データベースシステム　109－5
（１９９６．７．２４）

広域検索のための関係代数の拡張

大塚真吾　宮崎収兄
千葉工業大学情報工学科

広域ネットワークでは不特定多数のデータベースが散在している。このような環境での問合せでは問合せたい内容によって対象データが変化するため、あらかじめスキーマの統合を行うことは困難である。我々はネットワーク上にサイトが多数存在する環境で全体を一つの大きなデータベースと考え、各データベースはその一部のデータを格納した不完全なデータベースと考えることにより広域検索を行う方法を提案した。本稿では、この方法を実現するために条件付タプル、マージなどの概念を導入し関係代数を拡張して広域データベース検索処理の具体化を検討する。

Extending Relational Algebra for Global Query Processing

Shingo Otsuka and Nobuyoshi Miyazaki
(otsuka, miyazaki)@mz.cs.it-chiba.ac.jp
Dept. of CS, Chiba Institute of Technology
2-17-1 Tsudanuma Narashino Chiba 275 Japan

It is difficult to process even simple global queries if many independent databases are scattered in a large network like the Internet, because there are various mismatches between databases. A collection of databases in a network can be regarded as a very large virtual database, where individual databases contain partial incomplete information. We proposed a method to process global queries based on the concept of incomplete database. This paper discusses extended relational algebra based on extended tuple concepts and introduces extended operations such as "merge" that are useful to realize global query processing.

1　はじめに

広域ネットワークでの検索はインターネット上でのWWWなどの各種情報ツールによって扱われている。検索ツールの背後にデータベース管理システムを用いることも行われているが、これらはデータベースの条件検索機能を直接広域検索に用いていない。WWWのようなリンクによるアクセスやサーチエンジンによるホームページの内容のキーワード検索だけでなく、データベースに対する質問処理が実現できればネットワークを通じた情報の高度利用が可能となる。従来のマルチデータベースではあらかじめ定められた少数のデータベースの統合的扱いが課題であり、広域ネットワークのように不特定多数のデータベースが散在する環境での統合的な検索はあまり研究されていない。このような環境での問合せでは問合せたい内容によって対象データが変化するため、あらかじめスキーマの統合を行うことは困難である。このため広域ネットワークでのデータベースの質問処理を可能とするための新しい枠組みが必要となっている。我々はネットワーク上に情報提供サイトが多数存在する環境で全体を一つの大きなデータベースと考え、各データベースはその一部のデータを格納した不完全なデータベースと考えることにより広域データベース検索を行う方法を提案した[宮崎９６]。本稿では、この方法を実現するため関係代数を拡張し広域データベース検索処理の具体化を検討する。以下、２章で不完全データベースによる広域検索、３章で拡張関係、４章拡張関係代数、５章で広域検索への適用について述べる。

2　不完全データベースによる広域検索

広域検索ではどこにどんなデータがあるか正確にはわからない場合でも問合せを可能としたい。このため関係データベースを拡張し不完全データベースの概念を導入する。不完全データベースでは分かっている情報のみを表現するので、タプルによって存在する属性が異なっても良い。また、問合せの条件にマッチするかどうか不明な時や条件付のデータを表現するために条件付タプルの概念を導入する。関係はこのように拡張されたタプルの集合である。不完全ベータベースによる広域検索では問合せは情報を持っている可能性のあるサイトに送られ問合せを受けたサイトは答えられるものだけ答える方式を取る。以下に例を挙げて説明する。

［例２．１］　以下のスキーマのデータがあるとする。

学生｛名前、住所、電話｝

また以下のようなタプルがある。

学生｛名前＝千葉太郎、住所＝習志野市、
電話＝0474-00-0000｝

千葉太郎の年齢と住所に関する問合せ

？学生(名前＝千葉太郎、年齢＝X、住所＝Y)

を考える。従来のDBMSではこの問合せはスキーマと不一致のためエラーになる。不完全データベースでは存在する情報から、

学生（名前＝千葉太郎、住所＝習志野市）

を解とできる。年齢に関するデータはないのでこのような解を部分解と呼ぶ。これに対し問合せた全ての情報を含む解を完全解と呼ぶ。

［例２．２］　複数のサイトに問合せを行う場合

？学生(名前＝千葉太郎、住所＝X、単位数＝Y)

を学生課と教務課に送りそれぞれ以下の部分解が得られたとする。

（学生課）

学生（名前＝千葉太郎、住所＝習志野市）

（教務課）

学生（名前＝千葉太郎、単位数＝８０）

この２つの部分解の自然結合を取れば完全解が得られる。

3　拡張関係

不完全データベースに対する問合せでは、問合せた属性がデータベースに存在しないということが多々ある。また演算に関しては、広域で問合せを行う場合、各リレーションの属性がまちまちで、和両立が成立する可能性が低く複数サイトからの解の和がとれない場合がある。このような状況を解決するために関係を拡張する。

