、 $L_{ji} = (L_{ij})^{-1}$ も同様に定義する。条件式グラフの枝 $\langle R_i, R_j \rangle$ は、 $|L_{ij}| > 1$ ならば多重枝と言う。条件式グラフが木となる質問と等価な質問を木型質問、それ以外の質問を巡回型質問と言う。

2つの組 $t_i (\in R_i)$ と $t_j (\in R_j)$ が $C_{ij} (c_{ij}^h)$ を満足するとき、それを $t_i \langle C_{ij} \rangle t_j (t_i \langle c_{ij}^h \rangle t_j)$ で表す。 $p = \langle R_i = R_{i0}, R_{i1}, \dots, R_{i_{k-1}}, R_{ik} = R_j \rangle$ を条件式グラフ中の経路、 X を $X \subseteq R_j$ なる属性集合とする。 $t_{i_{h-1}} \langle C_{h-1 h} \rangle t_{ih} (1 \leq h \leq k)$ なる組 $t_{ih} (\in R_{ih}) (1 \leq h \leq k-1)$ が存在するときまたそのときに限り組 $t_i (\in R_i)$ は p に沿っ

て組 $t_j[X]$ (ただし $t_j \in R_j$)と結合可能であると言う。以後、2,3節では、 C を R_i と R_j の間で定義された結合節の論理積とする。 C 上の R_i と R_j の結合は $R_i \underset{C}{\bowtie} R_j$ で表現する。 c を R_i と R_j の間で定義された結合節とすると、 c 上の R_j による R_i の単数属性準結合は $R_i \underset{c}{\bowtie} R_j$ で表され $R_i \underset{c}{\bowtie} R_j = (R_i \underset{C}{\bowtie} R_j)[\underline{R}_j]$ で定義される。

3. 複数属性準結合

C 上の R_j による R_i の(複数属性)準結合 $R_i \underset{C}{\bowtie} R_j$ は単数属性準結合の自然な拡張として $R_i \underset{C}{\bowtie} R_j = (R_i \underset{c}{\bowtie} R_j)[\underline{R}_j]$ のように定義される。複数属性準結合を用いれば、任意の木型質問及び多重閉路質問[4]は多重枝を含むものでも処理可能となる。

3.1 極小組集合 単数属性不等号準結合(たとえば $R_i \underset{A}{\bowtie} R_j$)を実行するためには、一個の値(即ち $\min(R_j[B])$)を知るだけでよい。一般の複数属性準結合 $R_i \underset{C}{\bowtie} R_j$ を実行するために必要十分な R_j の結合属性値集合を以下で定義する。

[定義3.1] C を関係 R_i と R_j の間で定義された結合条件式とし、 X を C に現れる R_j の属性集合とすると、

(i) for R_i ; $R_i \underset{C}{\bowtie} R_j = R_i \underset{C}{\bowtie} R_j'$ かつ

(ii) for $R_i, R_j' (\subsetneq R_j)$; $R_i \underset{C}{\bowtie} R_j \neq R_i \underset{C}{\bowtie} R_j'$

が成立するような $R_j[X]$ の部分集合 R_j' を R_j のC-極小関係と言う。 ■

3.2 不等号射影 結合条件式の比較演算子に" \neq "を含まない場合は、以下に定義する不等号射影がC-極小関係を与える。

[定義3.2] A_1, A_2, \dots, A_k を属性の集まり、 $\theta_1, \theta_2, \dots, \theta_k$ を対応する比較演算子の集まり(ただし"≠"は含まない)とすると、 $(A_1 \theta_1, A_2 \theta_2, \dots, A_k \theta_k)$ を射影条件式と言う。

[定義3.3] t, t' を属性 A_1, A_2, \dots, A_k を含む関係の組、 $(A_1 \theta_1, A_2 \theta_2, \dots, A_k \theta_k)$ を射影条件式とすると、

$$\bigwedge_{h=1}^k (t[A_h] \theta_h t'[A_h])$$

が成立するときまたそのときに限り、 t は t' より $(A_1 \theta_1, A_2 \theta_2, \dots, A_k \theta_k)$ -小と言う。

