データベースシステム 114－16
（1998. 1. 20）

DBSENA：マルチデータベース環境における情報資源管理と検索方式

星野 隆, 綱川 光明, 町原 宏毅
NTT 情報通信研究所
{hoshino, tunakawa, hiroki}@dq.isl.ntt.co.jp

ネットワークに複数のデータベースが接続されているマルチデータベース環境で情報を検索するとき，データベースの異種性に起因して，データの所在が不明，データ検索手段が不明，データベース毎にデータ表現が異なるなどの問題が生じている．従来は統合スキーマを作成してこれらに対応していたが，多数のデータベースが存在するマルチデータベース環境で，統合スキーマを作成，維持管理するには非常に大きな稼動がかかってしまっていた．本稿ではこれら異種性の問題を，情報資源管理技術を用いて解決し，統合スキーマを作成せずにマルチデータベース環境にあるデータベースを検索する検索エンジンとしてDBSENA を提案する．

DBSENA: Information Resource Management and Retrieval Engine for Multi-Database Environment

Takashi HOSHINO, Mitsuaki TSUNAKAWA, Hiroki MACHIHARA
NTT Information and Communication systems Laboratories

To retrieve information from different database systems, a Multi-database environment, it was necessary to analyze the metadata of all database systems, make an integrated data model. A Multi-database environment is a confederation of preexisting, autonomous, and possibly heterogeneous, database systems. Since the open environment links so many databases together, it has become too expensive to make a complete integrated data model and to ensure that the integrated data model matches each database scheme. In this paper, we propose DBSENA, Database Semantic navigation system, for Information Resource Management (IRM) and Information Retrieval (IR) in a Multi-database environment.

1 はじめに

ネットワークのオープン化に伴い，これまで個別に運用されていた企業のデータベースをイントラネット，エクストラネットに接続した情報共有形態の実現が加速している．

ネットワークに複数のデータベースが接続されたマルチデータベース環境において，その複数のデータベース中の情報を効率よく検索しようとした場合，これまでは個々のデータベースのメタ情報を分析し統合した，統合スキーマを作成する必要があった．この統合スキーマを用いた検索システムでは，統合スキーマに対する検索を考えれば複数のデータベースに対する検索が可能となり，物理的に分散された情報の所在やそれぞれのデータベース上のデータ構造などは意識する必要がなく，利用者はあたかもそれが 1 つのデータベースであるかのように利用できた．

しかし，このためには多くのデータベースを分析して統合スキーマを作成し，個々のデータベースとの対応関係を定義する必要があり，これを提供することは管理者に高度なスキルが必要であり，その実現は非常に困難であった．また，個々のデータベースにおいてメタ情報が変更になった場合の統合スキーマとの整合性確保についても多くの問題があった．

さらに，個々のデータベースを横どおしで見ると，同じ意味のデータ項目に別の名称がついていたり，同じ意味を示すデータであっても，そのデータ値の表現が異なっているなど，データベースの異種性があり，複数のデータベースに対する検索ではそれらを解消しておく必要があった．

本稿では，これらの問題を解決しマルチデータベース環境での情報検索を実現するために，統合スキーマを作成する方法ではなく，個々のデータベースのメタ情報を各データベースから自動的に収集し，それらに対して情報を付加し，情報資源として一元的に管理する辞書を構築する情報資源管理方式とそれを用いた検索システムであるDBSENAを提案する．

以下，2章ではマルチデータベース環境における情報検索の現状と問題点を指摘し，3章ではそれらの問題を解決するために我々が開発したDBSENAの概要を述べる．4章ではDBSENAの異種性解消方式を示し，5章では情報資源辞書の構造と管理ツールを，6章ではアプリケーションプログラムがDBSENA機能を使うためのSENA-APIを，7章ではDBSENAを用いた異種DB検索例を示す．最後に8章でまとめとして，今後の課題などを示す．

2 マルチデータベース環境での異種データベース検索の問題点

マルチデータベース環境にあるデータベースに対して，簡単に，かつデータベースが複数あることを意識せず検索し，情報を取得したいという要求がある．これまでデータベースは基幹業務を効率的にこなすという目的のため，部門，組織など業務遂行に適した形態で独自に構築されてきた．

