データベースシステム　106－18
（１９９６．１．２５）

不完全データベースと広域データベース検索

宮崎収兄
千葉工業大学情報工学科

ネットワークに散在するデータベースからの情報の検索では種々の不整合や重複、欠落などの問題があり単純な問い合せでも処理することが困難である。マルチデータベースでは異なったデータモデルを用いた複数のデータベースへのアクセスが研究されているが、インターネット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困難である。WWWの様なリンクによるアクセスだけでなく、散在するデータベースに対する統合的な条件検索が可能になればネットワーク利用の可能性は飛躍的に高まる。全体を仮想的なデータベースとし各サイトはその一部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実現する方法を提案する。

Global Query Processing based on Incomplete Database Concept

Nobuyoshi Miyazaki
miyazaki@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 internet, because there are various mismatches between databases. The applicability and usefulness of networks will be enormously increased if integrated conditional retrieval on various databases can be performed. A collection of databases in a network can be regarded as a very large virtual single database, where individual databases contain partial incomplete information. We propose a method to process global queries based on the concept of incomplete database.

1　はじめに

コンピュータネットワークに散在するデータベースの全体を１つの大きなデータベースと考えた時、従来のデータベースでは扱いきれない問題が発生する。各データベースが独立に設計されているため、データモデル、データの場所、スキーマ、意味、データの重複や欠落、不整合などの問題があり単純な問い合せでも処理することが困難である。

情報検索の分野では類似の問題はインターネット上でのWWWなどの各種の情報検索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るアプローチはとっていない。データベース分野ではこのような問題はマルチデータベースで扱われている[KS90][BE96]。マルチデータベースでは異なったデータモデルを用いた複数の自律的なデータベースへの統合的なアクセスが可能である。

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。しかし、インターネットにおけるWWWによる情報検索とその応用が急速に普及した現在、各組織のデータベースが公開されリンクによるアクセスだけでなく複数のデータベースに対する統一的な条件検索が可能になればネットワーク利用の可能性は飛躍的に高まると考えられる。このような環境では問い合わせたい内容によって対象データベースが変化するため、あらかじめスキーマの統合を行うことは困難である。たとえば、商品の価格が注文数量によって変わる会社のデータベースや、納入時期によって変わる会社のデータベースなどが散在している場合、特定の商品の価格の問い合せを行い最も安い価格を探すのは簡単ではない。また、一人の人のデータが市役所、勤務先、病院、出身校などのデータベースに散在している場合、この人について知りたいことを仮想的なスキーマを考えＳＱＬで書けたとしても、結果を得るのは容易ではない。

このように情報提供サイトが多数存在する環境では全体を仮想的なデータベースとし、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が実現できる。本稿では散在する不完全なデータベースへの問い合せという観点から、この問題を検討する。以下、２節で広域情報検索の現状と課題、３節で不完全データベースとその問合せ、４節で不完全データベースによる広域問合せ処理について述べる。

2　広域情報検索の現状と課題

広域ネットワークに散在する情報の検索や問合せを実現するためのアプローチには大きく３つがある。

（１）データベース

組織化、統合化された情報を格納し、スキーマに従った問合せを用いて条件検索を行う。条件検索を用いて複数のデータベースから必要なデータを検索することができる。

（２）情報検索システム

文献検索システムなどは情報提供側の機能に合わせて設計されているので一般にはシステムごとに異なった記述形式をとっている。したがって、複数のデータベースから統一的な形式で情報を得るのは容易ではない。

（３）WWW

WWWでは統一した形式で情報が提供され散在するサイトから情報を検索できる。しかし、条件にあった必要な情報を見つけるにはリンクをたどって情報の内容を見ないと分らない。

これらの方法のうちインターネットのように情報提供サイトが多数ある環境での広域情報検索で広く用いられているのはナビゲーションに基づいたWWWである。また、各サイトのデータベースを提供する場合にもWWWのバックエンドにＤＢＭＳをおくアプローチが主流である。ＤＢＭＳは標準化されたデータモデルと問合せ言語を持っているので、イン

