

XMLに基づく異種情報源メディエーションシステム： MediPresto/XM

鈴木源吾 小西一也 林孝志 小林伸幸 芳西崇

日本電信電話株式会社 NTT サイバースペース研究所

ネットワーク化・データベース化の進展により、Web ページ・関係データベース・マルチメディアデータベースなど様々な情報源がインターネット・イントラネット上に氾濫している。一方、情報流通のフォーマットとして XML が、進展していく方向にある。そこで、本稿では、XML のモデルに基づき、異種の情報源の統合検索・横断検索を実現するメディエーションシステムを提案する。本システムは動的な構造異種性の解消を特徴としている。本システムのアーキテクチャ・問い合わせ言語を提案する。さらに、プロトタイプシステムを作成し、その有効性を検証した。

Mediation System for Heterogeneous Information Sources based on XML: MediPresto/XM

Gengo Suzuki, Kazuya Konishi, Takashi Hayashi, Nobuyuki Kobayashi,
and Takashi Honishi

NTT Cyber Space Laboratories,
NTT Corporation

Development of networking and databases leads vast mount of information sources such as web pages, relational databases, and multimedia databases on the Internet and Intranets. On the other hand, XML has been accepted as a standard format for information delivery. In this paper, we propose a new mediation system based on XML model, which enables integrated retrieval for heterogeneous information sources. A main feature of this system is a dynamic structural conflict resolution. We propose system architecture and a query language of this mediation system. We implemented a prototype system, and show effectiveness of this system

1. はじめに

ネットワーク化・データベース化の進展により、様々な情報がインターネット・イントラネット上に氾濫している。情報源は HTML・XML などの Web ページ・関係データベース・画像・音声・動画などを含むマルチメディアデータベースなど様々である。そのような情報源に対する

横断的な検索や、統合的な検索のニーズは今後ますます増えると予想される。

一方、情報流通・情報発信のフォーマットとして、XML 技術が今後は進展していく方向にあり、インターネット上には今後 XML に基づくコンテンツ・サービスが増大すると予想されている。また、XML の持つ柔軟な情報表現能力は、異種情報源の統合を行うモデルとして有効で

あると考えられる。そこで本稿では、XML モデルに基づいて異種情報源の統合を行うメディアータシステムを提案する。以下、本稿の構成を示す。2章で、異種情報源の統合のためのメディアータ技術について議論する。3章において、本システムの特徴とアーキテクチャを述べ、4章で問い合わせ言語と問い合わせ処理について述べる。5章でプロトタイプシステムの実装について紹介し、6章でまとめを行う。

2. メディアータ技術

複数のデータベースを統合する研究としては、古くから、マルチデータベース・連邦データベース技術と呼ばれる分野が研究されていた。この分野においては、異種分散環境下での、問い合わせ最適化・更新問題など様々な研究課題が議論されてきた。しかし、近年のWWWの急速な普及・マルチメディア情報の増大により、これまでとは違った観点を含めて、異種情報源の統合検索技術が研究されている。その主な観点を以下に挙げる。

(1) 情報源の大量性

対象はインターネットを含むために、膨大な情報源に対して検索を行うことが求められる。そのための問い合わせ最適化アルゴリズムなどが重要である。

(2) 情報源のモデルの多様性

Web ページのような半構造的な情報から、関係データベースのような固い構造まで様々なモデルの統合を行う必要がある。

(3) 情報源の能力の多様性

古典的なマルチデータベース技術においては、情報源は関係モデルなり、ネットワークモデルなりで規定される検索能力を事前に仮定することができたが、現在では、静的 Web ページのように全く能力のない情報源もあれば、動的な Web ページ（フォーム形式で検索する）のように限定された能力を持つ情報源もある。それらの制約を考慮しつつ問い合わせプランを生成する必要がある。

これらの観点に基づく、新しいタイプの異種情報源を統合検索するシステムは、メディアータ、メディアーション・システム等と呼ばれる[1]。これまでの研究としては、TSIMMIS[2]・Information Manifold[3]などが代表的である。

