L-15

地域WWWページの更新調査とディレクトリ集への表示
Investigation of update of WWW pages of a regional community and its presentation on a directories site

山内慎祐　白川正知　古川泰男
Shinsuke Yamauchi　Masatomo Shirakawa　Yasuo Furukawa

1．はじめに

地域におけるインターネットの活用と情報流通の活発化のために，地域コミュニティ発の WWW ページを生活情報のカテゴリ別に分類し，そのタイトル，概要，URL 等をデータベース化したディレクトリ集を WWW サイトとして公開している。ここには豊橋地域発の約 500 のページを収容し，豊橋コミュニティ・ナビゲータと呼んでいる[1,2]。

情報流通の活発化を図るためには，どの程度新鮮な情報が発信されているかが重要である。その目安としてページの更新頻度がある。また更新されたページをディレクトリ上でそのように表示することも閲覧の促進を図る上で役に立つであろう。

そこで，ディレクトリ集に更新状態を毎日定時に巡回調査するロボット[3]を連携させ，更新調査を行うと共に更新日時をデータベースに書き込み，閲覧日時との差が所定期間内であれば，更新表示をするシステムを構築した。

2．地域WWWページのディレクトリ集

豊橋コミュニティ・ナビゲータでは豊橋地域から発信されている約 500 の WWW ページを 13 の生活情報のAカテゴリに大分類し，さらにこれらを細分化した全体で 60 のBカテゴリに分類している。OS に Linux，WWW サーバに Apache，データベースに PostgreSQL などのフリーソフトを用いて構築した。PostgreSQL と Apache は Tomcat(JavaServlet & JSP エンジン)によって結合した。

クライアントはトップ画面から所望のカテゴリを選択し，サイトのタイトルと概要の配列から，所望のサイトを閲覧することができる。

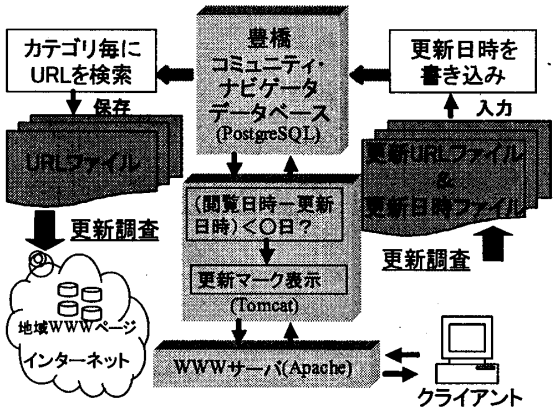
3．ページの更新調査方法

内容が更新されたかどうか，正確に推定するためには，本文部分を抽出し内容を比較することが必要である[4]。ここでは，最も簡便な目安である WWW サーバからの HTTP Request に対する Response のヘッダ情報中の Last-modified に記載された更新日時から調査することとした。しかし更新日時が省略されるという問題がある。その補完として，次に簡便なデータサイズの変化を見ることとした。ヘッダ情報中の Content-length からデータサイズが変化していれば，更新と判断するものである。この場合もデータサイズが保持されない場合や内容変化と直接関係しないなどの問題がある。

4．ディレクトリ集への更新の表示

クライアントの閲覧日時と更新日時とが所定期間内であれば，地域 WWW ページのタイトルに更新マークを表示することが合理的であろう。そのために，図1に示すように更新日時をデータベースに書き込み，閲覧日時との差異によって，更新の有無を表示させることとした。

すなわちコミュニティ・ナビゲータに書き込まれた更新日時とクライアントからの閲覧日時との差異が所定期間内



かを JavaServlet プログラムにより判定する。そうであれば更新マーク付加した HTML 出力ページを生成してWWWサーバに返す。

図1．更新マークの表示の方法

5．更新調査ロボット

図2に更新調査ロボットのフローを示す。データベースの調査対象 URL の取得から，更新判断，更新日時の書き込みまでの一連の過程を自動で行う。

更新状態の統計処理をAカテゴリ毎に行うこととし，データベースからカテゴリ毎に URL を取得する。HEAD メソッドにより取得したヘッダ情報から Last-modified を抽出し，前回のそれと比較して更新されていれば，最新情報ファイルを更新する。Last-modified が取得できない場合は，Content-length を抽出し，前回と変化していれば最新情報ファイルを更新する。

ロボットの一連のプログラムは Perl 言語によって作成した。プログラムは大別して，(a) データベースから URL を読み込む機能，(b) WWW へアクセスする HTTP 通信機能，(c)500 余りの URL に HEAD Request を出し，Response のヘッダ情報から更新の有無を判断する機能，(d)更新日時をデータベースへ書き込む機能，からなる。

図3に(a)(d)の機能を示す。データベースには収録ページの登録番号，URL，名称，更新日時などの属性を記録するサイトテーブルがある。さらにページをカテゴリに分類するための，登録番号，カテゴリ ID などを記録するサイトインデックステーブルがある。二つのテーブルに分離したのは，複数カテゴリへの登録の便利等のためである。

URL をAカテゴリ毎にまとめるため，サイトインデックステーブルにおいてAカテゴリに対応する登録番号を検索

豊橋技術科学大学　未来技術流動研究センター
Research Center for Future Technology, Toyohashi University of Technology

し，サイトテーブルにおいて登録番号から URL を得て，更新調査 URL ファイルとする。ロボットが得た更新 URL から更新日時をサイトテーブルに書き込む。

ロボットが得る日時文字列形式を localtime($epoc_sec)（例えば Wed May 14 06:02:55 2002）としてサイトテーブルの更新日時カラムに与え，PostgreSQL の datetime 型に入力可能とした。