ターネットからデータベースへの問合せ（条件検索）を可能にすれば情報利用の可能性は飛躍的に高まると予想される。このとき、各データベースへの問合せを個々に可能とするだけではなく、１つの問合せを複数のデータベースに発行し情報を検索できることが望ましい。このような機能を実現することにより利用者は統一的な方法で全世界にある情報を検索・利用できる。

例　商品提供業者と価格の検索

パソコン雑誌には大量の商品広告が掲載されている。これを電子化し、各業者が自社の販売情報を公開する。これを実現する方法とその特徴を以下に示す。

WWW

・業者ごとにホームページをおきブラウザなどで検索する。補助的に業者情報を情報提供業者などのサイトにおく。

・業者ごとの情報の構成はそれぞれの業者によって異なる。

・業者は迅速に情報を更新できる。

・利用者は情報のありそうなサイトにアクセスし必要な情報を探す。

・業者側では電子ショップなど色々工夫できるが、利用者の欲しい情報が必要な形式で見つかるとは限らない。

・利用者の検索処理は自動化しにくい。

ＤＢＭＳを用いデータベース化

・情報提供業者、業者団体などにデータベースをおき統一したスキーマの元に各社の情報を集め格納する。

・有力ユーザでは自社でデータベースを設計しそれに合わせたデータを業者から集めることも可能

・業者ごとに自社の情報をデータベースとして検索可能にし、マルチデータベースとして検索する方法も有る。

・内容に関する条件検索が可能

・業者の自律性が少なくとも部分的に失われる。

・新しい業者が参入しにくい。

・情報の更新に問題。

例　個人情報

交通事故で患者が病院に運ばれたとする。持っている書類から名前や住所が分ったが、家に電話をかけても誰も出ない。既往歴や家族の連絡先などを至急調べたい。現在は患者に関係ありそうなところに次々に電話をかける位しか方法がない。

勤務先、市役所、学校、税務署、警察、病院などの情報を電子化し公開可能なものは公開し検索可能とすれば迅速に必要な情報を得ることができる。プライバシイ保護に注意が必要である。このような情報検索はマルチデータベースによっても従来の方法では実現困難である。

広域ネットワークに散在するデータベースへの統一的な条件検索を実現するためには以下のような問題を解決する必要がある。

１．サイトの自律性、独立性の確保。このため、情報の形式や内容のミスマッチがおきる。

２．どこにどんな情報があるか分らない。各サイトはお互い別のサイトが何を持っているかを知らない。

最初の問題はマルチデータベースでは従来から研究されてきた[KS90][BE96]。たとえば、各データベースの上位に共通的なビューを定義することによって差異をある程度吸収できる。従来の方法ではマルチデータベースの対象となるデータベースがあらかじめ分っていることを前提にしているので、スキーマの統合を事前に設計しておくことができる。しかし、広域検索では少数の決められたデータベースの統合ではなく、不特定多数のデータベースへの統合的アクセスが課題である。たとえば、問合せによって関連するデータを持つサイトが変化するので、どこへ問合せを送るかという問題も含め、２番目の問題の解決も重要である。

広域検索の実現のために用いることのできる技術には以下のようなものがある。

・ディクショナリ、シソーラス

・情報マップサーバー

・ビュー機能
・モデル変換
・予備問合せによる対象絞り込み
・コントラクトネットなどの協調プロトコル
・エージェント技術

構文の違い、名前の違いなどはある程度ディクショナリ、シソーラスなどで吸収可能である。しかし、情報の有無や概念的な違いなどは変換しきれない。従って、情報の有無などはなんらかの統一的表現を用いる必要がある。あるデータベースには存在し別のサイトには存在しない属性などを表現するには何らかの形で不完全情報[AHV95]を表現する方法が必要である。情報の欠落はNullで表現するのが代表的であるが、Nullの意味論には各種の問題があり[Date90]、１つの対象の情報が複数のデータベースに散在する場合の検索に用いるのは困難である。

