情 報 学 基 礎 50－3
（1998．7．17）

マルチデータベース環境におけるジャンル検索方式

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

近年、インターネット上の情報検索インタフェースとして、HTML のハイパーリンク機能を利用したジャンル検索が一般化してきている。ジャンル検索インタフェースは、情報検索に特別な知識を必要としないため、インターネット上の不特定多数のユーザに対する検索インタフェースとして広く受け入れられ、利用されている。一方、従来、関係データベースに対するジャンル検索を実現するには、行を分類する列を管理し、アプリケーション・プログラムでジャンル項目に応じた SQL を発行する方式が一般的であった。当該方式では、データ分類にはシステム毎の異種性が存在するため、システムを跨るジャンル検索の実現が困難であった。

本稿では、データ分類の異なる複数のデータベースを跨ったジャンル検索を実現する仕組みとして、情報資源管理技術を用いたジャンル検索方式を提案する。

Genre Retrieval Method in the Multi-database Environment

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

In recent years, the genre retrieval using the hyperlink function of HTML is generalized as an information retrieval interface on the Internet. The users don't need special knowledge for information retrieval to use the genre retrieval system. So, the genre retrieval is widely used as retrieval interface to many and unspecified users on the Internet. On the other hand, in general, if we developed the genre retrieval system using a relational database, we needed to make and store the data classification for record, and develop application program issuing SQL corresponding to the genre item. In this method, it was difficult for plural databases to achieve the genre retrieval system, because of the heterogeneity of data classification in the system.

In this paper, as genre retrieval mechanism for plural databases with a different data classification, we propose the genre retrieval method using the information resource management technology.

1.はじめに

近年、インターネット上の情報検索インタフェースとして、HTML のハイパーリンク機能を利用したジャンル検索が一般化してきている。ジャンル検索は、システムが保有しているデータのデータ分類を階層的に示すことにより、ユーザが所望しているデータの容易な絞り込み検索を可能とするため、情報検索に特別な知識を必要とせず、インターネット上の不特定多数のユーザに対する検索インタフェースとして広く利用されている。

一方、従来、関係データベースに対するジャンル検索を実現するには、行を分類する列を管理し、アプリケーション・プログラムでジャンル項目に応じた SQL を発行する方式が一般的であった。当該方式では、一般に、システム毎にジャンル情報を構成するジャンル項目の分類や表現形式が異なっているため、システムを跨るジャンル検索の実現が困難であった。

本稿では、データ分類の異なる複数のデータベースを跨ったジャンル検索を実現する仕組みとして、情報資源管理技術を用いたジャンル検索方式を提案する。

2.ジャンル検索とマルチデータベース

2.1 ジャンル検索

(1)ジャンル検索の概要

インターネットにおいて、一般に、以下の条件を満たすジャンル検索が広く用いられている。

- ●システムは、保有するデータのデータ分類をジャンル情報として木構造でユーザに示す。
- ●ユーザは、ジャンル項目間を上や下に辿り、ジャンル項目を探索するとともに、目的の情報を検索する。

ジャンル検索のイメージを図 1 に示す。

ユーザにとってジャンル検索は、システムが所望の情報が保持しているか容易に識別可能である。また、システムにとって、無効な検索を削減するのに適している。

(2)ジャンル検索の普及度

ジャンル検索は、WWW における情報提供の普及に伴い、急速に発展した情報検索インタフェースの 1 つである。

インターネット上では、主に以下の用途に用いられている。

- ●URL サーチエンジンの URL 情報の検索
- ●EC の取扱商品情報の検索
- ●店舗等の情報提供プロバイダの情報検索

実際に、著名な URL サーチエンジン 20 サイトを調べた結果、1998 年 5 月現在、80%の高い割合でジャンル検索インタフェースがサポートされていた。

2.2 マルチデータベース環境

(1)マルチデータベース環境の必要性

個別に蓄積されてきた企業データベースをイントラネットやエクストラネットに接続し、当該データベースを社内や社間等で共有し、地域別の商品の売れ筋情報を求めるなど、経営戦略的な情報を得るためのデータベースの利用が加速している。経営戦略的な情報を得るためにデータベースを用いる場合、1つのデータベースの情報だけでは不十分な場合があり、例えば顧客情報データベースと売り上げ情報データベースを組み合わせるといったように、複数のデータベースの情報を結合して利用する必要がある。

