

マルチデータベース環境における時間的・空間的関連性評価によるデータ結合方式

細川 宜秀^{†1} 石橋 直樹^{†2}
八代 夕紀子^{†3} 清木 康^{†4}

広域ネットワークに連結された独立なデータベース群を対象とした検索・統合機構の実現によって、それらのデータベース群の利用価値は飛躍的に増大する。

従来の異種データベース間検索・統合は、データ間の等価性、類似性、包含性などの関連性の評価によって異種データベース群を連結するために、異種データベースに含まれるデータ間の関連性をパターン表現の対として静的に記述し、その記述中のパターン表現と異種データベースに含まれるパターン表現のパターンマッチングによって実現されている。異種データベース間の関連性記述においては、各データベース間において各々の関連性に関する記述が必要となり、その組み合わせ数は多く、関連性記述の生成および更新のオーバーヘッドは大きい。

また、既存のデータ検索・統合機能（関連性評価機能）群の、適用範囲、評価対象関連性、および、実現方式は互いに異なっているので、それらを直接組み合わせて利用するには、それらの間のデータ形式変換を行う必要がある。

本論文では、関連性評価機能の適用範囲を互いに補完する異種データベース間検索・統合機能の実現方式を提案する。本方式の特徴は、異種データベース間検索・統合機能を実現するために、異種データベース間の多様な関連性を包括的に評価するメタレベルシステムを実現している点にある。さらに、関連性評価機能間でのそれらに独立な共通データ表現の受け渡しによって、データベース間での直接的なデータ表現変換機能の構築を不要とする。本方式により、たとえば、時間表現、および、空間表現を対象とした検索・統合機能群の連結による、異種データベース間に含まれるデータ群からのプラン生成が可能になる。本方式を実現した実験システム上で実験を行い、本方式の有効性を明らかにする。

A Data Integration Method Realizing Evaluation for Temporal and Spatial Relationships in a Multidatabase Environment.

YOSHIHIDE HOSOKAWA,^{†1} NAOKI ISHIBASHI,^{†2} YUKIKO YASHIRO^{†3}
and YASUSHI KIYOKI^{†4}

In a wide area network environment, facilities for retrieving and integrating heterogeneous legacy databases are very important to increase the values of those databases.

Traditional methods for integrating different databases are based on combining pattern matching techniques and static knowledge information for relationships between them. In the wide area network environment, heterogeneity among a number of legacy databases makes it difficult to create the knowledge information on relationships among those databases. Furthermore, differences among domains, relationships and implementations among underlying retrieval and integration systems make it difficult to implement data integration with comprehensive evaluation of a variety of relationships such as equality, similarity and topology.

In this paper, we present a new data integration method for complementing domains of the underlying systems each other. The feature of this method is to realize a meta-level system with comprehensive evaluation mechanisms for a variety of relationships between different legacy databases. The meta-level system is independent of those databases. This method makes it possible to connect the underlying systems by passing common data structures in the meta-level. By using this method, for example, we can implement a planning application with combining the underlying systems for temporal and spatial data representations in a heterogeneous database environment. We clarify feasibility and effectiveness of the method by showing some experimental results.

1. はじめに

ネットワークの広域化に伴い、ネットワーク上でアクセス可能なデータベース数は年々増大している。これらの異種データベース群を対象としたデータ検索・統合のためのマルチデータベースシステム^{2),14),19)}を実現することにより、それらのデータベース群の利用価値は飛躍的に増大する。そのような環境では、それらのデータベースは各々独立に作成されており、それらに含まれるデータ群は、同じ内容を表すが異なる表現形式を持つ、あるいは、異なる内容を表すが同一表現形式を持つという状況が発生する。

従来の異種データベース間の検索・統合は、データ間の等価性、類似性、包含性などのデータ間に内在する性質（本論文では、これらの性質を関連性と総称する）の評価によって異種データベース群を連結するために、異種データベースに含まれるデータ間の関連性をパターン表現の対として静的に記述し、その記述中のパターン表現と異種データベースに含まれるパターン表現のパターンマッチングによって実現されている。しかし、現在、多数のデータベース群が広域ネットワークに連結されており、さらに、アクセス可能なデータベース数が増加している中で、関連性記述の生成、および、更新のオーバーヘッドは大きい。

現在、データ間に内在する多種の関連性を包括的に評価するための機構によるデータ間検索・統合は実現されておらず、既存のデータ検索・統合機能群は、適用範囲、評価対象関連性、および、操作系、対象データ構造などの実現方式は異なり、既存のデータ検索・統合機能群を直接組み合わせて利用するためには、データ検索・統合機能間のデータ表現形式変換機能を実現する必要がある。文献1), 15)の方式は、1次元ユークリッド空間上の時間データ表現を対象とした検索・統合システムとして実現され、包含性、方向性、および、連結性を評価する。文献⁴⁾の方式は、3次元ユークリッド空間上の空間データ表現を対象とした検索・統合システムとして実現され、包含性、方向性、および、連結性を評価する。単語データ表現を対象とした意味的連想検索機能では、

1単語表現は、多義性を持つという前提の上で、文脈や状況に応じて、その意味を確定し、多義性を排除した単語表現間の同義性を評価する^{11),12),23)}。文献^{11),12),23)}の対象は、約2000次元の正規直交空間上のベクトルとして定義される。ここで、データ間の同一性、類似性、包含性などの評価を行う機能を関連性評価機能と定義する。

関連性評価関数群を組み合わせて利用する方式に関しては、 n 個の関連性評価関数群を直接組み合わせる場合には、 $n(n-1)$ のデータ変換機能を実現する必要がある。広域ネットワークを介してアクセス可能なデータ表現の種類は多く、それらの間に内在する関連性の種類も多いので、広域ネットワーク環境においては、 n が大きくなり、そのデータ変換機能群の作成、および、変換処理のオーバーヘッドは無視できない。

さらに、既存データベースに含まれるデータの表現形式、および、データ操作の単位（データ粒度）の相違は、関連性評価機能群のそれと異なり、既存データベースに含まれるデータを、直接的に、関連性評価機能群に適用することはできないので、異種データベース群を対象とした検索・統合機能の実現を困難にしている。

本論文では、関連性評価機能の適用範囲を互いに補完する異種データベース間検索・統合機能の実現方式を提案する。さらに、提案方式によって、マルチデータベース環境を対象とした検索・統合機能の適用範囲の拡大を示すために、前提として、広域ネットワーク介してアクセス可能なデータ表現形式に、文字列表現、時間データ表現、および、空間データ表現を設定し、文字列表現、時間データ表現、および、空間データ表現を対象とした関連性評価機能群を提案システムへ段階的に連結することによって、問合せ結果の精度の変化を評価するための実験の結果を示す。

本論文では、次の理由により、主にデータ統合操作の実現方式について述べる。利用者によって与えられるデータと既存データベースに含まれるデータ間の関連性評価を伴う検索操作、および、異種データベースに含まれるデータ群の関連性評価を伴う統合操作は、操作対象データの配置位置が異なるだけで、同一のデータ間関連性評価方式によって実現可能である。

本方式の特徴は、関連性評価機能群、あるいは、既存データベースシステム群を本システムへ連結する場合において、本システムの変更を行うことなく、それらを連結可能とする機構を実現している点にある。その手順は、次のとおりである。(1) 既存データベースシステムの実現方式に依存しない一段抽象度の高いメタレベルに、関連性評価を統一的に行うための共有計算空間を設

University of Tsukuba

†2 慶應義塾大学大学院政策・メディア研究科
Graduate School of Media and Governance,
KEIO University

†3 お茶の水女子大学大学院理学研究科
Faculty of Science, Ochanomizu University

†4 慶應義塾大学環境情報学部
Faculty of Environmental Information,
KEIO University

定する。(2) メタレベル共有計算空間を関連性評価機能群の計算空間によって構成し、各部分共有計算空間上の評価結果群を合成する機構を実現する。関連性評価機能群の計算空間群、および、それ上の値の構造は異なり、それらの評価結果群を直接合成することは困難であるので、提案方式の各計算空間で生成された評価結果群の合成機構は、各計算空間上の評価対象となる値にメタレベルで有効な名前を割り当て、その名前を各計算空間間で受け渡しする機構として実現する。これにより、関連性評価機能群の実現方式に依存せずに、関連性評価機能群の連結を実現する。(3) 既存データベースに含まれるデータ群をメタレベル共有計算空間上の対応する値に写像する機能を実現することによって、既存データベースシステムの本システムへの統一的連結を可能にする。メタレベルにおける関連性評価機能群の実行制御を行うシステムをメタレベルシステムと定義する。さらに、本方式では、関連性評価機能間における、それらに独立な共通データ表現の受け渡しによって、関連性評価機能群を直接連結する場合に必要な関連性評価機能間のデータ表現形式変換機能を不要とする。