我々は広域に散在するデータベースの全体を仮想的な１つのデータベースと考え、各データベースはその一部の不完全な情報を保持すると考える。このようなデータベースでは不完全なデータベースに対する問合せとその解を用いてできるだけ完全な解を得ることが主な課題となる。従って、本稿ではまず不完全データベースとその問合せ及び解の表現方法を検討する。名前の不一致などの問題は何らかの方法で解決されたと仮定する。全体データベースには多くの矛盾が存在する可能性もあるが、情報を得ることを優先し矛盾の問題はここでは検討しない。

３．　不完全データベースとその問合せ

問合せは何らかのスキーマを想定してなされる。問合せの前提とするスキーマと対象のデータベースのスキーマとの間に変換しきれない不一致が有るときどうするかがここでの主題である。たとえばＳＱＬでは、

Select 属性リスト
From 関係リスト
Where 条件リスト

の形で問合せを表現する。これらのリストでの指定と対象ＤＢのスキーマが不一致の場合はどうするか？従来のマルチデータベースではあらかじめスキーマの不一致を基に統合的なスキーマを設定しておき、統合スキーマに基づいて問合せが対象データベースのスキーマに変換して処理される。不完全データベースによるアプローチでは問合せはそのまま対象に送られデータベースは答えられるものだけ答える方法をとる。

３．１　不完全データベースの表現

演繹データベースの述語にラベルを導入し、引数の数が異なっても良いとするレコード記法（オブジェクト記法）を用いる[YM94]。この記法は演繹オブジェクト指向データベースQuixote[Yokota92]などでも用いられているが、関係データベースとの互換性を重視しここではオブジェクト指向の概念は導入しない。

（１）単純ファクト（タプル）

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

ここでａは拡張述語記号（関係名）、l_i はラベル（属性名）である。値の分らないラベルは書かない。

（２）条件付ファクト

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

これは if 以下が真なら$a(l_1=b_1, l_2=b_2, \ldots, l_i=b_i)$ が真であることを表す。また、if 以下に変数を用い不等号条件を表すこともできる。ここで、ファクト部と条件部は同一の対象に関するデータを表現しており、演繹データベースのルールとは異なった意味を持っている

$(a(l_1=b_1, l_2=b_2, \ldots, l_i=b_i, l_{i+1}=X_1, \ldots l_m=X_m)$
if $X_1=c_1, \ldots X_m=c_m$ と同じ意味である）。

（３）データベース

単純ファクトと条件付ファクトを総称しファクト（レコード）と呼ぶ。データベースはファクトの集まりである。

（４）問合せ

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

定数は条件、変数は求める対象を表す。また、複

数の拡張述語を書くことができる。引数の順序は任意である。比較記号に不等号を許す。本記法は述語記法と比較しＳＱＬとの親和性が良いことに注意されたい。

３．２　データベースへの問合せとその解

問合せ：

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

に対し、データベースに以下のファクトがあれば解が生成される。

・拡張述語名が一致する。

・ラベルの一致する条件を満たす。さらに、条件付ファクトの場合は問合せ中の値がファクトの同一ラベルの条件を満足する。

解の形式：

・問合せ中の条件部分がファクトにより満足されればそれを解のファクト部に入れ、対象ラベルの変数にファクトの同一ラベルの値を代入し解に入れる。

・問合せの条件部分のラベルがファクトになければ、その条件を解の条件部に記述する。

・対象ラベルと一致するラベルがなければ解では対象ラベルの部分を削除する。

・条件付ファクトの条件を（問合せの条件で満足された部分を除き）解の条件に付加する。

・対象ラベルの一部が出力されない場合は部分解、条件が付いているときは条件付解[注1]と呼ぶ。また、条件付ファクトの条件のラベルが問合せの条件に表れないとき問合せを不完全問合せと呼ぶ。

