技術ノート

分散するWebコンテンツの効果的活用技術

山田一徳*　　土屋敏夫*

Technology for effective and practical use of the dispersed Web contens

１．はじめに

通常，情報は一元管理することが望ましいが，コンピュータネットワークが広範囲に渡るにつれ，データは分散化し，それが困難になってきている。そのため蓄積されている情報が十分に活用されてない傾向がある。

このことはWebのコンテンツにも該当する。分散するWebコンテンツを効果的に活用する一般的な方法として，ひとつに<A HREF>タグを利用したリンクがあげられる。しかし，これのみではコンテンツの量やディレクトリの階層の増加に対応できなくなる。例えば，閲覧者が欲する情報が掲載されているページまで到達しにくくなることなどがあげられる。

この問題を解消するために，「検索エンジン」の導入があげられる。<A HREF>タグを利用しただけのリンクよりは有効であるが，検索用のデータテーブルが必要となり，その生成はWebサーバが分散するほど煩雑さを増す。また，Webのコンテンツ数が多くなるにつれ，それらを１つずつマニュアルにて取得し，検索用のデータテーブルを作成することは不可能となる。このような課題を解決できるのであれば，分散したWebコンテンツを効果的に活用するひとつの手段として，検索エンジンは有効である。ここでは，この課題解決として次の①から④までの手法を検討した。

① 検索用のデータテーブルをあらかじめ生成し，それを元にユーザによる検索が行う。

② 検索により該当したページを表示する場合は，実際のURLへリンクさせる。

③ 「②」より，検索用のデータテーブルはユーザが入力したキーワードに該当するURLが存在することのみ判別できれば良いので，HTMLファイルなどのテキストファイルのデータのみで構成する。

④ 「③」より，収集対象はHTMLファイルなどのテキストファイルに限定できる。

２．HTMLファイルなどを収集する手法

HTMLファイルなどを収集する一般的な方法にFTP

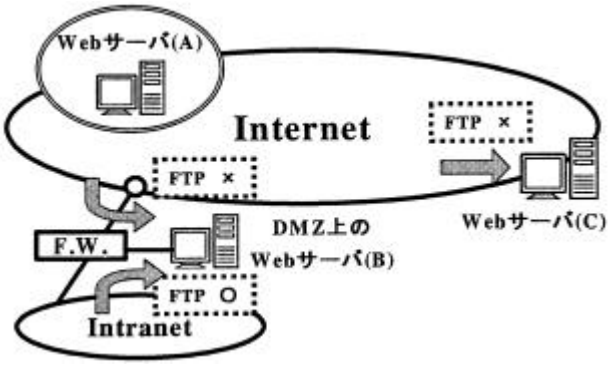

図１　インターネット上のWebサーバ

がある。しかし，インターネットでは不正アクセスを防止するため，ファイヤウォールやWebサーバでFTPのサービスを止めている場合が多い（図１）。特定のサーバからのみFTPのアクセスを許可する方法もあるが，セキュリティホールとなる危険がある。また，ドメイン名やサーバ名を変更すると，送信側と受信側双方のサーバの設定を修正する必要が生じる場合がある。

よって，FTPよりHTTPを利用する手法が有効である。特に，Perlでは１つのHTMLファイルなどを取得するモジュールHttpget. plが公開され，様々なサイトあるいは書籍にて紹介されている[1)]。これを応用し，HTMLファイルなどを収集する。指定したURLから得られた情報をprint出力する部分を，配列変数に格納し，以降の処理に使用するよう変更する（表１）。

表１　Httpget. plの変更点

Httpget. plの変更前 while(<SOCK>){ print; } #応答をすべて標準出力へ ↓ Httpget. plの変更後 @data_list = <SOCK>; #応答をすべて配列変数に格納

また，HTMLファイルなどを収集するためのアルゴリズムは，当該ファイルのソースに<A HREF>タグが存在する限り継続される（図２）。同じファイルを繰り返し収集し続けたり，目的外のファイルを収集することも考えられる。これを解決し実用化するためには，収集を停止する条件を付与する必要がある。ここでは次のように

