社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－FI－73 (8)
2003／11／13

スクリプト細分割記憶によるWebアーカイブ方式

遠藤裕英†

Webページをアーカイブする必要性が指摘されているが，Webページは増大を続けているので，アーカイブに必要とされる記憶容量は膨大になる．そこで，Webページを圧縮して記憶する方法が有効と考えられる．本稿では部分的に更新されるというWebページの特徴に着目して，更新前のWebページのスクリプトを再利用することによって記憶データ量を削減する．HTMLスクリプトを内容的にまとまりのある単位でブロック化し，ブロック単位に再利用を図る．

提案方式をドキュメント主体のWebページと，テーブル構成のWebページについて評価した結果，記憶データの削減効果が見込める見通しが得られた．

Web Page Archiving by Segmented Scripts Filing

Hirohide Endo*

Archiving Web pages is important issue. The storage capacity required to archive web pages is skyrocketed because web pages are still growing. This leads an idea of data reduction of web pages for archive. The point of this paper is reuse of a previous webpage's scripts. This comes from the fact that most web pages are partially renewed. The scripts are segmented by article blocks so as to increase reuse ratio.

The proposed reduction method is evaluated by both a document type web page and table form web pages, and it is obtained the data are reduced to some extent.

1. まえがき

Webページをアーカイブする必要性が指摘されている[1]．一方，アーカイブの対象となるWebページは増大を続けているので，アーカイブに必要とされる記憶容量は指数的に増大する．そこで，Webページを圧縮して記憶するのが有効と考えられる．

圧縮方式には，ビット系列を圧縮する方式や，文字系列を圧縮する方式，2次元データを圧縮する方法，時系列フレームを圧縮する方法などが開発されている[2][3][4]．本稿の方式は時系列フレームの圧縮に属し，ビット系列の圧縮方式などとの併用が可能である．

Webページには，（1）部分的に更新が繰り返される更新型Webページ（差し替え型と追記型）と，（2）一旦作成されれば廃棄されるまで変更されない完結型Webページとがある．本研究では，（1）を対象としている．

WebページはHTMLスクリプトで記述されたhtmlファイルと，HTMLスクリプトから参照される画像等の外部ファイルからなる．本研究では，htmlファイルを取り上げる．

† 立命館大学
* College of Science and Engineering, Ritsumeikan University

Webページを内容的にまとまりのあるブロックに分割することを考える．Webページの更新では，更新に関係するブロックは変更されるが，更新に関係のないブロックはそのまま据え置かれる．そこで，ブロックごとにアーカイブファイルを作ることにすれば，更新ページのアーカイブに対して，更新ブロックだけをアーカイブファイルして，それ以外は前回のアーカイブファイルを再利用すればよく，アーカイブファイルのデータ量を削減できると考えられる．

Webページは視覚的に区画化されて表示される場合が多く，各区画にはある種の性格付けがある．たとえば，Webページの上端にはWebページのロゴが配置され，下端には，連絡先や著作権表示などのフータが配置される．そして，これらと本文との間には視覚的に区別できるように，水平線や空行等のスクリプト記述がある．そこで，HTMLスクリプトの記述上の特徴を見つけてWebページをブロック化する方式を提案する．

以下，2．で代表的なWebページの構成方法を検証し，Webページをブロック化するのにどのような特徴に注目すべきかを検討する．3．では2．の検証結果を踏まえ，これらの特徴を識別するWebページのブロック化方式を提案し，4．でいくつかのWebページに本提案方式を適用したときのアーカイブデータ量の削減効果について述べる．

2. Webページの構成

2.1 Webページの構成例

いくつかのWebページを見てみる．図1は表形式でWebページを構成している例である．

このWebページは第1階層に4個の<TABLE>要素を，第2階層に4個の<TABLE>要素を，第3階層に2個の<TABLE>要素を使用してページ全体を構成している．

