差分分析を用いたWebページの有益情報抽出手法*

3X－02

樋口 毅　釜坂 等†
三菱電機株式会社 情報通信システム開発センター§

1. はじめに

インターネットの普及に伴い、その上に公開される情報の量は膨大になっている。また、これらの情報は日々変化をしている。したがって、インターネット上の情報を閲覧するユーザが、過去に見たページが変化したのかを知りたいというニーズが高まっている。現在、Web ページが変化したことを知るには、以下のような手段がある。

- サイト側での New フラグの表示
- Webdiff.com や AT&T の TopBlend に見られるような変更通知とビューワの提供

しかしこれらはユーザにとっての本当の新規の変更情報ではない、ユーザにとって本当に必要な変更情報ではない、あるいは変更部分がわかりにくいといった問題がある。

そこで本稿では、これらの問題を解決するWeb ページの構造をベースに有益度に基づく変化情報の抽出を行う手法について提案する。また、抽出した有益な変化情報を視覚化するツールについてもあわせて報告する。

2. Web ページのブロック化

HTML 文書は構造化文書であることから、意味のあるブロック化を行うことが可能である。例えば、<OL>で始まり、</OL>で終わる部分を一つのブロックとみなす。ただし、すべてのタグを対象とすると、細かくなり、全体構造が分からなくなってしまうため、ブロック化する対象のタグを選択した。また、i-mode ページのように表示領域に制限がある場合には、<HR> タグを用いて横線でブロックを分割するような場合もある。その場合は、そこがブロックの切れ目とみなした。入れ子構造にも対応し、構成名に対応するブロックの内容は入れ子構造を意識し、外側のブロックは中のブロックの内容を除くようにした。図 1のような Web ページをブロックで分割し、その構成を表示した例が 図 2である。

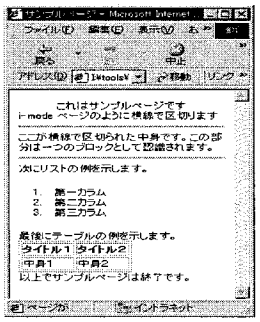

図 1 Web ページの例

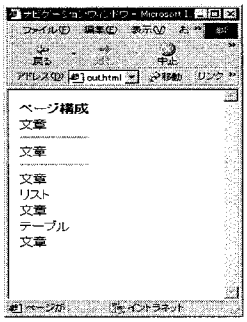

図 2 構成リストの例

構成リストを作成するために利用するタグの例を以下に示す。

	開始タグ	終了タグ
テーブル	<TABLE>	</TABLE>
リスト	<OL>	</OL>
横線	<HR	>
フォーム	<FORM>	</FORM>
…	…	…

* The method of useful information extraction based on analysis of Web pages diffrence.
† Tsuyoshi Higuchi, Hitoshi Kamasaka
§ Mitsubishi Electric Corporation. Information & Communication Systems Development Center

3. 有益情報の抽出

新旧の Web ページを単にブロックごとに比較し、変更の通知、表示を行うだけでは、ユーザにとって有益な情報を通知できないという問題は解決しない。この問題を解決するために、新旧の Web ページの内容の比較を行う際に、有益度を用いることとした。ここでいう有益度とは「特定のブロック内の変更は有益である」、「ユーザが指定したキーワードを含んだブロックの変更は有益である」、「特定のタグで囲まれた部分の変更を含んだブロックの変更は有益である」という情報を数値化したものである。例えば、オークションサイトや掲示板といったページは商品の情報や内容、トピックス等がテーブル構造で記載されている。このようなページの場合、テーブルというブロック内の変更が有益な情報となる。

したがって、以下のようなテーブルをユーザごとに用意し、Web ページの変更の有益度を計算し、ユーザごとに指定された閾値を越えたら変更を通知する。この結果、ユーザにとって有益な変更があった場合のみ通知が行われるようになる。また、変更があってもそれを変更とみなさないものも定義する。例えばコメントやカウンタ等の変更はほとんどのユーザにとっては変更とはみなされない。そういうものは有益度を 0 として変更とみなさないようにする。

		有益度
テーブル	<TABLE>	15
フォント	<FONT>, <B> 等	5
打ち消し	<S>	5
キーワード	i-mode	25
コメント	<!-- -->	0
それ以外		1
…	…	…

この有益度のリストは、ユーザごとに定義されるものである。最初はデフォルト値をセットしておき、あるユーザにとってこの数値が妥当なものでないと判断された場合には、そのユーザの有益度リストに登録されている有益度の値を変えられるようにしている。これにより、ユーザごとに異なる有益な情報を確実に通知、表示するというパーソナライズを実現している。

4. 視覚化ツール

Web ページの構成をリストに表わすことが出来、その構成に対応するブロックがわかっているため、新旧の Web ページを比較する際、ブロックごとに比較を行い、この構成リスト上で変更情報を表示することが可能となる。この構成リストによる変更情報をナビゲーションウィンドウとし、これを表示することで、変更箇所を認識することができる。図 3はナビゲーションウィンドウ付きの比較結果の表示例である。

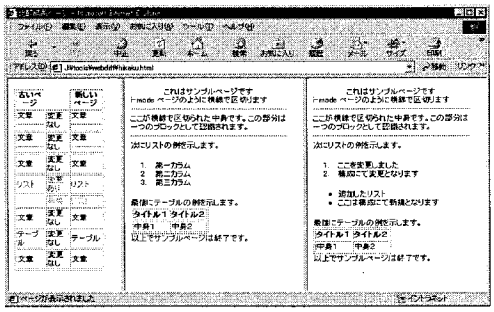

図 3 比較結果の表示例

図 3では、左側にナビゲーションウィンドウを、真ん中に古い Web ページの内容を、右側に新しい Web ページの内容を表示している。ナビゲーションウィンドウの表示から、どのブロックが新規に追加されたのか、また、どのブロックが変更になったのかという全体的な変更情報を見ることができる。この内容に従い、さらに詳細が知りたい場合には、ナビゲーションウィンドウ内の構成名をクリックすることで実際のページの対応する部分がわかるようになっている。このようなナビゲーションウィンドウを提供することで、ユーザにとって変更箇所がわかりやすいものになる。

5. おわりに

本手法を用いる事によって、Web ページのユーザにとっての有益な変更情報を通知・表示することができることを示した。本手法のページ単位の差分分析を用いることでサイト全体の有益な変更マップを作成することも可能である。