このようなマルチデータベース環境を考えるとき，構成要素となるデータベースが多種多様であることから，以下のような意味的異種性を考慮しなくてはならない[1]．

- データの名称の違い
- データ構造の多様性
- データ表現の違い

これら異種性に起因する，マルチデータベース環境上で情報検索を行うための問題は以下のとおりである．

- スキーマ名からだけでは，情報の所在を特定することは困難

  スキーマ名が同じ名称であっても内容が異なっていたり，逆に異なった名称であっても同じ内容であったりすることがある．また，テーブル名，列名などに意味を表さない記号（通番や略称）が使用されている場合などはその内容を類推することすらできない．
- データ構造が異なるため，情報取得のための検索命令を生成するのが困難

  データ構造がデータベース毎に異なるため，ほしい情報が複数テーブルにまたがる場合の結合条件などが異なってしまう．そのため，検索要求に応じ対象のデータベース毎に異なる検索命令(SQL)を発行しなければならない．
- 情報の表現形式がデータベースにより異なり，取得方式を特定することは困難

  データベース中の値がコード化されている場合がある．しかもそのコードがデータベースによって異なる場合がある．また「男/女」と「male/female」のように，同じ意味のデータでも表現が異なる場合もある．このため，検索対象となるデータベースのデータの表現形式を考慮して条件の指定をしないと，データは存在するにもかかわらず，「Not Found」となり検索できない場合がある．

これらに加え，ユーザインタフェースとして，「自分の持っているデータベースと同じ表現形式でデータを取得したい[2]」という要求があり，

検索対象データベースとともに，検索者が望んでいるデータの表現形式も考慮する必要がある．

これまでこのような異種性の問題を解決し，複数のデータベースに対する検索を行うためには，個々のデータベースのスキーマを統合した統合スキーマを作成していた．しかしこの統合スキーマを構築するためには，その要素となる各データベースのスキーマを整理し統合する必要があり，高度なスキルを必要とするるため，その実現は非常に困難なものであり，かつ多くの工数を必要としていた．さらに一度作成した後，データベースに変更が生じたり，あらたにデータベースを追加する場合，統合スキーマの変更に非常に多くの工数が必要となっていた．

3 情報検索システム DBSENA

3.1 DBSENA の概要

我々は，このような異種性の問題を解決しマルチデータベース環境にある複数のデータベースを検索するために，情報資源管理技術を用いた情報検索エンジンである，DBSENA(DataBase SEmantic NAvigation system)のプロトタイプを開発した．このプロトタイプは Windows NT 上で動作する．

DBSENA の構成を図 1に示す．DBSENA は以下の要素から構成される．

- 異種データベース検索機能
  ユーザが指定した検索要求に従い，異種性を解消し，DB アクセス機能を介して複数のローカルデータベースを検索する．この詳細は4章で述べる．
- DB アクセス機能
  ローカルデータベースとの間で，検索要求，検索結果，スキーマ情報等のやりとりを行う．
- 情報資源辞書
  検索対象となるデータベース（ローカルデータベース）のデータ構造，データの表現情報，データ値の範囲情報と，それぞれのローカルデータベース間での表現形式の関係などを管理する．詳細を5.1節にて示す．

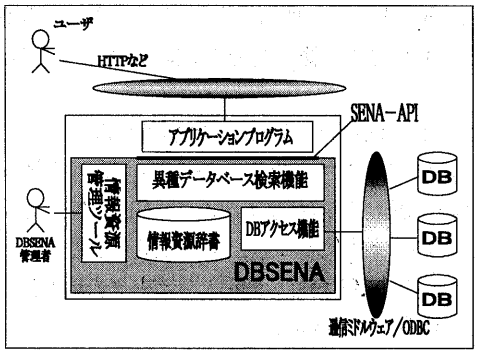

図 1：DBSENA 構成図

- 情報資源管理ツール
  情報資源辞書のメンテナンスを行う．5.2節で詳細を示す．
- SENA-API
  DBSENA 機能をアプリケーションプログラムで使用するため，SENA-API を提供する．詳細は6章で述べる