図1　表構成のWebページ例

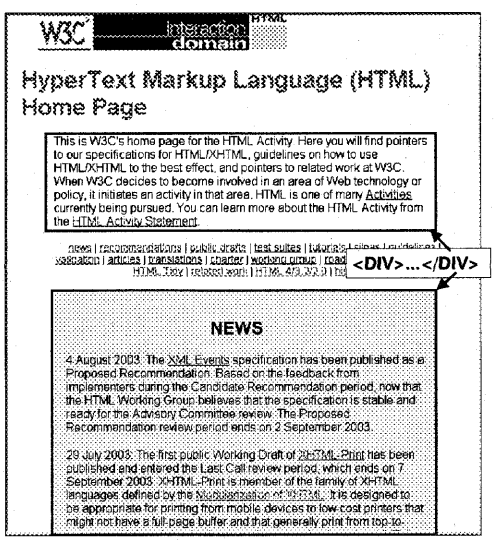

図2　表構成を使わないWebページ例

（1）第1階層の<TABLE>要素

上段にロゴ，中段にインデックス・トピックス・サービス等，下段上部に関連サイトへのリンク部，下段下部にフータ（連絡先，著作権表示等）が配置されている．

（2）第2階層の<TABLE>要素

中段のテーブル要素の中に，4個の<TABLE>要素がある．①左側面のインデックス部，②中央の記事の見出し部，③右側面上部に写真入のトピック（Pick up），④右側面下部に過去の Pick up とのリンクと MoreNews である．

（3）第3階層の<TABLE>要素

中段右側面の上部と下部の<TABLE>要素に，それぞれ1個の<TABLE>要素が含まれている．

この例のように，ホームページのトップページをテーブル形式で構成しているサイトは極めて多い．whitehouse.gov，acm.org，ibm.com，microsoft.com，などは一例である．

図2はテーブル要素を使わずに Web ページを構成している例である．

この Web ページはテキスト主体で，記事を次のようなタグでまとめている．

① 見出し<DIV>....</DIV>

② <DIV>見出し....</DIV>

③ <DIV>....</DIV>

② 見出し<DL>....</DL>

③ 見出し<UL>....</UL>

Web ページの種類や Web ページの作成ツールによって使用されるタグに相違が出てくることが考えられる．

それでは，Web ページをまとめるのにどのようなタグが HTML で定義されていのか見てみる．

2.2 ページ構成に関連する HTML タグ

Web ページを構成するのに使用されるタグをまとめた表を，表1に示す．

（1）テーブル　テーブルには2種類のタイプがある．ひとつは文書中にデータを表形式で整理して示すために用いられるテーブルであり，もうひとつは，Web ページをレイアウトするのに用いられるテーブルである．文章中のテーブルは文章の従たる要素とみなすことができる．

テーブルには行と列があるが，<TR>タグで行を，<TD>で各列のデータを記述する．
一行のデータは<TR>～</TR>であり，一区画のデータは<TD>～</TD>である．

（2）Web ページのレイアウトを<FRAMESET>で行う．各フレームは，画像ファイルや HTML ファイルなどから構成される．このタイプの Web ページでは，主ページのスクリプトは短く，<FRAMESET>を構成する<FRAME>が主たる内容になる．

（3）リストは，単独で用いられる場合，テーブルの区画内で用いられる場合と文章中に用いられる場合とがある．単独で用いられる場合以外は，テーブルや文章の従たる項目とみなし，ここでは，単独に用いられる場合を考える．

表1　Webページのレイアウト区分に使用されるタグ

分類		タグ
テーブル	見出し	<CAPTION>……</CAPTION>
	表	<TABLE>……</TABLE>
	行、データ	<TR>……</TR>、<TD>……</TD>
フレーム		<FRAMESET>……</FRAMESET>
リスト	定義型	<DL>……</DL>
	ﾗﾍﾞﾙ付ﾘｽﾄ	<UL>……</UL>, <OL>……</OL>
	ﾃﾞｨﾚｸﾄﾘ ﾒﾆｭｰ	<DIR>…</DIR>, <MENU>…</MENU>
文章	見出し	<Hn>……</Hn>
	仕切り	<HR>,
	分割	<DIV>……</DIV>
	囲み	<BLOCKQUOTE>……</BLOCKQUOTE>
	整形文	<PRE>……</PRE>
その他	マップ	<MAP>……</MAP>

