

集合制約質問の記述方法および計算量

Representation and Complexity of Set Constraint Queries

岩井原瑞穂

上林 彌彦

九州大学工学部

1. まえがき

データベースに対する質問の答えは、条件を満たす組またはオブジェクトの集合であることが多い。これに対して、答えが条件を満たす集合、または組合せの集合であるような質問が考えられる。集合の個々の要素が条件を満たすだけではなく、集合全体として条件を満たすような質問である。このような条件を満たす集合を求める質問を、本稿では集合制約質問と呼ぶ。

例えば、次のような質問である。

【例 1.1】 次のような組み合わせを求める。

- ・ 各人の時間帯が重ならない作業の割り当て
- ・ 互いの距離が r 以下の部品の集合
- ・ 年齢が 5 才離れている人のペア $\square$

データベース理論において、集合が陽に構成要素として扱われる例として、オブジェクト指向データベースや、非正規関係などが挙げられる[13]。このような場合、集合の構造はスキーマとして静的に固定されて扱われることが多く、集合制約質問のように、動的に集合を生成する機能はあまり考慮されていない。

正規形から非正規形への変換、すなわち集合を生成する操作は NEST/UNNEST [7] など、基本的に同じ値を持つ組を 1 つにまとめる操作のみで、組合せの条件を記述するには、表現力が不足する[12]。

論理データベースにおいては、再帰質問[2]を用いて推移的閉包や、親子関係での同世代の集合などを求めることが出来る。これらの質問は、最小不動点が存在することにより、答えは 1 つの集合となるが、組合せを求める問題では、解となる集合は複数存在するのが普通である。集合型の変数を導入することが自然であると考えられる。また、集合制約質問のクラスは推移的閉包を含む。

論理データベースへの集合型の導入に関する研究がある[1][9][10]。これらの文献では、述語論理の強力な記述力を用いた、大変表現力の高いデータベース言語が提案されている。本稿で考察する質問も、これらの言語で表すことができる。

これらの言語の処理系を実現する場合に、集合を扱うときに生じる特有の問

題がある。例えば、答えの数が指数的になる場合や、集合のネストの深さを任意にすると、少なくとも指数領域を必要とする質問が書けること[5]などである。

このような困難な質問も言語で記述できることは問題があり、除外されるように制限することが望ましいが、そのためには容易に求められる質問のクラスが明確にされねばならない。本稿では、集合制約質問で現実的な時間で求められる質問のクラスについて考察している。

集合制約質問の機能を備えるためには、集合全体を制約するような、条件を明確に記述できる手段が必要となる。本稿では、反射的・対称的である二項関係を与えられた条件とし、この二項関係の数学的性質より定義される集合を、質問の答えとする方法をとる。

例 1.1 においては、二項関係 γ_r を用いて、答え S は次のように表される。

$$t_1 \gamma_r t_2 \text{ iff } \|t_1[\text{座標}] - t_2[\text{座標}]\| \leq r$$

$$\forall t_1, t_2 \in S \quad ; \quad t_1 \gamma_r t_2$$

以下 2 節では基本的事項を説明し、3 節では集合制約質問の定義及び質問の分類を行う。そしてその部分クラスである、FD 型集合制約質問について述べる。4 節ではその処理法を検討する。質問の条件の FD が 1 つ及び 2 つの場合の多項式アルゴリズムを述べ、3 つの場合は NP 完全になることを示す。また、データベースで保持されている JD を用いた処理の効率化を検討する。

2. 基本的事項

まず関係データベースについて説明する。

属性集合 $X = \{A_1, A_2, \dots, A_n\}$ の各属性 A_i は、定義域 D_i を持つ。属性集合 X からなる関係 $R(X)$ は、 $t : \{A_1, A_2, \dots, A_n\} \rightarrow D_1 \times D_2 \times \dots \times D_n$ で表される写像 t の有限集合であり、写像により各 A_i は D_i の要素に写されるものとする。 X を関係スキーム、 t を R の組と呼ぶ。関係 R は列を属性、行を組とした、表の形で表現できる。本稿では、属性を $A, B, C, \dots$ で表し、属性集合を $X, Y, \dots$ で表す。属性集合の接続は、その和集合を表す。