昨今の急速な技術革新に追従し、時代に即したシステムを構築するため、マルチデータベース環

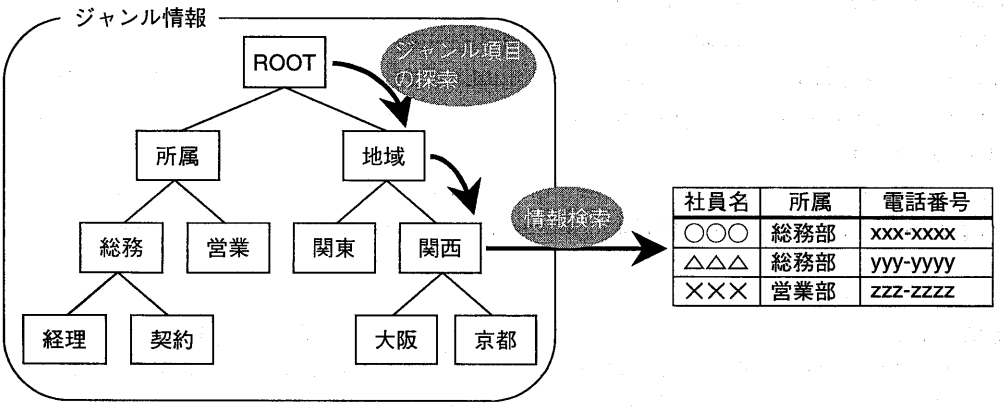

図 1.ジャンル検索のイメージ

境の構築や維持管理(データベースの追加/修正/削除)を柔軟かつ容易に実現可能な技術が必要とされている。

(2)データベースの異種性

既存の独立したデータベースをもとにマルチデータベース環境を構築/運用するには、以下のようなデータベースの意味的異種性[1]を考慮しなくてはならない。

(a)データ名称の相違

データの名称だけでは、その所在や内容を特定することができない。同じ名称で内容が異なっていたり、異なる名称で内容が異なっていたりする場合がある。

(b)データ構造の相違

同様の内容を管理対象とするデータベースでもデータ構造は一意ではない。アプリケーションに特化させ、正規化のレベルが異なったりする場合がある。

(c)データ表現の相違

同じ内容を表すデータでも、その表現形式は一意ではない。数値データの単位, データ型, コード化の有無, 等、データの表現形式には様々な相違がある。

これらの異種性を解消した従来のマルチデータベースの実現方式として、統合モデルを用いる方式があげられる[2]。当該方式は、マルチデータベースを構成する全てのデータベースを解析し、統合モデルを設計し、個々のデータモデルと統合モデルのマッピング定義を行うことで、データベースの異種性を解消する。当該方式は、構築時にシステム開発者の負担が大きく、維持管理面でも急速なデータベースの増減に時間的, 稼働的に対応しきれないという問題点があった。

2.3 マルチデータベース環境におけるジャンル検索の問題点

(1)前提条件

マルチデータベース環境におけるジャンル検索の検討に際し、以下を検討の前提とした。

[前提 1] 関係データベースに対するジャンル検索を実現するにあたって、列や表の追加などのデータベースのカスタマイズを行わない。つまり、既にデータベース上で管理されているデータ分類情報を利用し、ジャンル検索を実現することとする。

[前提 2] ジャンル情報として利用するデータ分類は、データの分類レベルは互いに異なっていても良いものとする(データの分類レベルの相違には、「部」単位と「課」単位, 「国」単位と「県」単位, などがある)。

[前提 3] 上位概念のデータ分類と合わせてデータ分類の一意性を保証するものは対象外とする。つまり、ジャンル情報において、同一のジャンル項目を複数登録することはできないものとする。

(2)マルチデータベース環境におけるジャンル検索実現のための課題

マルチデータベース環境において、ジャンル検索を実現するには、以下の課題を解決しなければならない。

[課題 1] ジャンル情報の柔軟な運用方法の確立

[課題 2] ジャンル項目とデータベース問合せ文のマッピング情報の効率的管理の確立

[課題 3] 検索結果の形式統一方法の確立

(3)従来のジャンル検索実現方式

これまで、マルチデータベース環境におけるジャンル検索を実現するための手順として、一般に、以下が用いられてきた。

[手順 1] データ分類を示す列をカテゴライズし、ジャンル情報を生成し、アプリケーション内で管理する。

[手順 2] ジャンル項目毎の各データベースに対する SQL をアプリケーション内で管理する。

[手順 3] データ項目の表現形式をアプリケーション内で管理し、検索結果の表現形式を変換して、返却する。

この手順により、ジャンル項目毎の柔軟なマルチデータベース検索を実現可能である。しかし、この方式では、データベースの追加/削除, 各データベースのメタ情報の変更, などの場合、SQL の変更や、ジャンル情報の改造など、アプリケーションを改造する必要があった。