（4）文章には，①見出し付の文章と，②見出しなしの文章とがある．見出しつきの文章は，見出しから文章の終わりまでをひとつのブロックと考えることにする．

（a）分割，囲み，整形文ではタグで示された範囲をひとまとまりの文章とみなす．

（b）（a）以外の文章の終端は，次の見出し，仕切り，分割，囲み，整形文タグとする．

（5）その他　　<MAP>タグなどがある．

2.3 用語の定義

HTMLスクリプトでは，< > に囲まれた範囲にタグ名や属性名，属性値を記述する．本論文では，< > に囲まれた範囲に記述される文字をインタグ文字と呼び，< > 外に記述される文字をアウトタグ文字と呼ぶことにする．アウトタグ文字には，①<STYLE>の属性記述スクリプト，②<SCRIPT>タグで記述される Jscript 文，③ブラウザで表示される文字等が含まれる．③のアウトタグ文字をメッセージ文字と呼ぶ．

<Hn>(n=1～6)，<HR>，

，を文仕切りタグと呼び，<DIV>，<BLOCKQUOTE>，<PRE>を文区画タグと呼ぶことにする．また，<CAPTION>を表仕切りタグと呼び，<TABLE>，<TR>，<TD>を表区画タグと呼ぶ．

タグは入れ子に記述できるので，最初のタグを第1階層のタグ，以下，入れ子に記述されるタグを第2階層のタグ，・・・という呼び方をし，番号の小さい階層を上位階層，番号の大きい階層を下位階層と呼ぶことにする．

3. スクリプト細分割記憶による Web アーカイブ方式

3.1 スクリプト再分割記憶方式

図 3 にスクリプト細分割記憶方式の概念図を示す.

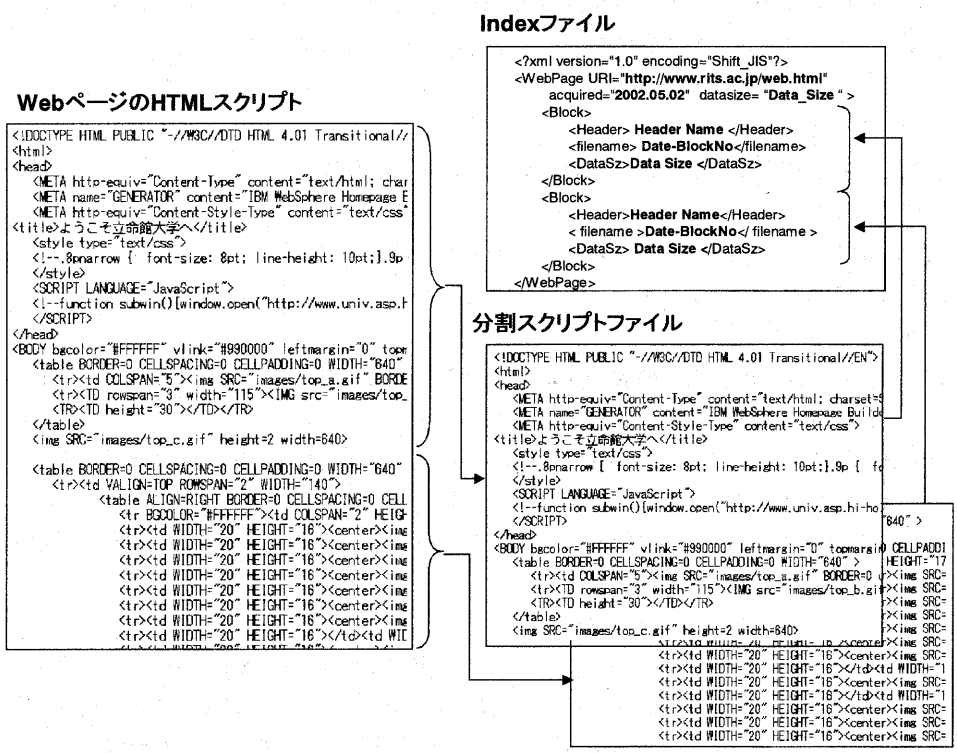

図 3　スクリプト細分割記憶方式の概念図

スクリプト細分割記憶方式は，細分化されたスクリプトを保存する分割スクリプトファイルと，分割スクリプトファイルから Web ページを復元するためのファイル構成情報を保存するインデックスファイルからなる．スクリプト細分割記憶方式のアーカイブ手順は次のようになる．

