

3P-01

## WebBeholder Personal Edition: Web の差分情報 提供システム

平 博司    Santi Saeyor    伊庭 斉志    石塚 満

東京大学工学部電子情報工学科

### 1 はじめに

1990年代に入って、インターネットは瞬く間に世界中に広く普及するようになった。このインターネット普及をもたらした最大の要因は、1989年に開発されたWWWにあるといえる。個人で容易にかつ瞬時に情報を送受信することを可能にしたWWWは、来るべき情報革命を担う中心的ツールとなっている。しかし、WWWはその広がりゆえに、内容も更新日時も各ホームページによって全くバラバラであり、それらを効率よく巡回するのはきわめて難しい。

そこで、ユーザがあらかじめ選んだホームページの、ある一定時間経過後の変化を調べ更新情報を知らせるシステムが、1997年に開発されたWebBeholderである[1]。本稿では、このWebBeholderを拡張し、個人単位で簡単に利用できるよう改良を施したWebBeholder Personal Edition(PE)について報告する。

### 2 WebBeholder PE

本システムの最大の目的は、ユーザが自らが求める情報を効率的に手に入れることができる事にある。この目的を達成するため、WebBeholderはユーザに対し、登録されたホームページの更新状況と、更新内容の2つを提供する。

同様のシステムとしては、ホームページの更新状況を知らせるソフトウェア[2]や、1ヶ月以内に登場/

更新したホームページについて検索することのできるサーチエンジン[3]があるが、本システムは更新の有無の他に、具体的にどこが変わったのかという差分情報を伝えることができるのが大きな特長となっている。Push技術がフロー型のニュース的信息を対象にするのに対し、本システムはストック型情報の更新部分を対象とする点で違いがある。

このためWebBeholderは、HP閲覧者には、1. 注目しているページの新情報を遅滞なく知ることができる、2. 更新されていないページを見に行く手間を省く、といった利益をもたらす。また、HP製作者にとっても、関心を持つ人(すなわち自分のHPに関する最新情報を知りたい人)に、素早く、的確に更新情報を伝えることができるというメリットがある。これによって、WWW上の新たな情報流通の形態を可能にする。

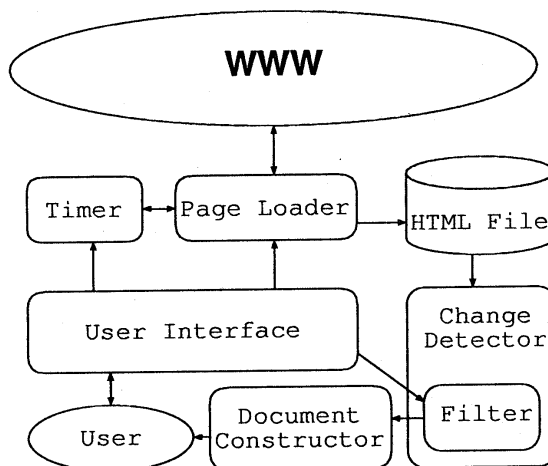


図1: WebBeholder PE システム概要

### 3 構成と機能

#### 3.1 動作の流れ

本システムの概要を図1に示す。ユーザは、まずプログラムに「見に行きたいホームページのURL」「どの程度の変化を更新と見なすか(Filter)」「見に行く時間間隔(Timer)」などを設定する。プログラムはその設定を受け、指定された時間にホームページを見に行く。この時、フィルタに指定されている更新が行われていたら、更新の有無を知らせ、差分情報ファイルを作成する。

これらの設定は、図2のようなGUIを用いて行われる。これは個人が簡単に使えるように配慮したものである。

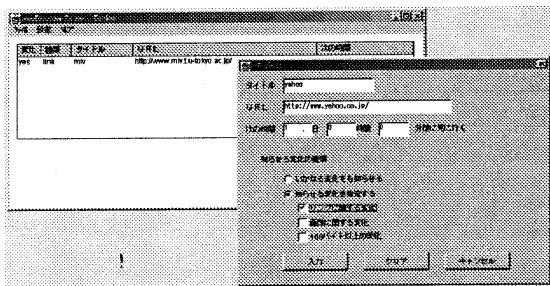


図2: WebBeholder PE の GUI

#### 3.2 ページ読み込み

ページ読み込みは、Timer に指定された時間間隔に従って、Page Loader によって自動的に行われる。Page Loader によって取り込まれたドキュメントは、HTML ファイルの形でデータベースに保存される。なお、一番最初のファイルは oldfile として、以降のファイルは newfile として保存され、比較作業の後 oldfile の内容は newfile の内容に置き換えられる。

#### 3.3 ページ比較

新旧のページ比較は、Change Detector によって行われる。この比較システムは Longest Common Tag Sequence Algorithm(LOCTAGS)[4] をベースとしている。これは HTML ドキュメントが、タグという独特の表記法を用いて構築されていることを利用する。

比較によって「リンクが変更されている」「画像が変更されている」など、変更の種類に関する情報が得られる。これを、ユーザが設定した Filter と比較し、もしそれに合致する変更がなされていれば「更

新なされている」というアナウンスをユーザに行う。

#### 3.4 差分情報出力

更新されているとシステムが判断したドキュメントについては、oldfile と newfile から差分ドキュメントが作成される。差分ドキュメントは、削除された部分は赤字で書かれた上に横線が引かれ、新たに加わった部分は太字で書かれた上にアンダーラインで表示されるようになっている。これによって、簡単に差分情報を手に入れることができるようになる。図3に一連の過程を示す。

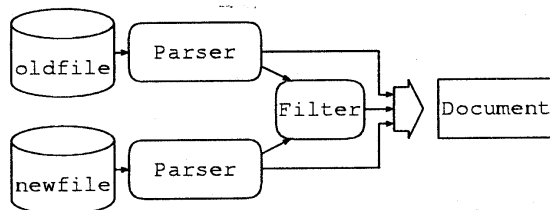


図3: 差分情報出力過程

### 4 おわりに

本稿では、WebBeholder を拡張した、各個人が手軽かつ効率的に新しい情報を入手できる Personal Edition について記した。ただ、現在のところ多くの人にβ版を配布しフィードバックを得るまで至っておらず、現段階では評価は難しい。今後は、Java の特性を利用した汎用性の拡大と、より使いやすい操作性を目指して改良を加えていきたい。

#### 参考文献

- [1] Santi Saeyor, 石塚: 「WebBeholder: エージェントコミュニティによる Web 情報変化の監視と指示」, 情処, 情報学シンポジウム, pp.41-47(1999.1)
- [2] 著名なものの中島 智秋氏の WWWC がある。  
<http://www.kinet.or.jp/naka/tomo/wwwc.html>
- [3] ロボット型検索エンジンの東芝フレッシュアイ。  
<http://fresheye.toshiba.co.jp/guide/intro.html>
- [4] S.Saeyor, M.Ishizuka: Longest Common Tag Sequence Algorithm for precise reviewing of changes in the WWW, 情処第56回全大, 3-313(1998.9)