3. XML に基づくメディアーション・システム：MediPresto/XML

3.1. XML に基づくメディアーション

情報流通・情報発信の共通フォーマットとして、近年 XML が非常に注目を浴びている。現在、インターネットでの情報発信は HTML によるものが主流であるが、将来的には XML によるコンテンツ・サービスの増大が予想されている。XML はツリー構造によって情報を表現し、柔軟なモデル表現能力を持つ。

メディアータ・メディアーションシステムでは、異種の異なるデータモデルを吸収するためのデータモデルが必要である。情報源毎に異なるモデルは、そのモデルに一度マッピングされ、問い合わせなどの処理が行われる。そのモデルは共通データモデルと呼ばれる。これまでのメディアータでは関係モデル・オブジェクト指向モデルなど様々なモデルが共通データモデルとして採用されていた。共通データモデルに対する要求条件は、様々なモデルの違いを吸収できるだけの十分な表現能力を持つことである。

そこで、我々は、XML のモデルを異種情報源のための共通モデルとして採用し、情報の統合を行うシステム MediPresto/XML (multiMEDIAspecific REtrieval and STORage system/XML Mediator) を提案する。以下の理由で、XML を採用している。

- XML のモデルの持つ（ある程度）豊かな表現能力を利用できる。
- 増大する XML コンテンツに対して最適に対応できる。
- XML に関する標準化が進展しつつあるために、標準的なインターフェースを採用できる。

3.2. MediPresto/XML の特徴

MediPresto/XML は、インターネット環境に適合したメディアーションを実現するために以

下の2つの特徴を持っている。

- ・ 動的な異種性の解消
 - ・ ビューの生成と操作
- 以下に説明する。

3.2.1. 動的な異種性の解消

今後、インターネットに流通するXML情報は、様々な構造を持ち、かつ大量に存在することになると予想される。インターネット環境における、そのような構造の多様性・情報源の大量性のため、あらかじめ、すべての情報源を分析し、メディエーションシステムに対して、その構造や表現の変換を記述しておくということは困難である。そこで、MediPresto/XMLでは、あらかじめ、事前に定義しておく知識（静的な変換ルール）は、最小限に留め、ユーザ・アプリケーションの問い合わせの指定に対して、それに合致する（ないしは類似する）情報の構造を探し出し、もし、同一の構造でなくとも、変換可能であれば、構造変換を行うという、動的な問い合わせ処理を志向している。

3.2.2. ビューの生成と操作

インターネットに対して、問い合わせを行い、上記のような動的な問い合わせ処理を行う場合、その結果は様々な形式・構造を持つ情報源から得られ、一意に決まるものではない。

MediPresto/XMLでは、その問い合わせ処理の結果として得られる、複数の情報源に対する問い合わせパターンを「動的ビュー」と呼ぶ。ユーザが指定した問い合わせに対して、対応する（仮想的な）XMLの要素を探し出し、動的ビューを求める機能を「動的ビューの生成」と呼ぶ。古典的なマルチデータベースシステムにおいては、事前に（静的な）「ビュー」を作成し、保存しておき、その「ビュー」に対して、ユーザは問い合わせを行う。MediPresto/XMLにおいては、「ビュー」は問い合わせの実行時に動的に生成されると解釈されているわけである。

動的なビューに対しては、

- ・ 保存
- ・ 和 (Union) ・ 選択 (selection) などの問い合わせ演算

などの操作を行うことが可能である。この操作を利用すれば、インタラクティブな情報探索が容易に実現できる。例えば、以下のようなアプリケーションを実現できる。インターネットに対して、ユーザがパソコンの価格に関する情報を調べる場合、

- ・ パソコンに関する価格・製品名などを調べる問い合わせを作成する。
- ・ その問い合わせから、MediPresto/XMLに対して検索を行い、パソコン業者の情報源に対する問い合わせパターンの候補