3.2 DBSENA での異種性解消の考え方

DBSENA では情報の所在を指定させないため，普遍関係によるインタフェース技術[3]をマルチデータベースに適用し，複数のデータベースへの検索を実現する．したがってユーザは検索要求時に，データベース名，テーブル名など所在を指定する必要がない．すなわち SQL 文の select 句に相当する検索したい項目名と，where 句に相当する検索条件となる項目名，比較演算子，条件値のみを指定することとなる．

この検索要求に対して，断片的に定義された異種性解消関数を組み合わせる動的な異種性解消方式である Fragment View[4]を利用し，データベース間の異種性を解消し検索を行う．

これにより，DBSENA で管理する情報資源は個々のデータベースの情報と異種性解消のために必要な表現形式間の関係，同義語辞書であり，統合スキーマを構築する必要はなくなる．したがってローカルデータベースの追加，変更に対しても，そのデータベースの情報を追加，変更して，表現

形式間の関係を追加，修正するだけとなり，構築時，維持管理での稼動の大幅な削減が可能となる．

4 DBSENAによる異種性解消方式

2章であげたマルチデータベースにおける情報検索の問題点を解消するために，DBSENAでは3.2節の考え方に基づき，情報の所在を推定し，データ構造の異種性を解消し，データ表現の異種性を解消し情報検索を行う．

4.1 情報の所在の推定

DBSENAでは，ユーザの検索要求にもとづき情報の所在を推定する．情報の所在の推定は2段階にわけて行う．まず第1段階では列の所在推定を行う．これにより検索要求に対応するローカルデータベースの列を決定し，検索候補となる列を確定する．つづいて第2段階では，検索条件として指定された条件値によって，検索対象候補のDBがその情報をもっているかどうかを判断し，検索対象となるデータベースを絞り込む．この結果，検索候補となるデータベース，テーブル，列が決定し，情報の所在を推定する．

4.1.1 列の所在推定

第1段階として列の所在推定を行う．ユーザが入力した検索項目に対応する，ローカルデータベース上の検索候補となる列を推定する．ここでは普遍関係インタフェースを用いるため，検索項目単位に所在の推定を行う．

検索要求として受け取った検索項目を，あらかじめ情報資源辞書に定義された同義語辞書を使用して，ローカルデータベース上の列名に対応づけることにより検索対象となる列を推定する．このように同義語辞書を用いることにより，入力項目名と列名は一致している必要はなくなる．

このようにして推定した列を組みあわせ，検索要求と対応づけることで，データベース，テーブルなどが指定された検索候補を生成する．

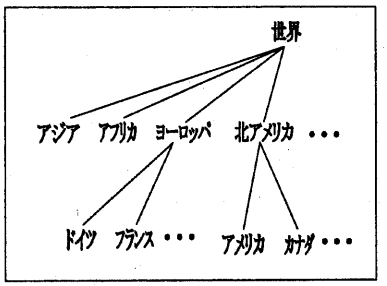

図2：レンジ情報ツリー

4.1.2 検索候補の絞り込み

第2段階では条件値による検索候補の絞り込みを行う．ここでは列のレンジ[5]を用いる．レンジとは値のとりうる範囲を表す概念である．

ここでは値のとる範囲を表現するため，データベース中に格納されている値を木構造で表す．この木構造をレンジ情報ツリーと呼ぶ．これにより，上位概念は下位概念を含むことを表現する．たとえば地域に関してレンジ情報ツリーを考えると，地域の情報は図2のように木構造として整理することができる．このとき世界はヨーロッパを含み，ヨーロッパはフランスを含むことがわかる．

ローカルデータベースの列がとりうるデータの範囲を列のレンジとして情報資源辞書に設定しておく．ユーザが検索条件値として指定した値と検索候補項目となっている列のレンジとを，レンジ情報ツリーを用いて比較することで，値の範囲情報をチェックし，その列が検索対象となるデータをもっているかどうかを判断する．データを持っていない列であれば，そのデータベースは検索対象外となり，このデータベースを検索対象としている検索候補を排除することで，検索候補の絞り込みを行う．

