データベースシステム 116－77
（1998．7．10）

マルチデータベース環境における関連情報推定と検索方式

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

これまで個別に構築されてきたデータベースを，オープン化されたネットワークに接続し，その情報を共有化しようという動きが活発化している．しかし，データベースの異種性が存在することから，複数のデータベースを検索するための統合スキーマの作成には非常に大きな稼動を必要とし，新規データベースの追加などその維持管理にも一層の稼動を必要としていた．さらに，オープンなネットワークでは非常に多くのデータベースを統合するため，データベース間の関連情報を発見することは，そのすべてを管理者の知識にゆだねなければならないこともあり困難なものとなっていた．また，関連情報を発見したとしても，データの異種性のために結合して検索することができない場合があった．

そこで本稿では，普遍関係インタフェースを用いたマルチデータベース検索システム DBSENA を用いて，これらの問題を解決する手法を提案する．

Inter-Database Relationships Discovery and Retrieval Method for Multidatabase Environment

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

Recently, the databases individually constructed are connected with the network for the information sharing. Since there are many heterogeneous databases, it has become too expensive to make a complete integrated data model and to ensure that the integrated data model matches each database schema. In addition, it occasionally has to be entrusted the all to developer's knowledge and is difficult to find related information between databases to integrate many databases. Moreover, it is not possible to retrieve by joining with the heterogeneity straightening of the database even if related information is found.

In this paper, we propose the method that solves those problems, using DBSENA and Global Information Resource Dictionary.

1 はじめに

ネットワークのオープン化に伴い，これまで個別に構築されてきたデータベースをインターネットや，イントラネット，エクストラネットに接続して社内や社間で情報を共有し，データを有効活用しようとする動きが活発化している．ところがこれら複数のデータベー

スは個別に開発，運用されてきたため，多種，多様な意味的，構造的異種性が存在している．このため複数のデータベースがネットワーク上に存在するマルチデータベース環境で情報を効率よく検索しようとした場合，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[1]．しかし，統合スキーマを作成するためには，多くのデータベースを分析，統合し，個々のデータベースとの対応関係を定義する必要があった．これには管理者に高度なスキルが要求され，データベースが多数存在する場合，その実現は非常に困難なものとなっていた．また，あらたにデータベースが追加された場合や，個々のデータベースにおいてスキーマ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．

われわれはマルチデータベース環境におけるデータベースの異種性の問題を解決し，頻繁に行われるデータベースの追加，変更に対しても，容易かつ安価にマルチデータベース検索を可能とするため，情報資源管理技術を用いたマルチデータベース検索エンジンDBSENA(図 1)を開発した[3]．

DBSENA は，検索対象となるそれぞれのデータベースの情報と，データベース間に存在する名称，データ構造，データ表現などの異種性を解消するための情報を管理する情報資源辞書(G-IRD:Global Information Resource Dictionary)，これら異種性を解消し検索を行う異種性解消検索機能，検索対象データベース(LDB:Local Database)とのやりとりを行うLDBアクセス機能から構成される．

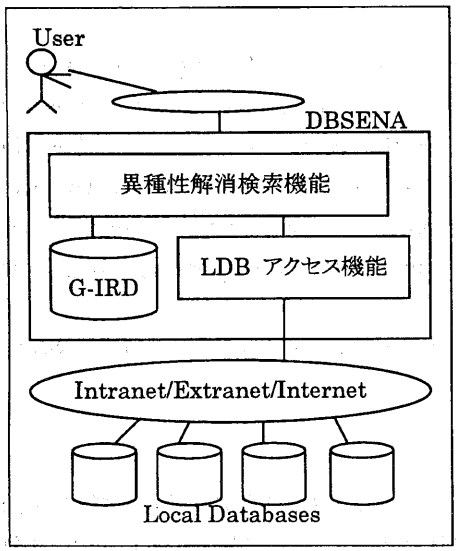

図 1:DBSENAを用いた検索システム

DBSENA は普遍関係インタフェース[2]を用いており，カラムを単位として異種性を解消し検索を行う．ユーザは検索をおこなうために検索対象の項目と検索条件を指定する．DBSENAはこの要求にしたがって，以下の 3 段階の操作を経て，データベースの検索を行う．

1) ユーザの検索要求に対して，G-IRDに定義された情報にもとづき，データの所在を推定し，外部スキーマを作成し，検索対象項目，検索条件に対応するデータベース，テーブル，カラムや，テーブル間の関連の情報とともに，検索候補としてユーザに提示する．
2) ユーザは提示された検索候補から実際に検索する候補を選択し，DBSENA へ検索を要求する．
3) DBSENA は選択された検索候補から各LDBに対する検索文を作成し，LDBを検索する．
4) LDB の検索結果を統合し，データの表現形式などをそろえて，ユーザに出力する．