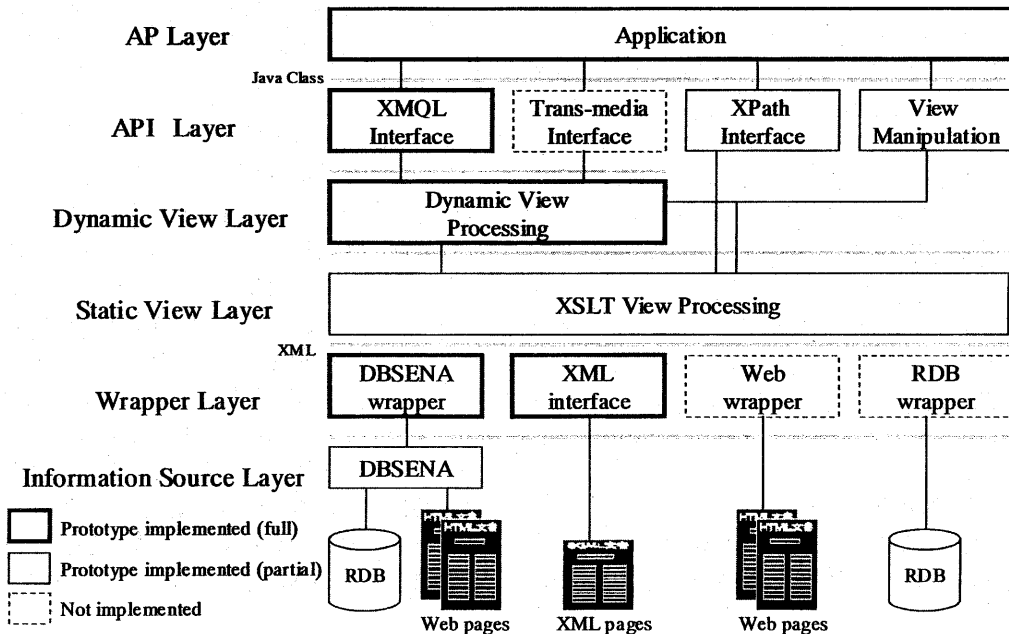


図1 Architecture of MediPresto/XML

(動的ビュー)が複数得られる。

- ・ 動的ビューがユーザの求めるものだったら、それを保存し、また見れるようにしておく(この保存とはデータの保存ではなく、問い合わせパターンの保存である)。
- ・ 複数のパソコン業者の価格の情報を一括して見るときは、動的ビューの和をとる。
- ・ パソコンの価格を調べるときに、価格によってさらに絞り込みを行うときには、動的ビューに対して選択処理を指定し、検索実行を行う。

3.3. MediPresto/XML のアーキテクチャ

以下、上記の特徴を実現するための MediPresto/XML のアーキテクチャを述べる。図1にアーキテクチャを示す。以下それぞれの層について説明する。

3.3.1. アプリケーション層

インターネットに対する検索を実現するアプリケーションを実現する層である。MediPresto/XML の提供する API を利用して、アプリケーションを作成する。

3.3.2. API 層

MediPresto/XML の提供する API を実現する層である。上記の動的な問い合わせ処理を行うための API 層として、問い合わせ言語 XMQML インターフェイスが提供される。4章において詳述する。保存・和などのビュー操作のインターフェイスも提供される。動的問い合わせ処理の上にメディア変換を行うような問い合わせを実現することも想定している[4]。また、動的問い合わせ処理をスルーさせることにより、XPath などのような問い合わせインターフェイスも実現できると考えている。

3.3.3. 動的ビュー層

上記の動的な問い合わせ処理を実現する層である。ユーザが指定した構造に対して、その

所在を探索したり、構造の異種性がある場合それを自動的に解消したり、などの処理を実現する。

3.3.4. 静的ビュー層

あらかじめ管理者により構造変換が規定できるような、静的な構造変換を行う層である。XSLT などにより変換を記述する。また、動的ビューの処理を行う場合、その処理を XSLT などに変換することによって、この層で実現する処理に委譲することも可能である。