例　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｝

また以下のようなファクトがある。

学生（名前＝千葉太郎、住所＝習志野市、
電話＝0474-65-0000）

問合せ１

Select 年齢、住所
From　学生
Where　名前＝千葉太郎

これをレコード記法で書くと、

？学生（名前＝千葉太郎、年齢＝Ｘ、住所＝Ｙ）

となる。従来のＤＢＭＳではこの問合せはエラーになる。不完全データベースでは解があり、

学生（名前＝千葉太郎、住所＝習志野市）

となる。年齢に関するデータはないのでこの解は部分解である。

問合せ２

？学生（住所＝習志野市、名前＝X,
学科＝情報）

データベースには学科のデータはないので従来の方式ではエラーになるが、本方式では解があり、

学生（住所＝習志野市、名前＝千葉太郎）
if 学生（学科=情報）

となる。これは条件付解である。

不完全問合せの例

条件付価格

１　時期（何日まで、何日以降など）
２　数量（1000以上など）
３　取引先

などにより価格が変わることは多い。しかも、業者により条件が異なる。

商品｛商品名、価格、条件｝の様なデータが有ったとき、関係ＤＢでこの種の情報を扱う場合は条件の内容はアプリケーションプログラムで処理することになる。このようなデータベースに、

Select　価格
From　商品
Where　商品名＝Mac LC630 and 価格<100000

の様な問合せをしても条件が考慮に入れられていないので正確な解が求められない。

不完全ＤＢでは条件付ファクトを用い、

商品（商品名＝Mac LC630, 価格=95000）
if 商品（発注日< 1231）
商品（商品名＝Mac LC630, 価格=90000）
if 商品（発注数>= 100）

の様に表すことができる。上記の問合せはレコード記法では、

[注1] 条件付解はQuixoteの仮説生成に対応する。

？商品（商品名＝Mac LC630，価格=X），X<10000となり、条件に対応するラベルが問合せ中にないので解は、

商品（商品名＝Mac LC630，価格=95000）
　if 商品（発注日< 1231）
商品（商品名＝Mac LC630，価格=90000）
　if 商品（発注数>= 100）

となる。

4．不完全データベースに基づく広域検索

不完全データベースの概念はNull値を用いる方法と類似しているが、以下のような点で異なる。

（1）情報の欠落をそのまま表現するため、スキーマの差異に対し柔軟に対応できる。

（2）Null値で3値論理を導入し問合せに答える方法と比較し、不確定要素を詳細に表現できる。

（3）条件付ファクトの追加

これらの特徴を利用し、不完全データベースの概念に基づき広域データベース検索を実現することができる。また、検索結果の部分的情報を基により完全な情報を得ることが可能となる。

4．1　解のマージ

ファクトの両立性

2つのファクトに共通に表れる各ラベルの値が以下の条件を満たす2つのファクトは両立するという。

（1）両方のラベルが単純ファクト部に表れるときはその値が一致する。

（2）一方が単純ファクト部に、他方が条件部に表れるとき単純ファクト部のラベルの値が条件部を満たす。

（2）両者が条件部に現れるときは、両者を満足する値が存在する。

解の両立性

解では値でなく変数が存在することがある。この場合は、上記（1）で共通ラベルの少なくとも一方が変数のとき両立する。（2）（3）については変数への値または変数の代入を行って条件を満たす場合に両立する。

両立するファクトまたは解は以下のようにマージすることができる。マージは関係データベースの自然結合を拡張した演算である。

（1）両者が単純ファクト部に表れるラベルは各々いつにまとめる。一方が変数のときは値を代入する。一方の単純ファクト部にのみ表れるラベルとその値はそのままマージ結果に含む。