反射的・対称的な二項関係を同調関係といい、推移的な同調関係を同値関係という。 A 上の同調関係 γ において、 A の部分集合 C で、 C の任意の 2 要素が γ を満たすものを同調類という。

3. 集合制約質問

3.1 集合制約質問の定義

同調関係により定まる同調類は、同調関係による制約を受けた集合であるといふことができる。その集合は同調関係の定義を反映した性質を持つ。これ

を用いて、任意の要素が互いに似ているとか、同じである、近くにある、互いに独立である、交換可能である、などの集合を求めることができる。関係の組を1つの実体とし、組の属性値を実体の性質とみなし、質問で得たい組集合の特徴を、属性値で定義された2組間の同調関係で表すことにより、解の組集合を定めることができる。

【定義 3.1】

関係 $R(X)$ の組集合を定義域とする、反射的かつ対称的な二項関係 γ が与えられたとき、 γ を $R(X)$ 上の集合制約質問と呼ぶ。質問の解は次の条件を満たす集合 S である。

$$\forall t_1, t_2 \in S \subset R(X); \quad t_1 \gamma t_2 \quad \square$$

定義 3.1 では、 S が解ならばその部分集合も解となり、解の数は指数関数的になる。そのため、次の方法が考えられる。

① 極大集合を解とする。

しかし n 個の点からなるグラフの極大クリークの最大数は、 $3^{n/3}$ であるため、全ての解を出力することは適当でない。

② 極大集合の解を1つ出力する。

これは、組を解に1つずつ加えてゆくことにより、極大にでき、多項式時間で可能である。しかし出力される解は組の入力の順序により異なったものとなる。

③ 最大集合の解を1つ出力する。

組合せを求める問題では、要素数最大の組合せが最適であることがよくある。しかし最大クリーク問題は NP 完全であり [11]、任意の質問に対する処理法を求めることは難しい。

3.2 質問の分類

集合制約質問は、集合の定義に二項関係を用いているため、その定義の方法により、以下のような特徴づけができる。

① 基本論理式

述語論理において、真偽が与えられる最小単位の述語である。例えば2組の距離が r 以下であるという、例 1.1 で定義した論理式 γ_r などである。他にも、2つの組のある属性の値が等しいという論理式等がある。

② 再帰的定義を含む場合

互いに連結している部品である、というようなことを表せる。

③ $\exists$, $\forall$ などの限量子で修飾されている場合

④ ブール論理式

基本論理式がAND・OR・NOTで結合された、すなわちブール論理式で定義されたクラスがある。これはかなり単純なクラスと考えられる。

ブール論理式で定義された質問の例を示す。

【例 3.1】

質問 Q₃: 図 3.1 に示す関係講義(科目, 時限)は行われている全ての講義の科目と時限を表す。同じ科目の講義がいくつかの異なる時限で行なわれており、また同じ時限でいくつかの講義が行なわれているとする。

ここである学生が自分の時間割をつくることを考える。学生は同じ科目を2つ以上受講しないし、また同じ時限で2つ以上の科目は受講できない、そしてできるだけ多くの講義を受講できる時間割をつくりたい可能な講義の組み合わせからなる講義の集合は次の条件を満たす。

- ① 各講義の科目はそれぞれ異なる。
- ② 同じ時限の講義は存在しない。
- ③ 集合の要素数は最大である。

①と②より、関係・講義上の同調関係 γ_c を定義する。

$$t_1 \gamma_c t_2 \text{ iff } t_1 = t_2 \vee (t_1[\text{科目}] \neq t_2[\text{科目}] \wedge t_1[\text{時限}] \neq t_2[\text{時限}])$$

質問の答は γ_c の要素数最大の同調類である。□

科目	時限
英語	火 2
英語	火 2
物理	月 2
物理	水 1
数学	火 2
数学	月 2
数学	水 1
化学	木 1
化学	金 2

図 4.1

質問 Q₃ は 3.1 節の③で述べた最大の集合を1つ求める質問である。

従来の関係データベースシステムでは、このような質問は応用プログラムによって、手続き的に記述されていた。

従属性はデータベース内の事実として、与えられている条件である。FD, MVDなどは、二項関係で表されることによって、集合制約質問と同じレベルで取り扱うことができる。例えば、関係 $R(X Y Z)$ における FD: $X \rightarrow Y$ は、以下のように表せる。

$$\forall t_1, t_2 \in R \quad ; \quad t_1[X] = t_2[X] \Rightarrow t_1[Y] = t_2[Y] \quad \square$$

