数理解析研究所講究録
593巻 1986年 295-303

非正規関係にもとずくデータベースの考察
－情報表現の最少化について－

盛屋邦彦*　三浦孝夫**　有澤　博***
Kunihiko Moriya　Takao Miura　Hiroshi Arisawa
* 日本情報処理開発協会　**三井造船㈱　*** 横浜国立大学工学部

１．まえがき

非正規関係（Non First normal form Relation NFR）は、関係モデルの前提である第１正規関係と比べて、いくつかのよい性質をもっている。冗長性の減少、直観性の向上、表現能力の差（集合値や関係値としての意味）などが、その例である［小林８０］。したがって、非正規関係はいろいろな立場で議論することが可能である（例えば、メインモデル、ユーザインタフェース等）。

ここでは、関係データベースの実現構造として、非正規形を用いることを考える。非正規関係を実現構造として用いることの最大の利点は、複数の関連を１つにまとめた情報表現として考え、いわゆるサーチスペース（情報単位の数）を減らせることである。

例

A	B	C
a 1	b 1	c 1
a 1	b 2	c 1
a 1	b 1	c 2
a 2	b 1	c 2
a 2	b 2	c 3

図１　第１正規形

A	B	C
a 1	b 1	c 1
a 1	b 2	c 1
a 2	b 1	c 2
a 2	b 2	c 3

図２　非正規形その１

A	B	C
a 1	b 1 b 2 b 1	c 1 c 1 c 2
a 2	b 1 b 2	c 2 c 3

図３　非正規形その２

ここでいう、実現構造とは、従来の記憶構造（例えば、B-Treeやハッシュなど）を指すのではなくそれと独立して、関係モデルにおけるタプルや属性値などが論理的に扱えるレベルのものと考えることにする。このレベルを設けることにより、実際の記憶構造に依存しない論理的な蓄積方法の最適化を考えることができる。最適化としてサーチスペースの減少化を考えれば非正規関係は有用なクラスである。図１の第１正規関係を非正規関係で表すとたとえば図２あるいは３のようになる。図２，３は、行を情報表現単位（いわゆる非正規関係におけるタプル）と考えれば、減少化が可能であることがわかる。これはもとの関係がある種の制約条件（ＭＶＤやＷＭＶＤ［ＦＶ８４］等）を満たしている時、非常に大きな減少をもたらすことがある。のみならず、局所的部分関係が制約条件を満たしている場合にも、その可能性が十分あると考えられる。

図２では属性値定義域に集合値を許したものであり、図３では関係値を許したものとなっている。図３のように、関係値を許したものは、その特定の関係値をもとにした操作を行なう場合には有用であると考えられるが、一般的な操作を行なう場合には構造が平坦でないため、属性の種類により、別の操作メカニズムを持たねばならない。そこでここでは、図２のような属性値として、集合値のみを許した非正規関係を用いる立場をとっている。さらにもとの正規関係での情報単位の数を減らすという意味で、次の例を考えてみる。

学生	得意科目
山本	物理
山本	数学
高橋	物理
高橋	数学
高橋	化学
田中	数学
田中	化学

図５　１ＮＦ

学生	得意科目
山本	物理　数学
高橋	物理　数学　化学
田中	数学　化学

図６　ＮＦＲ　その１

学生	得意科目
山本　高橋	物理　数学
高橋　田中	数学　化学

図７　ＮＦＲ　その２

図６は、図５の第１正規関係を非正規関係として表現した場合の最小

形（情報表現の数が最少）となっている。図７は図５とそのままの意味では等価ではないが（重複した情報表現（高橋　数学）がある）、集合論的には同じものとして考えることができる。すなわち、第１正規関係の実現構造として考えて、重複したタプルを同一視すれば、さらに少ない情報表現で表すことができる。もともと、関係モデルは集合論を基にしたものであり、情報が損失する意味での不都合さはない。このように重複した表現を許した場合、情報単位の数が大きく減少することが考えられる（図６が３に対して図７は２）。さらに、これは従来の記憶構造（例えば、B-Treeやハッシュなど）と独立に議論できそれらと組み合わせれば、サーチスペースをより減少させることができる。また物理的な圧縮技法（例えば、コード化によるデータ圧縮）と共存することも可能であり、より小さな情報表現を得ることも可能である。