3.3.5. ラッパー層

情報源のモデルの違いを隠蔽する層である。関係モデル・HTML モデルなど様々なモデルを XML をベースとするモデルに変換する。

3.3.6. 情報源層

インターネット・イントラネットに存在する関係データベース・Web ページなど様々な情報源であり、それぞれの固有の処理を行う層である。

4. 問い合わせ言語 XMQML

上記アーキテクチャにおける動的ビュー層を実現するための問い合わせ言語 XMQML を提案する。

4.1. XMQML の特徴

既存の XML 問い合わせ言語 (XPath[5]・Xquery[6]・XML-QL[7]など)は、基本的に XML 文書・XML データベースから、検索するための対象を特定して抜き出し、それを組み合わせ加工し、検索結果を得るという特徴を持っている。それは、XML の構造を知っている人が、細かく指定を行い、あらゆる問い合わせ要求を実現できるということを目指して設計されている。しかし、インターネット環境におけるメディアータを想定する場合、検索する側から、必ずしも正確な構造を指定しなくとも、情報源側の構造を意識させない検索を実現する必要がある。

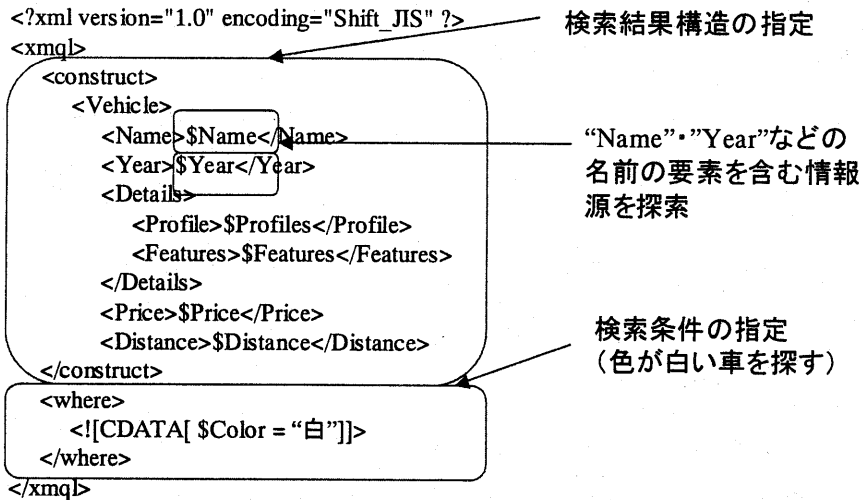


図2 XMQの問い合わせ

そこで、基本的に情報源の構造を意識せず、結果構造の規定と条件指定のみを行う問い合わせ言語 XMQ (XML based Mediator Query Language) を提案する。図2に XMQ の記述を示す。

XMQ は、XML で記述され、

- ・ CONSTRUCT 句
- ・ WHERE 句

の2つの句を持つ。CONSTRUCT 句は、結果構造を規定する句である。\$Name という形式で変数を指定する。指定した変数名は、基本的には情報源に存在する要素の名前もしくは同義語と解釈される。変数名から発見される情報源の要素を組み合わせ、CONSTRUCT 句に沿う形で情報を返却する。WHERE 句は条件式の規定である。

4.2. XMQ 問い合わせ処理

本節では XMQ 問い合わせの処理の概要について説明する。3.2.2 節でも述べたように、XMQ 問い合わせによって、情報源に対する問い合わせは一意に決定するとは限らず、その複数の情報源に対する問い合わせパターンの候補を動的ビューと呼んでいる。問い合わせの処理は、動的ビューの生成・ビューの実行という大きく2つの処理に分割される。それぞれの概要について説明する。

4.2.1. 動的ビューの生成

XMQ 問い合わせから、どの情報源のどの要素に対する問い合わせを行うかまでを確定させる機能である。以下の機能により実現される。

(1) 要素の所在探索

指定語 (\$ で指定される語) に対して、その名前・またはその同義語から、その要素の所在を見つける。

(2) 組み合わせ生成

要素の所在探索で見つけた要素に対して、その組み合わせを作成する。問い合わせに指定した指定語が複数の情報源に分割され、それらを統合するようなパターンをビューとして生成する場合もある。