4.2 データ構造の異種性解消

DBSENAでは普遍関係インタフェースを用いることで，データ構造の意識をなくしている．そのため，検索しようと思う項目，条件とする項目のみを意識することとなる．

所在の推定の結果，検索対象となる項目がローカルデータベースのスキーマ上では複数のテーブルに分散している場合がある．このとき，情報資源辞書に設定されている，対応するテーブルどうし，列どうしのつながりである関連情報を用いて，それらの中で対象テーブル間の最短経路をつなぐ関連をテーブル間の関連とする．

4.3 データ表現の異種性解消

データ表現の異種性はドメインという概念を用いて解消する[4]．ドメインを用いることで，データベース間の表現の異種性を解消するとともに，ユーザの望む表現形式での出力を可能とする．

ドメインとは値の表現形式を表す概念である．たとえば，「1,000,000 円」は「金額円単位カンマ区切りあり」表現ということができ，「金額円単位カンマ区切りあり」ドメインであるといえる．図 3にドメインの例を示す．

同じ意味をもつドメインの集まりをドメイングループと呼ぶ．「1,000,000 円」という表現（ドメイン）と，「100 万円」は同じ意味をもち，表現形式が異なっているだけである．これらは同一のドメイングループに属しており，「金額」ドメイングループということができる．このドメイングループを代表するドメインをグローバルドメインと呼ぶ．これはドメイングループの中から 1 つのドメインを選択する． 図 3では「円単位（1000000 円)」となる．

個々のデータベースの列で用いられているドメインをローカルドメインと呼ぶ（千円単位，百万円単位)．また，ユーザが検索結果として受け取りたいドメインを，ドメイングループの中から 1 つ選択する．これをユーザドメインと呼ぶ（円単位カンマ区切りあり)．

ドメインの変換は，1 つのドメイングループの中でグローバルドメインを介して変換する機能を提供する．この変換を行う関数をドメイン変換関数と呼ぶ．この変換関数は，DBSENA で提供するものと，ユーザが独自に作成するものとがある．

検索時は，まずユーザが入力した検索条件値

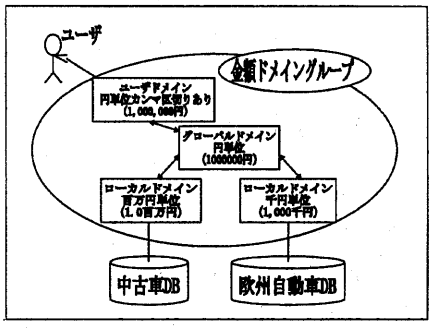

図 3：ドメイン

のドメインを推定する．この推定されたドメインと検索対象列のローカルドメインとの変換を行うことで，ローカルデータベースに適合した検索条件値となる．これを用いて検索を行い得られたローカルドメインで表現されている検索結果を，ユーザが指定したユーザドメインに変換し出力する．

5 情報資源管理

5.1 情報資源辞書

これまで述べたように，DBSENA ではいろいろな情報を使用してデータベースの異種性を解消する．これらの情報は情報資源辞書で管理する．ここで管理している情報の概要を以下に示す．

A) 個々のローカルデータベース毎の情報

- データベースのスキーマ情報
  テーブル名，列名，データ型，意味，テーブル間の関連などのスキーマ情報
- 列のドメイン情報
  列の表現形式を示すドメイン情報
- 列のレンジ情報
  列がとる値の範囲を示すレンジ情報
- DBMS 種別
  ローカルデータベースの DBMS 情報
- データベースアクセス手段
  通信ミドルウェア，ODBC 等によるローカルデータベースアクセスに必要な情報

B) 異種性を解消するための情報

- 同義語辞書
  データ項目名に対する同義語辞書

・ドメイン，ドメイングループ情報
　ドメイングループ，グローバルドメインなどドメインに関する情報
・ドメイン変換情報
　ドメイン変換を行う関数の情報
・レンジ情報
　レンジ情報ツリーに関する情報

C) ユーザに関する情報
・アカウント情報
　ユーザ名，パスワードなど
・アクセス可能データベース
　ユーザがアクセス可能なデータベース
・ユーザドメイン
　ユーザが望むデータ表現形式