DBSENA のように外部スキーマを作成する場合に限らず，これまでのように統合スキーマを作成する場合であっても，テーブル間の関連情報が必要となる．しかしながら，このような関連情報の発見は，LDBを分析し，他のLDBとの整合性を考慮する必要があり，設計者もしくは管理者の大きな稼動を必要としていた．また，関連が定義されていたとしても，異なるデータベースのテーブルを結合しようとする多くの場合，データの表現形式の違いからテーブルを結合した検索ができないという問題がおきている．

そこで，本稿では，容易にデータベース間を結合

する検索を可能とするため，データベース間の関連情報の推定を行い，データベース間を結合するために障害となる異種性を解消し検索する方式を提案する．

以下，2 章ではデータベースを連携するための関連設定および検索時の問題を指摘し，3 章では DBSENA が異種性を解消するための管理情報を説明する．つづいて 4 章ではデータベースの結合を行うための関連情報の推定方式を提案し，5 章でデータベース間の結合時の異種性解消方式を述べ，6 章では検索候補の提示方式を述べる．7 章ではこれらを具体例で説明し，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における関連

2.1 関連情報定義の問題

マルチデータベース環境で，複数のデータベースの情報を統合するためには，それぞれのデータベースを結ぶ結合条件となる関連情報を規定する必要がある[5]．1つのデータベースに閉じる関連情報であれば，カラムの名称が同じであることやデータベースの設計情報などから，結合条件となるカラムを決定し関連情報とすることは可能である．しかしながら，複数のデータベースにまたがる関連については，同じ意味を持つデータであっても別の名称がつけられていたり，同じカラム名であってもまったく別のデータを持っているなどの異種性が存在する．たとえば，あるデータベースに“電話番号”というカラムがあり，もう一方のデータベースには“連絡先”という電話番号をあらわしているカラムがある場合や，おなじ“連絡先”という名称のカラムであっても一方のデータベースでは電話番号で，もう一方では住所である場合などである．

このとき，この2つのカラムを関連づけ，結合条件とすることができるかどうかは，カラム名からだけではわからない．これらを結合条件とすることができるかどうか判断するためには，カラムの名称，設計情報，実際のデータ，設計者の知識などを用いて類推することになってしまう．

このような場合であっても，マルチデータベース環境に存在するデータベースが比較的少数であれば問題ないが，多数のデータベースが検索対象となっている場合には，有効な関連情報を発見し定義するのは困難なものとなってしまう．

2.2 結合データ値の異種性

マルチデータベース間で関連が定義され結合条件となるカラムが設定されたとしても，異なるデータベースを結合しようとすると，その多くの場合データ表現形式の異種性のために直接結合することができない[4]．それは同じ意味を表すデータであったとしても，データ表現形式が異なっているためである．

たとえば，同じ“電話番号”というカラムが異なるデータベースにあったとき，一方のデータベースでは“0468-59-0000”と表現されており，もう一方では“0468590000”と表現されている場合，結合条件としてこのカラムどうしを等号で結んだとしても，その比較結果は不一致となってしまい，実際には結合処理を行うことができない．したがって，定義された結合条件を用いたとしても，検索実行時にはデータ値が一致するレコードが存在しないため，検索結果は返却されないこととなってしまう．

2.3 普遍関係インタフェースをマルチデータベース環境に適応した情報検索の問題

普遍関係インタフェースを用いた主な検索システムでの関連の扱う手法として，1 つのカラムに複数の関連がある場合，1つのカラムを関連毎に複数のカラムに分離することでユーザからは関連を意識せず，異なるカラムとして見せる手法[7]や，関連情報，関数従属性などから maximal object を検索候補として決定し，ユーザにどの maximal object を使うかを選択させる手法[8]などがある．しかしながら，こ

れらの手法もデータベースの追加，削除,変更が頻繁におこるマルチデータベース環境に適応した場合，そのモデルの修正，再定義などには大きな稼動を必要とする．

また連邦データベースに普遍関係インタフェースを適応した場合，関連については外部スキーマのみを考慮するのでその数は減少すると考え，ユーザの検索要求に対応するすべての検索候補をその所在情報，関連情報などと一緒にユーザに返却している手法がある[9]．これは検索対象となるデータベース数が少ない場合にはそれほど問題とはならないが，データベース数が増加した場合には，大量の検索候補が作成され，ユーザ選択が難しいものになると考えられる．