現在提案されている異種データベース間データ統合方式は、次の3方式に分類される。本方式は、第3方式として位置づけられる。

方式1 ローカルデータベースに含まれるデータの標準データ形式への変換による異種データベース間データ統合方式

この方式による検索・統合のための操作は、次の手順によって行われる。(1) 標準データを対象としたスキーマ設計者は、ローカルデータベースに含まれるデータの標準データ表現形式、および、標準スキーマへの変換・写像操作を定義する。(2) マルチデータベースシステムの利用者は、異種データベース間検索・統合操作を行うための問合せを標準スキーマに対する検索・統合操作として記述し、マルチデータベースシステムに発行する。(3) マルチデータベースシステムは、(2)において発行された問合せを、(1)において定義されたデータ表現形式に関する変換・写像定義にしたがって、ローカルデータベースのデータ表現形式、および、スキーマ構造を対象とした部分問合せ群に展開する。(4) (3)において生成された部分問合せ群に対応するローカルデータベースシステム群に渡し、実行する。それらの結果を、(1)において定義されたスキーマ構造に関する変換・写像定義にしたがって、標準データ形式に変換・写像する。(5) (4)において生成された部分問合せ結果群を統合する。

標準データ形式の種類に応じて、主に次のデータ統合方式に分類される。

方式1.1 構造型データ形式

この方式は、異種データベースに含まれるデータ間の関連を、スキーマ構造を用いて間接的に記述する方式である。Garlic^(6),7)、および、DISCO^(20),21)は、ODMGを拡張したオブジェクト・モデルに基づいたグローバル・スキーマ構造により、異種データベース間データ統合を実現している。

方式1.2 半構造型データ形式

この方式は、方式1.1と異なり、異種データベース間のデータ間関連を直接記述することによって、異種データベース間のデータ統合を実現する方式である。この方式では、方式1.1の方式と比べて、構造型データとして記述できないデータをデータ統合の対象とすることが可能である。TSIMMIS^(3),5)は、オブジェクト識別と入れ子の概念のみをサポートしたオブジェクトモデルOEM⁽¹⁸⁾を標準データ形式として用いているシステムである。

方式2 マルチデータベース言語による異種データベースを対象とした統合的アクセス方式

SchemaSQL⁽¹³⁾は、異種リレーショナル・データベース間のスキーマ相違性を解決するために、問合せの記述対象を、検索者の想定するリレーショナル・スキーマとし、実行時に、問合せに記述されたリレーショナル・スキーマ構造とローカルデータベースのリレーショナル・スキーマ構造の間の写像を動的に行う機構を実現した。さらに、検索者の想定するリレーショナル・スキーマ構造を対象とした問合せを記述可能にするために、従来のSQLにおいて個別に記述されたスキーマ操作、および、データ操作を総合的に記述可能にするSchemaSQL言語を開発した。この方式では、方式1と異なり、常に、検索者の想定するスキーマ構造を用いて、異種データベースに含まれるデータにアクセスすることが可能である。

方式3 関連性評価機能を有する既存システム群を利用した異種データベース間データ統合方式

この方式は、複数の異種データ構造を融合したデータ表現を標準データ形式として用いることによって、ローカルデータベースに含まれるデータに対して、標準データ形式を構成する部分データ構造毎に定義されるデータ操作系を組み合わせる利用することが可能になる。NR/SD+⁽¹⁶⁾は、リレーシヨナ

ルデータベースに含まれるデータと SGML ドキュメント・データを対象とした検索・統合のための統一的機構を実現するために、SGML 形式、および、ネステッド・リレーション形式を融合したデータ形式を標準データ形式として設定し、SGML ドキュメントを対象としたリージョン代数演算、および、ネステッド・リレーションを対象とした演算の連結を実現している。提案システムは、文字列表現、時間データ表現、および、空間データ表現を対象とした関連性評価機能を有する既存システム群を組み合わせることで利用する機能を実現する。

空間データベースシステムに関する研究分野において、OGIS¹⁷⁾が提案されている。OGIS は、異種地理情報システム間の相互運用を実現するためのシステムに関する標準仕様を提供する。OGIS によって実現されたシステムは、地理的関連性評価機能を有しており、具体的には、それらの機能群は、異種地理情報システムにおいて定義された異形式の地図上のデータを対象とした検索・統合機能として使用される。提案システムは、時間データ表現、および、空間データ表現に加え、それら以外の広域ネットワークを介してアクセス可能なデータ群を対象とした異種データベース間データ検索・統合機能を実現するために、時間的関連性評価機能、および、空間的関連性評価機能に加え、それら以外の関連性評価機能群を連結するための統一的手法を実現する。これによって、時間的、および、空間的関連性評価機能が生成する情報と他の関連性評価機能群が生成する情報を統合した新しい情報の生成を可能にする。

2. 時間的、および、地理的関連性評価機能の連結による異種データベース間データ統合例

提案方式は、関連性評価機能を有する既存システム群を組み合わせることで利用することによって、異種データベース間の関連性評価機能群の連結を伴うデータ統合を実現する方式として位置付けられる。関連性評価機能を有する既存システムを、関連性評価システムと定義する。

図 1、図 2、および、図 3 は、それぞれ、2 種類のローカルリレーショナルデータベースシステム（鉄道データベース、および、航空データベース）間の地理的関連性評価機能、時間的関連性評価機能、および、それらの連結によって、列車から飛行機への乗り継ぎプラン生成例を表している。図 1 では、2 ローカルデータベース間の地理的関連性のみを評価対象としているために、列車 T3 から飛行機 P1 への乗り継ぎ不可能なプラン（列車が空港に到着したところには、飛行機はすでにそ

の空港を飛び立っている）が生成される。図 2 も同様に、列車 T1 から飛行機 P1 への乗り継ぎ不可能なプラン（列車が到着する空港と飛行機が飛び立つ空港が異なる）が生成される。このため、図 3 に示すように、関連性評価機能群の連結を実現することは重要である。

ここで述べたプラン生成例に関して、時間的連結性評価機能、および、空間的包含性評価機能のみを使用してプランを生成するためには、プラン作成者は、プランを構成する経路に関する情報を保持していないとプランを作成できない。例えば、この例では、プラン作成者は、列車、ならびに、飛行機などの経路上の移動手段、および、移動手段間の接続地点に関する情報を保持しているものとしている。一方、経路上の移動手段、および、移動手段間の接続地点に関する情報の検索者による保持を前提としない高度な経路探索の手法は、文献 9) において述べられている。この文献では、経路探索の手法として、出発地点からの累積コストが最小となるように経路の探索範囲を同心円状に広げることによって、出発地点から目的地までの最小コスト経路を探索する Dijkstra 法、人工知能的アプローチである A* アルゴリズム、ネットワーク階層化による経路探索の高速化、探索された経路群から類似した経路群を除去するためのアルゴリズム、および、巡回経路探索問題において、厳密な解を求めるのが困難な場合に対して、ヒューリスティクスに基づいた近似解を遺伝的アルゴリズムなどによって求める方法について述べられている。提案方式は、関連性評価機能群の連結のための統一的操作を実現することを目的としているので、文献 9) において述べられた経路に関する関連性評価機能群を提案システムに連結することによって、広域ネットワークを介してアクセス可能なデータ群を対象とした経路探索を伴う検索・統合操作が可能となる。さらに、経路に関する関連性評価機能群と他の関連性評価機能群との連結によって、経路に関する関連性評価機能が生成する情報と他の関連性評価機能群が生成する情報を組み合わせた新しい情報が生成される。

3. 関連性評価機能群の連結による異種データベース間データ結合方式

マルチデータベース環境では、次の状況によって、異種データベース間のデータ群の連結は困難である。(a) 関連性評価機能群の適用範囲、および、評価対象となる関連性の種類、操作系、操作対象データ構造などの実現方式は、互いに異なる。(b) 異種データベースに含まれるデータの表現形式、および、データ粒度は、関連性評価機能群のそれと異なる。

メタレベルシステムにおいて、次の 3 機能を実現する

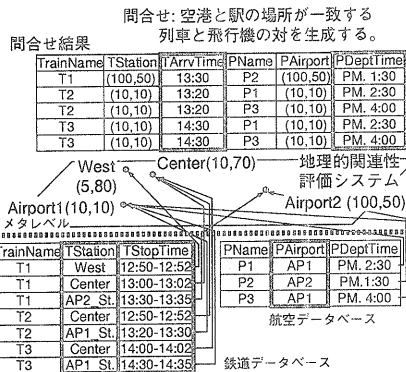


図1 地理的関連性評価システムによる異種データベース間データ結合例

Fig. 1 Data integration for evaluating geographical relationships between heterogeneous databases.



図2 時間的関連性評価システムによる異種データベース間データ結合例

Fig. 2 Data integration for evaluating temporal relationships between heterogeneous databases.



図3 地理的関連性、および、時間的関連性評価システムによる異種データベース間データ結合例

Fig. 3 Data integration for evaluating geographical and temporal relationships between heterogeneous databases.