” $\Rightarrow$ ”の部分を書き換えることにより、FDは属性の等式による述語と、そのブール論理式で表せることがわかる[3]。

【例 3.2】 質問 Q₃ の条件は、以下に示す FD の集合に書き換えることができる。

F_1 : 科目 $\rightarrow$ 時限

F_2 : 時限 $\rightarrow$ 科目

□

3.3 FD型集合制約質問

ここでは、集合制約質問のある部分クラスについて考察する。

【定義 3.2】 $R(X)$ の組 t_1, t_2 について、属性 $A \subset X$ についての基本論理式 $t_1[A] = t_2[A]$ を正リテラルと呼び、単に A で表す。正リテラルの否定を負リテラルと呼ぶ。正または負リテラルの AND・OR・NOT の結合で定義された R 上の二項関係をブール型二項関係と呼ぶ。□

ブール型二項関係は必ずしも同調関係ではない。同調関係であるためには、主乗法標準形に変換したとき、各和項に少なくとも 1 つ正リテラルを含まねばならない。

ブール型二項関係の部分クラスを定義する。

【定義 3.3】 ブール型二項関係 γ で表される集合制約質問について、 γ を主乗法標準形にしたとき、各和項に 1 つ正リテラルを含むものを、FD型集合制約質問と呼ぶ。□

FD型同調関係の各和項 $A_i \vee \neg B_{i1} \vee \neg B_{i2} \vee \cdots \vee \neg B_{ik(i)}$ は、

FD: $B_{i1} B_{i2} \cdots B_{ik(i)} \rightarrow A_i$ を表す。

FD は同調関係であるが、同値関係ではない。

ブール型二項関係質問は、論理関数の場合と同じ手法で簡単化できる [3]。

4. FD型集合制約質問の処理法

本節では、FD型質問の解を求めるアルゴリズムについて考察する。条件の FD 集合は、極小被覆に変換されているとする。FD 集合の要素数により、場合分けをする。

4.1 FDが1つの場合

まず 3.1 節の①に対応する 極大の解を求めるアルゴリズムを示す。

【アルゴリズム 1】

入力 : 関係 R (組数 m)、 R 上の FD F_1

ステップ 1 : 関係 R を FD の左辺が等しい組の集合 $S_1, \dots, S_p$ に分ける。

ステップ 2 : 各 S_i について S_i を FD の右辺が等しい組の集合 $S_{i1}, \dots, S_{iq(i)}$ に分ける。

ステップ 3 : $S = S_{1r(1)} \cup S_{2r(2)} \cup \cdots \cup S_{pr(p)}$ ($1 \leq r(i) \leq q(i)$) とする。

S は極大集合である。□

アルゴリズム 1 において、解の個数が m の多項式に比例しない場合がある。そのため、FD型集合制約質問においても、全ての極大の解の個数は莫大なも

のとなり、すべて求めるのは適当でない。3.1節の② ③の、極大または最大の解を1つ求めるのが適当である。

4.2 F Dが2つのときの最大の解

F Dが複数になると、1つの場合のように、単純なアルゴリズムを設計することは難しい。以下において、次の条件を満たすF Dについて考察する。

【条件4.1】

$R(X)$ 上のF D型集合制約質問の条件のF D: $Y \rightarrow Z$ について、

$Y \cup Z = X$ となる。 $\square$

条件4.1を満たすF Dの左辺は、 $R(X)$ のキーとなる。このため、解はF Dの左辺の値が皆異なる組からなる集合である。これにより、質問 Q_3 のような、ある属性の値が皆違う、といった集合を表せる。

条件1を満たすF Dにおいて、 R を左辺の値が等しい組集合に分け、その各集合をそれぞれ一つの超枝で覆った超グラフを考えると、最大の質問の解は、その超グラフの最大独立点集合に対応する。ここでは、各超枝からは1つしか点を選べないことを表す。

2つのF Dの場合、超グラフに対し、超枝 $\rightarrow$ 点、点 $\rightarrow$ 有向枝の変換を行えば2部グラフが得られる。並行枝を除けば、最大解となる最大独立点集合は、2部グラフの最大マッチングに対応する。

