異種情報源統合利用環境におけるインデックスサーバの活用

6K－1

森嶋 厚行　北川 博之
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が発達し，各種情報源の利用が容易になるにしたがい，それらの統合利用が重要な課題となっている．我々は，構造化文書リポジトリ，World Wide Web (Web)，リレーショナルデータベース (RDB) 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 (図1)．本環境では，ラッパーが各情報源を統合データモデル WebNR/SD [1] に変換し，メディエータが統合スキーマと操作系をユーザに提供する．ユーザは視覚的操作系を用いて本環境を利用する．

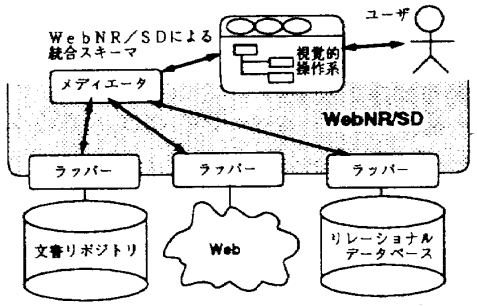

図1. WebNR/SD を利用した統合利用環境

本環境により，構造化文書，Web，RDB の統合利用が可能になる．例えば，これらの上に，リレーショナルビュー，構造化文書ビュー，ハイパーテキストビューを構築するといったことができる．

これまで，Web を統合するために，本環境では次のような機構を提供していた．すなわち，特定の Web ページ群を起点とするハイパーテキストリンクをたどり，条件を満たす Web ページやパスを探索する機構である．しかし，この機構は，ある程度探索の範囲が限定された状況では有効であるものの，次の問題点がある．(1) 起点となるページ群から到達不可能なページは探索することができない．(2) より広い範囲の探索を行う際に，ハイパーテキストリンクをたどるだけでは効率が悪い．

従来から Web では，インデックスサーバが広く利用されている．これは，あらかじめロボットなどを用いて Web ページの情報をデータベース化しておき，ユーザが単語群などの条件を投入すると，条件にマッチする Web ページ群を検索するものである．最近では，goo [2] のように，ある特定のドメインの範囲で，条件を満たす Web ページ群を検索するインデックスサーバも存在する．

本稿では，本統合利用環境にインデックスサーバを利用するための機構を導入することによって，上で述べた問題点の解決を図る．

2 WebNR/SD

WebNR/SD のデータ構造は，入れ子型リレーションに，抽象データ型である構造化文書型 (SD 型) と，ハイパーテキストリンク型 (Hlink 型) を導入したものである．図2では，属性 B が SD 型，属性 E が Hlink 型である．WebNR/SD のデータ操作系である WebNR/SD 代数では，入れ子型リレーショナル代数演算子と，SD 型に付随する文書検索関数に加え，統合利用のための演算子群を用意している．以下では，Web 統合のための演算子である，Import と Navigate について説明する．

A	B	C	
		D	E
abc	<table><dep> Department...	1	link
		2	...

図2. WebNR/SD のリレーション例

Import

Import (I) は，ハイパーテキストリンクが参照する Web ページの内容を，リレーション中に SD 型の値として取り込む．リレーション r_1 が Hlink 型の属性 A を持つ時，次式の結果 r_2 は，r_1 に属性 U,L,G を追加したものになる．r_2 の属性 A には，参照先の Web ページの内容が格納される．U にはハイパーテキストリンクが指していた URL が，L には，リンクのタグで囲まれていた文字列が，G にはそのタグ自身が，それぞれ格納される．

$$r_2 := \mathrm{I}_{A,U,L,G}(r_1)$$

Navigate

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 Web のリンク構造をたどり，条件を満たすリンクのパスの集合を求めるものである．リレーション r_3 が Hlink 型の属性 A を持つ時，次式が適用可能である．

$$r_4 := \mathrm{N}_{A \Rightarrow B[CAD] \| A \Rightarrow B(\rightarrow .)^* \rightarrow .[CAD],E}(r_3)$$