ことが重要である。

- (1) 評価する関連性が異なる関連性評価機能群の統一的操作機能
- (2) 評価する関連性が異なる関連性評価機能群の連結機能
- (3) データ表現形式の変換、および、データ粒度の制御機能

本節では、上記3機能の実現方式について述べる。

3.1 評価する関連性が異なる関連性評価機能群を対象とした統一的操作機能の実現方式

メタレベルシステムにおいて、関連性評価機能群の実現方式に依存することなく、それらの関連性評価機能群を統一的に操作するために、関連性評価システム群に対する共通のインタフェースを設定する。メタレベルに連結された k 番目の関連性評価システム $SCR(k)$ の関連性評価機能、および、評価対象データ構造を、次のように定義する。

(E1) 計量対象空間 $CS(k, s)$

この空間は、関連性評価システム $SCR(k)$ における関連性評価を行う s 番目の空間を表し、次のデータ構造により構成される。

(E1-1) 計量対象空間構造 $SSC(k)$

この構造は、計量対象空間のデータ構造を表す。

(E1-2) 計量対象空間上の値構造 $VSC(k, sn)$

この構造は、関連性評価システム $SCR(k)$ の計量対象空間上の1つの値構造を表す。

sn は、値構造識別子を表す。

(E2) 関連性評価演算 $PF_{[k,p]}$

この関連性評価演算は、 $SCR(k)$ の計量対象空間上の値に対して適用される。 p は、演算識別子を表す。

3.2 評価する関連性が異なる関連性評価機能群の連結機能の実現方式

3.2.1 メタレベルにおけるデータ表現形式

関連性評価機能群の連結による異種データベース間データ結合機能を実現するために、関連性評価機能の計算空間をメタレベル共有計算空間の部分空間として定義し、各部分共有計算空間における評価結果を合成する枠組みを実現する。メタレベルシステムにおいて、次の2項目の実現が重要である。

- (1) 関連性評価機能群の共通の入出力データ構造 (標準データ構造)
- (2) 関連性評価機能における標準データの統一の評価方式

本方式では、関連性評価機能が処理可能な構造を有する

部分共有計算空間上の値に名前を割り当て、名前と値の対表現形式を、メタレベル共有計算空間上の標準データ形式として定義し、関連性評価機能間において、名前を受け渡す機構を実現する。これによって、実現方式が異なる関連性評価機能群における値の統一的評価機構を実現する。さらに、名前は、関連性評価機能群への共通入力データ形式であり、各機能間のデータ受け渡しを統一手法によって実現するために、関連性評価機能によって出力される値間の関連を、値に割り当てられた名前の対によって表現する。

メタレベルにおける値表現、および、2 値間の関連表現は、次のとおりである。

(D1) メタレベルにおける値表現形式

メタレベルにおける値表現を、計量対象空間上の値と対応する名前の対（名前付き値）によって記述する。

$$[name, v_{[k,s,sn]}]$$

$v_{[k,s,sn]}$ は、計量対象空間 $CS(k, s)$ 上の $VSC(k, sn)$ という構造を有する値を表す。

(D2) 2 値間の関連表現形式

関連性評価機能によって出力される 2 値間の関連を、名前の対によって表現する。次の対表現は、名前 $name_1$ を持つ値が名前 $name_2$ を持つ値と関連があることを表す。

$$(name_1, name_2)$$

3.2.2 メタレベルにおけるデータ操作関数群

メタレベルのデータ表現形式は、値表現、および、対表現によって定義されるので、メタレベルにおけるデータ操作関数として、値表現を対象としたデータ操作関数、対表現を対象としたデータ操作機能、および、値表現、ならびに、対表現を対象としたデータ操作関数を定義する。

$$(MP1) \text{joiner}(V_1^{in}, V_2^{in}, \{PF_{[k,p]}, CS(k, s)\}) \rightarrow P^{out}$$

この関数は、関連性評価関数 $PF_{[k,p]}$ を用いて、同一計量対象空間上の値から構成される 2 つの値集合 V_1^{in}, V_2^{in} それぞれに属する 2 値間の関連性を評価し、関連性が存在すると判断された 2 値の対から構成される対集合 P^{out} を生成する。対集合 P^{out} を、次のように定義する。

$$P^{out} = \{(n_1, n_2) | (\exists v_1, v_2) (([n_1, v_1] \in V_1^{in} \wedge [n_2, v_2] \in V_2^{in}) \wedge (V_1^{in}, V_2^{in} \subseteq CS(k, s)) \wedge (PF_{[k,p]}(v_1, v_2) = true))\}$$

$$(MP2) \text{union}_V(V_1^{in}, V_2^{in}) \rightarrow V^{out}$$

この関数は、2 値集合の和集合を生成する。値集合 V^{out} を、次のように定義する。

$$V^{out} = \{[n, v] | (([n, v] \in V_1^{in}) \vee ([n, v] \in V_2^{in}))\}$$

$$(MP3) \text{difference}_V(V_1^{in}, V_2^{in}) \rightarrow V^{out}$$

この関数は、2 値集合の差集合を生成する。値集合 V^{out} を、次のように定義する。

$$V^{out} = \{[n, v] | (([n, v] \in V_1^{in}) \wedge (\neg([n, v] \in V_2^{in})))\}$$

$$(MP4) \text{reconstructor}(P_1^{in}, P_2^{in}) \rightarrow P^{out}$$

この関数は、対集合 P_1^{in} に属する対の第 2 要素と対集合 P_2^{in} に属する対の第 1 要素が一致するとき、 P_1^{in} の対の第 1 要素と P_2^{in} の対の第 2 要素によって構成される対を生成する。対集合 P^{out} を、次のように定義する。

$$P^{out} = \{(n_1, n_2) | (\exists n') ((n_1, n') \in P_1^{in} \wedge ((n', n_2) \in P_2^{in}))\}$$

$$(MP5) \text{selector}_{PVV_1}(P^{in}, V^{in}) \rightarrow V^{out}$$

この関数は、対集合 P^{in} に属する対の第 1 要素であり、値集合 V^{in} に属する値を抽出する。値集合 V^{out} を、次のように定義する。

$$V^{out} = \{[n, v] | (\exists n') ((n, n') \in P^{in}) \wedge ([n, v] \in V^{in})\}$$

$$(MP6) \text{selector}_{PVV_2}(P^{in}, V^{in}) \rightarrow V^{out}$$

この関数は、対集合 P^{in} に属する対の第 2 要素であり、値集合 V^{in} に属する値を抽出する。値集合 V^{out} を、次のように定義する。

$$V^{out} = \{[n, v] | (\exists n') ((n', n) \in P^{in}) \wedge ([n, v] \in V^{in})\}$$

$$(MP7) \text{selector}_{PVP_1}(P^{in}, V^{in}) \rightarrow P^{out}$$

この関数は、値集合 V^{in} に属する値を第 1 要素に持つ対を対集合 P^{in} から抽出する。対集合 P^{out} を、次のように定義する。

$$P^{out} = \{(n_1, n_2) | (\exists v) ((n_1, n_2) \in P^{in} \wedge ([n_1, v] \in V^{in}))\}$$

$$(MP8) \text{selector}_{PVP_2}(P^{in}, V^{in}) \rightarrow P^{out}$$

この関数は、値集合 V^{in} に属する値を第 2 要素に持つ対を対集合 P^{in} から抽出する。対集合 P^{out} を、次のように定義する。

$$P^{out} = \{(n_1, n_2) | (\exists v) ((n_1, n_2) \in P^{in} \wedge ([n_2, v] \in V^{in}))\}$$

$$(MP9) \text{union}_P(P_1^{in}, P_2^{in}) \rightarrow P^{out}$$

この関数は、2 対集合の和集合を生成する。対集合 P^{out} を、次のように定義する。

$$P^{out} = \{(n_1, n_2) | ((n_1, n_2) \in P_1^{in}) \vee ((n_1, n_2) \in P_2^{in})\}$$

$$(MP10) \text{difference}_P(P_1^{in}, P_2^{in}) \rightarrow P^{out}$$

この関数は、2 対集合の差集合を生成する。対集

合 P^{out} を、次のように定義する。

$$P^{out} = \{(n_1, n_2) | ((n_1, n_2) \in P_1^{in}) \wedge (\neg((n_1, n_2) \in P_2^{in}))\}$$

3.3 メタレベルにおけるデータ粒度制御方式

本方式では、メタレベルのデータ表現形式は、名前と値によって構成されるので、メタレベルにおけるデータ粒度制御機能については、次の2機能を組み合わせることによって実現する。

- (C1) ローカルデータベースに含まれるデータを、メタレベルの計量対象空間上の値に変換する機能
 (C2) (C1)によって生成された計量対象空間上の値に名前を割り当てる機能

3.2.2節で述べたデータ操作機能群の操作対象は、値、および、名前であり、それらは、値の生成方式、および、名前の割り当て方式と独立であるので、データ粒度制御機能を、それらと独立に実現することが可能である。これより、ローカルデータベースシステムの連結の度に、新しいデータ粒度制御機能が本システムに組み込まれても、3.2.2節のデータ操作機能群の変更を行う必要はない。

