4Aa－5

WWWのハイパーテキストの構造を利用した検索機能の向上に関する研究*

尾山 譲一† 上原 徹三† 石川 知雄†
武蔵工業大学大学院工学研究科†

1 はじめに

現在，特定のWebサイトではWWW（World Wide Web）検索サービスをおこなっており，一般のWebブラウザから利用することができる．なかでも分野ごとに分類されているサイトのページは目的となるリンク先を見つけやすくたいへん便利である．これらの検索サービスのほとんどはそのサイト外のサイトにあるページへのリンクを提供しており，そのための専用のデータベースを保有している．[1]

ここで一般のWebサイトに検索システムを導入することを考える．一般のサイトではそのサイト内のページを紹介することが目的であると考えられるため，検索するリンク先をサイト内のページとすると，専用のデータベースを保有する必要もなく，利用者に全文検索サービスを提供することができる．また一般のサイトにおいても分類がなされている方が利用者にとって便利ではあるが，分野ごとの分類は手作業であるため一般のサイトの管理者にとってたいへん負担となる．

そこでこの負担を文書制作者に分散させる方法として，ハイパーテキストに含まれるマークを，分野を表す項目として文書制作者がハイパーテキストの構造上に新たに定義・記述し，その構造を検索に利用できる全文検索システムを一般のWebサイトに適用するための手法を提案する．

2 構造を利用するための手段

WWWで取り扱うハイパーテキストにはHTML（HyperText Markup Language）という書式が用いられており，Webブラウザにおける情報表示のための書式がマーク付けされている．HTMLは文書記述の標準規格SGML（Standard Generalized Markup Language：ISO8879）に従って定義された書式であり，その記述例は次のようになる．

<HTML>HTMLで記述された文書の簡単な例</HTML>

ここで<HTML>はHTMLの開始を，</HTML>はHTMLの終了を表すマークである．

SGMLの仕様には文書の論理構造を定義・記述するDTD（Document Type Definition）という文書のクラスの定義方法と，そのインスタンスとしての構造化された文書の記述方法とが含まれているが，現在のハイパーテキストにDTDは利用されていない．

ハイパーテキストの構造上に新たにマークを定義・記述し，その構造を検索に利用する上で，それら定義をまとめておく必要がある．そこでDTDを利用することが考えられる．以下にHTMLの書式を定義・記述したHTML DTDの一部を示す．

```
<!ENTITY % HTML.Version
                    "-//MIT//DTD HTML//JA">
 ...
<!ENTITY % body.content "(%heading | %text |
                %block | HR | ADDRESS)*">
<!ELEMENT BODY O O %body.content>
<!ELEMENT HTML O O "HEAD,BODY,PLAINTEXT?">
```

これを簡単に説明すると，まずこのDTDを参照するための識別子をHTML.Versionという実体に宣言している．途中は省略してあり，最後の部分ではHTMLの1つ下の階層にHEAD，BODYなどのマークを，BODYの1つ下にADDRESSなどのマークを定義している．

3 本手法における定義・記述方法

本手法では複数の文書制作者を対象としており，複数の文書構造に対応する必要があるため，1つのDTDでは対応することができない．しかし情報表示は一般のWebブラウザでおこなうためHTMLの書式は継承する必要がある．そこでHTML DTDのサブクラスとなる文書制作者レベルのDTDを用いる．このDTDの定義方法はSGMLに従う．

このDTDを記述する上で必要な事項として，まずシステムがDTDを識別するために，HTML.Versionという実体に識別子を宣言する．そしてDTDを継承する場合，それをシステムに認識させるために，DTD

*Improvement of Retrieval Function Using WWW Hypertext Structure
†Jyoichi Oyama , Tetsuzo Uehara , Tomo Ishikawa
Musashi Institute of Technology

の最後の部分で継承元の識別子を参照するための記述をおこなう．以下に新しくマークを定義した文書制作者レベルの DTD の例を示す．

```
<!ENTITY % HTML.Version
          "-//MIT//DTD ORIGINAL HTML//JA">
 ...
<!ELEMENT AUDIO - - (%text)*>
<!ENTITY % new.mark
           "(%computer | %music | AUDIO)*">
<!ENTITY % body.content "(%new.mark |
%heading | %text | %block | HR | ADDRESS)*">
<!ENTITY % html PUBLIC
                    "-//MIT//DTD HTML//JA">
%html;
```