【アルゴリズム2】

入力: 関係 $R(X)$ 及び条件4.1を満たすF D F_1, F_2

出力: F_1, F_2 を満たす $R(X)$ に含まれる最大の組集合

ステップ1: $R(X)$ を、 F_1 の左辺の値が等しい集合 $S^1_1, \dots, S^1_r$ に分け、同様に F_2 について $S^2_1, \dots, S^2_s$ にわけ、すべての S^1_i, S^2_j について、それを超枝として持つ超グラフ H をつくる。

ステップ2: ある2つの超枝の共通部分に点が2つ以上含まれているならば、共通部分が1つの点になるようにする。

ステップ3: H の F_1 の超枝の集合に点集合 U 、 F_2 の超枝の集合に点集合 V 、超枝の共通部分の点の集合に枝集合 E がそれぞれ1対1に対応する2部グラフ $B = (U, V, E)$ を作る。

ステップ4: B の最大マッチング M を求め、それに対応する H の点を独立点とする。 $\square$

【定理4.1】

条件4.1を満たす2つのF Dからなる、 R (組数 n)上の集合制約質問の最大の解は、 $O(n^{1.5})$ で求まる。

証明 アルゴリズム 2 のステップ 1 の超枝の個数は $O(n)$ であり, ステップ 3 までは $O(n \log n)$ でできる. ステップ 4 の最大マッチングは $O(n^{1.5})$ のアルゴリズムが知られている [11]. $\square$

【例 4.2】

質問 Q_3 に対して, アルゴリズム 2 を適用する. ステップ 1 の結果は 図 4.3 の超グラフとなり, ステップ 3 の 2 部グラフは 図 4.4 となる. この最大マッチングを求めると, 例えば枝 (u_1, v_1) , (u_2, v_3) , (u_3, v_2) , (u_4, v_4) がマッチングとなる. 各枝はそれぞれ組 t_1, t_4, t_5, t_8 に対応し, 超グラフの独立点となる. 図 4.5 が質問 Q_3 の解であり, F_1, F_2 を満たしている. $\square$

科目	時限		科目	時限	
t1 英語	火 2	u1	t1 英語	月 2	v1
t2 英語	火 2		t6 数学	月 2	
t3 物理	月 2	u2	t3 物理	月 2	v2
t4 物理	水 1		t2 英語	火 2	
t5 数学	火 2	u3	t5 数学	火 2	v3
t6 数学	月 2		t4 物理	水 1	
t7 数学	水 1	u4	t7 数学	水 1	v4
t8 化学	木 1		t8 化学	木 1	
t9 化学	金 2		t9 化学	金 2	

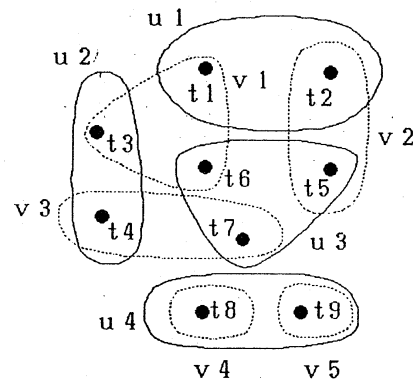


図 4.3 超グラフ H

図 4.1 F_1 の適用

図 4.2 F_2 の適用

	科目	時限
t1	英語	月 2
t4	物理	水 1
t5	数学	火 2
t8	化学	木 1

図 4.5 最大の解

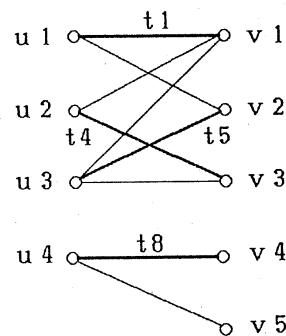


図 4.4 2 部グラフ B

4.3 FD が 3 つ以上のときの最大の解

ここでは, FD が 3 つの場合の集合制約質問の最大の解を求める問題が, NP 完全であることを証明する. これより, 3 つ以上の FD の場合は, 大変困難

になると考えられる.

3つのFDの問題はNPに含まれる. NP完全である3次元マッチング問題を集合制約質問に帰着する.