図4、および、図6は、2節において述べた異種データベース間データ統合問合せ例に対する本方式による2実行例を示している。図4は、1つの名前が、3値に割り当てられる場合の問合せ実行例である(図5)。図6は、1つの名前が、1値に割り当てられる場合の問合せ実行例である(図7)。すなわち、これは、データ粒度が最小に設定された場合の例である。対集合 Pt , Pp , および、値集合 Vtn , Vtg , Vtt は、データ粒度制御後のメタレベルにおけるデータ表現を表している。また、Allenによって提案された13時区間を対象とした時間的関連性¹⁾を評価する機能、および、Egenhoferによって提案された2次元ユークリッド空間上のデータ間の幾何学的関連性⁴⁾を評価する機能を連結しているものとする。図4、および、図6の下線は、その実行結果を表す。メタレベルにおけるデータの粒度の相違性に関わらず、同じ結果が得られていることが確認される。

4. メタレベルシステム構成方式

本節では、メタレベルシステムの構成方式、および、問合せ処理方式について説明する。メタレベルシステムの構成方式、および、問合せ処理方式の詳細については、文献8)において述べた。

4.1 システム構成

図8に示すように、異種データベース環境を対象とした検索・統合システムは、次の3システムによって実現される。

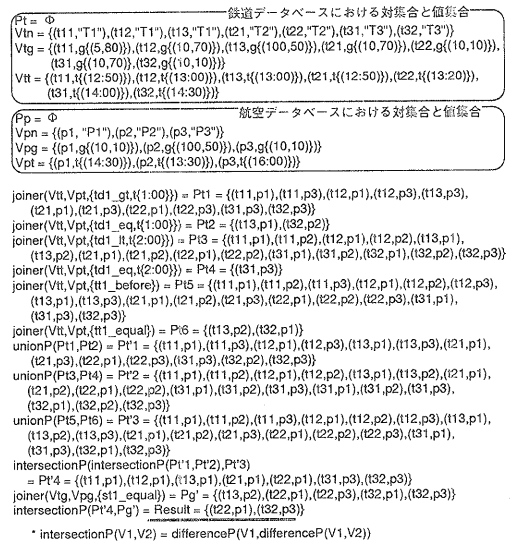


図4 問合せ実行例1

Fig. 4 Query execution (Example 1).

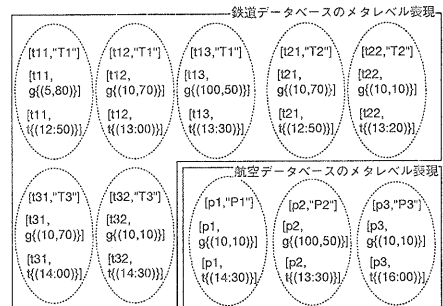


図5 問合せ実行例1におけるローカルデータベースに含まれるデータのメタレベル表現

Fig. 5 Meta-level representations of local databases for Example 1.

- メタレベルシステム
- 関数インタプリタ
- ローカルデータベースシステム

メタレベルシステムは、メタレベル問合せ処理系、関連性評価システム、および、その関数インタプリタによって構成される。メタレベル問合せ処理系は、ユーザからメタレベルシステムに与えられた問合せを解釈・実行するシステムである。関連性評価システムは、データ間の関連性を評価するシステムである。関連性評価システムの関数インタプリタは、メタレベルシステムへの問合せに記述された関連性評価機能、および、操作対象に付けられた名前により、その名前に対応する値構造を関連性評価関数に適用するための制御を行うシステムである。

関数インタプリタは、メタレベルに設定されたロー

表現を対象とした関連性評価システム $SCR(1)$, 時間データ表現を対象とした関連性評価システム $SCR(2)$, および, 地理データ表現を対象とした関連性評価システム $SCR(3)$ を段階的に連結することによって, 問合せ結果の精度の変化を定量的に評価する。

さらに, 異種データベース間検索・統合機能の適用範囲に関して, ローカルデータベースに含まれるデータの標準データ形式への変換, および, 標準データ形式を対象としたパターンマッチングの組み合わせによる従来の異種データベース間データ検索・統合方式 (従来方式) に対する提案方式の有効性を示すために, 従来方式と提案方式における問合せ結果の精度を評価する。従来方式では, ローカルデータベースに含まれるデータに割り当てる標準データ形式のパターンを変化させることによって, 標準データ形式を対象としたパターンマッチングによる検索・統合の精度は変化する。たとえば, ローカルデータベース 1 に含まれる文字列表現“A 市 B 町”とローカルデータベース 2 に含まれる文字列表現“A 市 C 町”に対して, データ変換時に, “A 市 B 町” から“A 市”と“B 町”を, “A 市 C 町” から“A 市”と“C 町”を生成し, 市名に関するパターンマッチングによって, “B 町”と“C 町”が“A 市”に属することを計量することが可能である。そこで, 従来方式において, 次の 2 データ変換機能を実装した。

DC1 ローカルデータベースに含まれるデータ項目の

分割, あるいは, 合成を伴わないデータ変換操作

DC2 アドレス情報に含まれるエリア情報抽出のよう

なローカルデータベースに含まれるデータ項目の分割, あるいは, 合成を伴うデータ変換操作

表 1 は, 本実験で使用した関連性評価機能を表す。ここで, メタレベルにおける時間データ表現, および, 地理データ表現を, それぞれ, 1 次元ユークリッド空間上のベクトル, および, 2 次元ユークリッド空間上のベクトルとして定義する。メタレベルにおける関連性評価システムの構成に応じて, 次の 4 メタレベルシステムを構築した。

Level 1 $SCR(1)$ を連結したメタレベルシステム

Level 2-1 $SCR(1)$, および, $SCR(2)$ を連結したメタレベルシステム

Level 2-2 $SCR(1)$, および, $SCR(3)$ を連結したメタレベルシステム

Level 3 $SCR(1)$, $SCR(2)$, および, $SCR(3)$ を連結したメタレベルシステム

Level1 は, ローカルデータベースに含まれるデータのメタレベルの標準文字列表現への割り当て, および, 標準文字列表現を対象としたパターンマッチングを組み合

表 3 評価対象データ統合方式

Table 3 Data integration methods for the experiments.

データ統合方式
(Level1, DC1)
(Level1, DC2)
(Level2-1, DC1)
(Level2-1, DC2)
(Level2-2, DC1)
(Level2-2, DC2)
(Level3)

わせによる異種データベース間検索・統合操作を実現する従来のマルチデータベースシステムに対応する。

本実験において実装されたメタレベルシステムには, 2 ローカルデータベースシステム (美術館情報を有するリレーショナルデータベース $ldbs1$, 鉄道情報を有するリレーショナルデータベースシステム $ldbs2$) が連結されているものとする (表 2)。美術館データベース, および, 鉄道データベースには, それぞれ, 100 タプル, および, 6210 タプルのデータが格納されている。2 ローカルデータベースシステムは, PostgreSQL6.0²⁴⁾ によって構築した。

本実験では, 関連性評価システムの構成, および, データ変換機能の構成に応じた 7 データ統合方式の結合精度を評価する (表 3)。ここで, Level2-1 は, 空間的関連性評価システム $SCR(3)$ を連結していないので, ローカルデータベースに含まれる空間データ表現から標準文字列表現への変換機能, および, 標準文字列表現を対象としたパターンマッチングによる検索・統合機能を組み合わせることによって, 空間的関連性に関する異種データベース間検索・統合操作を実現した。同様に, Level2-2 は, 時間的関連性評価システム $SCR(2)$ を連結していないので, ローカルデータベースに含まれる時間データ表現から標準文字列表現への変換機能, および, 標準文字列表現を対象としたパターンマッチングによる検索・統合機能を組み合わせることによって, 時間的関連性に関する異種データベース間検索・統合操作を実現した。

本実験では, 表 3 の 7 データ統合方式による結合結果を次の指標によって評価した¹⁰⁾。

$$\begin{aligned} \text{適合率} &= \frac{\text{システムの検索結果に含まれる正解数}}{\text{システムの検索結果数}} \\ \text{再現率} &= \frac{\text{正解に含まれるシステムの検索結果数}}{\text{正解数}} \end{aligned}$$

本実験では, (Level3) の結果を正解として使用した。

本実験で使用した問合せは, 鉄道を利用して美術館を訪問するスケジュールを生成する次の 5 問合せである。

GQ1 美術館データベースと鉄道データベースを時間的包含性, および, 地理的包含性による結合問合せ

表1 実験で使った関連性評価機能

Table 1 Functions of *SCR*(1), *SCR*(2) and *SCR*(3)

関連性評価システム	関連性評価機能	真を返す条件
<i>SCR</i> (1)	<i>sp_equal</i> (<i>s</i> , <i>s1</i> , <i>s2</i>)	<i>CS</i> (1, <i>s</i>) において, 文字列表現 <i>s1</i> , <i>s2</i> は, 一致する.
<i>SCR</i> (2)	<i>ttl_before</i> (<i>s</i> , <i>tp1</i> , <i>tp2</i>)	<i>CS</i> (2, <i>s</i>) において, 時点 <i>tp1</i> は, 時点 <i>tp2</i> より過去にある.
<i>SCR</i> (2)	<i>ttl_contain</i> (<i>s</i> , <i>ti1</i> , <i>tp2</i>)	<i>CS</i> (2, <i>s</i>) において, 時点 <i>tp2</i> は, 時区間 <i>ti1</i> に含まれる.
<i>SCR</i> (2)	<i>td1_lt</i> (<i>s</i> , <i>tp1</i> , <i>tp2</i> , <i>td</i>)	<i>CS</i> (2, <i>s</i>) において, 2 時点 <i>tp1</i> , <i>tp2</i> は, 期間 <i>td</i> 以内にある.
<i>SCR</i> (3)	<i>gdcl_north</i> (<i>s</i> , <i>gp1</i> , <i>gp2</i>)	<i>CS</i> (3, <i>s</i>) において, 地点 <i>gp2</i> は, 地点 <i>gp1</i> の北に位置する.
<i>SCR</i> (3)	<i>gdt1_lt</i> (<i>s</i> , <i>gp1</i> , <i>gp2</i> , <i>gd</i>)	<i>CS</i> (3, <i>s</i>) において, 2 地点 <i>gp1</i> , <i>gp2</i> 間は, 距離 <i>gd</i> 以内にある.