ここでは，まずHTML.Version に新しい識別子を宣言している．次にBODY の 1 つ下の階層にAUDIO などのマークを新たに定義している．最後にhtml という実体を用いて，継承元の識別子を参照するための記述をおこなっている．この DTD のインスタンスであるハイパーテキストの例を以下に示す．

```
<!DOCTYPE HTML PUBLIC
          "-//MIT//DTD ORIGINAL HTML//JA">
<HTML> ...
<AUDIO>オーディオに関する記述</AUDIO>
 ... </HTML>
```

ここでは，まず参照元となる DTD の識別子を SGML に従い DOCTYPE 宣言を用いて最初の部分に記述している．本文ではオーディオに関する記述を挟む形で，AUDIO のマークを新たに記述している．

4 システムの概要

このシステムは，検索システムのページを中心に前後 2 つの処理段階に分かれている．図 1 にシステムの全体構成図を示す．

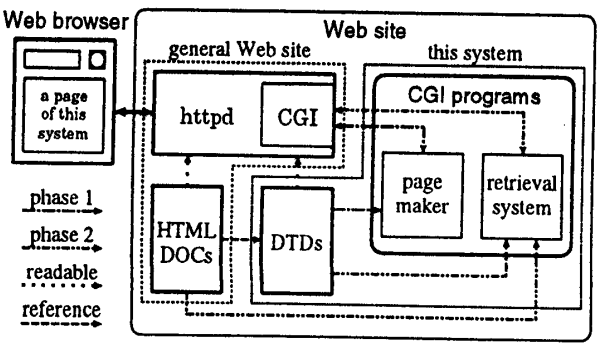

図 1. システムの全体構成図

まず第 1 段階として，Web ブラウザから CGI (Common Gateway Interface) プログラムである page maker というファイルへの接続要求を，WWW サーバである httpd に送信し，page maker に接続する．

page maker では，新たに定義されたマークを含む文書構造を各 DTD から取り出し，図 2 に示すような検索システムのページを生成する．

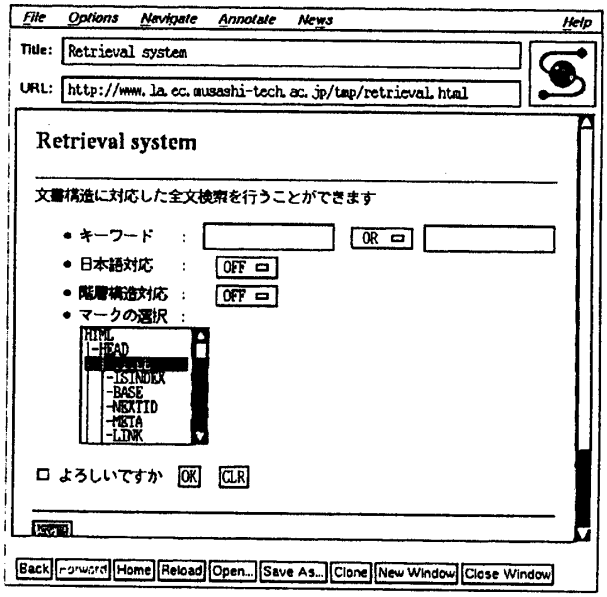

図 2. 生成された検索システムのページ

検索システムのページでは，利用者がキーワードの入力とマークの選択をおこない，検索を実行する．

そして第 2 段階として retrieval system というファイルに接続し，選択されたマークを定義している DTD のインスタンスであるハイパーテキストのみを対象とした全文検索をおこなう．文章とキーワードとの照合も選択されたマークの範囲内のみを対象とするため，すべての文書を対象とした全文検索や構造を利用しない全文検索にくらべ検索時間が短い．

最後に検索結果をハイパーリンクを用いて表示する．

5 まとめ

本手法により一般の Web サイトにおいて管理者の負担も少なく，利用者に検索範囲として分野を表した分類の明確なマークの選択がおこなえる全文検索サービスを提供することができる．

参考文献

[1] “WWW サーバー検索サービス国内で本格始動，量·質を競う”，日経コミュニケーション 1995.11.20 No.210