これら情報資源を管理する情報資源辞書は市販のRDBを用いて運用されている．

5.2 情報資源管理ツール

情報資源辞書を管理するために情報資源管理ツールを提供し，管理稼動の削減を図っている．DBSENA管理者はこのツールのGUI画面上で，異種性解消のために必要な，レンジ，ドメイン，テーブル間の関連情報など各情報の設定を行う．

このツールは，データベース名，DBMS種別，データベースアクセス手段を指定することで，ローカルデータベースのスキーマ情報を自動収集し，情報資源辞書に情報を蓄積することが可能である．このスキーマ情報はDBMS毎にメタ情報の内容，構造などが異なっているが，その異種性を解決している．したがってDBSENA管理者は，収集されたスキーマ情報に，異種性解消のための情報を設定する作業のみを行うこととなる．

6　アプリケーションインタフェース

DBSENAの異種データベース検索機能をアプリケーションプログラム（AP）から利用するために，JDBCライクなSENA-APIを提供する．APからこのSENA-API関数を呼び出すことにより，DBSENAを利用した情報検索APを構築することができる．現在，SENA-APIはC言語の関数ライブラリとして提供している．

SENA-APIは，検索要求作成，検索候補作成，検索実行の3種類に分類される．まず検索要求作成API を用いて検索要求構造体を作成する．次に検索候補作成 API を呼び出すことで，データベースの異種性を解消し，検索対象となるデータベースなど，検索候補の情報をユーザに提示する．最後に検索実行 API を用いて，ローカルデータベースを検索しユーザに検索結果を出力する．

7　異種DB検索例

図 4のようなスキーマを例にとり，異種性解消方式を説明する．2つのデータベースがそれぞれ異なるデータ構造を持っており，同じ意味を持つ列のデータ項目名が異なっている．「中古車DB」は，日本国内で販売している全世界の車の情報が格納されており，3テーブルで構成されている．一方「欧州自動車DB」は日本国内で販売

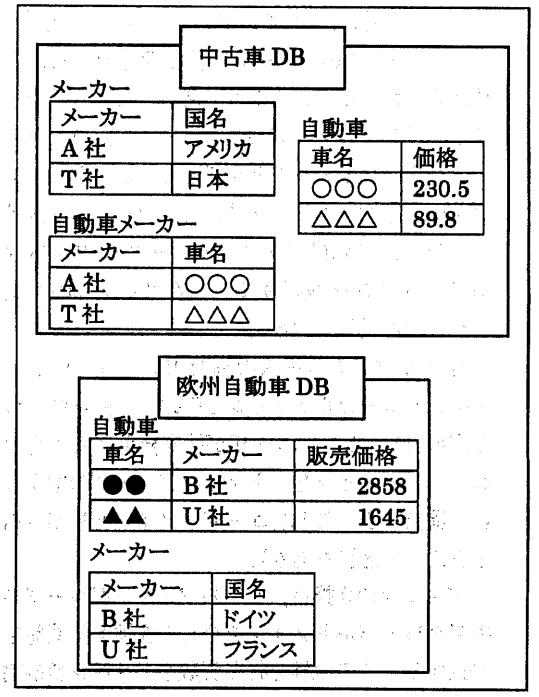

図 4：検索対象データベース

しているヨーロッパ製の乗用車が格納されており，2テーブルで構成されている．この2つのデータベースに対して異種性を解消して検索を行うために，先に図 2，図 3で示したレンジ情報，ドメイン情報を情報資源辞書に登録しておく．またドメイン，レンジ，同義語辞書，関連情報を以下のように設定する．

- ドメイン
  - 中古車DB.自動車.価格：金額百万円単位
  - 欧州自動車DB.自動車.販売価格：金額千円単位
- レンジ
  - 中古車DB.メーカー.国名：世界
  - 欧州自動車DB.メーカー.国名：ヨーロッパ
- 同義語辞書
  - 販売価格｜価格｜Price
- 関連情報

