構造化文書とデータベースの統合利用方式の研究 — WWW 環境への適用のための拡張 —

5F－7

森嶋 厚行† 北川 博之††

† 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，ネットワークの普及に従って異種分散情報資源が容易に利用可能となり，それらの統合利用が重要な課題となっている．特に，WWW の普及により，WWW とデータベースの統合利用への要求が高まっている．WWW のページは 通常HTML 文書として記述されている．HTML 文書ではDTD (Document Type Definition) と呼ばれるタグ付け規則が固定されているが，DTD を文書ごとに指定できる XML[3] の制定が現在進行中である．XML 文書では，一般の SGML 文書と同様にタグを用いて文書の論理構造をより直接的に表現することができるので，文書の論理構造を利用したデータ操作が容易となる．

我々は，構造化文書リポジトリとリレーショナルデータベースを対象とした，統合利用環境の研究開発を行なっている [1][2]．この統合利用環境では，各情報資源を統合データモデルである NR/SD+ に変換し，その上で統合操作を行なう．NR/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) を導入したものであり，各構造化文書は SD 型の値として扱われる (図1)．NR/SD+ は通常の入れ子型リレーショナル代数演算子の他に，NR/SD+ の特徴である "コンバータ" を持つ．コンバータは構造化文書と入れ子型リレーション構造を，動的かつ部分的に相互変換するための演算子である．これらの演算子を組み合わせることにより，構造化文書とリレーショナルデータベースの統合利用，構造化文書の構造変換やビューの作成などが可能である．

A	B
1	⟨ b:**seq**(c, d:**rep**(c)), "<c>T1</c><d><c>T2</c><c>T3</c></d>" ⟩

図1. SD 値を含むリレーション r_1(属性 B が SD 型)

本稿では，WWW を本統合利用環境の情報資源の一つとして利用するために必要な，NR/SD+ の拡張について述べる．XML に従って，WWW のページごとに DTD が指定できると仮定する．WWW は通常の文書リポジトリと同様に構造化文書の集合体であるにも関わらず，NR/SD+ を用いた統合利用環境で，通常の文書リポジトリと同様に扱うことには以下のような問題がある．(1) WWW 中の全ての構造化文書 (WWW ではページと呼ぶ) 群を完全にリレーション構造の中に取り込むことは事実上不可能である．(2) NR/SD+ ではページ間のリンク構造を利用した操作が不可能である．NR/SD+ を拡張しこれらの問題点を解決した統合データモデルを WebNR/SD と呼ぶ．WebNR/SD を用いた統合利用環境を図2に示す．まず，ラッパーが各情報資源中のデータを WebNR/SD に変換し，メディエータが WebNR/SD に基づいた統合操作の機構を提供する．

WebNR/SD では，ページ間のリンク構造を扱うためのデータ型である Hlink 型を導入し，リレーション構造を通じた WWW ページ群の操作やナビゲーションを実現する．これらとコンバータを組み合わせることにより，WWW とデータベース，文書リポジトリ上にハイパーテキストビューを構築することが可能である．以下では，WebNR/SD の基となった NR/SD+ の特徴であるコンバータと，WebNR/SD への拡張について述べる．

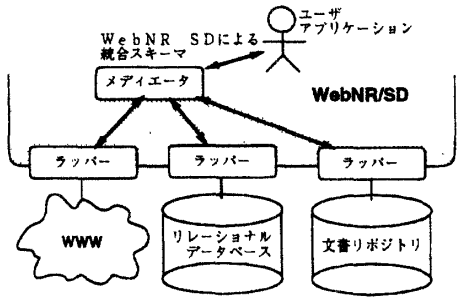

図2. WebNR/SD を利用した統合利用環境

2 コンバータ

SD 型の値 (SD 値) は，文書構造を表す DTD と，その DTD に従ったタグ付きテキストから構成される．テキスト中で同じ名前のタグで区切られた部分を要素と呼ぶ．コンバータは，SD 値と入れ子型リレーション構造の相互変換を行なう演算子である．プリミティブなコンバータとして，Unpack (**U**) と Pack (**P**) がある．Unpack は，構造化文書中の要素群を含む副リレーション構造を作成する．Pack は，副リレーションを構造化文書に変換する．図3のリレーション r_2 は r_1 に (1) 式を適用した結果である．

$$r_2 := \mathbf{U}_{B \to C(O, D[c \cap 1] \ as \ x)}(r_1) \quad (1)$$

$$r_1 := \mathbf{P}_{C(O, D \ as \ x) \to B}(r_2) \quad (2)$$

r_2 の属性 C の各副リレーションには，要素型 c を持つ要素のうち，任意の要素の第一副要素として現れるものを持つ SD 値が含まれる．r_2 に (2) 式を適用すると，各副リレーションを，r_2 の属性 B の SD 値をテンプレートとして構造化文書に変換し，r_1 を得る．