一方、重複を許した場合、更新や削除における不都合さが問題となる。これは重複したものを同時に更新または削除しなければならないことから生じるが、例えば辞書や図書の目録のように読み込みまたは、挿入のみであるような応用に対しては十分意味を持っていると考えられる。さらに、この場合、一度最小形を求めてしまえばあとは、最小形を考える必要がないということが大事な点である。一般に重複を許さない場合の非正規関係の最小形を求める問題は、特殊な場合でもＮＰ完全であることが知られている［武田８５］。しかし、重複を許した非正規関係の最小形を求める問題について論じられたものはなく、非正規関係の一般的な性質を考える上でも意味があると思われる。

２．　非正規関係の最小化問題

ここでは、いわゆる関係モデルにおける属性欄（カラム）に集合値のみを許す非正規関係を定義し、さらに重複表現を許した場合の最小化問題を考える。

２.１　諸定義

定義　非正規関係（ＮＦＲ　Non First normal form Relation）

Ｅ１，．．．．，Ｅｎを単純領域とする。

ｒがＥ１，．．．Ｅｎ上の非正規関係（ＮＦＲ）とは、

$r \subseteq 2^{E_1} \times \cdots \times 2^{E_n}$ の時にいう。

したがって、非正規関係におけるタプルｔは次のように表す。

$t = [E1(e1) \ldots En(en)]$

ここで、$ei = \{e_1^i, \ldots, e_{m_i}^i\}$　$m_i \geq 1$

このタプルに対して、

$t' = [E1(e_1) \ldots En(e_n)]$　$1 \leq i \leq n$

で $e_i \in e_i$ なるタプルを単純タプルといい、t'は tに含まれるという。

非正規関係rの各組に含まれる単純タプルより重複を除いて構成される第1正規関係（1NF）をr*で表し、この時、r とr*は等価であるという。

定義　合成（Composition）

r内の2つのタプルt，sに対して、

t = [E1（e1）　...　En（en）]

s = [E1（f1）　...　En（fn）]

tとsのE1上の合成（Composition）とは次で定義される。

$$\begin{cases} [E1(e1+f1)\ E2(e2)\ \ldots\ En(en)] \\ \qquad ei = fi \text{ の時 } i = 2, \ldots, n \\ t \qquad \text{その他} \end{cases}$$

これをv（E1，t，s）と書く。特に前者の場合、合成可能という。

定義　既約

rが既約であるとはr内のどの2つのタプルに合成操作を加えても、もはや不変となるものをいう。

定義　最小

非正規関係rが最小であるとは、rと等価な第1正規関係r*に対して、これと等価でrよりもタプル数が少ないような他の非正規関係がない時にいう。

性質1

rが最小ならば、既約である。

なぜなら、既約でないとするとさらに合成可能であるから、r*と等価でかつrよりもタプル数の少ない非正規関係が得られてしまう。

2.2　非正規関係と完全n部グラフ

2.1で定義した非正規関係を前提として、第1正規関係r*より、これと等価な最小なものを求める問題を考える。

最小の非正規関係は、その各タプルがより多数の単純タプルを含む形となっている。そこで、問題を単純化するために、各タプル自身の表現や，各タプルが単純タプルを含んでいることを表す方法としてここではグラフを用いるアプローチをとることにする。

すなわち、タプルの各属性値に対して、グラフにおけるノード（点）

を対応させ、タプルの表す関連を、そのノードを結ぶ辺として考えることができる。このとき、非正規関係のタプルは複数のグラフを含むような完全n部グラフとしてみることができ、最小形の問題をグラフの性質の問題としてとらえることができる。