3.マルチデータベース環境におけるジャンル検索

3.1 システム概要

我々は、データベースの様々な意味的異種性を、情報資源管理技術を用いる方式で解消し、マルチデータベース検索システムを実現するDBSENA を開発した[3]。DBSENA のシステム構成を図 2 に示す。

DBSENA は SQL インタフェースであたかも 1 つのデータベースにアクセスしているかのように扱う従来のマルチデータベース管理システムとは異なり、情報の所在を指定させない普遍関係によるインタフェース技術[4]をマルチデータベース環境に適用し、マルチデータベース検索を実現している。ユーザは、検索要求時に、データベース名，表名，などの所在を指定する必要がなく、検索項目と検索条件とジャンル項目を指定するのみで良い。この検索要求に対し、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[5]を利用し、データベース間の異種性を解消した検索を実現している。

具体的に、DBSENA は、以下の要素から構成される。

(a)情報資源辞書

検索対象となるデータベースのデータ構造，データの表現形式，列の取り得る値の範囲，等のメタ情報や、ジャンル情報を管理する。

(b)ジャンル検索インタフェース機能

アプリケーションに対し、ジャンル項目情報へのアクセスを制御する。また、検索要求で指定されたジャンル項目をデータベースのデータ分類レベルに展開する。

(c)異種データベース検索機能

ジャンル検索インタフェースで展開された検索要求に対し、情報資源辞書を用いて、データベースの異種性を解消し、データベース毎の SQL を生成し、データベース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し、検索結果をユーザ毎の表現形式に変換する。

(d)データベースアクセス機能

実際にデータベースにアクセスし、SQL の送信，検索結果の受信，等を行う。

(e)情報資源管理ツール

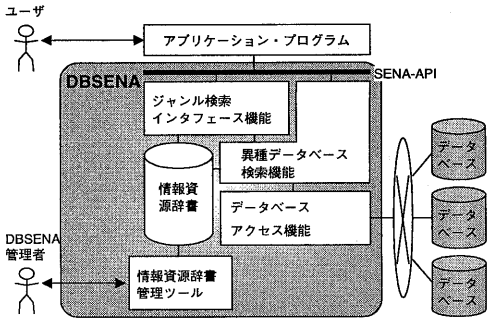

図 2.DBSENA システム構成

情報資源辞書の維持管理を行う。

(f)SENA-API

DBSENA が提供するマルチデータベース検索機能をアプリケーションから利用するためのインタフェースを提供する。

3.2 アプローチ

DBSENA では、マルチデータベース環境におけるジャンル検索実現のための各課題を以下のアプローチで解決している。

[課題 1] ジャンル情報の柔軟な運用方法の確立

ジャンル情報をデータ項目と対応させるのではなく、データ項目の表現形式と対応させて管理することにより、データベースの変更に影響しないジャンル情報の管理方法を実現した。

[課題 2] ジャンル項目とデータベース問合せ文のマッピング情報の効率的管理の確立

ジャンル項目に対し、データベースの情報資源を用いて、動的にデータベース問合せ文を生成する機構を確立することにより、ジャンル項目とデータベース問合せ文の独立性を確保した。

[課題 3] 検索結果の形式統一方法の確立

各データ項目の表現形式と、表現形式間の変換関数を管理することにより、ユーザ毎に検索結果の表現形式の統一を実現した。

3.3 ジャンル検索実現方式

DBSENA は、情報資源辞書を用いて、マルチデータベース環境におけるジャンル検索機能を

実現する。情報資源辞書では、個々のデータベースの情報(スキーマ), 列の同義語, 異種性解消のために必要な表現形式に関する情報(ドメイン), 値の取りうる範囲(レンジ), などを管理する。ドメインとレンジについて、概要を以下に示す。

(1)ドメイン

DBSENA はデータの表現形式をドメインという概念を用いて、管理している。特に個々のデータベース上のドメインをローカルドメイン, ユーザが検索要求や検索結果の指定に用いるドメインをユーザドメイン, ドメイン内で標準的に用いるドメインをグローバルドメインと呼ぶ。また、同じ内容を示すドメインの集まりをドメイングループと呼ぶ。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行う。ドメインの概念を図 3 に示す。

