数理解析研究所講究録
第556巻 1985年 81-92



# ネットワークデータベースにおける
# 選択・射影・結合質問の処理

京都大学工学部 古川哲也(Tetsuya Furukawa)
九州大学工学部 上林彌彦(Yahiko Kambayashi)

## 1．まえがき

ネットワークデータベースシステムは、効率がよくある程度の自由度の高さを有しており広く普及している。本稿では、ネットワークモデルにおける質問処理について、ＳＰＪ質問と呼ばれる質問を巡航操作で処理する場合について考察している。

ネットワークデータベースにおける質問処理は、巡航操作によりレコードの対応を求めることを基本にしている。巡航操作のみで処理できない場合は、巡航結果に対して関係演算などの親言語による処理を行わなければならない。巡航結果が質問の解になれば、質問処理は親言語による処理を必要とせず容易になる。

関係データベースにおいて“属性集合Ｘの値がｘである属性集合Ｙの値を求めよ”という質問は選択，射影，結合演算のみを用いて解くことができ、ＳＰＪ質問と呼ばれる。このクラスの質問は木質問と巡回質問に分類され、ネットワークモデルにおける木質問は巡航操作のみで解くことができ、対応する関係モデルでも木質問となる[1]。文献[2]では、木質問の処理における最適な巡航順序について検討しており、文献[3]では巡航結果に対して関係演算が行えるときの木質問の処理法を示している。

ネットワークデータベースの構造は、関係を非正規化する基本操作と対応しており、これを用いてより効率のよい処理を行うことができる。文献[2]での処理法はレコードのすべての対応を求めるものであるが、質問処理結果の非正規関係による表示は巡航操作から直接得るこ

とができ、この表示を行うときはレコードの対応をすべて求める必要がなく、レコードの検索数を大幅に減らすことができる。また、この表示は利用者にとっても見易いものとなる。

本稿では、ネットワークデータベースの構造と非正規関係の基本操作との対応を示し、それを用いたＳＰＪ質問の処理法をまず木質問に対して，さらにその結果を巡回質問へ拡張して検討する。

## ２．基本的事項

ネットワークモデルは、同じデータ項目（本稿では関係モデルとの対応のため属性と呼ぶ）からなるレコードの集合であるレコード型と２つのレコード型間のレコードの１対多の対応を表す親子集合型の集合である。属性集合Ｘから成るレコード型Ｒを R(X) で表す。あるレコード型（親レコード型）の１つのレコード（親レコード）に、別のレコード型（子レコード型）の任意個のレコード（子レコード）を関連づけたものを親子集合という。親子集合型は親子集合を個々のレコードではなくレコード型間の関係として記述したもので、親レコード型，子レコード型がそれぞれR,Q である親子集合型を〈R,Q〉で表す。

ネットワークモデルの構造はバックマン線図と呼ばれる有向グラフ $B(N,E)$ で表される。Ｎは各レコード型に対応する節点集合，Ｅは各親子集合型に対応し、親レコード型に対応する節点から子レコード型に対応する節点に向かう有向枝の集合であり、親子集合型名のラベルを付けることがある。

値が定まればレコード型 R(X) の対応するレコードがただ１つ定まるような最小の属性集合Ｋを R(X) のキーと呼ぶ。必ずしも $K \subseteq X$ である必要はなく、Ｒが子レコード型となるすべての親子集合型の親子集合が定まれば（即ち親レコード集合が定まれば）、その子レコード間で属性集合 $K \cap X$ の値によってレコードが一意に定まればよい。バックマン線図で、レコード型のキーを下線を付けた属性集合（レコード型に含まれない属性は括弧を用いる）で表す。

本稿で扱う質問は、属性集合Ｘの値を指定し対応する属性集合Ｙの

値を求めるというもので、 Qr(X;Y) で表す。このような質問は、関係モデルにおいてＳＰＪ（選択・射影・結合）質問と呼ばれるものである。ＳＰＪ質問は、選択，射影，結合のみで表現できる質問で、

$(R_1\ [X_1 = 'c_1\ ']*R_2\ [X_2 = 'c_2\ ']*\cdots)[Y]$

ここで、$X_1 \cup X_2 \cup \cdots = X$.

