WebNR/SD 異種情報源統合利用環境の研究 — メディエータの設計と開発 —

2Aa－7　根本 剛[†]　森嶋 厚行[††]　北川 博之[†††]

[†]筑波大学 理工学研究科 [††]筑波大学 工学研究科 [†††]筑波大学 電子・情報工学系

1 はじめに

近年，コンピュータネットワークの発達に伴い，異種情報源の統合利用が重要な課題となっている．各種情報源の中でも，構造化文書は World Wide Web(WWW) やディジタル図書，電子出版などにおいて多大な役割を果たしており，もっとも重要な情報源の一つである．構造化文書の利用が拡大されるに従い，構造化文書とデータベースの統合利用の必要性が高まっている．

我々は，WWW，構造化文書リポジトリ，リレーショナルデータベースを対象とした異種情報源統合利用環境の研究開発を行なっている [1][2]．本統合利用環境のアーキテクチャを図1に示す．ラッパーは各情報源の情報を統合データモデル WebNR/SD に変換する．メディエータは統合スキーマと操作系を提供し，ユーザは GUI を通じて統合利用環境を利用する．本稿では，本統合利用環境のプロトタイプシステムについて述べる．特に，メディエータの設計と開発を中心に説明する．

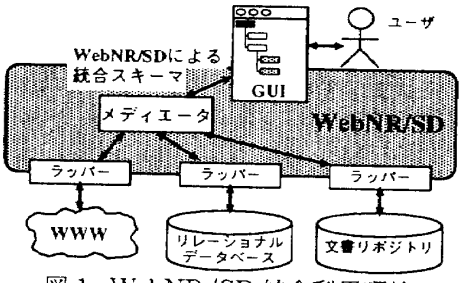

図 1. WebNR/SD 統合利用環境

2 WebNR/SD

我々が提案している統合データモデル WebNR/SD は，入れ子型リレーショナルモデルに構造化文書を扱うための抽象データ型 "構造化文書型"(SD 型) を導入し，構造化文書を SD 型の値 (SD 値) として扱う．SD 値は，文書構造を表す DTD(*Document Type Definition*) と，それに従ったタグ付きテキストから構成される (図2右)．

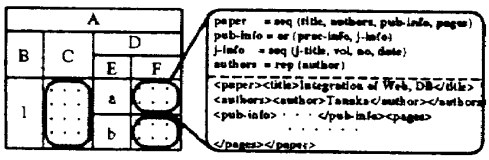

図 2. WebNR/SD のデータ構造 (右が SD 値)

WebNR/SD の特徴は，(1) 構造化文書と入れ子型リレーション構造を，動的，双方向，部分的に相互変換する演算子 "コンバータ" と，(2)WWW 上の情報を扱うための演算子群にある．

コンバータ

コンバータにより，同一のデータに対して入れ子型リレーショナル代数系と文書検索操作の両者が適用でき，次のようなことが可能である．(1) 入れ子型リレーショナル代数系を利用して，構造化文書データの構造変換操作等を行なう．(2) 入れ子型リレーション構造を構造化文書に変換し，型と独立した検索やメタデータを手がかりとした検索等を行なう．(3) 異なる構造のデータを適当な抽象度で抽象化して統一的に操作する．コンバータには Unpack と Pack がある．Unpack は SD 値中の要素群を値とする副リレーション構造を作成し，Pack は副リレーション構造中の要素を持つ SD 値を作成する．

WWW 上の情報を扱う演算子

WebNR/SD で WWW 上の情報を扱うためには，数ある WWW ページの中から操作対象となるページを選択し，それを SD 値として含むリレーションを作成する必要がある．そのために用意されているのが，操作対象ページを選択する Navigate 演算子と，実際にページに対応する SD 値を含むリレーションを作成する Import 演算子である．また，それとは逆にリレーション中の SD 値を WWW ページとして WWW の世界に作成する演算子 Export も用意されている．

3 統合利用環境プロトタイプシステム

本プロトタイプシステムでは，各モジュールを Java で記述する．モジュール間通信には HORB[3] を使用する．

GUI

ユーザとのやりとりを行なう GUI モジュールは，ユーザに対してブラウジングと問合せを融合したインタフェース (視覚的問合せ言語) を提供する．ユーザが入力した問合せを WebNR/SD 代数式に変換し，メディエータに送信する．また，問合せ結果をユーザに表示する．