(2)レンジ

DBSENA は列の値の取りうる範囲であるレンジ[6]を管理している。レンジを用いることにより、検索要求に対する検索先のデータベースの限定を実現する。また、データベースに跨って格納される値をレンジ情報ツリーと呼ぶ木構造で管理し、ジャンル検索のジャンル情報に用いる。レンジ情報ツリーの概念を図 4 に示す。

レンジ設定では、レンジ情報ツリーにおいて、包含する上位ノードを一括して指定する。なお、各列の取りうる値を個別に指定することも可能である。図 4 において、ある列の取りうる値の範囲が「経理」と「契約」のみの場合、当該列のレンジには、「所属/総務」と一括して指定する。

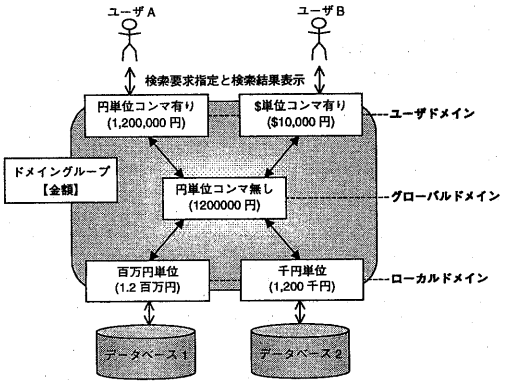

図 3.ドメインの概念

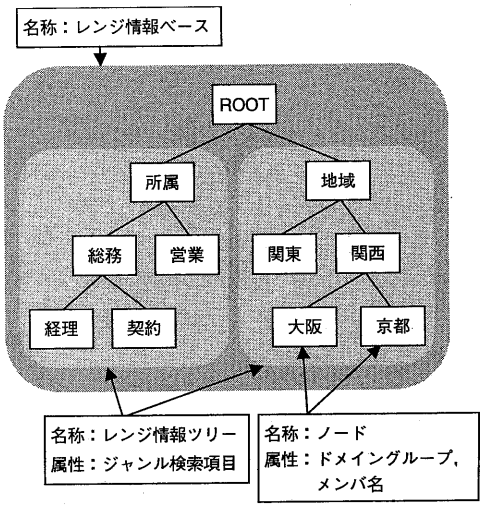

図 4.レンジ情報ツリーの概念

(3)レンジ情報ツリー

レンジ情報ツリーは、組織構成, 書籍分類, 商品種別, 等の同様の内容毎に生成する。また、これらレンジ情報ツリーのルートノードの上位に共通のルートノードを設け、レンジ情報ツリーを統合したものをレンジ情報ベースと呼ぶ。実際にはレンジ情報ベースをジャンル情報に、ノードをジャンル項目にマッピングする。

(a)レンジ情報ツリーの属性

各レンジ情報ツリーは、以下の属性をもつ。

・ジャンル検索項目

レンジ情報ツリー毎の、検索要求に対して用いられる検索項目である。例えば、図 4 のレンジ情報ツリー「所属」に対し、「社員名, 入社年次, 電話番号」と指定する。

(b)レンジ情報ツリーの各ノードの属性

レンジ情報ツリーの各ノードは、以下の属性を持つ。

・ドメイングループ

ノードの属するドメイングループ名である。

・メンバ名

各ノードの指定ドメイングループのグローバルドメインを用いた値である。

4.ジャンル検索方式

本章では、DBSENA を用いたマルチデータベース環境におけるジャンル検索方式について、図5 のデータベースを例に説明する。データベース「ショップ A」と「ショップ B」は行のデータ分類を示す列として、それぞれ「商品種別」を保有しているが、その分類レベルが異なっている。つまり、データベース「ショップ B」の列「商品種別」は、データベース「ショップ A」の列「商品種別」に対し、上位の概念で分類されている。

4.1 各種管理情報の設定

ジャンル検索を実現するため、先にドメイン, レンジ情報ツリー, レンジ, 列の同義語を以下のように情報資源辞書に登録しておく。

(1)ドメイン

各ドメインを以下の通り、登録する。

(a)ローカルドメイン

列名	ドメイン名
ショップ A.取扱商品.商品種別	商品小分類用語
ショップ A.取扱商品.価格	千円単位価格
ショップ B.取扱商品.商品種別	商品大分類用語
ショップ B.取扱商品.価格	万円単位価格

(b)ドメイングループ

ドメイン グループ名	ドメイン名	グローバルドメイン フラグ
商品大分類	商品大分類用語	YES
商品小分類	商品小分類用語	YES
価格	千円単位価格	NO
価格	万円単位価格	NO
価格	円単位価格	YES