3次元マッチング問題[4]

インスタンス: 集合 $M \subseteq A \times B \times C$. ここで A, B, C は互いに素な q 要素からなる集合である.

問題: $M' \subseteq M$, $|M'| = q$ となるマッチング M' が存在するか? ここでマッチング M' とは M' の各要素の A, B, C の値が皆異なることを表す.

3次元マッチング問題は以下のようにして集合制約質問の解を求める問題に帰着できる.

3次元マッチング問題のインスタンス $M \subseteq A \times B \times C$ に対応する3つの属性からなる関係 $R(A B C)$ を作る. R の組は M の要素に対応する. 次の3つの(条件4.1を満たす)FDからなる R 上の集合制約質問を Q_2 とする.

質問 Q_2 : $A \rightarrow B C, B \rightarrow C A, C \rightarrow A B$

要素数 q のマッチング M' が存在するかという問題は, 質問 Q_2 に q 個の要素からなる解が存在するかという問題に帰着される. 以上により次の定理を得る.

【定理 4.2】

3つの異なるFDによる集合制約質問は, NP完全である. □

4.4 JDによる質問の変換

3.3節で述べたように, データベース管理のために, 常に保持されているFD, MVD(JD)などの従属性をFD型集合制約質問の処理に利用することを検討する. FDについては, 前述のように質問のFD集合をデータベースのFD集合で削減することが考えられる. JDについては以下のような方法が考えられる.

関係 $R(X Y Z)$ が $JD: *[X Y, Y Z]$ を満たすとは, 以下の条件が成立するときである.

$$\forall t_1, t_2 \in R \quad ; \quad t_1 = (x_1 \ y_1 \ z_1), \ t_2 = (x_1 \ y_2 \ z_2) \quad \Rightarrow \\ \exists t_3 \in R \text{ s.t. } t_3 = (x_1 \ y_1 \ z_2)$$

直感的には, X の1つの値に Y, Z のそれぞれの値の集合が直積的に対応することを意味する. 2つの関係 $R(X Y), R(X Z)$ を結合, すなわち X の値で2つの表を突き合わせて得られる, $R(X Y Z)$ は $JD: *[X Y, Y Z]$ を満たす. これより, 複数の関係を結合して得られた関係上での集合制約質問を処理する場合に, JDが利用できる.

4.3節で用いた, 3つのFDからなる質問 Q_2 について考える. 質問の対象である $R(A B C)$ 上で, $JD: *[A B, B C]$ が成り立つとする. このとき, R

(A B C)は図4.6に示すような非正規表現が可能である。属性Aの点からCの点へのパスの集合が $R(A B C)$ の組集合に対応する。図4.6の場合21個の組が存在する。属性A, Cの部分には、重複した値が存在している。

質問 Q_2 により、A, B, Cの値がすべて重複のない組集合を求める。

J D: $*[A B, B C]$ が存在するとき、最大の解は2つのFDのときに2部グラフの最大マッチングを利用するのと同様に、最大流量問題を利用することができる。

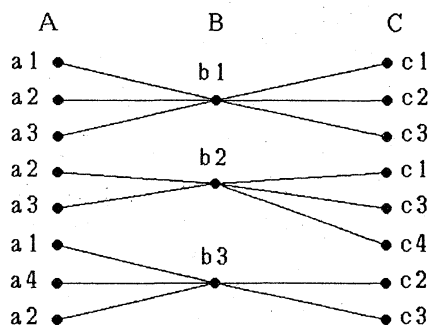


図4.6 J Dのある関係 $R(A B C)$

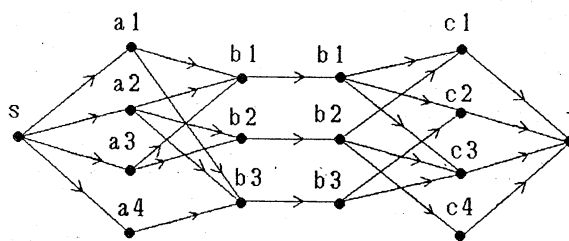


図4.7 最大流問題による解法

$R(A B C)$ を図4.7に示すネットワーク $N = (s, t, V, A, b)$ に変換する。sを流入口、tを流出口とし、全ての枝の容量を1とする。AとBの間の枝は、分解された関係 $R(A B)$ 、同様にBとCの間の枝は $R(B C)$ に対応する。