A	B	C	
		O	D
1	⟨ b:**seq**(c, d:**rep**(c)), "&x.1; <d>&x.2;<c>T3</c></d>" ⟩	1	⟨ c, "<c>T1</c>" ⟩
		2	⟨ c, "<c>T2</c>" ⟩

図3. リレーション r_2

コンバータを用いることにより，SD 値と副リレーション構造の変換の他にも，SD 値と入れ子型リレーションの属性部分列の相互変換が可能である．

3 WebNR/SD への拡張

WebNR/SD は，NR/SD+ にページ間リンクを扱う Hlink 型といくつかの演算子群を追加したものである．

Hlink 型

Hlink 型の導入に先立ち，まず SD 型の新たな要素構造として **hlink** 構造を導入する．**hlink** 構造を持つ要素は，要素属性 href を持ち，内部構造 (副要素) を持たない．次に，SD 値のうち **hlink** 構造を持つ要素ただ一つからなるものを，Hlink 型の値 (Hlink 値) と定義する．⟨ a:**hlink**, " T-Univ " ⟩ は，Hlink 値の例である．

Integration of Structured Documents and Databases
— Application to WWW Data Integration —

Atsuyuki Morishima† and Hiroyuki Kitagawa††

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

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

WebNR/SD 固有の演算子

Export (**E**) はリレーション中の SD 値の内容をもつページを WWW 中に作成し，そのページを参照する Hlink 値を得る．逆に，Import (**I**) はリレーション中の Hlink 値が参照する WWW 中のページを SD 値としてリレーションの中に取り込む．図 4 は以下の式に対する，Export と Import の実行例である．

$$r_4 := \mathbf{E}_{B,U,L,G}(r_3)$$

$$r_3 := \mathbf{I}_{B,U,L,G}(r_4)$$

A	B	U	L	G
1	⟨ e:seq(f:rep(g),h), "<e><f><g>T4</g></f> <h>T5</h></e>" ⟩	http://T.ac.jp/p1.xml	T Univ	a

A	B
1	⟨ a:hlink, " T Univ " ⟩

図 4. リレーション r_3(上) と r_4(下)

Navigate (**N**) はパラメータで指定されたパス正規表現に基づいて WWW のリンク構造をたどり，条件を満たすページ群を求めるものである．

$$r_6 := \mathbf{N}_{B(\rightarrow.)*\rightarrow C\rightarrow D,E}(r_5).$$

ここで，$B(\rightarrow .)* \rightarrow C \rightarrow D$ がパス正規表現，E は新たな属性名である．また，B は r_5 の Hlink 型の属性名と一致しなければならない．パス正規表現では，アルファベットとピリオドがページを，$\rightarrow$ がリンクをそれぞれ表す．* は 0 回以上の繰返しを表す．上式は，r_5 に新たな属性 E を追加したリレーション r_6 を生成する．属性 E は下位属性 C と D を持ち，リレーション値を格納する．このリレーション値の各タプルには，属性 B 中の Hlink 値が参照するページから始まるリンク構造をたどる各パスのうち，パス正規表現が受理可能なパスの C と D に対応するページを参照する Hlink 値が格納される．また，パス正規表現では，各ページの直後に "[ページの内容に関する選択条件]" を記述可能である．

URL generator $\mathbf{URL}_U(r)$ は，r に属性 U を追加する．U には，他と重ならない新たな URL 群が格納される．これらは新たな WWW ページの作成に利用される．

4 WWW とデータベースの統合操作例

WWW 環境に DB 研究室のホームページが存在すると仮定する．その研究室は 3 グループに分かれており，各グループごとに年度別の公表論文リストページが用意されている (図 5(a))．論文リストページの要素 "pl-title" には文字列 "paper-list" が必ず含まれており，かつ各論文の書誌情報はそれぞれ図 6 で定義される要素 "paper" として表現される．一方，リレーショナルデータベースにはその研究室の構成員リレーション "Member" が格納されている (図 5(b))．WebNR/SD を用いた統合利用環境における統合スキーマでは，WWW は Hlink 型の属性を持つ単項リレーションとして表現される．このリレーションには，問合せの手がかりとなるページ群への Hlink 値が格納される．この例では，DB 研究室のホームページを参照する Hlink 値が格納される (図 7)．この時，問合せ **Q1** に対する WebNR/SD によるデータ操作を示す．