ここで，'$A \Rightarrow B[CAD] \| A \Rightarrow B(\rightarrow .)^* \rightarrow .[CAD]$' がパス正規表現，E は新たな属性名である．パス正規表現では，アルファベットとピリオドが Web ページを，$\rightarrow$ が同一の Web サーバ上にあるページ間のローカルリンクを，$\Rightarrow$ が異なる Web サーバ上にあるページ間のグローバルリンクを，それぞれ表す．各ページの直後には "[ページの内容に関する条件]" を記述可能である．このパス正規表現は，「属性 A のリンクが参照するページからグローバルリンクを一つたどり，そのページ B 自身もしくは B からローカルリンクで到達可能ないずれかのページに単語 "CAD" を含むようなパスの集合」を表す．上式は，r_3 に新たな属性 E を追加したリレーション r_4 を生成する．属性 E は下位属性 B 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，このパス正規表現が受理可能な各パスの B に対応するページを参照するリンクが格納される．

3 インデックスサーバの利用

本環境にインデックスサーバの利用機構を導入するにあたっては，2種類の利用を想定する．すなわち，(1) ユーザにインデックスサーバの存在が見える「Explicit な利用」と，(2) その存在が見えないが自動的に利用されている「Implicit な利用」である．(1) を実現するために，ユーザが明示的にインデックスサーバを利用するための Search 演算子を導入する．(2) では，Navigate 演算子の処理過程で，インデックスサーバの機能を活用することにより，処理の効率化を図る．

Utilization of Index Servers in an Integration Environment for Heterogeneous Information Sources
Atsuyuki Morishima and Hiroyuki Kitagawa
Institute of Info. Sci. and Elec., Univ. of Tsukuba

3.1 Search 演算子の導入

インデックスサーバを明示的に利用するために Search 演算子 (S) を導入する．図 3 は次式の実行例である．

$$r_6 := \mathbf{S}_{goo,[B \text{ and } C],D(O,E)}(r_5)$$

ここで goo はインデックスサーバの ID である．属性 D の副リレーションには，インデックスサーバの検索結果 (URL 群) が格納される．この例では，goo に対して問合せ "tsukuba AND database" が投入され，その結果が属性 D に格納される．

A	B	C
1	tsukuba	database

A	B	C	D	
			O	E
1	tsukuba	database	1	http://...
			2	...

図 3. リレーション r_5(上) と r_6(下)

3.2 インデックスサーバを用いた Navigate 処理の効率化

本環境では，Navigate 演算の処理は，次のように行なわれる．(1) パス正規表現から有限オートマトンを作成する．(2) ハイパーテキストリンクをたどりながら，このオートマトンに受理されるパスを探索する．例えば，Navigate のパス正規表現が $A \Rightarrow B[CAD]||A \Rightarrow B(\rightarrow .)* \rightarrow .[CAD]$ の時，図 4(左) のオートマトンが作成される．状態遷移は，グローバルリンクやローカルリンクが存在する場合や，ページ内容の条件を満たす場合に行なわれる．このオートマトンによるパスの探索は次のように行なわれる．まず，状態 A から開始し，起点となるページとは異なる Web サーバ Y 上に存在するページへのリンクを見つけて状態 B に遷移する．その後は Y 上の Web ページを探索しながら "CAD" を含むページに到達するパスが存在するかどうかを判定する．

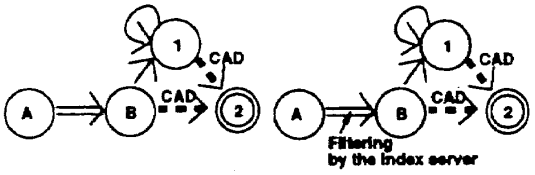

図 4. Navigate 処理のための有限オートマトン (左) とインデックスサーバによるパスのフィルタリング (右)