の形で表される。関係モデルにおいては、結合操作の処理が特にコストが高いこともあり、それに注目して質問全体を木質問と巡回質問に分類する[4]。この分類は、関係の結合を表す質問グラフによって行われ、質問グラフが木で表せる質問を木質問と呼びそれ以外を巡回質問と呼ぶ。

ネットワークモデルにおける質問処理は、巡航操作によりレコードの対応を求めることを基本としている。属性集合Ｘの値が指定された値と等しくなるレコードの対応での属性集合Ｙの値が質問 Qr(X;Y) の解となる。レコードの対応を求める経路がバックマン線図上で有向性を無視したときに木となる質問を木質問，それ以外の質問を巡回質問と呼ぶ。ネットワークモデルの構造が表す結合従属性を対応する関係モデルとしたとき、ネットワークモデルにおける木質問は対応する関係モデルにおける木質問の部分クラスとなり、ネットワークモデルにおける木質問は巡航操作のみで解くことができる[1]。

## ３．非正規関係による質問処理結果の表示

巡航操作によって得られた木質問の解をすべて書き並べると、量も多くなり見にくいものとなってしまう。より見やすく、効率的な解の表示は、質問処理の上で重要な要素となる。関係モデルにおいては、１つの方法として非正規形による表示がある。ネットワーク構造は関係を非正規形へ変換する操作と対応しており、それを用いた表示が可能である。ネットワークモデルにおける質問処理で非正規関係による表示を用いれば、すべてのレコードの対応を求めなくてもよく、レコードの検索回数を減らすことができる。本節では、バックマン線図が木で表されるネットワーク構造と非正規形の操作の対応を考察し、非

正規関係での選択，射影演算、巡航操作から直接その表示を行う方法を示す。

関係モデルにおける正規形から非正規形への変換のうちRow-nest, Group-by, Relation-nest の各操作はネットワーク構造と対応させて考えることができる。各操作の内容は文献[5]に詳しいが、簡単には、

Group-by: 値の組をある属性の値ごとにまとめる.

Row-nest: 属性の値を集合値とする.

Relation-nest: 局所関係をくくり出す.

という操作である。図1(a)の関係を属性Aで Group-by し、属性集合BC,DEF で Row-nest したものが図1(b)，それを Relation-nest したものが図1(c)となる。

ネットワーク構造がレコード型Rを根とした木構造Tのとき、親子集合によるレコードの対応と非正規形の各操作とを比較する。RはTにおいてn個の子レコード型$P_1, P_2, \cdots, P_n$ ($1 \leq i \leq n$)を持ち、それぞれを根とする部分木を$T_{P_1}, T_{P_2}, \cdots, T_{P_n}$とする。各 $T_{P_i}$のレコードの対応はRのレコードごとにまとまっており、 Group-by 操作に対応していると見ることができる。関係モデルにおける Group-by 操作との違いは、関係モデルでは属性集合の値による Group-by であるのに対し

(a)

| A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| 1 | 2 | 1 | 2 | 2 | 2 | 2 |
| 2 | 2 | 2 | 3 | 2 | 2 | 3 |
| 2 | 2 | 2 | 3 | 2 | 3 | 4 |
| 2 | 2 | 3 | 3 | 2 | 2 | 3 |
| 2 | 2 | 3 | 3 | 2 | 3 | 4 |

(b)

| A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 2 | 1 | 2 | 2 | 2 | 2 |
| 2 | 2 | 2 | 3 | 2 | 2 | 3 |
| | 2 | 3 | 3 | 2 | 3 | 4 |

(c)

| A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 2 | 1 | 2 | 2 | 2 | 2 |
| 2 | R1 | | R2 | | | |

R1

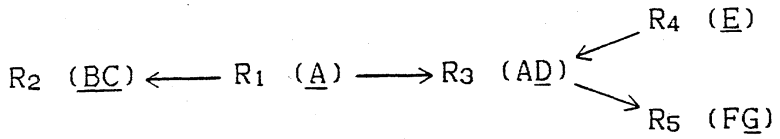
| B | C |
|---|---|
| 2 | 2 |
| 2 | 3 |

R2

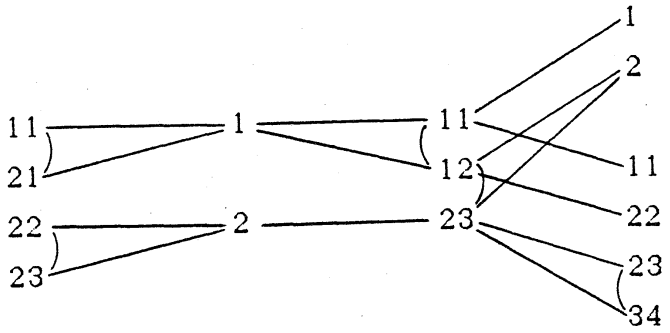
| D | E | F | G |
|---|---|---|---|
| 3 | 2 | 2 | 3 |
| 3 | 2 | 3 | 4 |