表2 実験ローカルデータベース

Table 2 Local databases for the experiments.

リレーション museum(l dbs1)			リレーション train(l dbs2)				
name	time	address	name	direction	station	arrival	departure
M000	09:30-19:30	lm-pr	T1	UP	S1	05:22	05:25
M001	11:30-19:30	ch-jm	T1	UP	S2	05:32	05:35
M002	10:00-17:00	gb-mo	T1	UP	S3	05:42	05:45
:	:	:	:	:	:	:	:
M097	09:00-19:30	ce-ph	T4	DN	S14	23:43	23:46
M098	10:00-16:30	zs-ji	T4	DN	S24	23:53	23:56
M099	12:30-16:30	oh-il	T4	DN	S25	24:03	24:06

この問合せは, 次の条件を満たす列車と美術館の対集合を生成する.

- (1) 美術館は, 駅を中心とした半径 2000m の円形領域に含まれる.
- (2) 美術館開館 20 分前に, (1) を満たす駅に到着する列車に乗車する.

GQ2 GQ1 に比べて, 2 ローカルデータベース間の時間的連結時点, および, 地理的連結地点の候補を縮小した問合せ

この問合せは, 次の条件を満たす列車と美術館の対集合を生成する.

- (1) 美術館は, 駅を中心とした半径 500m の円形領域に含まれる.
- (2) 美術館開館 5 分前に, (1) を満たす駅に到着する列車に乗車する.

GQ3 GQ1 に比べて, 2 ローカルデータベース間の時間的連結時点, および, 地理的連結地点の候補を拡大した問合せ

この問合せは, 次の条件を満たす列車と美術館の対集合を生成する.

- (1) 美術館は, 駅を中心とした半径 6000m の円形領域に含まれる.
- (2) 美術館開館 60 分前に, (1) を満たす駅に到着する列車に乗車する.

GQ4 GQ1 に比べて, 時間的連結時点の候補を拡大し, 地理的連結時点の候補を縮小した問合せ

この問合せは, 次の条件を満たす列車と美術館の対集合を生成する.

- (1) 美術館は, 駅を中心とした半径 1000m の円形領域に含まれる.
- (2) 美術館開館 30 分前に, (1) を満たす駅に到着する列車に乗車する.

GQ5 GQ1 と異なり地理的方向性評価を含む問合せ

この問合せは, 次の条件を満たす列車と美術館の対集合を生成する.

- (1) 美術館は, 駅を中心とした半径 5000m の円形領域に含まれる.
- (2) 駅は, 美術館の北にある.
- (3) 美術館開館中に, (1), および, (2) を同時に満たす駅に到着する列車に乗車する.

図 9, 図 10, および, 図 11 は, それぞれ, (Level1, DC2), (Level2-1, DC1), および, (Level3) による問合せ GQ1 に対応する問合せ記述を表す.

5.1.2 関連性評価システム群の連結による異種データベース間データ検索・統合処理時間に関する評価方式

操作対象表現形式の異なる関連性評価機能群を連結するための方法は, それらの間のデータ表現形式変換操作の実現方式の相違によって, 次の 2 方式に分類される.

(M-1) メタレベルに標準データ形式を設定し, それと関連性評価機能の処理可能なデータ形式間のデータ変換操作によって, 評価関数群の間接的な連結を実現する方式 (間接的連結方式)

この連結方式は, 関連性評価機能の実行時にデータ変換操作を行うか否かによって, 次の 2 方式に分類

```

data_conversion("museum",{NameToName,"name",CS(1,1),V1n},
{IntervalToStartHour,"time",CS(1,2),V1sh},
{IntervalToStartMinute,"time",CS(1,2),V1sm},
{IntervalToEndHour,"time",CS(1,2),V1eh},
{IntervalToEndMinute,"time",CS(1,2),V1em},
{AddressToArea,"address",CS(1,3),V1ga},
{AddressToPoint,"address",CS(1,3),V1gp})->Pm
data_conversion("train",{NameToName,"name",CS(1,1),V2tn},
{DirectionToDirection,"direction",CS(1,1),V2td},
{StationToArea,"station",CS(1,3),V2ga},
{StationToPoint,"station",CS(1,3),V2gp},
{TimeToHour,"arrival",CS(1,2),V2ah},
{TimeToMinute,"arrival",CS(1,2),V2am},
{TimeToHour,"departure",CS(1,2),V2dh},
{TimeToMinute,"departure",CS(1,2),V2dm})->Pt
joiner(V2ga,V1ga,{sp_equal,CS(1,3),V2ga,V1ga})->mPg
joiner(V2sh,V1ah,{sp_equal,CS(1,2),V2sh,V1ah})->mPt
intersectionP(mPt,mPg)

```

図 9 (Level1,DC2) による問合せ GQ1 記述

Fig. 9 The query description for GQ1 by using (Level1,DC2).

```

data_conversion("museum",{NameToName,"name",CS(1,1),V1n},
{IntervalToStartPoint,"time",CS(2,1),V1sp},
{IntervalToEndPoint,"time",CS(2,1),V1ep},
{AddressToAddress,"address",CS(1,2),V1gp})->Pm
data_conversion("train",{NameToName,"name",CS(1,1),V2tn},
{DirectionToDirection,"direction",CS(1,1),V2td},
{StationToAddress,"station",CS(1,2),V2gp},
{TimeToPoint,"arrival",CS(2,1),V2ap},
{TimeToPoint,"departure",CS(2,1),V2dp})->Pt
joiner(V2ap,V1sp,{td1_it,CS(2,1),V2ap,V1sp,t{00:20}})->mP20
joiner(V2ap,V1sp,{ti1_before,CS(2,1),V2ap,V1sp})->mPb
intersectionP(mP20,mPb)->mPt
joiner(V2ga,V1ga,{sp1_equal,CS(1,2),V2ga,V1ga})->mPg
intersectionP(mPt,mPg)

```

図 10 (Level2-1,DC1) による問合せ GQ1 記述

Fig. 10 The query description for GQ1 by using (Level2-1,DC1).

```

data_conversion("museum",{NameToName,"name",CS(1,1),V1n},
{IntervalToStartPoint,"time",CS(2,1),V1sp},
{IntervalToEndPoint,"time",CS(2,1),V1ep},
{AddressToGPoint,"address",CS(3,1),V1gp})->Pm
data_conversion("train",{NameToName,"name",CS(1,1),V2tn},
{DirectionToDirection,"direction",CS(1,1),V2td},
{StationToGPoint,"station",CS(3,1),V2gp},
{TimeToPoint,"arrival",CS(2,1),V2ap},
{TimeToPoint,"departure",CS(2,1),V2dp})->Pt
joiner(V2ap,V1sp,{td1_it,CS(2,1),V2ap,V1sp,t{00:20}})->mP20
joiner(V2ap,V1sp,{ti1_before,CS(2,1),V2ap,V1sp})->mPb
intersectionP(mP20,mPb)->mPt
joiner(V2gp,V1gp,{gd1_it,CS(3,1),V2gp,V1gp,g{20}})->mPg
intersectionP(mPt,mPg)

```

図 11 (Level3) による問合せ GQ1 記述

Fig. 11 The query description for GQ1 by using (Level3).

される。

(M-1.1) あらかじめ、関連性評価機能群において処理可能な対象データ群にメタレベルの名前を割り当て、関連性評価機能群の表現形式に依存しないデータ受け渡し機構を実現する方式

(M-1.2) 関連性評価機能の実行時に、標準データ形式とその機能群のデータ表現形式間の変換操作を行う方式

(M-2) 関連性評価機能群のデータ表現形式の直接的な変換機構を実現することによって、それらの関

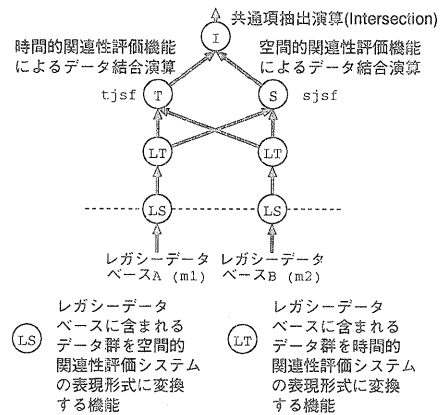


図 12 提案方式による関連性評価機能群の連結によって実現された問合せ木

Fig. 12 The query tree by the proposed method.

