6L－6

クライアントベースのWWW文書検索の検討*

田中 久士† 上原 徹三† 石川 知雄†
武蔵工業大学大学院 工学研究科 電気工学専攻†

1 はじめに

WWW上で公開されている情報は，ユーザの意志で自由にアクセスでき，情報を収集することができる．この情報収集[1]にはWWWクライアント(Netscape, Mosaicなど)を利用して行われ，情報を収集している際に，ユーザが興味のある(求めていた)情報を見つけた場合，もう一度同じ手順でアクセスする手間を考えたら，その情報を保存するというのは，当然の流れであろう．

しかし，保存した情報が増えれば，ユーザ自身が保存した情報であっても，どのファイルに何の情報が記載されているのか，探すのが困難になる．このことを解決するためには，保存したファイルの中から，ユーザが見たい情報を探し出す検索システムが必要となる．

また，現在，WWW上をリアルタイムに全文検索するシステムは，最新の情報が入手できるという反面，検索時間がかかり，検索するWebサイトのURL(Uniform Resource Locator)は，ユーザの入力に頼っている．これらを解決するのに，ローカルに保存した文書を検索した結果を利用する方法が考えられる．

そこで本研究では，ローカルに保存された文書をクライアント側で検索する機能とその結果を基にWebサイト上の文書を検索する機能の二つを融合した検索システムを提案し，その有効性を確認する．

2 システム概要

システムの全体構成は，WWW文書の蓄積，ローカルに保存された文書の検索(以下，これをローカル検索と呼ぶ)，ローカル検索の結果をもとにサイト上の文書に対して実行する検索(以下，これをサイト検索と呼ぶ)，の3つの機能から成り立つ．

ローカル検索は，ローカルに保存したWWW文書を検索するために必要である．また，この保存された文書だけでは，データベースとして小規模であるので，サイト検索を行い，データ量の少なさを補う．

上の機能の利点は，1. 検索エンジンは小規模なものでよく，システム負荷が少なくてすむ，2. 検索範囲の削減による検索時間(検索を開始し，結果が表示されるまでの時間)の短縮[2]，3. ユーザが取り込んだ情報を対象としているので，ユーザにとって興味のある情報が得られる，が挙げられる．

3 本検索システムの機能

3.1 WWW文書の蓄積

蓄積を行いたいファイルを見つけたらローカルに置いてある蓄積用のJava Appletを実行する．実行するとURLは画面に表示され，ユーザが入力したディレクトリに指定されたファイル名でファイルをダウンロードする．この時，ファイルの先頭にHTMLタグのコメントとしてダウンロードもとのURLが埋め込まれる．

3.2 ローカル検索

ローカル検索は，保存したファイルを検索する機能である．検索を実行する場合は，まず，ローカルに置いてあるJava Appletを実行する．

ユーザは，検索キーワード，検索を行うディレクトリ名，検索条件を入力する．検索方法は，指定した検索キーワードを含む文書を探し出す検索である．

検索を行う際，全文からの検索だけでなく，検索範囲を限定した検索を行えるようにした．この検索範囲を指定するのが検索条件である．用意した検索条件を以下に示す．

TITLE ···TITLEタグの開始タグから終了タグに囲まれた部分の検索．

HEADLINE ···H1からH6タグの開始タグから終了タグに囲まれた部分の検索．

ITEM ···DL, UL, OLタグの開始タグから終了タグに囲まれた部分の検索．

TABLE ···TABLEタグの開始タグから終了タグに囲まれた部分の検索．

ADDRESS ···ADDRESSタグの開始タグから終了タグに囲まれた部分の検索．

ANCHOR ···Aタグの開始タグから終了タグに囲まれた部分の検索．

FULL ···これはHTMLタグと関連はなく，全文検索．

*Client-Based WWW Document Retrival

†Hisashi Tanaka , Tetsuzo Uehara , Tomoo Ishikawa
Department of Electrical Engineering , Research Division in Engineering , Musashi Institute of Technology

以上をもとに検索を行う．検索結果として，検索でヒットしたファイルへのリンク，文書が存在したURL，キーワードに対する得点を表示する．最後のキーワードに対する得点とは，検索でヒットした文書からリンクされ検索キーワードと関連のある文書が存在するかどうかの可能性の強さを示す．例えば，TITLEタグの間に検索キーワードが見つかった場合，TITLEタグ以下の文書には，キーワードに関係する内容が書かれている可能性が高いので，アンカータグがあれば，このファイルに対して得点を与えるといったものである．この得点をもとにサイト検索に移る．得点付けの方法を以下に示す．

