HTML文書の個人ビュー生成のための操作言語の研究

1N－3

篠原 誠† 田島 敬史†† 田中 克己†

†神戸大学大学院自然科学研究科 ††神戸大学工学部

1 はじめに

現在のWWWシステムは、発信者が作成したページを他のユーザがそのままブラウズするだけの物であり、データベースシステムで行われるような、検索によって必要な情報だけを抽出し、それらを利用者の要求に応じて再構成できるような利用形態をあまり考慮していない。しかし、イントラネット上などで、組織体として保有するWeb文書群の共有と再活用を図るためには、利用者の側がWeb文書群を編集・カスタマイズし、Web文書群の「個人ビュー(personalviews)」を生成できれば非常に有用である。そこでこの研究では、HTML文書から個人ビューを生成する操作を定義するための操作言語の設計を行った。

この操作言語の特徴は以下の通りである。

- HTML文書に対して、タグの階層構造にしたがった検索、操作が行える。
- 既存のページから得た情報を個人ビューとして合成して再利用できる。
- 個人ビューの定義に必要なリンク先の変更、情報の追加・削除を定義できる。

2 基本的事項と関連研究

HTML(HyperText Markup Language)

HTML(HyperText Markup Language)とは、WWWの文書の文字やレイアウト等を設定するための言語で、SGML(Standard Generalized Markup Language)を基礎としたものである。またHTML文書はタグ付きの構造化文書であり、そのタグ情報がタグの入れ子構造においてどの深さのレベルに出現するかがあらかじめ特定できないという意味において「半構造化データ(semi-structureddata)」である。最近、このような半構造化データに関して活発な研究が開始されており[1]、中でも特に、Web文書の検索と再構造化(restructuring)を目的とした操作言語が多く提案されている[2,3,4]。

3 HTML文書の個人ビュー

HTML文書を、利用者個人がカスタマイズして「個人ビュー」を生成できれば、次のような用途に用いることができると考えられる。

- 既存のWeb文書(URL)集合からの動的な目次ページ(ディレクトリ)の生成
- 既存のWeb文書群に含まれるリンク情報の新規Web文書への再活用
- 組織で保有するマニュアル情報や電子教科書等のためのWeb文書群から利用者の必要な部分だけを適宜抽出し再構成した個人用の文書の作成

本研究では、前述のようにイントラネット上の電子教科書のような、ある程度利用範囲の狭められたHTML文書群を想定しているので、「個人ビュー」の生成は、基本的には既存の情報を抽出し再利用して行い、またその際に、利用者は利用する情報の存在するURLを知っているものと想定している。よってWeb文書群からの個人ビューを生成するために必要となる操作は、大別すると、URLで指定されたHTML文書内のある部分を抽出する操作と、それらを合成する操作である。更には以下の操作が考えられる。

- HTML文書の合成操作
  - 収集された文書データの連結
  - 必要な部分文書の追加、挿入
  - 不要な部分文書の削除

以下、これらの操作について詳しく述べる。

3.1 HTML文書の部分抽出操作

HTML文書の構造は、以下のような要素に分けて考えることができる。

- タグ
- タグ内に記述される属性
- タグに挟まれ表示される文字列

A language for Personal View Specificaton of HTML documents
Shinohara Makoto†, Keishi Tajima††, Katsumi Tanaka†
†Graduate School of Science and Technology, Kobe University
††Faculty of Engineering, Kobe University

よって、これらの要素を検索により特定し抽出する操作が必要である。前述のタグの構造より、タグ名を中心とし、これに属性名と、タグに挟まれている表示される文字列という構造があると考える。そこで、HTML 文書の構造を、図１のようなエッジラベル付グラフとして扱うこととした。また個人ビュー生成の際に、タグの順序が必要になるので順序付グラフとする。図で、*Visible* とされているのが、タグに挟まれ、ブラウザで表示される文字列である。HTML 文書をこのようにグラフで表現することにより、以下のようなパス表現を使った SQL-like な構文によって HTML 文書の特定部分を指定することにする。