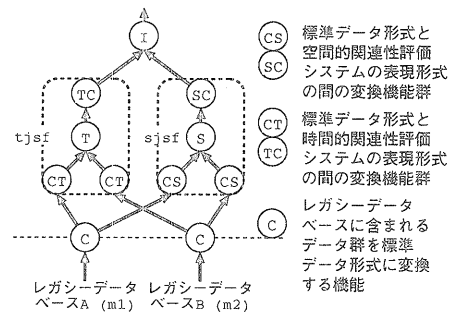


図 13 方式 (M-1.2) による関連性評価機能群の連結によって実現された問合せ木

Fig. 13 The query tree by the method (M-1.2).

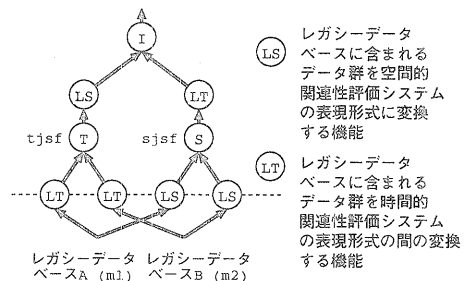


図 14 方式 (M-2) による関連性評価機能群の連結によって実現された問合せ木

Fig. 14 The query tree by the method (M-2).

連性評価機能群の連結を実現する方式 (直接的連結方式)

提案方式は、方式 (M-1.1) に対応する方式として位置づけられる。

図 12, 図 13, および, 図 14 は, 2 レガシーデータベースに含まれるデータ群を時間的関連性評価機能群, および, 空間的関連性評価機能群を用いて連結するため

の問合せ木を示している。これらの木構造は、マルチデータベースシステムにおける異種データベース間データ統合操作において共通に見られる基本構造を表す。そこで、本実験では、単一基本構造を有する問合せに対する3方式の実行時間を比較する。ここで、図14において、結合演算適用前に、レガシーデータベースに含まれるデータ群を時間的、および、空間的関連性評価システムのデータ表現形式に変換することによって、方式(M-1.1)の問合せ木を対象とした処理が可能であると考えられる。しかし、関連性評価システムが他の関連性評価システムのデータ表現形式を定義、および、処理する機構を備えておらず、さらに、方式(M-2)では、異なる関連性評価システム間でデータ群を共有するための機構が実現されていないので、方式2は、方式(M-1.1)の問合せ木を操作対象とすることは困難であり、異なる関連性評価システム間で操作対象データ群を受け渡す際に、それらのシステムが処理可能なデータ表現形式に変換する操作が必要になる。

本実験において比較する3連結方式の問合せ基本構造におけるデータ変換操作に要する計算量は、式によって表現することが可能である。ここで対象とする問題を単純にするために、各方式のデータ変換操作において、レガシーデータベースに含まれる1データは、各関連性評価システムの計量対象空間上の1つの値に写像されるものとし、レガシーデータベースに含まれる複数のデータ群を合成し、それに対応する1つの値に写像するというような複雑なデータ変換操作は行わないものとする。統合対象レガシーデータベースに含まれるデータ数をそれぞれ $m1, m2$ とする。空間的関連性評価機能によるデータ結合演算の結合率を $sjsf$ とする。時間的関連性評価機能によるデータ結合演算の結合率を $tjsf$ とする。提案方式における計量対象空間上の値に対する名前割り当て機能の計算量を $C(\rightarrow name)$ 、方式(M-1.2)、および、方式(M-2)におけるレガシーデータベースに含まれるデータをメタレベルのスキーマに写像するのに要する計算量を $C(\rightarrow schema)$ として表現する。次の式は、基本問合せ構造を対象とした3方式におけるデータ変換操作に要する計算量を表す。

(M-1.1)

$$\{C(LS) + C(LT) + 2 \cdot C(\rightarrow name)\} \cdot (m1 + m2)$$

$C(LS)$ 、および、 $C(LT)$ は、それぞれ、レガシーデータベースに含まれる1つのデータを空間的関連性評価システムの表現形式、および、時間的関連性評価システムの表現形式に変換するのに

要する計算量を表す。 $C(\rightarrow name)$ は、1つの値に対する名前割り当て操作の計算量を表す。

(M-1.2)

$$(m1 + m2) \cdot \{C(C) + C(CS) + C(CT) + 2 \cdot C(\rightarrow schema)\} + m1 \cdot m2 \cdot (tjsf \cdot C(TC) + sjsf \cdot C(SC))$$

$C(C)$ は、レガシーデータベースに含まれる1データをメタレベルの標準データ形式に変換するのに要する計算量を表す。 $C(CS)$ は、メタレベルの標準データ形式を空間的関連性評価システムの表現形式に変換するのに要する計算量を表す。 $C(CT)$ は、メタレベルの標準データ形式を時間的関連性評価システムの表現形式に変換するのに要する計算量を表す。 $C(SC)$ は、空間的関連性評価システムの計量対象空間上の1データをメタレベルの標準データ形式に変換するのに要する計算量を表す。 $C(TC)$ は、時間的関連性評価システムの計量対象空間上の1データをメタレベルの標準データ形式に変換するのに要する計算量を表す。

(M-2)

$$\begin{aligned} & \{C(LS) + C(LT) \\ & + 2 \cdot C(\rightarrow schema)\} \cdot (m1 + m2) \\ & + \min\{m1 \cdot m2 \cdot \{C(LS) \cdot tjsf \\ & + C(LT) \cdot sjsf\}, \\ & + m1 \cdot m2 \cdot \{C(SC) \cdot tjsf + \\ & C(TC) \cdot sjsf\}\} \end{aligned}$$

$C(LS)$ は、レガシーデータベースに含まれる1データを空間的関連性評価システムの計量対象空間上の値に変換・写像するのに要する計算量を表す。 $C(LT)$ は、レガシーデータベースに含まれる1データを時間的関連性評価システムの計量対象空間上の値に変換・写像するのに要する計算量を表す。 $C(SC)$ は、空間的関連性評価システムの計量対象空間上の値を標準データ形式に変換するのに要する計算量を表す。 $C(TC)$ は、空間的関連性評価システムの計量対象空間上の値を標準データ形式に変換するのに要する計算量を表す。この方式では、時間的、および、空間的データ結合演算の結果を連結するために、結合結果のドメインをそろえるためのデータ変換操作が実行される。

各式によって表現された計算量の比較において、提案方式における名前割り当て機能は、方式(M-1.2)、お

および、方式 (M-2) のレガシーデータベースに含まれるデータ群をメタレベルのスキーマに写像する操作に対応する。本実験では、3 方式において使用されるレガシーデータベースのスキーマ構造は同一なので、スキーマ構造に関する変換操作のオーバーヘッドは同一であると見なすことが可能である。さらに、提案方式におけるデータ変換操作に要する計算量を表す式は、他の 2 方式のそれらにも現れる。それゆえ、提案方式におけるデータ変換操作に要する計算量は、他の 2 方式に比べて少ない。

これらの式より、3 方式の計算量は、レガシーデータベースに含まれるデータ数、および、メタレベルにおける結合演算の結合率に依存するので、それらを変化させた場合の 3 方式におけるデータ変換に要する計算量を測定する。

本実験では、方式 (M-1.1)、方式 (M-1.2)、および、方式 (M-2) によるメタレベルシステムにおけるデータ検索・統合機構を UniSQL/X²²⁾ を用いて実装した。メタレベルシステムにおける結合演算は、ネステッド・ループ方式によって実装した。時間的、および、空間的関連性評価機能群を、UniSQL/X のメソッドとして連結し、JS5/170、SunOS 4.1.4 上において実行させた。

さらに、2 レガシーデータベースに含まれるデータ数を 100、200、および、300 と変化させた。このデータ数は、一般的なマルチデータベース環境と異なり小さいが、本実験の目的は、3 方式の比較にあるので、それに十分なものを設定した。時間的、および、空間的関連性評価機能群によるデータ結合演算の結合率を 0.01、および、0.1 と変化させた。

5.2 実験結果と考察

5.2.1 関連性評価システム群の連結による異種データベース間データ検索・統合操作の精度に関する実験結果と考察

図 15、図 16、図 17、図 18、および、図 19 は、それぞれ、問合せ GQ1、GQ2、GQ3、GQ4、および、GQ5 の問合せ結果の適合率、ならびに、再現率を表している。これらの実験結果より、ローカルデータベースに含まれるデータの標準文字列表現へのデータ変換方式を DC1、あるいは、DC2 に固定した場合、Level2-1、および、Level2-2 は、従来のパターンマッチングによる異種データベース間検索・統合機能を実現するマルチデータベースシステム Level1 に比べて、問合せ結果に含まれる正解数が多く、その問合せ結果に含まれるノイズが小さいことが確認された。すなわち、Level2-1、および、Level2-2 は、Level1 に比べて、時間的関連性、および、空間的関連性に関する高精度の異種データベ

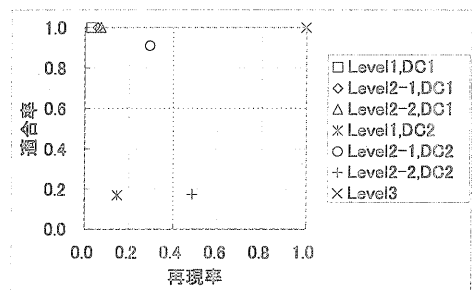


図 15 問合せ GQ1 の適合率と再現率
Fig. 15 Precisions and Recalls for GQ1.