[定義3.4] $(A_1 \theta_1, A_2 \theta_2, \dots, A_k \theta_k)$ を $A_h \in R$ ($h = 1, 2, \dots, k$)なる射影条件式とするとき、 R の $(A_1 \theta_1, A_2 \theta_2, \dots, A_k \theta_k)$ 上の不等号射影 $R[A_1 \theta_1, A_2 \theta_2, \dots, A_k \theta_k]$ は $\{t' \mid (R(t) \wedge t'[X] = t \wedge t'$ は半順序 " $(A_1 \theta_1', A_2 \theta_2', \dots, A_k \theta_k')$ -小"において極小}で定義される。ただし、 X は A_1, A_2, \dots, A_k から重複を除去して得られる属性集合、また θ_h' は以下の規則に従って θ_h を置き換えたものである。 ($h = 1, 2, \dots, k$)

θ_h	=	<	>	≦	≧
θ_h'	=	≦	≧	≦	≧

上記の定義は、我々が文献[4]で与えた定義とは、主に θ_h を θ_h' に置き換える点で異なり、改良された結果となっている。

C を結合節 $(R_j A_h \theta_h R_j B_h)$ ($1 \leq h \leq k$)の論理積結合とすると、以下のよ
うに結合条件式を用いて射影条件式を表すものとする。

$$R_j[C] = R_j[B_1(\theta_1)^{-1}, B_2(\theta_2)^{-1}, \dots, B_k(\theta_k)^{-1}]$$

一般に R_j の C -極小関係は $R_j[C]$ となる。したがって、 $R_i \times R_j = R_i \times R_j[C]$ が成立する。

4. 一般化準結合を用いた弱1重閉路質問の処理

C を $at(C, R_j) \subseteq R_j$ なる結合条件式とするとき、 C 上の R_j による R_i の 一般化準結合 $R_i \times R_j$ は $R_i \times R_j = R_i \times R_j[C]$ で定義される。ただし、ここで C' は、 C の結合節の中で $at(c_h, R_i) \subseteq R_i$ を満足するものの論理積とする。

一般に、条件式グラフの閉路の部分进行处理するためには、閉路中の各関係 R_i の各組 t_i に対し、 t_i が閉路に沿って t_i 自身と結合可能か否かを調べる必要がある。閉路を $R_1-R_2-\dots-R_{n-1}-R_n-R_1$ とすると、このための方法として以下の2つが考えられる。

(I) [2つの隣接した関係の結合属性の比較]

以下の3条件を満足する2つの組 $t_1 (\in R_1)$ 及び $t_n (\in R_n)$ が存在するか否かを調べる (図1(a)参照)。

- (i) t_1 は経路 $\langle R_i, R_{i-1}, \dots, R_2, R_1 \rangle$ に沿って t_1 と結合可能である。
- (ii) t_n は経路 $\langle R_i, R_{i+1}, \dots, R_{n-1}, R_n \rangle$ に沿って t_n と結合可能である。
- (iii) t_1 と t_n は C_{1n} を満足する。

(II) [1つの関係の結合属性の2つの属性値集合の比較]

以下の2条件を満足するような組 $t_j (\in R_j)$ が存在するか否かを調べる (図1(b)参照)。

- (i) t_j は経路 $\langle R_i, R_{i+1}, \dots, R_{j+1}, R_j \rangle$ に沿って t_j と結合可能である。

(ii) t_i は経路 $\langle R_i, R_{i \oplus 1}, \dots, R_{i \ominus 1}, R_j \rangle$ に沿って t_j と結合可能である。

ただし、ここで \oplus 及び \ominus はそれぞれmodulo n の加算及び減算であるとする。

[方法(I)に基づく処理] 図2(a)に示される条件式グラフを考える。ある枝(たとえば $\langle R_1, R_5 \rangle$)を他のすべての枝に埋め込み、 $L_{i-1} \cup i$ ($i=1,2,3,4$)と $(L_{51})^{-1}$ の和集合を求めることにより、同図(b)のように条件式グラフを木に変換することができ、変換後の木に沿って同図(c),(d)に示されるような順で一般化準結合を実行することにより、すべての関係の部分解が得られる。即ち、図2(c)に示される一般化準結合列を実行することにより、 $R_1[C_{15}]$ の部分集合で $t_i (\in R_i)$ が経路 $\langle R_i, R_{i-1}, \dots, R_2, R_1 \rangle$ に沿って結合可能なものを求めることができ、同じく図2(d)に示される一般化準結合列を実行することにより、 $R_5[C_{15}]$ の部分集合で t_i が経路 $\langle R_i, R_{i+1}, \dots, R_{n-1}, R_n \rangle$ に沿って結合可能なものを求めることができる。これら2つの集合を C_{15} に基づいて比較することにより、 t_i が R_i の部分解に含まれるか否かが判定できる。実際には、図2(c)及び(d)の一般化準結合列の任意のシャフルを実行すればよい。