(1) 検索条件が **TITLE** の場合 ···TITLEタグの終了タグ以下の文書中にアンカータグがあれば2点を与える．

(2) 検索条件が **HEADLINE** の場合 ···次に別の見出しタグのような区切りがある所までに，アンカータグがあれば3点を与える．

(3) 検索条件が **ITEM** の場合 ···<OL>と<UL>の場合，検索キーワードが見付かった項目 (<LI>) と同じ行にアンカータグがあれば，2点を与える．<DL>の場合，見出し (<DT>) に検索キーワードがあった場合，説明文 (<DD>) の中にリンクがあれば3点を与える．

(4) 検索条件が **ANCHOR** の場合 ···検索キーワードが見付かった時点で5点を与える。

(5) 検索条件が **FULL** の場合 ···上のもの全てを考慮に入れた得点付けを行う．

(6) その他 ···検索条件がどれであっても，検索キーワードに対してアンカータグがついている場合は，5点を与える．

以上のようにして計算した結果をローカル検索でヒットした個々のファイルに対する得点として表示する．

3.3 サイト検索

この検索は，保存したファイルが更新されたのかどうかの判定や新たな情報の検索に役立つ．サイト検索では，CGIを使用する．CGIを使用するには，HTTPサーバにアクセスし，検索システムに接続する必要がある．

サイト検索のためにユーザが何かを入力する必要はない．検索キーワードは，ローカル検索で入力した値，検索条件は全文検索が選択される．ユーザは，ローカル検索の結果の中から一つのURLを選ぶ．このURLに存在する文書を中心文書と名付ける．サイト検索はこの文書を中心に実行する．サイト検索の実行過程を図1に示し，説明を以下に述べる．

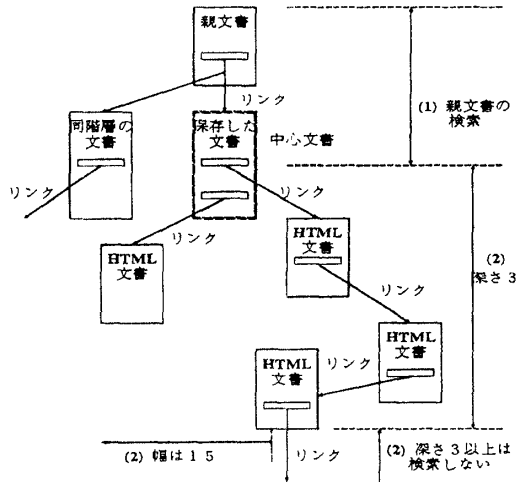

図1: サイト検索の実行過程

まず，中心文書の検索を行う．次に検索するサイト内で，中心文書の親文書の検索を行う．親文書とは，中心文書にリンクをしている文書のことである．次に中心文書と同じ階層にある文書の検索を行う．同階層の文書とは，親文書が存在する場合，その中からリンクされている文書のことである．次に中心文書からリンクされている文書を検索する．検索する文書数を約50と決め，中心文書から測って深さは最大3そして幅を15までとし，システム側で制限を設ける．このような制限を設けなければ，検索に時間がかかり，ネットワークに多大な負荷をかけることになる．最後に結果をWWWクライアントに表示する．その表示内容は，検索キーワードが出てきたファイルへのリンク，そのファイルのタイトルを表示する．

4 まとめ

本手法による検索システムでは，ローカルに保存された文書を中心に検索が行われるため，情報の絞り込みが容易である．そして，ローカル検索の結果は，効率よく検索できるサイトを示唆する．しかも，ローカル検索でのデータ量の少なさをサイト検索が補える．

本システムの適用先としては，仲間が取込んだファイルを互いに参照することが意義を持つような共通の関心を持つ複数人のグループ，例えば，大学の研究室や企業の部署などが有効と思われる．

参考文献

[1] 高田:“インターネット上の情報の集め方”bit. vol.27.No.2, p4-12(1995)

[2] 石上，箱崎:“WWW上の個人用情報検索システムの提案”，情報処理学会第52回全国大会(1996)