図1　非正規関係

(a)




(b)




(c)

| A | R2 | R3 |
|---|---|---|
| 1 | R2 (1) | R3 (1) |
| 2 | R2 (2) | R3 (2) |

(d)

R2 (1)

| B | C |
|---|---|
| 1 | 1 |
| 2 | 1 |

R2 (2)

| B | C |
|---|---|
| 2 | 2 |
| 2 | 3 |

e)

R3 (1)

| A | D | R4 | R5 |
|---|---|---|---|
| 1 | 1 | R4 (11) | R5 (11) |
| 1 | 2 | R4 (12) | R5 (12) |

R3 (2)

| A | D | R4 | R5 |
|---|---|---|---|
| 2 | 3 | R4 (23) | R5 (23) |

f)

R4 (11)

| E |
|---|
| 1 |

R4 (12)

| E |
|---|
| 2 |

R4 (23)

| E |
|---|
| 2 |

R5 (11)

| F | G |
|---|---|
| 1 | 1 |

R5 (12)

| F | G |
|---|---|
| 2 | 2 |

R5 (23)

| F | G |
|---|---|
| 2 | 3 |
| 3 | 4 |

(g)

| R1 | R2 | | R3 | | R4 | R5 | |
|---|---|---|---|---|---|---|---|
| A | B | C | A | D | E | F | G |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 2 | 1 | 1 | 2 | 2 | 2 | 2 |
| 2 | 2 | 2 | 2 | 3 | 2 | 2 | 3 |
| | 2 | 3 | | | | 3 | 4 |

図２　巡航操作と非正規関係の対応

ネットワークモデルではレコードによる Group-by となる点である。各 $TP_i$のレコードの対応はRのレコードを介して直積的に対応している。これは、Tにおけるレコードの対応をRによって Group-by した後の各$TP_i$についてのRow-nest 操作に対応する。また、各 $TP_i$のレコードの対応はRelation-nest 操作による局所関係のくくり出しに対応している。

［例１］　図２(a) のバックマン線図で表されるネットワーク構造で各親子集合は図２(b) であったとする。これは図１(a) と同じ値の対応を記憶している。このとき、 $R_1$ を根として非正規関係との対応を考える。 $R_1$ のレコードで Group-by 操作を行い、その値ｃによる部分木 $R_i$ の局所的なレコードの対応を $R_i$ (c) で表すと図２(c) となる。 $R_2$ (1), $R_2$ (2) は親子集合型$\langle R_1, R_2 \rangle$のレコードの対応から図２(d) となる。$R_3$ (1), $R_3$ (2) は $R_3$ のレコードでGroup-by し、$R_4$ , $R_5$ の局所的な対応をくくり出すことによって図２(e) となり、その局所的対応は図２(f) である。これらをまとめて１つの非正規関係で表すと図２(g) となる。

非正規関係をある属性集合に射影するとき、 Group-by 操作を行った属性が削除される場合はこの操作を表す横線を残すようにする。選択の場合、その属性値が集合値になっていれば選択の条件を満たすもののみを残す。空集合になる場合，又は集合値でない属性で条件を満たさないものについてはその値を含む組を削除する。図２(g) の非正規関係を属性集合BCF に射影すると図３(a), E= 2 による選択は図３(b), その両方を行うと図３(c) となる。巡航操作で木質問を処理するとき、目的属性のみの値を非正規形で出力すればよいが、その出力結果は選択の条件を満たす対応を目的属性に射影したものとなる。巡航は根の各レコードに対し部分木のそのレコードに対応するレコードを検索するというように行う。質問Ｑｒ(E(＝２),BCF) の処理では、まず $R_1$ のレコード１を読み、 $R_2$ (1), $R_3$ (1) を求める。次にレコー

