ＷＷＷサービスに対する構造をもった全文検索サポート

3Ｇ－9

川島宏仁、斎藤大輔、芝野耕司
（東京国際大学　商学部　経営情報学科）

1　はじめに

Mosaicにおける情報検索ではＵＲＬ（Uniform Resource Locator）を使った検索とハイパーテキストリンクによる情報検索を用いることができる。一般的にはこのリンク構造による情報検索が多く使われている。

ハイパーテキストリンクをもとにした検索では、テキスト中に埋め込まれたアンカーポイントクリックすることによって、関連する次のテキストに飛ぶことができ、極めて分かり易いユーザインターフェースを提供することができる。

一方、ハイパーテキストリンクを用いた情報検索では、webの迷路に落ち込み、たとえ目的の情報が見つけることができても、二度とその場所に行き着けない場合がある。

ＷＷＷを用いた多くのアプリケーションでは、この問題をＵＲＬを直接指定可能にすることによって、回避している。

しかし、最近のようにＷＷＷサービスが爆発的に普及してくると、この二つの仕組みだけでは、存在するであろう情報にたどり着くことは難しい。

また、たとえＵＲＬを用いて直接欲しい情報にたどり着けるとしても、新しい情報を得ようとすれば、やはりハイパーテキストリンクを一度はたどってその情報のありかにたどり着く必要がある。しかし、現在の“爆発的な”状況では、この方法すら実際的ではなく、ＷＷＷで提供されているサービスで、本来役に立たつであろうサービスであっても、なかなか見つけられないのが現状である。

この論文では、ＷＷＷで提供されている検索サービスを検討し、ＷＷＷで十分な情報検索サービスを提供することのできる方法を検討する。

すなわち、ＨＴＭＬの文書構造をサポートし、その構造を活かした全文検索をサポートするシステムを提案する。

2　現在の検索システムの問題点

一般の情報検索システムは、書誌情報及びキーワードをもとにした検索サービスを提供する。しかし、この検索方法では、書誌情報及びキーワードの選択に依存し、適切な情報がこれらの中で提供されない限り、目的の情報が文章中に含まれていても、その情報を検索することはできない。こうしたこれまでの情報検索システムの問題点は、数多く報告されている（例えば、［1］参照）。

一方、これまでの全文検索システムは、本来欲しい情報を見落とすことのある情報検索システムの問題は、解消されるが、ノイズにあたる関係のない情報も検索してしまうという問題がある（例えば、［1］参照）。

ＷＷＷにおける検策システム［2］についても、上に指摘した情報検索システム及び全文検索システムの問題点を克服していない。

事実、ＷＷＷでの情報検索を試みた報告［3］でも、同じ問題点を指摘できる。

［3］によれば、Netscapeの検索エンジンを用いた検索結果を報告している。この報告では、NetscapeのNetSearchを用いて、前ＮＨＫ会長の島桂次氏が作ったInternet上の週刊誌Shima Mecia Networkを「Shima」という検索語で探すと、ヒットした数がＷＷＷ catalog（検索エンジン）では６つあったが、その中には探している情報が見つからなかった。他の検索エンジンも調べて見たが同じようなものが見受けられた。ここで問題となるのがいくら「Shima」という文字を検索してきても探したい情報がない場合、また他の検索エンジンを使って探さなければならない。

逆に「Bookstore」という検索語で探すとヒットした数がWWW catalog では３０もあった。他のエンジンに行けば１００以上もあるところがある。こうなるとその中から見たい情報を探すのに時間がかかる。これではいつまでたっても見たい情報を引き出すことはできない。

すなわち、これまでの一般の情報検索方法及び全文検索法では“適切な”情報を得ることは難しい。

3　ＷＷＷ上での構造を考慮した全文検索

ＷＷＷサービスでは、HTML［4］で規定されるタグを用いて文書のマーク付けをおこなう。HTMLは、SGMLの文章構造を継承している。すなわち、HTML文書にも、SGML文書に対する検索と同様の構造を活かした検索を行うことができる。