(c)ユーザドメイン

ユーザ名	ドメイングループ名	ユーザドメイン名
user1	商品大分類	商品大分類用語
user1	商品小分類	商品小分類用語
user1	価格	円単位価格

(2)レンジ情報ツリー

レンジ情報ツリーの例を図 6 に示す。

(3)レンジ

各データベースの商品種別列のレンジを以下の通り、登録する。

データベース:ショップ A
テーブル:取扱商品

店名	商品種別	商品名	価格
ショップ A	デスクトップ PC	DPC-2	90 千円
ショップ A	デスクトップ PC	DPC-3	180 千円
ショップ A	ノート PC	NPC-1	270 千円
ショップ A	ノート PC	NPC-2	360 千円
ショップ A	ノート PC	NPC-3	450 千円

データベース:ショップ B
テーブル:取扱商品

店名	商品種別	商品名	価格
ショップ B	PC 本体	DPC-1	7.8 万円
ショップ B	PC 本体	DPC-2	9.8 万円
ショップ B	周辺機器	DSP-1	7.8 万円
ショップ B	周辺機器	PRT-1	15.8 万円
ショップ B	周辺機器	DC-1	4.8 万円

図 5.データベース例

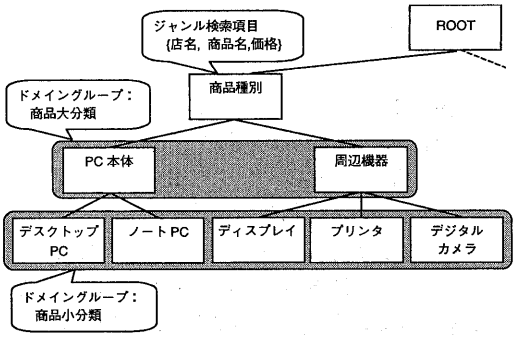

図 6.レンジ情報ツリー例

列名	レンジ
ショップ A.取扱商品.商品種別	/商品種別/PC 本体
ショップ B.取扱商品.商品種別	/商品種別

(4)列の同義語

データ名称の異種性を解消するため、列の同義語を以下の通り、登録する。

データベース名	表名	列名	同義語
ショップ A	取扱商品	店名	店名
ショップ A	取扱商品	商品種別	商品種別
ショップ A	取扱商品	商品名	商品名
ショップ A	取扱商品	価格	価格
ショップ B	取扱商品	店名	店名
ショップ B	取扱商品	商品種別	商品種別
ショップ B	取扱商品	商品名	商品名
ショップ B	取扱商品	価格	価格

4.2 ジャンル検索処理フロー

次に、ジャンル検索処理フローについて、ユーザ「user1」からの以下の検索要求に対する処理を説明する。

ジャンル項目	:/商品種別/PC 本体
検索条件	:価格 ＜100,000 円

(1)列レンジリストの生成

指定されたジャンル項目とレンジが以下の条件を満たす列を探索する。

[条件 1] 指定ジャンル項目配下のいずれかのジャンル項目をレンジとする列

[条件 2] 指定ジャンル項目の上位ジャンル項目から ROOT ノードまでのいずれかのジャンル項目をレンジとする列

[条件 3] レンジ未設定かつ、指定ジャンル項目配下のいずれかのジャンル項目と同じドメイングループである列。

結果として、以下が探索される。

・ショップ A.取扱商品.商品種別

・ショップ B.取扱商品.商品種別

(2)ジャンル項目の展開

ジャンル項目を(1)で探索された列毎に、ジャンル項目配下、かつ、列のレンジ内、かつ、列のローカルドメインのドメイングループ内のメンバ名に展開する。

・ショップ A.取扱商品.商品種別

ローカルドメインが商品小分類用語であり、指定されたジャンル項目「/商品種別/PC 本体」をドメイングループ「商品小分類」のレンジ内のメンバ名(「デスクトップ PC」と「ノート PC」)に展開する。

・ショップ B.取扱商品.商品種別

ローカルドメインが商品大分類用語であり、指定されたジャンル項目「/商品種別/PC 本体」のドメイングループと一致しているため、そのメンバ名(「PC 本体」)を引き渡す。