[方法(II)に基づく処理] 図2(a)で示される質問において、 a_m (a_M)を $t_i (\in R_i)$ ($i=1,2,3,4$)が経路 $\langle R_i, R_{i-1}, \dots, R_1, R_5 \rangle$ ($\langle R_i, R_{i+1}, \dots, R_5 \rangle$)に沿って結合可能な組 $t_5[A_5]$ のうち最小(最大)のものとする。枝 $\langle R_1, R_5 \rangle$ に対応する比較演算子は" $<$ "であるため、 $a_m \leq a_M$ が成立するときまたそのときに限り、組 t_i は R_i の部分解に含まれる。よって、以下の一般化準結合列(G1),(G2)の任意のシャフルを実行することにより全関係の部分解を求め得る。ただし、質問処

理前に R_5 は A_5 と同じ値を持つ属性 A_5^m, A_5^M を含むように (仮想的に) 拡張されているものとする。

$$\begin{array}{ccccc} R_1 \bowtie_{c'_5} R_5, & R_2 \bowtie_{c'_1} R_1, & R_3 \bowtie_{c'_2} R_2, & R_4 \bowtie_{c'_3} R_3, & R_5 \bowtie_{c'_4} R_4 \end{array} \quad (G1)$$

$$\begin{array}{cccc} R_4 \bowtie_{c'_4} R_5, & R_3 \bowtie_{c'_3} R_4, & R_2 \bowtie_{c'_2} R_3, & R_1 \bowtie_{c'_1} R_2 \end{array} \quad (G2)$$

ただし、
 $C_1': (R_2.B_2 = R_1.A_1) \wedge (R_2.A_5^M \geq R_1.A_5^m)$
 $C_2': (R_3.B_3 \geq R_2.A_2) \wedge (R_3.A_5^M \geq R_2.A_5^m)$
 $C_3': (R_4.B_4 > R_3.A_3) \wedge (R_4.A_5^M \geq R_3.A_5^m)$
 $C_4': (R_5.B_5 \leq R_4.A_4) \wedge (R_5.A_5^M \geq R_4.A_5^m)$
 $C_5': (R_1.B_1 < R_5.A_5) \wedge (R_1.A_5^M \geq R_5.A_5^m)$

一般の巡回型質問処理のためには、文献[3]のように条件式グラフ G_q の spanning tree T を求め、 $G_q - T$ に属する枝をすべて T に埋め込んだ後、本節で示した一般化準結合列を適用すればよい。

5. 分散データベースにおける質問処理への適用

本節では、4節までの結果をもとに、分散データベースの質問処理について基礎的な考察を行う。各関係は互いに異なる地点に存在するものとし、利用者はある属性集合 A_1, A_2, \dots, A_k と対応する比較演算子の集まり $\theta_1, \theta_2, \dots, \theta_k$ に対し

$$(\sigma_q(R_1 \times R_2 \times \dots \times R_n))[A_1 \theta_1, A_2 \theta_2, \dots, A_k \theta_k] \quad (5.1)$$

を答として要求するものとする。例えば、2つの関係、学生(学生名, 学生番号)と成績(学生番号, 科目, 点数)に対し、 $(\sigma_{q1}(\text{学生} \times \text{成績}))[\text{学生名} =, \text{点数} \geq]$ 、ただし、 $q1: (\text{学生. 学生番号} = \text{成績. 学生番号})$ は、各学生ごとに(科目に無関係に)最高

点を求める質問である。これはSQLのGROUP-BY操作に対応するが、不等号射影を用いることによりこのような質問も簡潔に表現することが可能となる。

次に、(5.1)式の質問及び利用者が答を要求している地点(S_u とする)が与えられたときの処理法を示す。

(1) S_u に関係が存在する場合(それを R_u とする)は、 $\{A_1, A_2, \dots, A_k\} - R_u$ を $\{A_1, A_2, \dots, A_h\}$ ($h \leq k$)とする。(一般性を失うことなくそのように属性の名前替えを行える。) S_u に関係が存在しない場合は $h=k$ とする。

(2) R_u が存在する場合は、 R_u を $R_u \times (\text{dom}(A_1) \times \text{dom}(A_2) \times \dots \times \text{dom}(A_h))$ で置き換え、 R_u が存在しない場合は、仮想的に R_u として $(\text{dom}(A_1) \times \text{dom}(A_2) \times \dots \times \text{dom}(A_k))$ が存在するものとする。