セレクト文 ::=
　select パス表現 **from** URL **where** 検索パターン
パス表現 ::= ⟨タグ名⟩.⟨*Visible*⟩ | ⟨タグ名⟩
　　| ⟨タグ名⟩.⟨タグの中の属性名⟩
URL ::= 既知の HTML 文書の URL
検索パターン ::= ⟨タグ名⟩[変数](文字列)
　　| ⟨タグ名⟩.⟨タグの中の属性名⟩[変数](文字列)
　　| ⟨タグ名⟩.⟨*Visible*⟩[変数](文字列)
⟨タグ名⟩ ::= H1 | IMG | A | TABLE | ⋯
{タグの中の属性名} ::= src | alt | href | ⋯

セレクト文は、**from** 節で指定されたページから、**where** 節で指定された検索パターンにマッチする部分グラフを検索し、これらの部分グラフについて **select** 節に記述された部分を返す。検索パターンは、パス表現で記述された構造を持ち、かつその末端のノードの内容に与えられた文字列を含むものにマッチし各変数に対応する位置のノードを束縛する。例えば、以下の記述は、あるHTML 文書から、href 属性の値の中に “ac.jp” という文字列を含むアンカータグ以下の部分グラフの集合を返す。

select X
from http://.../index.html
where ⟨A⟩[X].⟨href⟩('ac.jp')

3.2　HTML 文書の合成操作

ここでは前述の方法で特定した HTML 文書の部分グラフを用いて、個人ビューとなる HTML 文書の合成操作について考える。

- 必要な部分文書の追加、挿入。
  Add(SELECT 文 1、SELECT 文 2)
  SELECT 文 1 と同一レベルの１つ前の順序の枝として SELECT 文 2 を挿入
- 必要な部分文書との置換。
  Rep(SELECT 文 1、SELECT 文 2)
  SELECT 文 1 を SELECT 文 2 と置き換える。from 節の要素は同じ。
- 不要な部分文書の削除。
  Del(SELECT 文)
  SELECT 文で指定した部分グラフを削除する。
- タグの効果をなくす。
  Mask(SELECT 文)
- アンカーを埋め込み、リンクを張る。
  Anc(SELECT 文 1、SELECT 文 2)

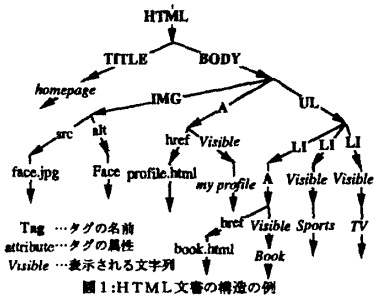

図１:HTML 文書の構造の例

4　おわりに

本研究では、HTML 文書中のタグによる内部構造をエッジラベル付きグラフとして扱い、個人ビューを定義するための操作言語を設計した。

今後の課題としては以下のようなものが挙げられる。

- HTML 文書中の他ページへのリンクをたどる操作を含む言語の設計
- システムの実装
- グラフィカルインターフェイスによる HTML 文書の直接操作を通して、検索操作文を生成するツールの開発

参考文献

[1] S.Abiteboul, *Querying Semi-Structured Data*, Proc. of ICDT'97, Vol.1186, LNCS, pp.1-18, Springer Veralg, Jan. 1997.

[2] D.Konopniki and O.Shmueli, *W3QL: A Query System for the World-Wide Web*, Proc. of the 21st VLDB Conf., Zurich, 1995.

[3] P.BUneman, *A Query Language and Optimization Techniques for Unstructured Data*, Proc. ACM SIGMOD'96, pp.505-516, 1996.

[4] M.Fernandez et al., *A Query Language and Processor for Web-Site Management System*, Proc. of Workshop on Management of Simistructured Data (in conjunction with ACM PODS/SIGMOD'97), 1997.