以下、グラフの定義及び性質を述べていく。

定義　n部グラフ

Gが（n部）グラフとは次をいう。

$G = (V1, \ldots, Vn, E)$

ここで、$Vi \neq \phi$ かつ $Vi \cap Vj = \phi$

$i > j \quad i, j = 1, \ldots, n$

$E \subseteq V1 \times \ldots \times Vn$

G'がGの部分グラフとは、

$G' = (V'1, \ldots, V'n, E')$ で、

$V'i \subseteq Vi$

$E' \subseteq V'1 \times \ldots \times V'n \qquad E' \subseteq E$

この時GはG'よりも大きいといい　$G \supseteq G'$ と書く。

さらに、Gが完全（Complete）とは

$E = V1 \times \ldots \times Vn$

の時にいう。

G'がGの完全部分グラフとは、

G'がGの部分グラフで、完全である時にいう。

Gの被覆表現（Covering）

$C = \{ (V\tilde{}1, \ldots, V\tilde{}n, Ei) \mid i = 1, \ldots m \}$ とは、

$Gi = (V\tilde{}1, \ldots, V\tilde{}n, Ei) \quad i = 1, \ldots m$ として

$Gi \subseteq G \qquad i = 1, \ldots, m$

$\bigcup_{i=1}^{m} Ei = E$

$\bigcup_{i=1}^{m} V\tilde{}j = Vj$　の時にいう。

Gの完全被覆表現（Complete covering CC）とは、

上の条件に加えて、

Gi（i=1,...,m）が完全部分グラフの時にいう。

以上の定義から、第1正規形をもとのn部グラフGとして考えれば、非正規関係をGの完全被覆表現として、みることができる。

すなわち、最小の非正規関係を求める問題は最も成分数が少ない完全被覆表現を求める問題に帰着できる。そこで完全被覆表現についてその性質について調べてみることにする。

性質２

完全被覆表現は常に存在する。

なぜなら、任意のグラフに対してその各要素を含むような完全部分グラフの集合を考えることができる。

定義　非冗長完全被覆表現

Ｇの完全被覆表現　Ｃ＝｛Ｇｉ｝を考える。

この時、Ｃが非冗長であるとは、ＣからどのＧｉを削除してもＣがＧの被覆表現にならない時をいう。

性質３

非冗長完全被覆表現は一意でなく、かつそれぞれが含む完全部分グラフの数も一定でない。

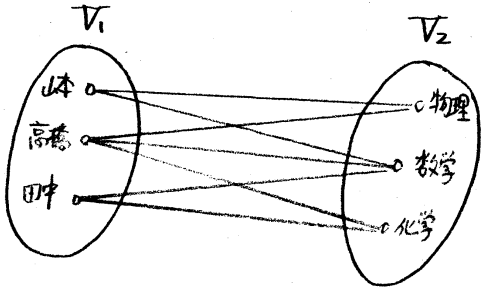
なぜなら、前の例がそれを示している。

すなわち、

Ｖ１＝｛山本，高橋，田中｝，Ｖ２＝｛物理，数学，化学｝として、

(1)　Ｃ１＝｛
｛（山本　物理），（山本　数学）｝
｛（高橋　物理），（高橋　数学），（高橋　化学）｝
｛（田中　数学），（田中　化学）｝
｝

(2)　Ｃ２＝｛
｛（山本　物理），（山本　数学），（高橋　物理），（高橋　数学）
｛（高橋　数学），（高橋　化学），（田中　数学），（田中　化学）
｝



性質４

成分数の最も少ない完全被覆表現は非冗長被覆表現である。
なぜなら、非冗長完全被覆表現でないならば、その完全被覆表現からある成分Ｇｉを削除でき、それよりも成分数の少ないものが作れる。

以上見てきたように、最小の非正規関係に対応する完全被覆表現は非冗長であり、一般の完全被覆表現において冗長なＧｉがどの程度存在するかが最小の非正規関係を求める鍵となっている。　これは、実際には指数べき個存在する。

さらに、最終的にはこの問題は、次の問題に帰着でき、効率よく解けないことがわかる。
［ＧＪ７９］より、

［ＧＴ１８］
２部グラフＧ＝（Ｖ，Ｅ），｜Ｅ｜≧Ｋ≧ｋ，Ｖ１，．．．，Ｖｋ⊆Ｖの時、各Ｖｉは完全２部グラフかつ、Ｅの被覆表現となることを問うことはＮＰ完全である。