（1）Web ページを記述する HTML スクリプトを 3.2 に述べるアルゴリズムによってブロックに分割する．

（2）各ブロックに対応する HTML スクリプトを抽出する．

（3）抽出された HTML スクリプトがアーカイブされているスクリプトと同一であるか否かを判定する．

（4）同一のスクリプトであった場合は，インデックスファイルにアーカイブされたファイルを参照するよう記述する．異なるスクリプトであった場合は，新しくファイルを作りスクリプトを保存する．そして，インデックスファイルにはそのファイルを参照するように記述する．

（5）HTML スクリプトの最後まで，（2）～（4）を繰り返す．

（6）html ファイルとともに取り込まれた参照ファイル（画像ファイルなど）から，すでにアー

カイブされているファイルと重複する参照ファイルを削除する．

Webページを加工することなく保存する場合，①htmlファイルと，②参照ファイルを一対のデータとして保存することになる．本提案方式では，①インデックスファイル，②更新された分割ファイルセット，③更新された参照ファイルを保存することになる．

3.2 ブロック化のアルゴリズム

前章における検討結果にもとづき，次のようなブロック化を考える．

（1）フレームセット

<FRAMESET>～</FRAMESET>を1ブロックとみなす．Webページをフレームセットで構成している場合，ページ全体が1ブロックとなる．

（2）文章ブロック

（a）文仕切りタグ間を第1階層のブロックとする．第1階層ブロック内にある文区画や表区画はブロック化しない．

（b）第1階層ブロックがない場合，最初に出現する表区画や文区画を第2階層ブロックとする．第2階層ブロックに含まれる文区画や表区画はブロック化しない．

（c）第1階層ブロック，第2階層ブロックがない場合，この段階でブロック化はしない．

（3）テーブルブロック

（a）ページレイアウト用のテーブル

最初の<TABLE>タグ出現前のメッセージ文字数や，最後の</TABLE>タグの出現後のメッセージ文字数から，ページレイアウト用に用いられているテーブルであるか否かを判定する．このテーブルでは，表区画内にメッセージ文字数が一定数以上であれば1ブロックとする．

（b）文章中のテーブル

（2）（a）（b）中に出現する場合は，文章ブロックのアルゴリズムに従う．

（c）その他のテーブル

タイトルを含めて，テーブル全体を1ブロックとみなす．

（4）リスト

（a）文仕切りタグ間に記述されるリスト，文区画タグ内に記述されるリスト，表区画タグ内に記述されるリストは（2）（3）のブロックに属する．

（b）その他のリスト

リストタグで指定される範囲を1ブロックとみなす．

（5）その他のタグ

特別な扱いはしない．

4. 評価

4.1 文章主体のWebページの評価

このWebページはW3Cのページで，8月5日更新のページ（図2，8. 5ページと記す）と9月24日更新のページ（図5，9.24ページと記す）である．3．で述べた方法でセグメント化した結果，以下のようになった．単位はバイトである．

	ブロック数	スクリプトサイズ（平均）		メッセージ文字数(平均)		画像ファイル
8. 5ページ	44	48,224	(1,096)	24,073	(547)	10,776
9.24ページ	44	48,929	(1,096)	24,341	(547)	10,776
（更新	1	4.252		2,213		0)

44ブロック中の3番目のブロックだけが更新（メッセージサイズ：1,945 バイト→2,213バイト）されている．そのほかのブロックの内容には変化がなかった．3 番目のスクリプトサイズは4,252 バイトなので，スクリプトの 91.3%は再利用が可能であると考えられる．しかし，今回収集したWebページでは，リンク先のサーバ名が，http//www.w3.org/ から，http://www.w3c.org/に変更されたため，各ブロックのスプリプトサイズは，44 ブロック中 13 ブロックしか一致しなかった.(メッセージサイズでは44ブロック中43ブロックが一致している).このような変更は，極めて頻度の低い特殊なケースと考えられる．通常はスクリプトの再利用率は高いものと考えられる．

4.2 テーブル形式のレイアウトを持つWebページの評価

情報系企業のホームページ（Web ページ B）のブロック化とブロックの更新状況を調査した．このWebページのスクリプトサイズは62,161，画像ファイルのサイズは45,103バイトである．メッセージ文字数は3,081バイトであった．セグメント数は11であった．セグメントあたりのスクリプトサイズとメッセージ文字数は以下のような結果が得られた．単位はバイトである．