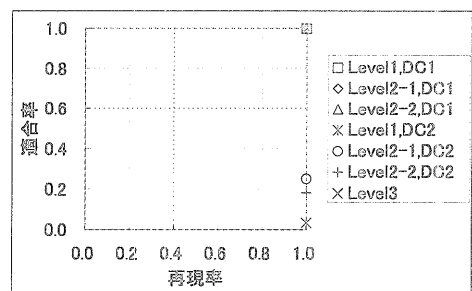


図 16 問合せ GQ2 の適合率と再現率
Fig. 16 Precisions and Recalls for GQ2.

ス間検索・統合機能を実現し、かつ、Level1 に連結されているすべての関連性評価機能群を有しているので、Level2-1、および、Level2-2 の異種データベース間検索・統合機能の適用範囲は、Level1 のそれに比べて大きいことが確認された。同様に、Level2-1、Level2-2、および、Level3 の比較において、Level3 の異種データベース間検索・統合機能の適用範囲は、Level2-1、および、Level2-2 のそれらに比べて大きいことが確認された。これらより、提案方式は、関連性評価システムのメタレベルシステムへの連結操作が統一的手法を実現することによって、関連性評価システムのマルチデータベースシステムへの連結の度にそのシステム変更を行うことなく、マルチデータベースシステムの異種データベース間検索・統合操作の適用範囲を拡大することを可能とした。

以上により、提案方式の有効性を明らかにすることができた。

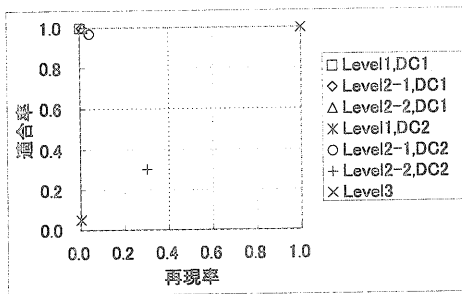


図 17 問合せ GQ3 の適合率と再現率
Fig. 17 Precisions and Recalls for GQ3.

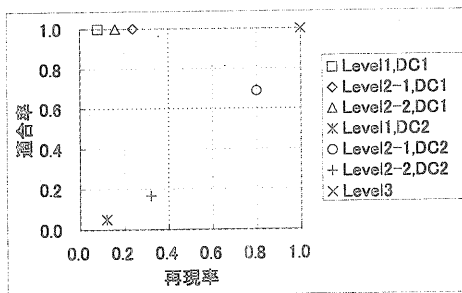


図 18 問合せ GQ4 の適合率と再現率
Fig. 18 Precisions and Recalls for GQ4.

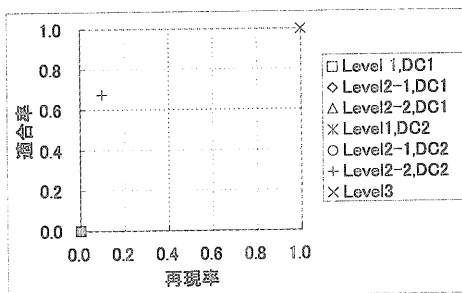


図 19 問合せ GQ5 の適合率と再現率
Fig. 19 Precisions and Recalls for GQ5.

5.2.2 関連性評価システム群の連結による異種データベース間データ検索・統合処理時間に関する実験結果と考察

図 20 は、2 レガシーデータベースに含まれるデータ数が 100 の場合における 3 方式のデータ統合問合せの実行時間を表す。図 21 は、2 レガシーデータベースに含まれるデータ数が 200 の場合における 3 方式のデー

タ統合問合せの実行時間を表す。図 22 は、2 レガシーデータベースに含まれるデータ数が 300 の場合における 3 方式のデータ統合問合せの実行時間を表す。表 4 は、3 方式におけるデータ変換操作に要した時間を表す。これらの実験結果より、提案方式は、他の 2 方式より高速に問合せを処理することが確認された。これは、方式 (M-1.1) (提案方式) は、方式 (M-1.2)、および、方式 (M-2) と異なり、問合せ中の結合演算の出力結果を対象としたデータ変換操作を伴わずに異種データベース間データ統合を実現するためである。

また、本実験では、小さいデータ数を持つレガシーデータベースを対象としたデータ統合操作に要する時間を測定したが、2 つの結合対象データベースのデータ数を 1000 とした場合、方式 (M-1.1) では、計算量の式によって、データ数 100 の場合の 10 倍になると予測される。方式 (M-1.2)、および、方式 (M-2) では、時間的、および、空間的結合演算の結合率が 0.01 のとき、計算量の式によって、データ数 100 の場合の約 10 倍になり、それらの結合率が 0.1 のとき、データ数 100 の場合の約 100 倍になると予測される。それゆえ、本実験に使用したデータ数よりも多くのデータ数を有する一般のレガシーデータベースを対象とした異種データベース間データ統合処理において、方式 (M-1.1) におけるデータ変換操作のための所要時間は、レガシーデータベースに含まれるデータ数に比例して増加すると予測され、方式 (M-1.2)、および、方式 (M-2) におけるデータ変換操作のための所要時間は、レガシーデータベースに含まれるデータ数と問合せに出現する結合演算の結合率に関連して増加すると予測される。さらに、ネステッド・ループ方式によって実装された時間的結合演算、および、空間的結合演算における 2 値の比較回数は、結合対象データ数 100, 200, および、300 に対して、それぞれ 10,000, 40,000, および、90,000 である。それらの結合演算を実行するのに要する所要時間は、2 値間の比較操作、および、二次記憶へのアクセス操作に要する所要時間によって決定されるが、2 値間の比較に要する計算量は大きく、二次記憶へのアクセスに要する所要時間は無視できるので、表 5、および、表 6 において 2 値間の比較回数の比が実行時間の比として現れた。それゆえ、メタレベルにおける結合演算の実行時間は、データ数の 2 乗に比例して増加すると予測され、データ変換操作に要する所要時間よりもメタレベルシステムにおける問合せ処理時間に影響を与えることが確認された。

以上により、実験結果、および、3 方式のデータ変換操作に要する計算量の式によって、本実験において使用したデータ数よりも大きい場合に対しても、提案方式の

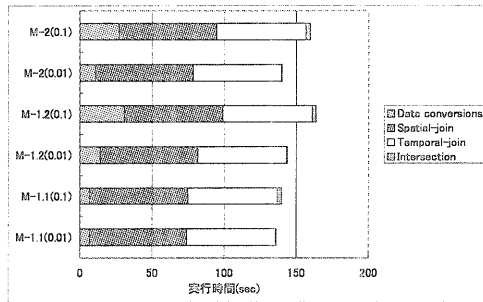


図 20 データ統合時間 (レガシーデータベースのデータ数=100)
Fig. 20 Query execution times (numbers of data items in two legacy databases are 100).

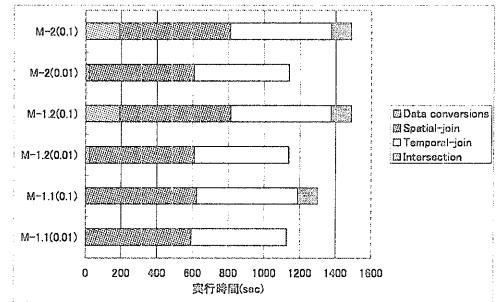


図 22 データ統合時間 (レガシーデータベースのデータ数=300)
Fig. 22 Query execution times (numbers of data items in two legacy databases are 300).

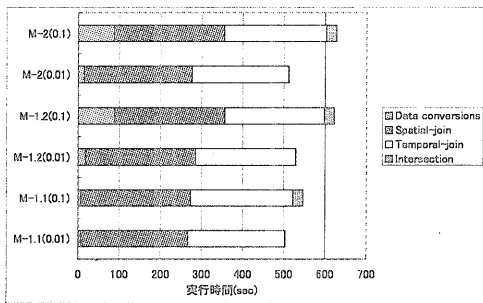


図 21 データ統合時間 (レガシーデータベースのデータ数=200)
Fig. 21 Query execution times (numbers of data items in two legacy databases are 200).