ここでは一般にｎ部グラフの問題であるから、最小の非正規関係を求めることは、ＮＰ完全問題よりも難しいクラスであることがわかる。
したがって、論理的な方法で最小形を求めることは非常に難しいことがわかる。しかし、更新がないような応用では、１度だけ最小形を求めればよいので、特定の分野では役に立つものと考えている。

３．　情報表現の減少化と最小形の適用問題
２の議論により、効率よく最小の非正規関係を求めることができないことがわかり、論理的なレベルでは実用の見通しがないことが明らかとなった。そこでここでは、より小さな非正規関係を求めることと、最小形が求まった後の問題について議論する。

３.１　情報表現の減少化
より少ない情報表現を求めるため、つぎのアプローチが考えられる。

(1)近似解を求める。（もしくはヒューリスティックアプローチ）
(2)ある制約条件を満たす場合の一意形、またその条件の性質。
(3)ある種の構造化を基にした高速化。

(1)としては例えば、次のような近似解が考えられる。
ａ．非冗長完全被覆表現の１つでもよいから見つける。
ｂ．完全ｎ部グラフのサイズ（その完全ｎ部グラフに含まれる辺の数）を一定にした時の最小形を求める。
ｃ．いわゆる非正規関係の標準形（たとえばcanonical form）に重複表現を許した上での最小形を求める。
(2)としては、ある制約条件のもとでの最小形の一意性や、それを求める計算量の問題である。現在のところ一意になる十分条件は見つかっているが、必要条件はまだわからない。
(3)はいわゆる、論理的または物理的な構造をあらかじめもたせておき、最小形を求める計算量を小さくしようとするものである。２での議論より、恐らく属性値を基にしたタプルの並べかえや（タプルに順序をもたせる）、組み合わせ等の操作の高速化が必要であると思われる。

３.２　最小形の適用問題

あらかじめ最小形が求まったものとして、次の問題が考えられる。

(1)　読み込み、挿入のみを許すデータベースを仮定した時、挿入による最小形の保存問題。
(2)　最小形のタプル数の予測と実際。
(3)　最適な記憶構造、物理構造での表現方法。

(1)は、最小形を崩さないようなデータの挿入方法の問題である。辞書のような応用分野では、読み込みまたは挿入だけが許されるものであり、計算量が小さくてすむ場合には有用であると思われる。
(2)は重複を許した場合の最小形の大きさの予測の問題である。これも応用分野に（すなわち、特定の一貫性制約などに）、強く依存しているが、統計量などをもとにして予測することができると思われる。これにより、最小形を求める妥当性を論じることができる。
(3)は非正規関係に最適な実際の記憶構造を考えることである。サーチスペースを減らす観点に立てば、値を基にした探索が容易にできるようなものが望ましい（たとえばハッシュやインデックスなど）。これらの記憶構造と組み合わせればかなりのサーチスペースを減らすことができる。また、コード圧縮のようなことを考えれば更に全体のデータ量を減らすことができる。

４．まとめ

当論文では、関係モデルの実現構造として非正規関係を用いることを考え、重複した表現を許すことで、より情報単位の数を減らすことに注目し、その最小形について議論し、実際には非常に難しいクラスの問題となることを示した。今後の展開としては、３であげた問題について議論していくつもりである。

参考文献

[小林８０] Kobayashi, "An overview of the DB Management Technology", Tech.Rep. TRCS 4-1 Sanno collage, 1980

[三浦８６] Miura, Moriya, Arisawa, "On the Irreducible Non First normal form Relations", Submitting.

[ＦＶ８６] Fischer, Van Gucht "Weak MVDs",PODS,1984

[有澤８３] Arisawa, Moriya, Miura, "Operations and the properties of NFR",PROC . of VLDB, 1983

[ＪＳ８２] Jaeschke, Schek, "Remarks on the algebra of NF2 relations", PODS, 1982

[武田８５] 武田、"非正規関係の一意性について"，情報処理学会全国大会、６０年度前期

[ＧＪ７９] Garey, Johnson, "Computers and Intractability", Freeman and Company, 1979.