3．1　タプルの表現

レコード記法[YM９４]を用いてタプルの定義を行う。レコード記法では存在する属性のみ扱うことができる。

（１）単純タプル

$a(l_1=b_1,l_2=b_2,....,l_n=b_n)$

ここではaは関係名、l_iは属性名である。値の分からない属性は書かない。

（２）条件付タプル

$a(l_1=b_1,l_2=b_2,....,l_i=b_i)$ if $a(l_{i+1}=c_1,....,l_m=c_m)$

これはif以下が真なら$a(l_1=b_1,l_2=b_2,....,l_i=b_i)$が真であることを表す。また、if以下が不等号条件を表すこともできる。

本稿では例に関してレコード記法の他に表によって表現する。表では存在する全ての属性を表記する。これ以降、レコード記法と従来の表形式を併用する。タプルによって存在しない属性は空欄とする。

3．2　拡張関係

拡張関係は同じ関係名の単純タプルと条件付タプルの集合として定義される。拡張関係に対して以下の２つの概念を定義する。

（１）共通属性制約：

・関係のすべてのタプルに少なくとも１つの共通属性が存在する。

共通属性制約は通常の関係の主キー制約に対応する。広域データベースで主キー制約を適用するのは困難である。したがって最小限の制約を導入した。

（２）和両立

・２つの関係に共通の属性があり、どのタプルにもその属性が存在する。

通常の関係の和両立の概念は制限が厳しいのでこのように変更した。

4　拡張関係代数

不完全データベースによる広域検索のために前節ではその基本的な部分の拡張を行ってきた。この節では、実現方法を検討するため、関係代数を拡張する。

4．1　集合演算

集合演算は従来の演算と同様に定義される。

和：二つの関係が和両立の時

$$A \cup B = \{x \mid x \in A \vee x \in B\}$$

と定義する。つまりAまたはBに属するタプルすべての集合。

差：二つの関係が和両立の時

$$A - B = \{x \mid x \in A \vee \neg(x \in B)\}$$

と定義する。つまりAに属しかつBに属さないタプルの集合。

共通部分：二つの関係が和両立の時

$$A \cap B = \{x \mid x \in A \wedge x \in B\}$$

と定義する。つまりAとBに属すタプルの集合。

直積：AとBを二つの拡張関係とする時

$$A \times B = \{(a,b) \mid a \in A \wedge b \in B\}$$

ここにa=(x1,x2,・・・,xm),b=(y1,y2,・・・,yn)とする時、(a,b)=(x1,x2,・・・,xm,y1,y2,・・・,yn)なるm+n項のタプルである。

［例４．１］　２つの拡張関係があるとする。

DB学会

会員名	所属	連絡先
菅原	K55	2011
佐藤	K55	2035
森	K55	2038
鈴木	K81	2201

AI学会

会員名	所属	住所
森	K55	町田
鈴木	K81	神戸
田中	K81	水戸

DB学会∪AI学会

会員名	所属	連絡先	住所
菅原	K55	2011	
佐藤	K55	2035	
森	K55	2038	
鈴木	K81	2201	
森	K55		町田
鈴木	K81		神戸
田中	K81		水戸

４．２　固有演算の拡張

ここでは固有演算の拡張の定義をする。

射影：

Xを属性の集合とする。関係Aの射影$\pi_X(A)$は、Aの各タプルからxに属する属性で属性値があるものだけを取り出したタプルの集合。また、Xに属する属性が１つも存在しないタプルの射影は属性なしのタプルとする。

選択：

関係Aの選択$\sigma_F(A)$は、Aのタプルのうち条件Fを真とするまたは真とする可能性のある全てのタプルの集合である。対象とする属性値が存在しない場合、結果は条件付タプルとなり、対象とする属性が条件付の場合はタプルの条件の値が選択条件を満たせば結果は条件付タプルとなる。

θ－結合：

二つの関係AとBのθ－結合$A\infty_F B$は$\sigma_F(A\times B)$で定義される。対象とする属性が存在しない場合条件付にして結合し、対象とする属性値が条件付の場合は条件の値を対象として結合する。

自然結合：

関係AとBの自然結合$A\infty B$は、AとBの直積から共通の属性の値が等しいものだけを選び、その結果から共通の属性の一方だけを残すように射影した関係である。一方の属性が存在せず条件部分に属性が現れる場合は条件を満たせば結合できる。通常の関係代数では自然結合はθ－結合と射影の組合せで表現できるが、拡張関係ではタプル対毎に共通属性が異なるので組合せでは表現できない。

無条件タプル選択：

関係Rのタプル内で条件付タプルを取り除く演算を無条件タプル選択と呼びunc(R)と書く。

４．３　解のマージ

［例２．２］のように複数のサイトから部分解が得られた場合、それらをもとに完全解が生成できる場合がある。例えば複数のサイトから得られた解の和が以下のようになったとする。

会員名	所属	連絡先	住所
菅原	K55	2011	
佐藤	K55	2035	
森	K55	2038	
鈴木	K81	2201	
田中	K81		水戸
森	K55		町田

