構造化文書とデータベースの統合利用方式の研究

5F－5 — 文書構造の多様性に対応するためのエレメント・フォーカシング機能 —

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，異種情報資源の統合利用が重要な課題となっている．各種情報資源の中でも，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資源の一つである．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，SGMLを想定した構造化文書とリレーショナルデータベースを対象とした異種情報資源の統合利用環境の研究開発を行なっている[1][2]．我々が提案している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，構造化文書と入れ子型リレーション構造の動的変換を利用したデータ操作を可能にしたものである．

構造化文書から入れ子型リレーション構造への変換等を行なう際，NR/SD+では，文書構造情報に基づくパラメータを指定する．従って，ユーザは文書構造をある程度知っている必要がある．しかし，一般には構造化文書の文書構造は要素定義が数十～数百行の大規模かつ複雑なものであり，ユーザがその構造を完全に把握することは困難である．さらに，異なる構造を持つ文書が複数存在する状況ではユーザの負担は多大なものとなる．

本稿では，構造化文書のテキスト要素に対するフォーカシング機能の概念を提案する．これは，ユーザが文書構造の全てに興味があるわけではないことに注目し，ユーザが興味を持つ要素のみに焦点をあて，文書の一部を隠蔽することでユーザによる文書構造の認識を助け，問合せ操作を支援するための枠組である．以下では，統合利用モデルNR/SD+とフォーカシング機能について述べる．

2 NR/SD+

NR/SD+は，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，そのDTDに従ったタグ付きテキストから構成される(図1)．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．

NR/SD+の特徴は，構造化文書と入れ子型リレーション構造を，動的，双方向的，部分的に相互変換する演算子"コンバータ"にある．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．

```
report  = seq(title, authors, body, ref)
authors = rep(author)
body    = rep(chapter)
chapter = seq(chaptitle, rep(section))
section = seq(sectitle, rep(para))
   ...
ref     = rep(refitem)
refitem = seq(title, authors)
   ...
```

```
<report><title>Element Focusing</title>
<authors><author>...</author>...</authors>
<body><chapter><chaptitle>prolog</chaptitle>
                    ....
```

図1. SD値の例(一部)

次式はUnpack(**U**)とPack(**P**)の適用例である．結果は図2に示す．

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

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

A	B
1	⟨a:seq(b,c:rep(b)),"<a>d1<c>d2d3</c>"⟩

A	B	C	
		O	D
1	⟨a:seq(b,c:rep(b)),"<a>d1<c>&x.1;&x.2;</c>"⟩	1	⟨b,"d2"⟩
		2	⟨b,"d3"⟩

図2. コンバータ**U**/**P**の適用例 r_1(上) と r_2(下)

r_2の属性Cの副リレーションには，(1)式のUnpackで指定しているリージョン代数式 $b \subset c$ に適合した要素，すなわち要素bのうち要素cに含まれるもの，を持つSD値が含まれる．r_2の属性BのSD値中の&x.i; をSDリファレンス(*SD reference*)と呼ぶ．(2)式のPackは，r_2の各副リレーション構造を，属性BのSD値をテンプレートとして構造化文書に変換する．Unpackを用いた複合演算子としてExtract(**X**)がある．$\mathbf{X}_{B \to C(O,D[b \subset c])}(r_1)$の結果は$r_2$の属性Bの値の代わりに$r_1$の属性Bの値そのものが格納されたものである．

3 要素に対するフォーカシング機能

2節の例のように，Unpackによって抽出されるべき要素群は，文書構造に基づくリージョン代数式を用いて指定するので，ユーザは文書構造を把握している必要がある．しかし，文書構造が大規模かつ複雑な場合，ユーザによる文書構造の把握は困難である．この問題を緩和するために，ユーザの興味に応じて文書の一部を隠蔽する演算子Focusing(**F**)と，隠蔽を解除する演算子Unfocusing(**UF**)を導入する．**F**はSD値をFSD値に変換する演算子である．FSD値とは，SD値の要素のうち，ユーザが興味を持たない要素に特別な印"マスク"を付けたものである．**UF**はFSD値からマスクを除く演算子である．

FSD値

まず，SD値中に注釈文字列を導入する．注釈文字列は"<!--"と"-->"に囲まれた任意の文字列であり，SD値中の任意位置に出現可能である．次に，FSD値(*Focused SD value*)を，"<!--#-->"という特殊な注釈文字列を要素に持つSD値として定義する．この注釈文字列をマスク(*mask*)と呼ぶ．図3のr_3の属性BのSD値はFSD値である．マスクは，その直後の要素がユーザにとって興味

Integration of Structured Documents and Databases — Element Focusing to Cope with Complex Document Data Structures —
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

がないことを表す. Unpack の適用時には, マスク直後の要素中の全リージョンがないものとして扱われる. (3) 式は FSD 値を含むリレーション r_3 に Unpack を適用した例である.

$$r_4 := \mathbf{U}_{B \to C(O,D[b] \text{ as } x)}(r_3) \quad (3)$$

A	B
1	⟨a:seq(b,c:rep(b)),"<a>d1<c>d2<!--#-->d3</c>"⟩

A	B	C	
		O	D
1	⟨a:seq(b,c:rep(b)),"<a>&x.1;<c>&x.2;<!--#-->d3</c>"⟩	1	⟨b,"d1"⟩
		2	⟨b,"d2"⟩

図 3. FSD 値の例 r_3(上) と r_4(下)

Focusing と Unfocusing

Focusing($\mathbf{F}_{Attr,e}(r)$) は, リレーション r の属性 $Attr$ 中の SD 値を FSD 値に変換する. リージョン代数式 e でユーザの興味対象となる要素を指定する. $\mathbf{F}$ は, SD 値中の要素のうち, リージョン代数式によって指定された要素を含まない, もっとも上位の要素の直前にマスクを挿入する (図 4). 図 4 では要素の階層構造を木構造で表す.