（2）一方が単純ファクト部に他方が条件部に表れるラベルで単純ファクト部が値のときは、単純ファクト部にラベルと値を含み条件部を削除する。単純ファクト部が変数のときは条件部の削除は行わわず変数を単一化する。

（3）両者が条件部に表れるときは合成した条件とする。

4．2　広域データベース検索

（1）問合せもとのサイトが（情報マップなどで対象サイトを決め）問合せを発行する。

（2）各サイトは問合せを受け取ったら自サイトのデータベースを検索し解を生成する。

（3）問合せもとで解を集め処理する。必要ならば複数サイトからの解のマージを行う。

（4）マージ処理後の条件付解はそのまま解とするか、条件部を問い合わせて無条件解にすることを試みる。

例1：部分解のマージで完全解が得られる場合

？学生（名前＝千葉太郎、住所＝X、単位数＝Y）の解が

（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）

の場合は、解をマージすれば以下の解が得られる。

　学生（名前＝千葉太郎、住所＝習志野市、
　　単位数＝８０）

これは関係データベースの自然結合と同様であ

る。

例２：マージにより条件がなくなる場合

？学生（名前＝千葉太郎、住所＝Ｘ、
　単位数＝Ｙ）、Ｙ＜９０の解が、
（学生課）
　学生（名前＝千葉太郎、住所＝習志野市）、
　　ｉｆ　学生（単位数＝Ｙ）、Ｙ＜９０
（教務課）
　学生（名前＝千葉太郎、単位数＝８０）
の時、マージすと以下の解を得る。
　学生（名前＝千葉太郎、住所＝習志野市、
　　単位数＝８０）

例３：条件付の解を基に問合せを発行し無条件解が得られる場合。

？学生（住所＝習志野市、所得＝X, 名前＝Y），
　X>= 1000000
学生（住所＝習志野市、名前＝千葉太郎）
　if 学生（所得>=100000）

となったとする。この場合は、別のところに所得の情報がある可能性があるので、たとえば、

？人（住所＝習志野市、名前＝千葉太郎、
　所得=X）、X>=1000000

の様に判明した情報を基に再度問合せを行えば、たとえば税務署から

人（住所＝習志野市、名前＝千葉太郎、
　所得=0）

という解が得られる可能性がある。この場合、収入が1000000を越えている場合は解がないが、データがないのか収入が条件を満足しないのか分らない。これを知るためには、

？所得（住所＝習志野市、名前＝千葉太郎、
　所得=X）

によって所得を求めてから条件をチェックする方法もある。

例４：解のマージを行わない場合

？商品（商品名＝ＬＣ６３０、価格＝Ｘ）の解が、
　商品（商品名＝ＬＣ６３０、価格＝８００００）
　商品（商品名＝ＬＣ６３０、価格＝８５０００）

の様に別々の業者から帰ってくれば、別の解として扱う。このような場合は特別なラベルとして「サイト名」などを付加することが考えられる。また、例１や２の場合でも「サイト名」ラベルを用いて情報の出所を示すことが考えられる。このような場合は「サイト名」ラベルはマージ処理で特別な扱いを必要とする。

４．３　実現方法

実現には検索エージェントを用いるのが容易である。既存のＤＢＭＳを用いて実現することもできる。関係ＤＢＭＳと検索エージェントを用いたシステムの処理概要は以下のようになる。ただし、条件付ファクトは扱わないと仮定する。