*情報システム技術グループ

条件を定めた。

① １つのファイルソースから取得したリンク先のURLを格納している配列@list_2の中で，重複を解消する。

② @list_2と既に取得しなければならないURLのリストが格納されている配列@listとを比較し，@listに「含まれている」URLは@list_2から削除する。

次に，URLの取得を特定のドメインやディレクトリに制限するためには，次の方法で処理する。

① ユーザが指定したURLのドメイン以外のドメイン，あるいはユーザが指定したURLのディレクトリより上位のディレクトリについてはリンクを追わないように，あらかじめ取得する範囲のリストを格納した配列@list_3を作成する。

② 前出の配列@list_2と配列@list_3を比較し，「該当しない」URLを削除する。

これらの条件を図２におけるa-a'間へ挿入する。

３．検索用データテーブルを生成する方法及び検索結果を表示する方法

収集したURL及びHTMLファイルなどのソースについて，改行コードを削除し，１件１行として整理する。これが，検索用のデータテーブルとなる（表２）。検索結果の表示は検索用のデータテーブルをもとに表示するが（図４），実際に検索に該当したページを表示するときは，検索用のデータテーブルの先頭フィールドに記述されているURLを参照する。

４．まとめ

今回提案した手法の利点は，分散したWebコンテンツをFTPに依存せずHTTPにより収集できることである。また，HTMLファイルなどを含む必要最低限かつ最上位のURLからリンクをたどって行くことが可能であれば，下位のディレクトリ構造に依存せず，収集を容易になっている点も分散するWebコンテンツの効果的な活用に貢献している。

この手法を利用し，関東甲信越静の各公設試験研究機関がWeb上で公開する設備や技術の情報を検索することができる「関東甲信越静バーチャル公設試」を構築した。この結果，今回提案した手法の有効性を確認できた。

参考文献

１）中島靖：Perl徹底活用インターネットダイレクトアクセス，情報管理，18－19 (1998).

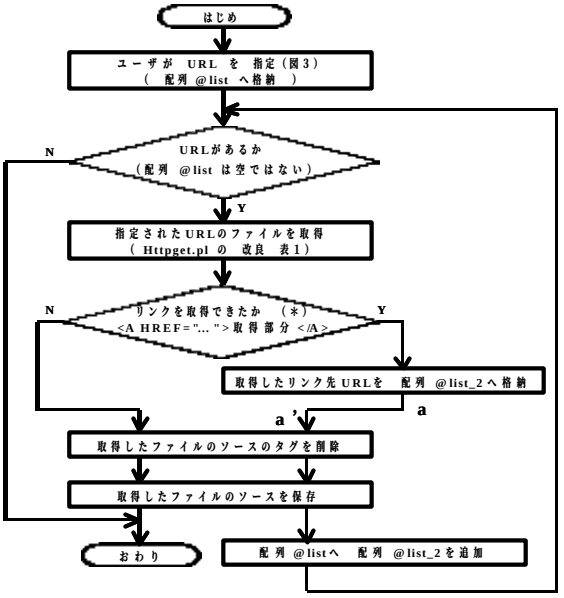

図２　コンテンツ収集のアルゴリズム

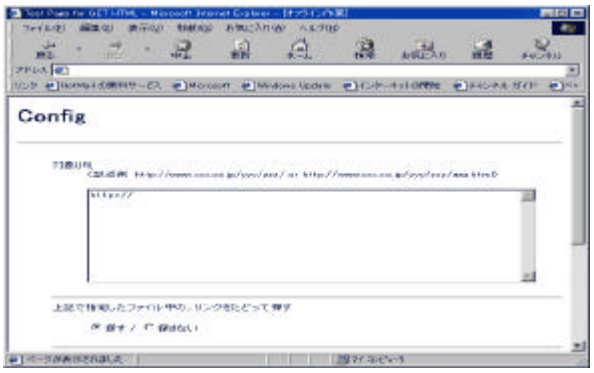

図３　ユーザが取得するURLを指定するページ

表２　検索用データテーブル

http://www...jp/test/dir_1/h10gijuken.html#tokken1;誤動作自己検知機能を内蔵・・・（改行） http://www...jp/test/dir_2/h9keijou1.html;膨張型消音器の音響特性・・・（改行） http://www...jp/test/index.html;研修スケジュールについて平成12年度・・・（改行） ・・・

図４　検索結果の表示