ド2を読み、 $R_2$ (2), $R_3$ (2) を求める。 $R_2$ (1), $R_3$ (1), $R_2$ (2), $R_3$ (2) も同様に根となるレコード型の各レコードごとにその部分木の対応を求めるようにする。レコードを検索したときに順次選択の条件を満たす目的属性を出力する。その際、$R_2$ (1), $R_2$ (2)が解に含まれるかどうかは対応する $R_4$ のレコードで E=2 となるレコードが存在するかによる。従って、まず選択条件を含むレコード型を巡航する必要がある。

## 4．木質問の処理効率

与えられたネットワーク構造に対して木質問を処理する際に、3節での操作を用いて効率化するとき、巡航順により処理効率が異なる。効率に関する要素として、

- スキーマ情報（親子集合の方向）
- 各レコード型のレコード数
- 各親子集合型で1つの親レコードに対応する子レコード数
- 選択条件を満たすレコードの割合

が考えられる。巡航の順序（巡航を開始するレコード型（木の根））をどれにするかは重要であり、これらの要素を考慮して決定しなければならない。そのためには、適当なコストモデルを考える必要があるが、本稿ではレコードを検索する回数をコストとして用いる。効率が最も良い巡航処理は、各レコード型Rについてそれを根にしたときの巡航のコスト（木構造Tでのレコード型Rのコストと呼び、C(T,R)

| B | C | F |
|---|---|---|
| 1 | 1 | {1} |
| 2 | 1 | {2} |
| 2 | 2 | {2} |
| 2 | 3 | {3} |

(a)

| A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 2 | {2} | {2 | 2} |
| | 2 | 1 | | | | |
| 2 | 2 | 2 | 3 | {2} | {2 | 3} |
| | 2 | 3 | | | 3 | 4} |

(b)

| B | C | F |
|---|---|---|
| 1 | 1 | {{2}} |
| 2 | 1 | |
| 2 | 2 | {{2} |
| 2 | 3 | {3}} |

(c)

図3　非正規関係の選択，射影

で表す）を求め、それが最も小さいものを木の根とした巡航である。レコード型のコストは次のようにして再帰的に求めることができる。

木構造Tでコストを求めたいレコード型Rのレコード数をr，Rでの選択条件を満たすレコードの割合をkとする。

・レコード型1個のとき、コストはそのレコード数となるので、

$$C(T,R) = r$$

・1個の子レコード型Pを持つとき、Pを根とする部分木を$T_P$とする（図4(a)）と、Rの各レコードに対してTのレコードの対応はkの確立で一度だけ検索されるので、

$$C(T,R) = r + k \cdot C(T_P, P)$$

・1個の親レコード型Qを持つとき、Qを根とする部分木を$T_Q$，親子集合型<Q,P>でのレコードの対応比を1対nとする（図4(b)）と、Rの各レコードに対して$T_Q$のレコードの対応はn回重複するので、

$$C(T,R) = r + k\{n \cdot C(T_Q, Q) + f(n)\}$$

$f(n)$：子レコードから親レコードをたどるときの余分なコスト

直接ポインタがあるとき　$f(n) = 0$

リンクをたどるとき、1つの親子集合でのi番目のレコードは余分にn-i個リンクをたどる必要がある。親子集合の数は親レコード型のレコード数であり、r/nとなるので、

$$f(n) = \sum_{i=1}^{n} (n-i) \cdot r/n = r(n-1)/2$$

・u個の子レコード型とv個の親レコード型を持つとき、子レコード型の部分木を$T_{P1}, T_{P2}, \cdots, T_{Pu}$，親レコード型の部分木を$T_{Q1}, T_{Q2}, \cdots, T_{Qv}$とする（図4(c)）と、

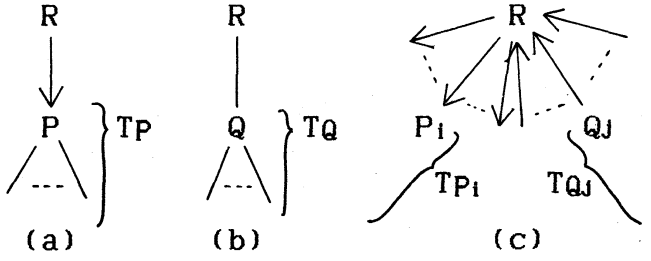



図4　巡航のコスト

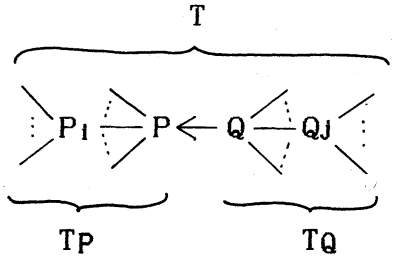



図5　最小コスト

$$C(T,R) = r+k\{\sum_{i=1}^{u} C(TP_i)+\sum_{j=1}^{v}(n_j \cdot C(TQ_j)+f_j(n_j))\}$$