3 DBSENAで管理する異種性解消情報

関連設定，検索の問題を解決するため，DBSENA を拡張し，管理している異種性解消情報を利用する．DBSENAで異種性を解消するために，G-IRD で管理している情報は以下のとおりである．

- 同義語辞書

DBSENA は普遍関係インタフェースを用いているため，ユーザは情報の所在を指定しない．そこでDBSENA は情報資源辞書にあらかじめ登録された同義語辞書を用いて，ユーザが入力した検索項目に対応する名称を持つカラムを推定する．カラムの名称は実際のデータベースに登録されている定義名とユーザがカラムの名称として G-IRD に登録したカラム名がある．同義語辞書はユーザが登録したカラム名に別名を定義することで構築する．

- 関連情報

DBSENAでは，普遍関係インタフェースを用いているため，ユーザは検索時に結合条件の指定を行わない．そこで結合条件となるカラムの組を関連として，あらかじめ G-IRD に登録しておく．

関連は，関連元のデータベース名，テーブル名，

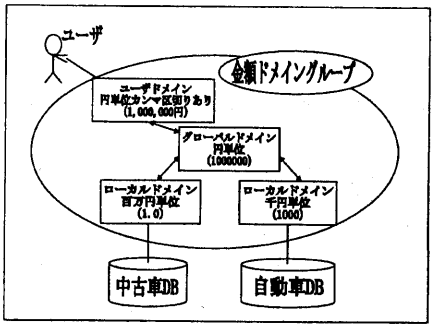

図 2:ドメイン

カラム名，関連先のデータベース名，テーブル名，カラム名，とその組み合わせを識別する関連名からなる．複数のカラムを用いた結合を実現するため，1つの関連の中で複数のカラムの組を設定することができる．

- ドメイン

データの表現形式の異種性を解消するために，DBSENA ではデータの表現形式をあらわすドメインという概念を用いる．このドメインのうち，同じ意味をもつドメインの集合をドメイングループと呼ぶ．

図 2は金額を表現するドメイングループをあらわしている．LDB のデータ表現形式をローカルドメイン(百万円単位，千円単位)と呼び，ユーザに出力するドメインをユーザドメイン(円単位カンマ区切りあり)と呼ぶ．それぞれのドメインはドメイングループの標準的なドメインであるグローバルドメイン(円単位)を介してデータの変換を行うことで，データの表現形式の異種性を解消する．

4 データベース間の関連情報の推定

2.1節で示したように，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は，特にデータベースをまたがる関連情報の発見，定義には大きな稼動を必要とする．これを削減するため，データベースをまたがったテーブル間の関連，結合条件の推定を行う．

G-IRD では，個々のデータベーススキーマに関する情報だけではなく，異種性を解消するための情報

を管理している．この情報を利用して，データベース間の関連推定を実現する．この推定した関連をDBSENA 管理者に候補として提示し，管理者がG-IRD に登録する．

なお，DBSENA 管理者は本方式を使用する以前に，LDB のスキーマ情報や異種性解消情報をあらかじめ登録しておく必要がある．

(A) 同義語辞書

まず同義語辞書の管理情報を用いて，関連となる候補を推定する．これはスキーマ統合を行う際のローカルスキーマと統合スキーマ間のカラムの対応づけと同様に考えることができる．スキーマ統合の際の手法として，知識ベースと名称の類似性を利用する手法が提案されている[10]．しかし知識ベースは適応できる領域が限られていたり，その作成に大きな稼動を必要となる．そこで，ここでは名称の類似性のみに着目し，同義語,類義語を管理する同義語辞書を用いる方式を提案する．

同義語辞書に登録されている用語は，カラム名の同義語，類義語となっている．そのため，同義語によっては1つの同義語から複数のカラムに対応づけられる場合がある．これらのカラムは異名同義であるので，関連の候補とする．たとえば"都道府県名"という言葉に対して，DB α のテーブル A の"県名"，DB β のテーブル B の"都道府県"という2つのカラムが対応づけられているとき，「α.A. 県名 $=\beta$. B.都道府県」という結合条件が候補となる．

(B) ドメイン

同義語辞書から推定された関連候補は，それを用いて結合しようとしても，データの異種性があるため必ずしも結合できるとは限らない．そこで，ドメイン情報を利用して結合可能かどうかを判定する．

同じドメイングループに属するドメインをカラムのドメインとしているそれぞれのカラムどうしの組み合わせはグローバルドメインを用いることで，相互に対応づけ可能であることから，結合処理を行うことができる．そこで，関連の候補となっている列のドメインが同じドメイングループに属している場合のみ関連の候補とする．