	スクリプトサイズ			メッセージ文字数		
	最少	平均	最大	最少	平均	最大
Webページ B	1,041	5,651	15,428	152	280	546

表2には，初回のデータと，第1更新日，第2更新日，第3更新日のデータが示してある．数値はブロックサイズである．更新日に前回と同じブロックが利用できる場合は数値の右側に * の表示がしてある．

表の合計行にはWebページのスクリプトサイズを，再利用行には再利用されたブロックのブロックサイズの合計値とページスクリプトサイズに対する比率を%で示した．また，画像行に画像ファイルの容量と再利用比率を示した．

このWebページの更新では，ページスクリプトサイズの69.4～94.8% ，43～58KBが再利用できる結果になっている．画像データは38～40KB再利用できる．この例では，スクリプトの再利用によって画像データのキャッシュ効果をやや上回る再利用効果が得られることがわかる．

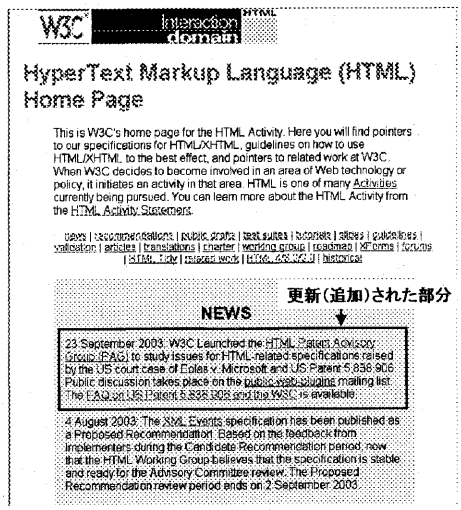

図5　9月24日更新ページ

表2 WebページBのブロック化と再利用率

		初 回	第1更新日	第2更新日	第3更新日
ブロックサイズ(バイト)	TOP	15,428	15,348	15,348 *	15,348 *
	LFT	9,395	9,395 *	9,395 *	9,395 *
	CNTR	1,041	931	931 *	858
		2,737	2,706	2,706 *	2,317
		2,628	2,628 *	2,462	2,462 *
		1,541	1,541 *	1,523	1,523 *
	RFT	6,692	6,692 *	6,692 *	6,692 *
		4,563	4,563 *	4,563 *	4,563 *
		11,959	11,959 *	11,959 *	11,959 *
	BTM	3,202	3,202 *	3,202 *	3,202 *
		2,985	2,985 *	2,985 *	2,985 *
	合計	62,171	61,950	61,766	61,304
	再利用		42,965 (69.4%)	57,781 (93.5%)	58,129 (94.8%)
画像	全画像	45,103	40,332	40,332	38,597
	再利用		29,742 (73.7%)	40,332 (100%)	34,583 (89.6%)

(単位：バイト)

4.3 記憶容量の削減率

スクリプトの記憶容量の削減率を，

削減率＝再利用率－（ブロック数×オーバヘッド）÷スクリプトサイズ

で計算すると，4.1のケースで71.3%，4.2のケースで第1更新日66.2%，第2更新日90.3%，第3更新日91.6%が得られた．ここで，オーバヘッドは1ブロックあたり180バイトとした．

5. まとめ

本稿ではWebページのアーカイブを念頭におき，WebページのHTMLスクリプトの記憶容量削減方式として，内容的にまとまりのある部分を識別できるタグでスクリプトをブロック化する方式を提案した．

提案方式を，更新型の文章主体のWebページとテーブル構成のWebページを用いて評価し，スクリプトデータ量の削減効果を確認した．

今後は文章主体のWebページの評価サンプル数を増やし，ブロック分割アルゴリズムの評価精度を向上させたい．

参考文献

[1] 廣瀬信己：国立国会図書館におけるウェブ・アーカイビングの実践と課題，情報処理学会研究報告 2003-DBS-130/2003-FI-71,p.95-111(2003)

[2] Ian H. Witten, Alistair Moffat, and Timothy C. Bell：Managing Gigabytes, Morgan(1999)

[3] Duane Wesels：Web Cashing,O'Reilly(2001)

[4] 横山昌平，大田 学，石川 博：要素圧縮によるXMLデータ圧縮手法の提案，DBWeb2000，p.331～377（2000）