4.2.2. ビューの実行

動的ビューを指定し、その検索結果を得る機能である。

(1) 検索結果の取得

動的ビュー生成の過程で見つけられた情報源に対して、情報源の形式にあった問い合わせを生成・発行し、検索結果を得る。

(2) 構造・形式変換

ユーザが指定した構造・形式と情報源の構

造・形式が違う場合、その違いの解消を行い、検索結果を得る。例えば、以下の構造変換が行われる。

- ・ 逆転
 - 例えば、情報源の構造では、
論文－著者
となっており、ユーザからは
著者－論文
と指定された場合、その構造を逆転させ検索結果を生成させる。また、指定を行っていない要素に関しては、その配下の要素に構造的な違いがある場合、逆転を

行い構造をあわせることができる [8]

- ・ 縮退
 - 例えば、情報源の構造では、
車情報－詳細情報－車色
となっていて、ユーザからは、
車情報－車色
と指定された場合、中間の「詳細情報」要素を縮退させ、ユーザから指定された構造を得ることができる。

5. プロトタイプシステムの実装

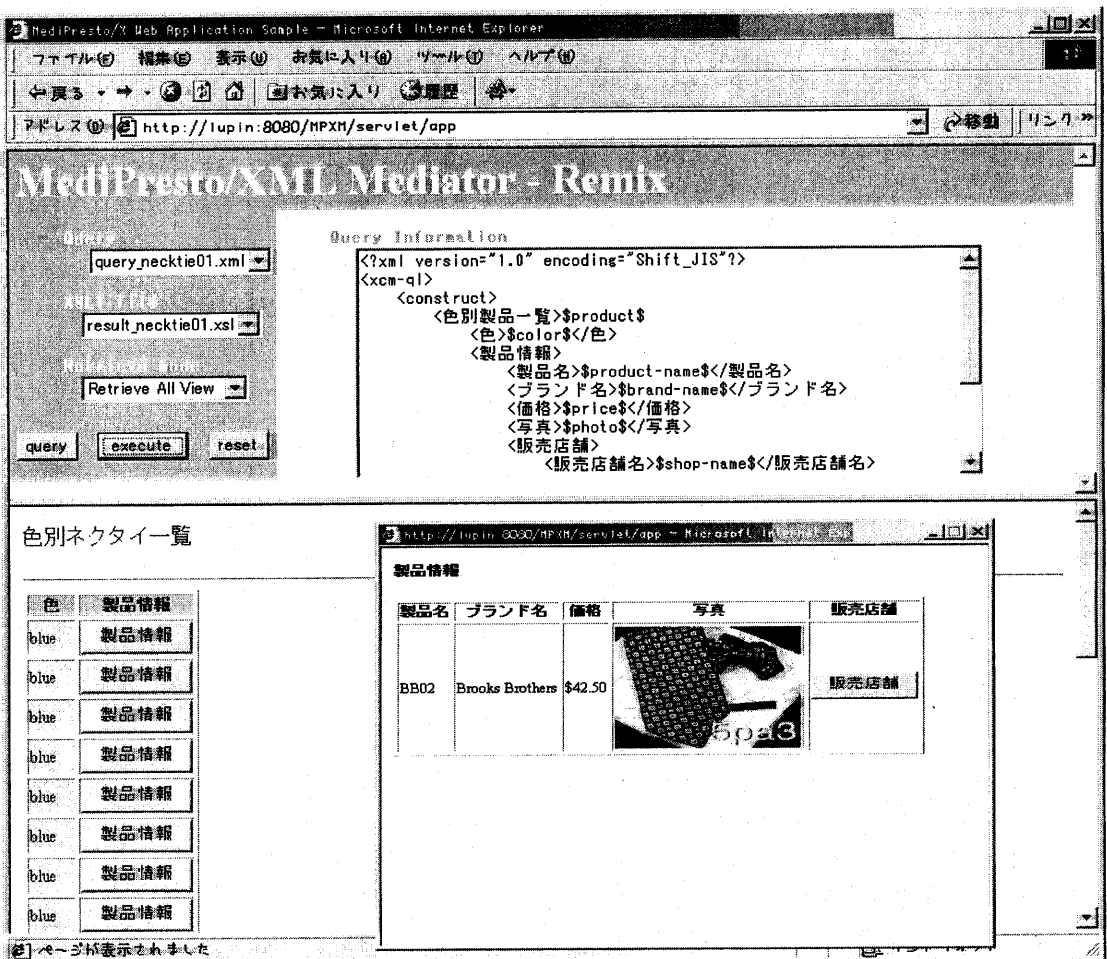


図3 プロトタイプアプリケーションの画面

上記アーキテクチャに従いプロトタイプを実装した。プロトタイプは、Java により実装され、Java による API が実現されている。Web ブラウザより、XMQL による問い合わせを行い、動的ビューを生成・表示し、選択したビューを実行するアプリケーションを、Servlet を用いて実現している。

動的ビュー層としては、XMQL を処理する機構を実現している。問い合わせ処理に含まれる構造変換としては、逆転・縮退が可能である。

静的ビュー層としては、XSLT によるビュー機能を提供している。XSLT で変換された XML 形式が、アプリケーション側からは見えている。

以下の情報源に対するラッパーを実装している。

(1) XML

URL で指定する XML ドキュメントに対するアクセスを実現している。

(2) 異種情報源統合システム DBSENA

普遍関係モデルをベースにした情報資源管理・異種情報源統合システムである DBSENA [9] [10] に対するラッパーを提供している。DBSENA は、Web 情報源・RDB 情報源など広範囲な情報源に対するアクセスが可能である。本プロトタイプでは、これらの情報源は DBSENA ラッパーを経由してアクセスしている。

プロトタイプアプリケーションの画面を図 3 に示す。この例では、ネクタイに対する検索を行っている。ショップ別に表されたネクタイや、メーカー別に表されたネクタイなど異なった表現がなされていても、一括検索が可能である。この例では、検索結果は XSLT によって変換し、HTML として表示している。生成されたビューは保存することができ、保存されたビューに対してさらに XPATH 検索を行うことが可能である。