メディエータ

各モジュール間の通信の要となる中心的なモジュールである．GUI から送信される問合せを各モジュールに割り振り，WebNR/SD に基づく演算を処理する．詳しくは次節で述べる．

ラッパー

プロトタイプシステムでは情報源としてリレーショナルデータベース，文書リポジトリ中の SGML 文書，WWW 中の XML 文書 [4] を想定しており，それぞれ RDB ラッパー，DR ラッパー，Web ラッパーを割り当てる．各ラッパーは，HORB を用いて各情報源のホスト上に配置される．各ラッパーは統一したインタフェースを持っており，メディエータはこれらの区別を意識することなく利用することができる．

RDB ラッパーは，WebNR/SD 代数式を SQL に変換し，データベースに問い合わせた結果を WebNR/SD に変換する．

DR ラッパーは，リージョン代数式に基づく問合せを処理する．問合せ処理の効率化をはかるために，構造化文書を問合せ処理に不必要な詳細部を抽象化した ASD 値 (*abstract SD value*)[1] に変換する．

Web ラッパーは，2 節で述べた WWW 上の情報を扱う演算を実行する．また，DR ラッパーと同様な ASD 値処理を行なう．

モジュール間のデータ通信

本統合利用環境のモジュール間でやりとりされる主要なデータ構造に，WebNR/SD モデルの入れ子型リレーション構造がある．実際には，それらにアクセスするため

Integration of Heterogeneous Information Sources with WebNR/SD — Design and Development of a Mediator —

Tsuyoshi Nemoto[†], Atsuyuki Morishima[††], Hiroyuki Kitagawa[†††]

[†]Master's Degree Program in Sci. and Eng., Univ. of Tsukuba

[††]Doctoral Degree Program in Eng., Univ. of Tsukuba

[†††]Institute of Info. Sci. and Elec., Univ. of Tsukuba

のハンドルがモジュール間を移動する(図3). 統合データを利用する側のモジュールは, ハンドルオブジェクトを通じてリレーションのデータを利用する. これにより, 以下の利点がある. (1) 一度に全てのデータを転送するのではなく, 必要になった時点で転送できる. GUIによる結果のブラウジング等では, 全てのデータが必要でない場合もある. (2) 問合せ処理の途中に結果を転送できる. 特にWebラッパーでは処理の対象となるドメインが巨大であるため, 全ての処理が完了するのに時間がかかる. あらかじめハンドルをメディエータに転送しておけば, メディエータはWebラッパーの終了を待つことなく, 処理の終ったものから順次利用することができる.

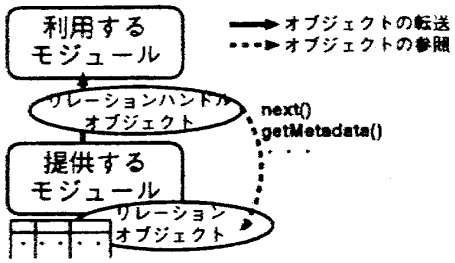

図3. リレーションとリレーションハンドル

4　メディエータの設計

3節で述べたように, メディエータは, ユーザの問合せ要求をGUIから受けとり, その要求を各ラッパー, またはメディエータ自身に振り分け, 各ラッパーからの問合せ結果をもとに, メディエータ自身が担当する演算の処理を行ない, GUIを通じてユーザに問合せ結果を提供する.

プロトタイプシステムでは, メディエータの演算処理過程で生成される中間リレーションの保持や, メディエータにおけるWebNR/SD演算機能を支援するためにInformix社のオブジェクトリレーショナルデータベース管理システムIllustraを用いる.

Illustraでは, ユーザが, 新しいデータ型とその型を扱う関数を定義することができる. この機能を用いて, WebNR/SDにおけるSD型とWebNR/SD演算処理を実現する. また, 通常のリレーショナルデータベースでは扱えない入れ子型リレーションはIllustraの集合型を用いて実現する.