たとえば，「α.A. 県名 $=\beta$. B.都道府県」という関連について，両方のカラムのドメインが"都道府県ドメイン"となっていれば関連の候補となる．または，"都道府県ドメイングループ"内に設定されている"都道府県コードドメイン"や，"都道府県略名ドメイン"がそれぞれカラムのドメインとなっている場合も関連の候補とすることができる．

5 結合データ値の異種性の解消

2.2節に示したように，関連が決定された場合であっても，結合対象となっているカラムにデータ表現の異種性があった場合，そのカラムどうしを結合した検索はできない．DBSENA ではドメインを利用して，この異種性を解消する．

結合を行うカラムは，ユーザの定義した関連情報にもとづき決定される．前述のように関連情報の定義を行う際に，結合を行うカラムのデータ値をドメインを用いて一致させることができるもののみを定義可能としている．そこでドメイン変換を行うことによりデータの表現形式を一致させる．

したがってユーザからデータの検索要求があったとき，LDBから検索したデータについて，結合対象となるカラムのドメインが異なっている場合には，データ値のドメイン変換を実施しその表現形式を一致させる．結合対象カラムのドメインはドメイングループが同一なので，標準形式であるグローバルドメインにデータ値を変換する．これにより結合カラムのデータ値の異種性を解消する．

このようにしてデータの表現形式が一致した結合カラムを用いて，ソートマージ方式で結合処理を実施し，検索結果として出力する．

6 複数の関連情報の提示

DBSENA は普遍関係インタフェースを用いているため，2.3節に示したような複数の関連に対する問

題が生ずる. DBSENA では1つの検索要求に対し, 関連が異なるものは別の検索候補として提示することで, ユーザによる関連の選択を可能とする.

前述のように DBSENA では, 検索要求に対して検索候補をユーザに提供し, ユーザがそれを選択することで検索対象を決定する. そのため, 検索要求に対するカラムの所在推定を行い, 関連情報の補完を実施する.

1) 同義語辞書を利用し, 検索要求の各項目に対応する, データベース, テーブル, カラムという検索する LDB 上のカラムの所在を推定する.
2) 一般に1つの検索要求項目に対して, 複数のカラムの所在が推定されるので, これら推定されたカラムを組みあわせ複数の検索候補を作成する.
3) この検索候補の各カラムをテーブル毎に分離し, G-IRD に定義された関連情報を用いて, 結合条件を補完する. 結合条件は 1 度通ったテーブルは複数通らないように選択する.
4) 検索要求項目がすべて結合できた検索候補のみがユーザに返却される.

検索候補は, 検索項目に対応する各カラムの所在情報と, その検索候補で対象とする関連情報で構成される. この関連情報の中には結合するカラムの所在情報とともに, 関連選択の指針となる関連名をあわせて持っている.

これにより, ユーザは検索候補を選択することで, 実際に検索する関連を選択することが可能となる.

7 データベースを結合する情報検索例

図 3に示すような, 電話の動作状況を管理している電話管理 DB と, 顧客の情報を管理している顧客 DB という2つのデータベースにについて, 関連情報の推定と, マルチデータベース間結合について例をあげて説明する.

電話管理 DB. 状態テーブル

電話番号	動作状況
0468590000	故障中
0468590001	正常
0468590002	正常

顧客 DB. 連絡先テーブル

顧客名	連絡先
星野隆	0468-59-0000
綱川光明	0468-59-0001
町原宏毅	0468-59-0002

図 3:LDB 例

同義語辞書

用語	DB	テーブル	カラム
電話番号	電話管理	状態	電話番号
電話番号	顧客	連絡先	連絡先

ドメイン設定

DB	テーブル	カラム	ドメイン
電話管理	状態	電話番号	電話番号
顧客	連絡先	連絡先	ハイフン電話番号

ドメイン定義

ドメイングループ	ドメイン	ドメイン種別
電話番号	電話番号	グローバルドメイン
電話番号	ハイフン電話番号	一般ドメイン

図 4:G-IRD 定義例

7.1 関連情報推定

それぞれのデータベースに対するスキーマ情報と異種性解消情報を登録したときの G-IRD の定義情報を図 4に示す. ここでは以下の説明に必要な同義語辞書情報, ドメイン設定情報, ドメイン定義情報のみを示している.

この情報を用いて関連情報の推定を行う. まず関連探索の対象となるカラムとして, "電話管理. 状態. 電話番号"を選択する. このカラムに設定されている同義語"電話番号"を同義語とするカラムとして, "顧客. 連絡先. 連絡先"がある. この組がまず関連情報の候補となる.