Nを流れる可能流について、sからtまで $a_i b_j c_k$ の点を通る流れがあれば、組 $(a_i b_j c_k)$ を質問 Q_2 の解の組の1つとすることができる。なぜならば、A, B, Cの各層において入る枝または出る枝が1本であり、また容量がすべて1であることより流れは常に1に飽和している。

そのためもし $a_i b_j c_k$ を通る流れがあれば、 a_i, b_j, c_k それぞれの点について、他にそこを通る流れは存在し得ないからである。

n組の関係 $R(A B C)$ から $O(n \log n)$ の手間でネットワーク $N = (s, t, V, A, b)$ を作ることができる。各節点に入る枝または出る枝が1本であり、全ての枝の容量が1である単純ネットワークの最大流を求める問題は、

$O(|V|^{1/2} |A|)$ で求められる[11]。

質問 Q_2 においては $|V|, |A|$ ともに $O(n)$ であるため、 $O(n^{1.5})$ で最大の解を求めることができる。

以上のようなJ Dによる質問の変換が考えられる。次にJ Dの持つ性質より、この変換の適用可能性を検討する。

J Dには巡回と非巡回の重要なクラス分けが存在する。J Dの要素を超枝とする超グラフが巡回であるか調べることにより、J Dの巡回性を判定できる。

上述の質問 Q_2 におけるJ D: $*[A B, B C]$ は、非巡回である。巡回であるJ

D の場合は質問処理に利用できるであろうか。

3次元マッチング問題は、インスタンスMが"pairwise consistent"に制限されていてもNP完全である[4]。"pairwise consistent"とは、任意のa, b, cについて $(a, b, w) \in M, (a, x, c) \in M, (y, b, c) \in M$ となるw, x, yが存在するならば、 $(a, b, c) \in M$ となることである。

これはJD: *[AB, BC, CA]が成立していることと同値である。このJDは巡回である。このため巡回であるJDが存在したとしても、質問処理は容易にならないと予想される。

5. まとめ

集合制約質問とその記述法について考察し、部分クラスであるFD型集合制約質問の処理法について検討した。他のクラスで効率的アルゴリズムが求められるものを見つけることが課題である。

参考文献

- [1] Beeri, C., Naqvi, S., Ramakrishnan, R., et al, "Sets and Negation in a Logic Database Language(LDL1)", ACM PODS 1987, pp.21-37.
- [2] Chandra, A.K., "Theory of Database Queries", ACM PODS 1988, pp.1-9.
- [3] Delobel, C., Casey, R.G., "Decomposition of a Data Base and the Theory of Boolean Functions", IBM J. Res. Develop. 17, pp.374-386, 1973.
- [4] Garey, M.G., Johnson, D.S., "Computers and Intractability", FREEMAN, 1978.
- [5] Hull, R., Su, J., "On the Expressive Power of Database Queries with Intermediate Types", ACM POD 1988, pp.39-51.
- [6] 岩井原, 上林, "二項関係によるデータベースの集合化質問", 昭和63年度電気関係学会九州支部連合大会論文集, pp.930.
- [7] Jaeschke, G., Schek, H.-J. "Remarks on the algebra of non-first-normal-form relations", ACM PODS, pp.124-138, 1982.
- [8] Kambayashi, Y., Tanaka, K., Yajima, S., "A Relational Data Language with Simplified Binary Relation Handling Capability", VLDB, pp.338-350, 1977.
- [9] Kuper, G.M., "Logic Programming With Sets", ACM PODS 1987, pp.11-20.
- [10] Kuper, G.M., "On the Expressive Power of Logic Programming Languages with Sets", ACM PODS 1988, pp.11-14.
- [11] Papadimitriou, C.H., Steiglitz, K., "Combinational Optimization", Prentice-Hall, 1982.
- [12] Paredaens, J., "Possibilities and Limitations of Using Flat Operators in Nested Algebra Expressions", ACM PODS 1988, pp.29-38.
- [13] Scholl, M.H. & Schek, H.J. (Eds.), "Proc. Int. Workshop on Theory and Applications of Nested Relations and Complex Objects", Darmstadt, 1987.