Unfocusing($\mathbf{UF}_{Attr}(r)$) は, FSD 値を SD 値に変換する演算子である. これは, リレーション r に属性 $Attr$ の属性値として格納されている FSD 値からマスクを消去する.

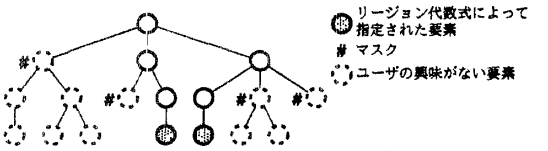

図 4. $\mathbf{F}$ によるマスクの付加

次式は $\mathbf{F}$ と $\mathbf{UF}$ の適用例である. 結果は図 5 に示す.

$$r_6 := \mathbf{F}_{B,b \subset c}(r_5) \quad (4)$$

$$r_5 := \mathbf{UF}_B(r_6) \quad (5)$$

A	B
1	⟨a:seq(c:rep(b),d:rep(b)),"<a><c>t1t2</c><d>t3t4</d>"⟩

A	B
1	⟨a:seq(c:rep(b),d:rep(b)),"<a><c>t1t2</c><!--#--><d>t3t4</d>"⟩

図 5. $\mathbf{F}$ と $\mathbf{UF}$ の適用例 r_5(上) と r_6(下)

統合利用環境における FSD 値の利用

統合利用環境では, ブラウザは FSD 値のマスクを利用することにより, ユーザに不必要な要素を隠蔽したタグ付きテキストや抽象化された DTD を見せることができる. 図 1 の SD 値がリレーション r の属性 A に格納されているとし, ユーザは author に関するデータ操作を行ないたいとする. この時, ユーザは $\mathbf{F}_{A,author}(r)$ を行なうことにより SD 値を FSD 値に変換する. ブラウザはこの FSD 値に基づいて必要な DTD だけを見せることができる (図 6(上)). この抽象化された DTD を手がかりとして, ユーザは author に関する NR/SD+ 演算を行なう. 場合によっては, ブラウザが DTD を抽象化することにより, 細部の異なる複数の DTD の差異を吸収することができる.

```
report  = seq(unknown, authors, unknown, ref)
authors = rep(author)
ref     = rep(refitem)
refitem = seq(unknown, authors)
```

```
report  = seq(unknown, authors, unknown, unknown)
authors = rep(author)
```

図 6. FSD 値に基づく DTD 表示の例

4 フォーカシング機能の適用操作例

統合利用環境に文書リポジトリとリレーショナルデータベースが存在し, 統合スキーマが図 7 であるとする. "Faculty" は T 大学の教員情報を格納したリレーションである. また, "Doc" には図 1 の DTD を持つ T 大学のテクニカルレポートが SD 値として格納されている. このレポートの DTD が大規模でユーザにとって認識し難いものであるとする. この時, "教授によって書かれたレポート" を求めたい. この操作では, レポートの著者要素を "Doc" 中の SD 値から抽出する必要がある.

シナリオ 1

まず, ユーザは author について関係のない要素群を隠蔽するために $Doc_1 := \mathbf{F}_{Report,author}(Doc)$ を行なう. ブラウザは結果の FSD 値に基づき図 6(上) のような DTD をユーザに示す. すると, レポートの著者の要素は, 全 author 要素のうち ref に含まれる author を除いたものであることがわかる.

$$r_7 := \mathbf{UF}_{Report}(\pi_{Report}(\sigma_{F\_title='Prof'}(Faculty) \bowtie_{Name=Author} (\mu_C(\mathbf{X}_{Report \to C(O_1,Author[\underline{author-author \subset ref}])}(Doc_1)))))$$

シナリオ 2

ユーザがあらかじめ参考文献欄にも author が出てくると見当をつけ, $Doc_2 := \mathbf{F}_{Report,author-author \subset ref}(Doc)$ を行なう. ブラウザは結果の FSD 値に基づき図 6(下) のような DTD をユーザに示す. すると, レポートの著者の要素を抽出するためには, 単に author 要素を抽出すればよいことがわかる.

$$r_8 := \mathbf{UF}_{Report}(\pi_{Report}(\sigma_{F\_title='Prof'}(Faculty) \bowtie_{Name=Author} (\mu_C(\mathbf{X}_{Report \to C(O_1,Author[\underline{author}])}(Doc_2)))))$$

この場合, 実際には ref 以下の部分構造に author が存在するにもかかわらず, FSD 値のマスクにより NR/SD+ 演算では ref 以下の author 要素は無視されるので, ユーザはそれを意識する必要がない.

Faculty					
FID	Name	DID	F_title	Cource	Speciality

Doc
Report

図 7. 統合スキーマ例

5 おわりに

本稿では, ユーザの興味に応じて文書の一部を隠蔽し, ユーザによる文書構造の認識と問合せ作成を支援する枠組である, テキスト要素に対するフォーカシング機能の概念について述べた. FSD 値は, 本稿で述べたユーザの問合せ作成支援の他にも, [1] で導入した, 問合せ処理の最適化のための ASD 値 (*Abstract SD value*) の利用の手がかりとして用いることも考えられる. 今後は, NR/SD+ に基づく統合利用環境における, ユーザの問合せ作成支援についてさらに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 Reposi tories and Relational Databases," *Proc. DASFAA '97*, pp.145-154, 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.