次にドメイン情報のチェックを行う. "電話管理.状

態. 電話番号"のドメインは"電話番号ドメイン"であり, ドメイングループは"電話番号ドメイングループ"である. また, "顧客.連絡先. 連絡先"のドメインは"ハイフン電話番号ドメイン"であり, もう一方のカラムのドメインとは異なるが, そのドメイングループはおなじ"電話番号ドメイングループ"である. したがって, このカラムどうしは互いに変換可能であるので, 関連候補となる. この関連情報を

関連名:顧客電話状態
結合条件:顧客. 連絡先. 連絡先
= 電話管理.状態. 電話番号

としてG-IRDに設定することで, この2つのデータベースを結合した検索が可能となる.

7.2 検索候補の提示

ユーザからの検索要求を

検索項目:顧客名
検索条件:動作状況="故障中"

とする. DBSENAは普遍関係インタフェースを用いているので, データの所在を指定する必要はない.

この検索要求に対して, 検索候補をユーザに提示する. 検索要求の各項目に対して, 同義語辞書を利用し, LDBのカラムの所在を推定する. この推定されたカラムを組みあわせ検索候補を作成し, それらカラムがテーブルをまたがっている場合はその間の結合条件を補完する. その結果できあがった検索候補をユーザに返却する.

この例では,

検索項目:顧客. 連絡先. 顧客名
検索条件:電話管理. 状態. 動作状況="故障中"
関連情報:
関連名:顧客電話状態
結合条件: 顧客. 連絡先. 連絡先
= 電話管理. 状態. 電話番号

という検索候補がユーザに返却される.

7.3 検索実施

検索実施時には, 検索候補を検索対象となるデータベース毎に分離し, それぞれのデータベースに対してデータベース間の結合条件となる項目をSELECT句に加えたSQL文を作成し, 検索を行う. したがってデータベースに対する検索文

- 顧客DB:
  select 顧客名, 連絡先 from 連絡先
- 電話管理DB:
  select 電話番号 from 状態
  where 動作状況 = "故障中"

がそれぞれ作成される.

結合を行うカラムである"顧客.連絡先. 連絡先", "電話管理. 状態. 電話番号"はドメインが異なっているので, それぞれのデータベースを検索した結果をドメイン変換する. この場合, "顧客.連絡先. 連絡先"の検索結果をグローバルドメインである"電話番号ドメイン"に変換する. すなわち"0468-59-0000"を"0468590000"へ変換する. これにより, 結合条件となっているカラムどうしのドメインが等しくなることから, DBSENAで結合処理を実施する.

したがって, ユーザは検索結果として

氏名: 星野隆

を得ることができる.

8 おわりに

以上, マルチデータベース環境において関連設定, 検索の問題点を示し, DBSENAによる解決手法を示した.

しかし2.3節で述べたように, 普遍関係インタフェースを用いて多数のデータベースに対する検索をおこなう場合, 1つの検索要求に対しても検索候補が大量になることがある. 今後はこれに対処するため, この検索候補数の絞り込み方式が必要となる.

また通常のマルチデータベース検索システムと同様に, 性能向上をはかるためのマルチデータベース間結合処理の最適化をおこなうこと考えている.

参考文献

[1] Y.Breitbart, "Multidatabase Interoperability", SIGMOD Record, 19(3), 1990.

[2] J.D.Ullman, "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.

[3] 星野 隆, 綱川 光明, 町原 宏毅, "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", 情報処理学会第 114 回データベースシステム研究会, 1998.

[4] 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.

[5] R.Elmasri, S.Navathe, "Object integration in logical database design", Int. Conf. on Data Engineering, 1984.

[6] S.B.Huffman, D.Steier, "Heuristic joins to integrate structured heterogeneous data", AAAI Sympo. on Information Gathering from Heterogeneous, Distributed Environments", 1995.

[7] R.Fagin, O.Mendelzon, J.D.Ullman, "A Simplified Universal Relation Assumption and Its Properties", ACM Trans. on Database Systems, 7(3), 1982.

[8] D.Maier, J.D.Ullman, "Maximal Objects and the Semantics of Universal Relation Databases", ACM Trans. on Database Systems, 8(1), 1983.

[9] J.L.Zhao, A.Sergev, A.Chatterjee, "A Universal Relation Approach to Federated Database Management", 11th Int. Conf. on Data Engineering, 1995.

[10] C.Yu, B.Jia, W.Sun, S.Dao, "Determining Relationships among Names in Heterogeneous Databases", SIGMOD Record, 20(4), 1991.