有効性を明らかにすることができた。

6. おわりに

本論文では、データ間の同一性、等価性、類似性、包含性、方向性、連結性などの関連性の評価を行う機能 (関連性評価機能) 群の連結による異種データベース間検索・統合方式を提案した。本方式の特徴は、関連性評価機能群、あるいは、既存データベースシステム群を本システムへ連結する場合において、本システムの変更を行うことなく、それらを連結可能とする機構を実現している点にある。さらに、提案方式を実現した実験システム上での問合せ処理実験の結果を示し、提案方式の有効性を明らかにした。

今後は、メタレベルシステムの間合せ最適化方式、マルチデータベース環境を対象としたデータ表現形式、および、データ粒度の相違性の解消を実現する方式の定式化、ならびに、異なるデータベースに含まれるデータ間の新しい結合方式を設計、実現する予定である。

参考文献

- 1) Allen, J.F.: "Maintaining Knowledge about Temporal Intervals", *Comm. of the ACM*, No.26, pp.832-843 (1983).
- 2) Bright, M.W., Hurson, A.R. and Pakzad, S.H.: "A Taxonomy and Current Issues in Multidatabase System," *IEEE Computer*, Vol.25, No.3, pp.50-59 (1992).
- 3) Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J. and Widom, J.: "The TSIMMIS Project: Integration of Heterogeneous Information Sources", *Proc. the 10th Meeting of the IPSJ*, pp.7-18, Japan (1994).
- 4) Egenhofer, M.J.: "Spatial Relations: Models, Inferences, and their Future Application", *Proc. Advanced Database Symposium '96*, separate volume, Japan (1996).
- 5) Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J. and Widom, J.: "The TSIMMIS Approach to Mediation: Data Models and Languages", *Journal of Intelligent Information Systems*, No.8, No.2, pp.117-132 (1997).
- 6) Haas, L., Kossmann, D., Wimmers, E. and Yang, J.: "An Optimizer for Heterogeneous Systems with Non-Standard Data and Search Capabilities", In Special Issue on Query Processing for Non-Standard Data, *IEEE Data Engineering Bulletin*, Vol.19, No.4, pp.37-43 (1996).
- 7) Haas, L., Kossmann, D., Wimmers, E. and Yang, J.: "Optimizing Queries across Diverse Data Sources", *Proc. the 23rd International Conference on Very Large Data Bases*, pp.276-285 (1997).

表4 データ変換操作に要した時間 (秒)

Table 4 Execution times of data conversions (sec)

(A のデータ数, B のデータ数)	(M-1.1)		(M-1.2)		(M-2)	
	結合率 0.01	結合率 0.1	結合率 0.01	結合率 0.1	結合率 0.01	結合率 0.1
(100,100)	6.9656	6.9071	14.304	31.059	11.022	27.287
(200,200)	6.9380	6.9105	19.231	89.569	15.080	88.173
(300,300)	6.7352	6.8193	24.193	191.57	21.027	190.17

表5 時間的結合操作に要した時間 (秒)

Table 5 Execution times of temporal joins (sec)

(A のデータ数, B のデータ数)	(M-1.1)		(M-1.2)		(M-2)	
	結合率 0.01	結合率 0.1	結合率 0.01	結合率 0.1	結合率 0.01	結合率 0.1
(100,100)	61.358	61.827	61.480	61.907	61.271	61.878
(200,200)	235.51	248.48	243.44	241.56	235.16	247.66
(300,300)	532.57	564.08	527.54	562.78	530.62	565.26

表6 空間的結合操作に要した時間 (秒)

Table 6 Execution times of spatial joins (sec)

(A のデータ数, B のデータ数)	(M-1.1)		(M-1.2)		(M-2)	
	結合率 0.01	結合率 0.1	結合率 0.01	結合率 0.1	結合率 0.01	結合率 0.1
(100,100)	67.280	68.054	67.456	67.965	67.357	67.581
(200,200)	260.58	266.61	265.88	266.98	261.19	266.75
(300,300)	586.17	616.83	584.47	620.75	586.42	619.92

- 8) 細川宜秀, 清木 康: “関数型計算によるマルチデータベースシステムの問合せ処理方式”, 情報処理学会論文誌 Vol.39, No.7, pp.2217-2230 (1998).
- 9) 加藤誠巳: “経路探索問題とその応用”, 情報処理, Vol.39, No.6, pp.552-557 (1998).
- 10) Kent, A., Berry, M., Leuhers, F. and Perry, J.: “Machine literature searching: VIII. Operational criteria for designing information retrieval systems”, *American Documentation*, Vol.6, No.2, pp.93-101 (1955).
- 11) Kitagawa, T. and Kiyoki, Y.: “The mathematical model of meaning and its application to multidatabase systems”, *Proc. the 3rd IEEE Int. Workshop on Research Issues on Data Engineering: Interoperability in Multidatabase Systems*, pp.130-135 (1993).
- 12) Kiyoki, Y., Kitagawa, T. and Hitomi, Y.: “A fundamental framework for realizing semantic interoperability in a multidatabase environment”, *Journal of Integrated Computer-Aided Engineering*, Vol.2, No.1, pp.3-20, John Wiley & Sons (1995).
- 13) Lakshmanan, L.V.S, Sadri, F. and Subramanian, I.N.: “SchemaSQL – A language for Interoperability in Relational Multi-database systems”, *Proc. the 22rd International Conference on Very Large Data Bases*, pp.239-250, India (1996).
- 14) Litwin, W., Mark, L. and Roussopoulos, N.: Interoperability of Multiple Autonomous Databases, *ACM Comp. Surveys*, Vol.22, No.3, pp.267-293 (1990).
- 15) Masunaga, Y.: “A temporal expansion to the multimedia object model in OMEGA”, *Proc. the 4th International Conference on Database Systems for Advanced Applications (DASFAA '95)*, pp.430-440 (1995).
- 16) 森嶋厚行, 北川博之: “構造化文書とデータベースの統合利用のためのデータモデル NR/SD+ とその問合せ処理”, 情報処理学会論文誌 Vol.39, No.4, pp.954-967 (1998).
- 17) Open GIS Consortium(OGC): “The OpenGISTM Guide, Introduction to Interoperable Geoprocessing, Part1.”, Available via WWW from <http://www.opengis.org/public/> (1996).
- 18) Papakonstantinou, Y., Garcia-Molina, H., and Widom, J.: “Object Exchange Across Heterogeneous Information Sources”, *IEEE International Conference on Data Engineering*, pp.251-260, Taiwan (1995).
- 19) Sheth, A. and Larson, J.A.: “Federated database systems for managing distributed, heterogeneous, and autonomous databases”, *ACM Computing Surveys*, Vol.22, No.3, pp.183-236 (1990).
- 20) Tomasic, A., Raschid, L. and Valduriez, P.: “A Data Model and Query Processing Techniques for Scaling Access to Distributed Het-

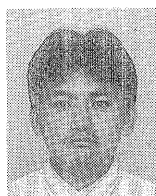
erogeneous Databases in Disco”, *Invited paper in the IEEE Transactions on Computers, special issue on Distributed Computing Systems* (1997).

- 21) Tomasic, A., Raschid, L. and Valduriez, P.: “Scaling Heterogeneous Distributed Databases and the Design of DISCO,” *Proc. the 16th International Conference on Distributed Computing Systems*, pp.449-457 (1996).
- 22) “*UniSQL/X User's Manual*”, NTT データ通信 (1995).
- 23) Yoshida, N., Kiyoki, Y., and Kitagawa, T.: “An Associative Search Method Based on Symbolic Filtering and Semantic Ordering for Database Systems”, *Proc. the 7th IFIP 2.6 Working Conference on Database Semantics (DS-7)*, pp.215-237, Switzerland (1997).
- 24) Yu, A. and Chen, J. (with the POSTGRES Group): “*The POSTGRES95 User Manual*”, Computer Science Div., Department of EECS, University of California at Berkeley (1995).

(平成1999年 7月8日受付)

(平成1999年10月4日採録)

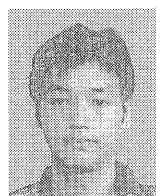
(担当編集委員 有澤 博)



細川 宜秀 (正会員)

1972年生。1995年筑波大学第三学群情報学類卒業。1999年同大学院工学研究科博士課程中退。1999年筑波大学電子・情報工学系に勤務。現在、同学系助手。データベースシ

ステムに関する研究に従事。感性工学会会員。



石橋 直樹

1973年生。1996年慶應義塾大学総合政策学部卒業。1998年同大学大学院政策・メディア研究科修士課程修了。現在、同大学院政策・メディア研究科博士課程在学中。データベースシステムに関する研究に従事。



八代タ紀子

1974年生。1996年お茶の水女子大学理学部情報科学科卒業。1998年お茶の水女子大学大学院理学研究科修士課程了。同年より、株式会社野村総合研究所に勤務。



清木 康 (正会員)

1978年慶應義塾大学工学部電気工学科卒業。1983年同大学大学院工学研究科博士課程修了。工学博士。同年、日本電信電話公社武蔵野電気通信研究所入所。1984~1995年筑波大学電子・情報工学系講師、助教授を経て、1996年、慶應義塾大学環境情報学部助教授、1998年同学部教授。データベースシステム、知識ベースシステム、マルチメディアシステムの研究に従事。ACM, IEEE, 電子情報通信学会, 日本ソフトウェア科学会各会員。