$f_j$ ：Rのレコードからj番目の親レコード型のレコードをたどるときの余分なコスト

$n_j$ ：j番目の親レコード型との親子集合の比

Rの各レコードに対して$TP_1, TP_2, \cdots, TP_u, TQ_1, TQ_2 \cdots, TQ_j$のレコードの対応を求める。これらは Group-by 操作後の Row-nest 操作で直積的に対応しているので、コストは和となる。

このようにして各レコード型についてそれを根としたときの巡航操作のコストを計算できる。コストが最小となるレコード型を求めたいとき、選択条件がなければ（すべてのレコードについて k=1 であれば）すべてのレコード型についてコストを求めなくても、トポロジカルに最小コストのレコード型となりえるものを選ぶことができる。

［定理］　選択のない木質問を巡航操作で処理しその結果を非正規関係で表示する場合、処理コストが最小となるときは、木の根（巡航を出発するレコード型）は他のレコード型の子レコード型ではないレコード型である。

証明　レコード型Pを根としたときで、その木構造Tでの子に親子集合での親レコード型Qがあったときの巡航のコストを、Qを根としたときの巡航のコストと比較する。P,Qのレコード数をそれぞれ p,q, Pを根としたときのTでのPの子の部分木を$TP_1, TP_2, \cdots, TP_u, TQ$ , Qを根としたときのTでのQの子の部分木を$TQ_1, TQ_2, \cdots, TQ_v, TP$ とする（図５）。このとき、

$$C(TP, P) = r+\sum_{i=1}^{u}(n_i \cdot C(TP_i)+f_i(n_i))$$

$$C(TQ, Q) = q+\sum_{j=1}^{v}(m_j \cdot C(TQ_j, Q_j)+g_j(m_j))$$

$f_i$（$g_j$）：P(Q)から$P_i$（$Q_j$）をたどるときのコスト

P(Q)が$P_i$（$Q_j$）の親レコード型であれば $f_i = 0$

$n_i$（$m_j$）：P(Q)は$P_i$（$Q_j$）の親レコード型→1

P(Q)は$P_i$（$Q_j$）の子レコード型→親子の比

親子集合型<P,Q>のレコードの比を1対n，PからQをたどるときのコストを$f(n)$とすると、

$$C(T,P) = p+\sum_{i=1}^{u}(n_i\cdot C(T_{Pi})+f_i(n_i))+n\cdot C(T_Q,Q)+f(n)$$
$$= C(T_P,P)+n\cdot C(T_Q,Q)+f(n)$$
$$C(T,Q) = q+\sum_{j=1}^{v}(m_j\cdot C(T_{Qj})+g_j(m_j))+C(T_P,P)$$
$$= C(T_Q,Q)+C(T_P,P)$$
$$\therefore C(T,P)-C(T,Q)=(n-1)\cdot C(T_Q,Q)+f(n)\geqq 0$$

従って、レコード型Pが親レコード型Qを持てば、Pを根とする巡航よりもQを根とする巡航の方がコストが小さい。　　Q．E．D．

選択のない木質問を巡航操作で解くとき、コストを最小にするには、定理より巡航を始めるレコード型を子レコード型とはならないものから選べばよい。即ち、定理は木の親子を親子集合の親子になるべく一致させた方がよいことを示している。親子集合は階層構造を表現しているとも見ることができ、その親から巡航を行うことは、巡航結果に階層構造を保存することとなり、利用者にとって理解しやすいものとなる。質問が選択を含む場合は、処理効率のよい巡航順は選択条件を満たすレコードの割合に左右される。

## 5．巡回質問への拡張

木質問に対しては、非正規関係による表示を用いて効率よい巡航処理ができることを示した。本節では、巡回質問の処理に対してその結果を応用する。簡単のため、バックマン線図上で1つの閉路Oからなる巡回質問について考察する。即ち、n個のレコード型$R_0, R_1, \cdots, R_n$ $(=R_0)$を巡航する必要があり、$\langle R_i, R_{i-1}\rangle$ $(1\leqq i\leqq n)$が存在する（図6）。