本プロトタイプにおいて、基本的な動的な問い合わせ処理のイメージの実現が確認できた。

6. まとめ

XML を共通モデルとするメディエーション・システムを提案した。本システムは動的な異種性の解消を行うこと・ビューの生成と操作が特徴である。アーキテクチャ・問い合わせ言語を提案し、プロトタイプを作成し、その基本的なイメージを実現を確認した。今後の課題としては以下があげられる。

(1) 問い合わせ最適化制御

現在のプロトタイプでは実現されていない。構造変換を考慮した検討が必要である。

(2) インターネット対応メタデータ管理

インターネットに対応するスケーラブルなメタデータの管理方式を開発する必要がある。

参考文献

- [1] G. Wiederhold: "Mediators in the architecture of future information systems, IEEE Computer, 25:38--49, 1992.
- [2] A. Levy, A. Rajaraman, and J. Ordille: "Querying heterogeneous information sources using source descriptions", Proceedings of the 22nd VLDB Conference, 1996.
- [3] <http://www-db.stanford.edu/tsimmis/tsimmis.html>.
- [4] 林他: "任意メディアにより問い合わせ可能な分散マルチメディア情報検索方式", 情報処理学会 データベースシステム研究会, May, 2001.
- [5] <http://www.w3.org/TR/xpath>
- [6] <http://www.w3.org/TR/xquery/>
- [7] <http://www.w3.org/TR/NOTE-xml-ql/>
- [8] 小西他: "構造的異種性を解消する XML 一括検索手法", 電子情報通信学会 データ工学研究会, June, 2001.
- [9] 星野他: "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 データベースシステム研究会, Jan, 1999.
- [10] 池田他: "連邦データベースシステムにおけるスキーマ構築の一方式", 情報処理学会論文誌, 40(SIG8(TOD4)), 29-40, 1999.