(b)HTTP 通信機能は Perl の通信モジュール LWP(Library for WWW access in Perl)を CPAN(Comprehensive Perl Archive Network)よりダウンロードして用いた。

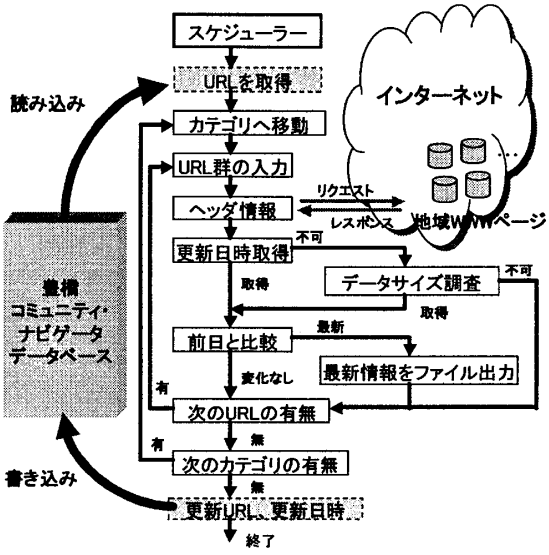


図２．更新調査ロボットのフロー

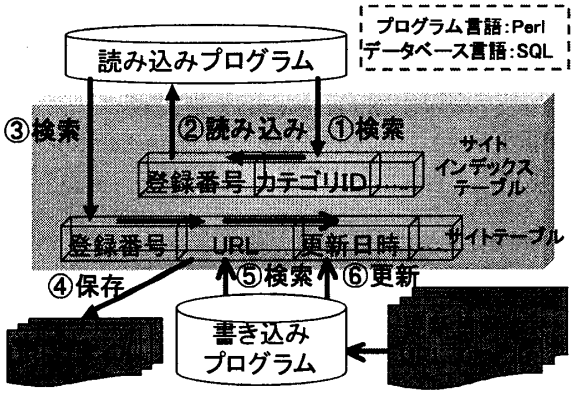


図３．データベースとの連携

６．更新調査とその結果

1 日で最も短時間で調査できる午前 5 時に毎日ロボットの巡回更新調査を開始させた[3]。本サーバは学内 LAN から SINET を通じて地域 WWW ページのサーバにアクセスする。約 500 の WWW ページからヘッダ情報を取得する時間は日によって数倍のバラツキを示し，平均は約 28 分であった。

Last-modified または Content-length で更新日時を把握できたページは約 85%であった。ただしこれは内容変更がどのようであるかについては，考慮していない。また 1 日当りの更新ページ数は平均 14 で，全体の約 3%であった。まだ人々の日々の生活行動に影響を強く与えるほど多くの新鮮な情報発信が行われているとは言い難いと思われる。

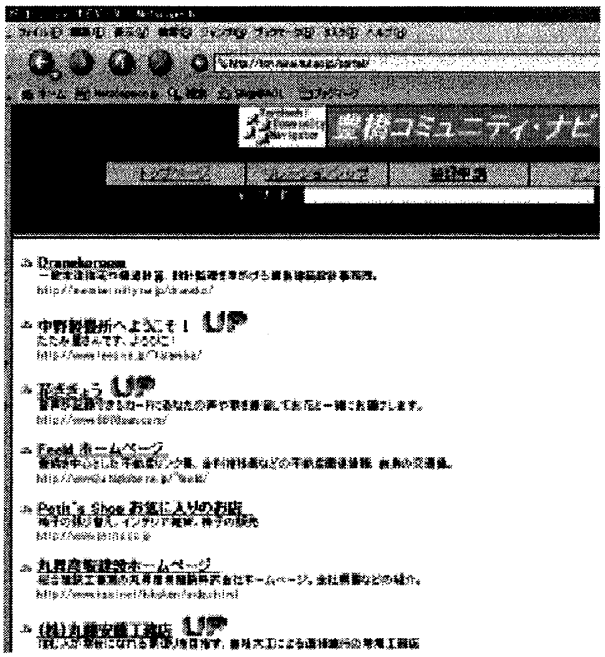


図４．更新マーク（UP）の表示

図４にブラウザの画面に表示されたディレクトリのうち更新されたページに更新マーク（UP:Update）が表示されていることを示す。更新日時と閲覧日時の差異が 1 週間以内であれば UP 表示することで運用している。

７．おわりに

地域 WWW ページのディレクトリ集，豊橋コミュニティ・ナビゲータに収録されている約 500 のページの更新調査を行うロボットを開発し，更新調査を行うと共に，その結果をブラウザの画面に更新マーク表示として反映させるシステムを構築した。

これによって，地域から発信されている WWW ページがどの程度活発に新たな情報を提供しているかが，形式的ではあるが，定量的に把握できる。また更新の自動表示は情報の利用者にとっても便利であろう。下層のファイルの更新や情報の内容にまで立ち入って，更新調査をすることが今後の課題である。さらに収録ページを増やしながら，このシステムを継続的に運用し，地域の情報流通の把握と活発化について研究を進めてゆきたい。

本研究は東海産業技術振興財団と旭硝子財団の助成研究の一部であり，記して感謝する。

参考文献

[1]古川：情報処理学会第 60 回全国大会, 4-243(2000).
[2]古川ほか：情報処理学会第 62 回全国大会, 4-425(2001).
[3]山中ほか：情報処理学会第 62 回全国大会, 4-413(2001).
[4]栗島ほか：情報処理学会第 64 回全国大会, 3-45(2002).