Oでのレコードの対応は次のように考えることができる。

$R_0$ $(=R_n)$のレコード $r_0$ に対しM:$R_0, R_1, \cdots, R_i$ の経路の親子集合でつながっているレコード集合とN:$R_n$ $(=R_0)$, $R_{n-1}, \cdots, R_i$ の経路の親子集合でつながっているレコード集合で、$R_i$ のレコ

ードは一致する。

関係演算を用いることができれば、M,N の経路でのレコードの対応を $R_1$ の属性で結合することにより０の対応を求めることができる。

巡航操作で対応を求めるときは、１つの変数ｘを必要とする。Ｍでのレコードの各対応に対しｘの値を $R_1$ のレコードのキー値とする。同じ $R_0$ のレコードを含むＮの対応で、$R_1$ にキー値がｘとなる選択条件をつければよい。このとき、$R_0$ を出発点，$R_1$ を結合点と呼ぶことにする。このような巡航操作でも非正規関係による表示を用いることができる。

［例２］　図７(a) のバックマン線図で親子集合が図７(b) のとき、$R_0$ を出発点，$R_2$ を結合点とすると巡航結果は図７(c) となる。

例２　では結果の非正規関係に直積的対応はない。一般に１つの閉路のみのときは Row-nest に対応する効率化はないが、閉路だけでなく他のレコード型も巡航する必要があれば巡航操作は Row-nest に対応させて効率化される。

## ６．あとがき

ネットワークデータベースにおけるＳＰＪ質問の非正規関係による

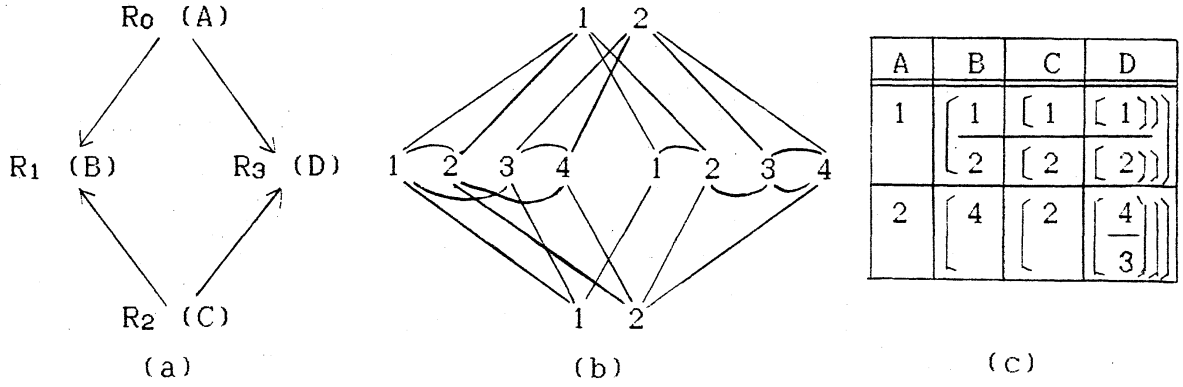



図６　巡回質問の処理

表示を用いた巡航操作による処理法を示した。この表示を用いれば、利用者にとっては見易くレコードの検索回数も少ない巡航を行うことができる。今後は親言語の処理も考慮した質問処理についても検討したい。文献[3]では巡航により得られた値の対応に関係演算を用いた処理方法について述べているが、最適な巡航の分割などの問題もある。

謝辞　御討論頂く京都大学矢島脩三教授，並びに矢島研究室の諸氏に感謝致します。

参考文献

[1] 古川，上林，“ネットワークデータベースにおける木質問”，電子通信学会技術研究報告，AL84-62，1985 年1 月．

[2] Dayal,U., Goodman,N., "Query Optimization for CODASYL Database Systems", Proc. of ACM SIGMOD Int. Conf. on Management of Data, pp.138-150, June 1982.

[3] Chen,H., Kuck,S.M., "Combining Relational and Network Retrieval Methods", Proc. of ACM SIGMOD Int. Conf. on Management of Data, pp.131-141, June 1984.

[4] Bernstein,T.A., Goodman,N., "Power of Natural Semijoins", SIAM J. Comput., Vol.10, No.4, pp.751-771, Nov. 1981.

[5] 上林，田中，武田，矢島，“関係データベースにおける意味制約を反映した非正規形の設計問題”，情報処理学会論文誌，第24 巻，第6 号，1983 年11 月．