このとき、ユーザが「会員名＝森」が同じものだと考え、下記のようにまとめて一つのタプルにしたいとする。

会員名	所属	連絡先	住所
菅原	K55	2011	
佐藤	K55	2035	
森	K55	2038	町田
鈴木	K81	2201	
田中	K81		水戸

この演算を“マージ”と呼ぶ。マージは“縮退”演算と自然結合の組合せで定義できる。

［タプルの順序］

同じリレーションの２つのタプルTiとTjを考える。以下の２つの条件が成立つ時Ti＜Tjと書く。"＜"は半順序関係である。

（１）Tiの属性名がすべてTjに存在し、その属性値すべてが等しい。

（２）Tiに条件付属性が存在する時は、

(a)Tjの属性がその条件を満たす。

または、

(b)Tjの条件付属性の条件がTiの条件を含む。

［縮退］

関係Rのタプルの中でTi ＜ Tjとなるタプルが存在するようなタプルTiをすべて消去する演算を縮退と呼びred(R)と書く。縮退の結果は消去する順番に依存せず一意に定まる。

A1	B		
A1	B	C1	D
A2		C2	

↓

A1	B	C1	D
A2		C2	

［マージ］

mrg(R)＝red(R∞R)

で定義される演算をマージと呼ぶ。すなわち同じ属性値のタプルの自然結合により情報を統合し、不要なタプルを消去するのがマージである。

5　広域検索への適用

問合せはレコード記法で以下のように表現する。

? $a(l_1=b_1,l_2=b_2,\ldots,l_i=b_i,l_{i+1}=X_1,\ldots,l_m=X_m)$

属性値の定数は条件を、変数は求める対象を表す。また、複数の拡張述語の論理積を書くことができる。

処理手順：

・問合せを各サイトへ送る。

・各サイトから解が帰ってくる。

・解の和をとる。

・場合によってユーザがマージ、無条件タプル選択を行う。

5．1　適用例

［例５．１］

学生課　　就職課

学生　　学生

名前	住所
千葉太郎	千葉市
大塚真吾	習志野市

名前	電話番号
千葉太郎	00-0000
大塚真吾	11-1111

「大塚真吾の住所と電話番号」を問合せをする。

?学生　(名前＝大塚真吾、住所＝Y、電話番号＝Z)

・拡張関係代数で表現すると

π {住所,電話番号}(σ 名前＝大塚真吾(学生)).

それぞれのサイトでこの問合せを処理し、その結果の和をとる。

名前	住所	電話番号
大塚真吾	習志野市	
大塚真吾		11-1111

これをマージすると

名前	住所	電話番号
大塚真吾	習志野市	11-1111

となる。

［例５．２］

学生課

学生

名前	住所
千葉太郎	千葉市
大塚真吾	習志野市

教務課

学生

名前	単位数
千葉太郎	85
大塚真吾	123

「単位数が１００以上の学生の名前と住所」を問合せする。

？学生 （名前＝X、住所＝Y、

単位数＝Z）、Z＞１００

・拡張関係代数では

π｛名前、住所、単位数｝(σ 単位数≧１００(学生)

結果の和は

名前	住所	単位数
千葉太郎	千葉市	if単位数≧100
大塚真吾	習志野市	if単位数≧100
大塚真吾		１２３

となる。マージすると

名前	住所	単位数
千葉太郎	千葉市	if単位数≧100
大塚真吾	習志野市	123

となる。ここで無条件タプル選択を行うと、

名前	住所	単位数
大塚真吾	習志野市	123

となり解が得られる。

５．２　問題点とその解決法

（１）［例５．２］で学生課へ問合せた解が全て条件付き解になってしまうという問題点がある。

これを解決するには以下のような方法がある。

・問合せの選択の条件を１つも満たしていないものは可能性解としサイト名だけ返す。

・他のサイトに問合せた解に部分解がある場合、それを基に可能性解があるサイトに再度問合せを行う。

（２）広域検索を行うときユーザは問合せをするサイトがどのような情報を持っているかを把握するのが困難である。問合せにSQLの＊（全属性指定）に相当する表現を導入すればユーザは問合せ先にある情報をすべて引き出す事ができ、その情報を有効に活用できる。

6　まとめ

広域データベースでは従来のマルチデータベースの方法のようにあらかじめスキーマ統合を行っておくことが困難である。スキーマ統合を行った場合でも、情報のミスマッチをいかに解消するかの問題がある。本稿では広域ネットワークに散在するデータベースに対し、問合せを行う方法を具体化するため関係代数の拡張を提案した。実現には多くの研究課題があるが広域データベース検索が可能になればネットワークの情報利用の可能性は飛躍的に高まるであろう。今後はリレーショナルデータベースシステムを利用したプロトタイプを作成し、その有効性を検討していく予定である。

参考文献

［宮崎９６］宮崎収兄，“不完全データベースと広域データベース検索”情報処理学会データベースシステム研究報告，９６－ＤＢＳ－１０６，pp. 131-138

［YM９４］　横田一正，宮崎収兄，“新データベース論”共立出版，１９９４