(3)各 $A_i \theta_i$ ($i=1,2,\dots,h$)に対し、以下の(4)を実行する。

(4)条件式グラフ G_q において R_u と A_i を含む関係(R とする)の間に枝を付加し、枝 $\langle R_u, R \rangle$ のラベルを $((A_i, A_i), \theta_i^{-1})$ とする。

(5) R_u を根とする条件式グラフの根付き全域木 T を求め、 $G_q - T$ の枝をすべて T に埋め込む[4]。

(6) T に沿って4節の一般化準結合を実行し、 R_u の部分解を求める。

上記のような条件式グラフの変形後、 R_u の部分解を求めることにより、それが(5.1)式の答となることがわかる。

6. あとがき

不等号を含む質問では、収束対[1],[2]の概念が重要であり、収束対を用いた効率的な処理法の開発が課題である。

謝辞 日頃御指導頂く京都大学工学部矢島脩三教授に感謝します。

参考文献

- [1] Bernstein,P.A. and Goodman,N., "Inequality Semi-Joins", CCA Report, No.CCA-79-28, Dec.1979.
- [2] Bernstein,P.A. and Goodman,N., "The Power of Inequality Semijoins", Information Systems, Vol.6, No.4, pp.751-771, Nov. 1981.
- [3] Kambayashi,Y., Yoshikawa,M. and Yajima,S., "Query Processing for Distributed Databases Using Generalized Semi-Joins", Proc. ACM-SIGMOD ICMOD, pp.151-160, June 1982.
- [4] Yoshikawa,M. and Kambayashi,Y., "Processing Inequality Queries Based on Generalized Semi-Joins", Proc. 10th International Conference on VLDB, pp.416-428 Aug. 1984.

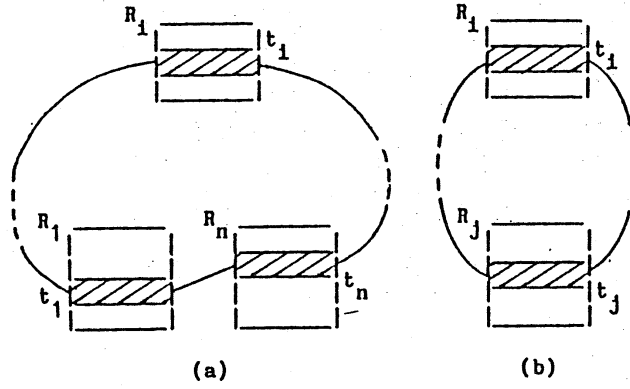
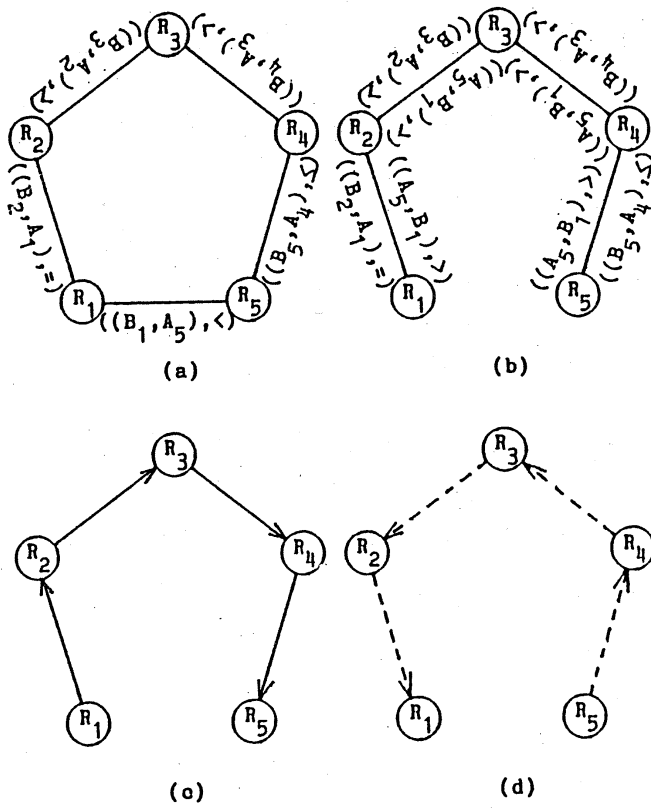


図1. 閉路の処理



注: (c), (d)のラベルは(b)と同じであるため省略されている。

図2. 一般化準結合の適用