HTML文書に適用可能な構造は、ＳGMLと、同様に、(1) グルーピング、(2) 連結子　(3) 出現標識及び(4) 識別子参照の基本的には、四つである。このＳGML文書ベースに対する文書ベース言語としては、DQL[1]などがあるが、HTMLという一つのDTDインスタンスを前提とすると、より簡略な検索で効

Structured Full Text Search Support for WWW
Hirohito Kawashima,Daisuke Saitou,Kouji Shibano
Tokyo International University

果的な検索が可能である。

すなわち、HTML文書でこれまでの情報検索及び全文検索の問題点を克服するためには、文書構造の包含関係（SGMLのグルーピングで指定される構造）を活かした検索を想定することができる。この検索とは、HTMLタグをもとにしたテンプレイトをインタフェースとして用いることによって、より適切な検索をより簡単に行うことができる。

主なHTMLタグの説明

<TITLE>～</TITLE>ページの表題
<Hy>～</Hy>章、ヘッディング
（yには１～６までの数字（階層）が入る）
<A>～</A>情報へのアンカー
<UL>～</UL>番号なしリストを宣言
<LI>～</LI>箇条書き
<OL>～</OL>番号付きリストを宣言
<LI>～</LI>
<DL>～</DL>説明付きリスト
<DT>～</DT>
<TT>～</TT>等幅タイプライタフォント
<B>～</B>ボールド
<I>～</I>イタリック
<U>～</U>アンダーライン
<IMG>インライン画像の表示
<P>文節区切りや改行

このテンプレイトベースの検索では、HTMLタグによって識別される文書要素の包含関係を検索に取り入れる。すなわち、章に対する検索指定では、章に含まれる節、段落などのすべての要素を検索対象として指定することになる。

一方、現在のSGML規格には、含まれていないが文書要素には、文書要素クラスを想定することができ、同時に、この文書要素クラスには、クラス階層を想定することができる。例えば、タイトルのすべての階層を一般化したタイトル階層を想定することは、意味的には、極めて自然なことであるが、このようなクラス階層に関しては、現行のSGMLには、含まれていない。ここでは、このクラス階層を活かした検索を可能にする。すなわち、タイトルクラスの下位クラスとして、章・節などのタイトルを想定し、タイトルという上位のクラス階層に対する検索指定は、すべての下位のクラス階層に適用する。

このように、ここで提案する検索システムの特徴を簡単にまとめると、次に三つになる。

（１）HTML（SGML）の文書要素の包含関係を活かした検索指定
（２）クラス階層の概念を導入した検索指定
（３）HTMLタグをもとにしたテンプレイトの提供

これらを導入することによって、より簡単にかつより効果的な検索の指定が可能となった。

4　残された課題

ここで検討してきた検索方法では、従来の情報検索及び全文検索システムの問題点の克服をHTMLを前提に検討してきた。

しかし、WWWは、もう一つ本質的な問題を提起している。すなわち、分散情報提供の問題である。

ここで提案したシステムの検索では、基本的には、WWWで提供される情報を収集し、一つのシステム上にデータベースを構築することを前提としている。この情報収集は、一時点に限って考えれば、かなりの程度網羅的に行うことが可能である。しかし、たとえある時点で網羅的な情報を収集したとしても、それぞれのサイトで提供される情報と、収集したデータベースの情報との整合性を保つことは、ほとんど不可能に近い。

今後は、HTTPの拡張を含めて、データベースと実際に提供される情報の整合性を採る方法を検討する必要がある。

参考文献

[１]　猪瀬他、文献の構造に基ずく全文データベースシステムの開発研究（研究課題番号02558007）平成４年度科学研究費補助金（試験研究（R））研究成果報告書、学術情報センター、１９９３－３
[２] WWW catalog
(URL：http://cui _www.unige. ch/w 3 catalog)
[３] 日経バイト、バイトセミナー　1994-1
[４] 初心者向きＨＴＭＬガイド
copyright 1993,1994 by the Boarc of Trustecs of the University of Illinois,but wegrant permission to freely disribute the document, provided you include this copyright
[５] 日本規格協会、JIS X 4151文書記述言語SGML、1993