本手法による Navigate 処理の効率化は，指定ドメイン内でのページ検索が可能なインデックスサーバを利用して，オートマトンによる探索を行う前にパスのフィルタリングを行うものである．上の例では，図 4(右) のように，グローバルリンクによって状態 B への遷移を行う時にインデックスサーバを利用し，ドメインを Y の範囲に限定して単語 "CAD" を含むページの検索を行う．その結果，そのようなページが存在しない場合には，その先のパスは条件を満たさないものとして，オートマトンによる探索を省略する．

4 応用例

ここでは 2 つの応用例を示す．どちらの例も，Web と RDB の統合利用を行うために，インデックスサーバを活用している．**Q1** では，データベースの処理結果に基づいて，不特定多数の Web ページ群を検索範囲とした Web ページ検索を行う．**Q2** では，インデックスサーバを利用し，Navigate による Web の探索を効率化する．

Explicit な利用の例: 筑波大学の事務データベースに，リレーション Student(SID, NAME, DEPT) が格納されていると仮定する．SID は学籍番号である．この時，以下の問合せを行う．

Q1. 学籍番号が 935290 である学生に関する Web ページの URL を求める．

Q1 は次式で表現される．ただし，'U. Tsukuba'$_{(UNIV)}$ は，文字列型の属性 UNIV を持ち，値'U. Tsukuba' のみをもつ単項リレーションを表す．

$$\pi_{O,B}(\mu_A(\mathbf{S}_{goo,[NAME \text{ and } UNIV],A(O,B)}(\sigma_{SID=935290}(Student) \times' \text{U. Tsukuba}'_{(UNIV)})))$$

Implicit な利用の例: 大学の就職課のデータベースに，企業の就職推薦枠情報のリレーション Offer(CNAME, NUMBER) がある．ここで CNAME は会社名，NUMBER は推薦枠の人数である．ある学生が，就職課にやってきて，以下の問合せを行うとする．

Q2. 自動車関連の企業のうち，CAD をやっている会社の推薦枠とホームページの URL を求める．

この問合せは，Yahoo[3] 等のディレクトリサービスの「自動車関連企業」に載っている各企業について，その企業の Web サーバ上のページの中に，単語 "CAD" を含むページがあるかどうかを調べることにより実現する．次式では，自動車関連企業のハイパーテキストリンク群が，リレーション *Root* によって表現されていると仮定する．*Root* は Hlink 型の属性 A を持つ．

$$\pi_{CNAME,NUMBER,URL}(Offer \bowtie_{CNAME=LABEL} \mathbf{I}_{B,URL,LABEL,G}(\mu_E(\mathbf{N}_{A \Rightarrow B[CAD]||A \Rightarrow B(\rightarrow .)* \rightarrow .[CAD],E}(Root))))$$

パス正規表現のオートマトンは図 4 になる．パスのフィルタリングは図 5 のように行なわれる．インデックスサーバによって，Web サーバ Z は "CAD" を含むページを持たないことがわかるので，そこにいたるパスの探索は行なわれない．

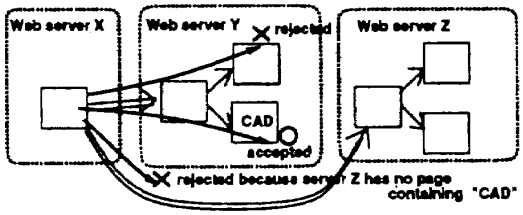

図 5. インデックスサーバを用いた Navigate 処理の例

5 おわりに

本稿では，異種情報源統合利用環境における Web インデックスサーバの活用について述べた．具体的には，インデックスサーバの Explicit な利用と Implicit な利用を行うための機構について述べ，その応用例を示した．今後の課題には，現在開発中のプロトタイプシステムにこれらの機構を実装することや，インデックスサーバをもちいた Navigate 処理の効率化の実験評価などがある．

参考文献

[1] A. Morishima and H. Kitagawa, "Integrated Querying and Restructuring of the World Wide Web and Databases," *Proc. DMIB'97*, Nov. 1997, pp. 262-271.
[2] "Goo ワーサーチ," http://www.goo.ne.jp/.
[3] "YahooJAPAN," http://www.yahoo.co.jp/.