(3)検索条件の結合

検索要求で指定された検索条件と(2)で展開したジャンル項目を AND で結合する。

・ショップ A.取扱商品.商品種別

商品種別 IN (‘デスクトップ PC’, ’ノート PC’)
AND 価格 ＜100,000 円

・ショップ B.取扱商品.商品種別

商品種別 IN (‘PC 本体’)
AND 価格 ＜100,000 円

(4)検索項目の補完

検索項目として、レンジ情報ツリーの属性「ジャンル検索項目」で設定されている「店名, 商品名, 価格」を補完する。

・ショップ A.取扱商品.商品種別

検索項目: 店名, 商品名, 価格
検索条件: 商品種別 IN (‘デスクトップ PC’, ’ノート PC’) AND 価格 ＜100,000 円

・ショップ B.取扱商品.商品種別

検索項目: 店名, 商品名, 価格
検索条件: 商品種別 IN (‘PC 本体’) AND 価格 ＜100,000 円

(5)項目の所在の探索

検索要求で指定されたジャンル項目以外の項目(店名, 商品名, 価格)を列の同義語として、所在を探索する。結果として、以下が探索される。

・店名 :ショップ A.取扱商品.店名
:ショップ B.取扱商品.店名

・商品名 :ショップ A.取扱商品.商品名
:ショップ B.取扱商品.商品名

・価格 :ショップ A.取扱商品.価格
:ショップ B.取扱商品.価格

(6)問い合わせデータベースの決定

(4)に対し、(5)で求めた所在情報より、すべての項目を含むデータベースを探索し、検索先データベースを決定し、SQL に展開する。

・データベース「ショップ A」

```
SELECT 店名,商品名,価格
FROM 取扱商品
WHERE 商品種別 IN (‘デスクトップ PC’, ’ノート PC’)
  AND 価格 ＜100,000 円
```

・データベース「ショップ B」

```
SELECT 店名,商品名,価格
FROM 取扱商品
WHERE 商品種別 IN (‘PC 本体’)
  AND 価格 ＜100,000 円
```

(7)問い合わせ文の変換

WHERE 句に含まれる列のローカルドメインとユーザドメインを探索する。

・ドメイン「価格」

ユーザドメイン	ローカルドメイン
円単位価格	・ショップ A.取扱商品.価格は千円単位価格 ・ショップ B.取扱商品.価格は万円単位価格

ローカルドメインとユーザドメインが異なる場合、ユーザドメインをローカルドメインに変換し、データベース毎の問い合わせ文を生成する。この例では、以下の問い合わせ文を生成する。

・データベース「ショップ A」

```
SELECT 店名,商品名,価格
FROM 取扱商品
WHERE 商品種別 IN ('デスクトップ PC', 'ノート PC')
  AND 価格 < 100 千円
```

・データベース「ショップ B」

```
SELECT 店名,商品名,価格
FROM 取扱商品
WHERE 商品種別 IN ('PC 本体')
  AND 価格 < 10.0 万円
```

(8)問い合わせの実行

問い合わせを実行し、以下の結果を得る。

・データベース「ショップ A」

店名	商品名	価格
ショップ A	DPC-2	90 千円

・データベース「ショップ B」

店名	商品名	価格
ショップ B	DPC-1	7.8 万円
ショップ B	DPC-2	9.8 万円

(9)検索結果の変換

ローカルドメインとユーザドメインが異なる場合、ローカルドメインをユーザドメインに変換し、検索結果の表示形式を統一する。これにより、本問い合わせの結果として、以下を得る。

店名	商品名	価格
ショップ B	DPC-1	78,000 円
ショップ A	DPC-2	90,000 円
ショップ B	DPC-2	98,000 円

5.まとめと今後の課題

本稿では、データ分類やデータの表現形式など、様々な異種性があるマルチデータベース環境において、ジャンル検索により、ユーザ所望の情報を探索するための処理方式について、提案した。

今後の課題として、以下があげられる。

[課題 1] 複数のジャンル項目を指定した検索を実現する。

[課題 2] 上位概念のデータ分類と合わせてデータ分類の一意性を保証するデータ分類を扱えるジャンル検索方式を確立する。

[課題 3] ジャンル情報の追加/更新/削除など、運用方式を確立する。

今後、様々なデータを用いて、本方式の検証を行い、適用可能なデータ分類など、当該システムの限界を見極めると共に、上記の課題について、検討を進めていく予定である。

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] Kim, W., et al., "On Resolving Schematic Heterogeneity in Multi-database Systems", Distributed and Parallel Databases, 1993.

[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", 第 114 回データベースシステム研究会, 1998.

[4] Ullman, 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.

[5] 鈴木源吾, 町原宏毅, 川下満"Fragment View－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

[6] 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