（１）各サイトに不完全ＤＢ問合せ処理のエージェントをもつ。ユーザから問合せを受け付けた検索エージェントはコーディネータになり、問合せ先を決定し問合せを送る。

（２）問合せを受信した検索エージェントは自サイトのスキーマを調べ該当する情報があるかどうか判断する。この際、名前やデータ型の変換をする。

（３）データベースにある属性のみが表れる形に問合せを変形しＤＢＭＳに問合せを行う。

（４）問合せ中の条件に表れる属性で自サイトにないものは得られた解に条件として付加する。解を問合せ元に送る。

（５）解を受信したコーディネータは必要に応じ解のマージや条件の再問合せを行い最終解を得てユーザに返す。

例

？学生（名前＝Ｘ、住所＝習志野市、収入＝Ｙ）、
　Ｙ＞１０００００
ＤＢには学生｛名前、住所、電話番号｝があるとすれば、この問合せは検索エージェントで、
　？学生（名前＝Ｘ、住所＝習志野市）
に変形される。
その解が、
　学生（名前＝千葉太郎、住所＝習志野市）
の場合は、問合せに使用しなかった条件を解に付加

し条件付解、

学生（名前＝千葉太郎、住所＝習志野市）、

ｉｆ　収入＝Ｙ　ａｎｄ　Ｙ＜１０００００

を得る。

４．４　問題点と検討課題

本方式の問題点として以下がある。

（１）問合せの発行対象サイトを絞り込まないと大量の無駄トラフィックが発生する恐れがある。

（２）ディクショナリやシソーラスで同義語の展開を行うと無関係のデータまで問合せを行う可能性がある。

（３）問合せの条件とデータとの間に共通のラベル（属性）が少ない場合、条件付解が大量に生成されることがある。条件付解と同様な仮説生成機能を持つQuixoteによる実験でもこの現象が観測された。この問題は条件の一定割合を満たす場合のみ条件付解を生成することにより対処できる。

また、以下のような検討課題がある。

（１）マージ処理を行うかどうかの判別方法。当面はユーザに指定さセルなどの方法がある。この問題は広域でのオブジェクト識別性の問題であり、将来重要な問題になろう。

（２）条件付き解が得られたときさらに問合せを行うかどうかの判別方法。

（３）複数の関係に対する問合せ

同一の情報があるサイトでは１つの関係に格納され、別のサイトでは２つの関係に格納されていることがある。問合せ元で仮想的なスキーマで問合せを記述するとき、複数の関係を想定するか１つの関係で記述するかの問題がある。ユニバーサル関係を用いたアプローチも考えられる。関係名より属性名の方が大域的には意味が可能性があり、ユニバーサル関係の概念は有効かもしれない。

（４）ルール（またはビュー）の広域分散

ルール（またはビュー）を用いたときルール本体の関係（または実関係）が同一サイトにない場合への拡張方法。他のサイトに問合せする方式を用いると分散協調処理と同様な問題になる。

５．　まとめ

インターネットなどの広域ネットワークに散在するデータベースへの問合せ（条件検索）はスキーマのミスマッチなどの問題があり実現が困難である。辞書などにより変換しても情報の有無は吸収できない。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はできない。本稿では不完全データベースに不完全解、条件付解などの概念を導入し、広域データベース検索を行う方法を提案した。また、既存の関係データベースを用いて広域検索を行う方法を示した。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。

参考文献

[AHV95] Abiteboul, S., Hull, R., and Vianu, V., Foundations of Databases, Chapter 19, Addison-Wesley, 1995

[BE96] Bukhres, O.A. and Elmagarmid, A.K., Object-Oriented Multidatabase Systems, Prentice Hall, 1996

[Date90]Date, C.J., An Introduction to Database Systems Volume I, Fifth Edition, Addison-Wesley, 1990

[KS90] Kim, W. and Seo, J., Classifying Scematic and Data Heterogenity in Multidatabase Systems, IEEE Computer, 24-12 pp.12-18, 1990

[KC95] Kong, Q. and Chen, G., On Deductive Databases with Incomplete Information, ACM TODS, 13-3 pp.354-369, 1995

[Ullman89] Ullman, J.D., Principles of Database and Knowledge-Base Systems Vol. II, Computer Science Press, 1989

[Yokota92] 横田一正、演繹オブジェクト指向データベースについて、コンピュータソフトウェア、5-4, July 1992

[YM94] 横田一正、宮崎収兄、新データベース論、共立出版、1994