Q1. DB 研究室の各構成員別に論文リストの WWW ページを作成せよ (図 5(c))．また，それらに対するインデックスページを作成し，そこには，各ページへのリンクに加えてリレーション "Member" 中の情報を格納せよ (図 5(d))．

$$r_7 := \mathbf{I}_{D,U,L,G}(\mu_C(\mathbf{N}_{Page(\rightarrow.)*\rightarrow D[\rho(D,\sigma['paper\text{-}list'](pl\text{-}title))],C}(WWW)))$$

$$r_8 := \mu_{As}(\mathbf{U}_{Authors\rightarrow As(O_2,Author)}(\mathbf{U}_{Paper\rightarrow(Authors,Title,Pub\text{-}Info,Pages)}(\mu_P(\mathbf{U}_{D\rightarrow P(O_1,Paper)}(r_7)))))$$

$$r_9 := \pi_{Name,Addr,Tel,Author,Title,Pub\text{-}Info,Pages}(r_8 \bowtie_{Author=Name} Member)$$

$$r_{10} := \mathbf{P}_{P(O_3:o,Pr)\rightarrow Papers:RC}(\nu_{P=(Pr)}(\mathbf{P}_{(Title,Pub\text{-}Info,Pages)\rightarrow Pr:SC,'paper'}(r_9)))$$

$$r_{11} := \mathbf{P}_{(Author,Papers)\rightarrow P\text{-}List:SC}(r_{10})$$

$$r_{12} := \mathbf{E}_{P\text{-}List,U_1,'papers','a'}(\mathbf{URL}_{U_1}(r_{11}))$$

$$r_{13} := \mathbf{P}_{Members(O_5:o,Member)\rightarrow Table:RC}(\nu_{Members=(Member)}(\mathbf{P}_{(Name,Addr,Tel,P\text{-}List)\rightarrow Member:SC}(r_{12})))$$

$$r_{14} := \mathbf{E}_{Table,U_2,'root','a'}(\mathbf{URL}_{U_2}(r_{13}))$$

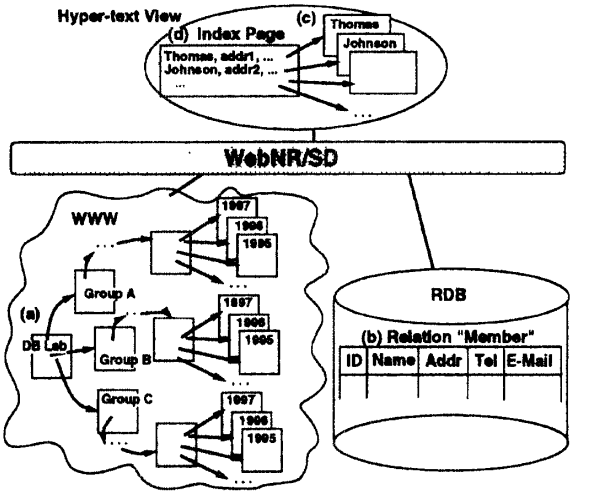

図 5.WWW とデータベース上のハイパーテキストビュー

```
paper       = seq(authors, title, pub-info, pages)
pub-info    = or(proc-info, j-info)
proc-info   = seq(proc-title, opt(location), date)
j-info      = seq(j-title, vol, no, date)
authors     = rep(author)
```

図 6. 論文リスト中の書誌情報の要素定義

ID	Name	Addr	Tel	E-Mail

Page
⟨ a:hlink, "" ⟩

図 7. 統合スキーマ ("Member"(上) と "WWW"(下))

5 おわりに

本稿では，構造化文書とデータベースの統合利用モデル NR/SD+ を拡張し，WWW を情報資源のひとつとして統合利用可能とした WebNR/SD について述べた．現在，WebNR/SD に基づく統合利用環境のプロトタイプシステムを開発中である．

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*, April 1997.

[2] 森嶋厚行, 北川博之, "参照の導入による構造化文書とデータベースの統合操作の検討," 第 113 回情報処理学会データベースシステム研究会, 1997 年 7 月.

[3] "Extensible Markup Language (XML)," World Wide Web Consortium Working Draft, http://www.w3.org/pub/WWW/TR/WD-xml.html.