✧ 中古車DB：
- メーカー.メーカー＝自動車メーカー.メーカー
- 自動車.車名＝自動車メーカー.車名

✧ 欧州自動車DB：
- 自動車.メーカー＝メーカー.メーカー

この環境において，「国名がアメリカで価格が1,000,000 円以上の自動車の，車名，価格が知りたい」という検索を DBSENA に対して行う．ユーザはDBSENAに対して，

```
検索項目：車名，価格
検索条件：国名 ="アメリカ"
          and 価格 ≧ 1,000,000円
```

と入力する．

まず DBSENA は，検索項目である「メーカー」，「車名」，「価格」，「国名」について，列の所在推定を行う．このうち「価格」は，同義語として定義された情報から，「中古車 DB：自動車.価格」，「欧州自動車 DB：自動車.販売価格」と，ローカルデータベースでの所在が推定される．

つぎに，検索条件に指定されている「価格」の条件値「1,000,000 円」のドメイン推定を行う．このドメインは「金額円単位カンマ区切りあり」であるので，個々のデータベースの対応する列のドメインにあわせて変換する．すなわち「中古車DB」に対する検索条件値は「1.0」，「欧州 DB」に対する検索条件値は「1,000」となる．

つづいて，もう 1 つの条件値である「国名」を用いて検索対象データベースの絞り込みを行う．「欧州自動車 DB」のデータはヨーロッパ製の自動車情報のみを管理しており，「国名」のレンジはヨーロッパの国名のみとなる．図 2のレンジ階層構造から，検索条件値「アメリカ」は「ヨーロッパ」には含まれないので，「欧州自動車 DB」は検索候補からはずれることになる．

残った「中古車 DB」に対して，関連情報を補い，以下のSQL文を生成する．

```
select 自動車.車名, 自動車.価格
from   自動車, メーカー, 自動車メーカー
where メーカー.国名 = "アメリカ"
  and 自動車.価格 >= 1.0
  and メーカー.メーカー
              = 自動車メーカー. メーカー
  and 自動車メーカー.車名 = 自動車.車名
```

この SQL 文を用いて，「中古車 DB」からデータを検索する．

検索結果については，「価格」についてはユーザドメイン（円単位カンマ区切り）が設定されている．ローカルドメインは（円単位）であるので，ローカルデータベースから検索した結果に対し，ユーザドメインへドメイン変換を行い，検索結果を出力する．

したがってこの検索結果として，

車名：○○○，価格：2,305,000円

という情報を得ることができる．

8　まとめと今後の課題

以上，マルチデータベース環境における異種データベース検索の問題点をあげ，その解決手段として DBSENA を提案し，DBSENA で実現した情報資源管理と検索方式についてのべた．

DBSENA は情報資源辞書に個々のデータベースの情報と，その情報間の関係を設定することに

より，マルチデータベース環境での異種データベースの検索を可能とする．個々のデータベースの情報を使用するので，統合スキーマを作成する必要がなく，情報資源辞書への情報設定も簡易化を行っているため，マルチデータベース検索の実現を容易にしている．

しかしながら，現在はプロトタイプ環境での動作のみを行っており，今後は実システムへ適用し，評価を行う予定である．さらに，現在は文字・数値データのみを扱うことが可能であるが，これに関しても静止画データなど，マルチメディアデータへの対応を行う予定である．

参考文献

[1] 上林弥彦，"マルチデータベースの研究開発動向"，情報処理，35(2)，1994.

[2] Raschid, L. and Chang, Y-H., "Interoperable Query Processing from Object to Relational Schemas Based on a Parameterized Canonical Representation", International Journal of Cooperative Information Systems, 4(1), 1995.

[3] Ulman, J.D., "ユーザインタフェースとしての普遍関係（第9章）", in データベースシステムの原理（第2版），1985.

[4] 鈴木源吾，町原宏毅，川下満，"Fragment View ーマルチデータベースにおける Global View を使わない異種性解消方式"，電子情報通信学会第96回データ工学研究会，1997.

[5] 鈴木源吾，町原宏毅，"データベースの値の範囲の管理法とその普遍関係ユーザインタフェースへの応用"，電子情報通信学会第 97 回データ工学研究会，1997.