メディエータはその働きから3つの部分に分けられる(図4). (分解部)GUIから送信されるWebNR/SD代数式を各モジュールが扱える演算に関する部分式群に分解する. メディエータは各モジュールオブジェクトがもつメソッドによってそのモジュールが扱える演算を知ることができ, それに基づいてWebNR/SD代数式を部分式群に変換する. (演算部) Illustraと通信し演算処理, 中間リレーション作成等を行なう. この際, 各ラッパーやメディエータ自身が作成した部分式に対する結果はリレーションハンドルを用いて扱う. Illustraとの通信は各演算に対応するC言語モジュールを介して行なう. (制御部) GUIから分解部への問合せ式の受け渡しや, 分解部によって生成した部分式を各ラッパーや演算部に振り分け, 結果を演算部に渡す. また, 最終結果をGUIに渡す.

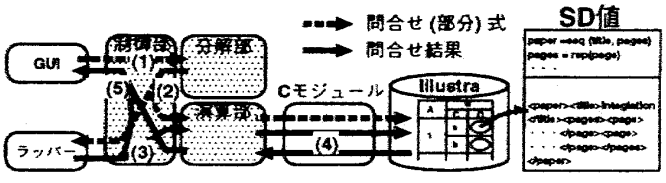

図4. メディエータモジュールの構成

5　メディエータ処理の流れ

ここでは, ある統合操作例を仮定した時の本プロトタイプにおける処理を, メディエータに焦点を当て説明する.

WWW上にT大学のホームページがあり, 講座別, 年度別論文リストにリンクがあると仮定する. また, リレーショナルデータベースにはT大学の教員リレーションがあり, 文書リポジトリには論文が構造化文書として格納されている. この時, 問合せ「論文概要を含む著者別論文リストページを作成し, それらに対する著者情報付きインデックスページを作成せよ.」に対する処理を示す. 以下の文中の番号のうち, (1) ～ (5) は図4中の番号に対応している.

(0) ユーザがGUIを用いて問合せを入力. (1)GUIは問合せをWebNR/SD代数式に変換してメディエータに送信. (2) メディエータはWebNR/SD代数式を各ラッパーが扱える演算からなる部分式群に分解して各ラッパーに送信. RDBラッパーには“教員リレーションから著者情報として必要なデータを取り出す”というリレーショナル代数式からなる部分式, DRラッパーには“T大学の教員が著者である論文データのうち, 論文概要を取り出す”という部分式, Webラッパーには“T大学のホームページからリンクをたどり, 講座別, 年度別論文リストページデータを取り出す”という部分式が送られる. (3) 各ラッパーはそれぞれの処理を行なうとともに, リレーションハンドルをメディエータに送信. (4) メディエータは受信したハンドルを用いてデータを受けとり, 演算を行なう. (4.1) コンバータを用いて論文リストデータ, 論文概要データを入れ子型リレーション構造に変換し, Join, Selection等による構造変換操作を行ない“論文概要付き著者別論文リストデータ”を作成, 再びコンバータを用いて構造化文書化する. これらは, Webラッパーを通じて実際のWWWページとして生成される. (4.2) 同様に“インデックスページ用著者情報データ”を作成する. コンバータを用いて構造化文書化した後, WWWページとして出力する. (5) インデックスページのURLをもつリレーションのハンドルをGUIに送信. (6)GUIが結果を表示.

6　おわりに

本稿では, 統合データモデルWebNR/SDに基づく統合利用環境のプロトタイプシステムの概要と, モジュール間通信の中心であり, WebNR/SD演算処理を行なうメディエータの設計について述べた.

今後はプロトタイプシステムの実装と評価, 問合せ処理の効率化, 最適化について研究を進めていく予定である.

謝辞

本研究の一部は文部省科学研究費補助金重点研究「高度データベース」の助成による.

参考文献

[1] A. Morishima and H. Kitagawa, “A Data Modeling and Query Processing Scheme for Integration of Document Repositories and Relational Databases,” *Proc. DASFAA'97*, pp.145-154, April 1997.

[2] A. Morishima and H. Kitagawa, “Integrated Querying and Restructuring of the World Wide Web and Databases,” *Proc. DMIB'97*, Nov. 1997.

[3] “HORB” HORB Home Page, http://ring.etl.go.jp/openlab/horb-j/WELCOME.HTM.

[4] “Extensible Markup Language (XML)”, World Wide Web Consortium, http://www.w3.org/TR/.