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

2005－FI－78（5）
2005－DD－49（5）
2005／3／25

ウェブアーカイブを目的としたHTMLスクリプトのブロック化と差分格納方式

福井　雅士[1]，遠藤　裕英[1]

ウェブページの数は指数的に増大しているが，その情報は空間的・時間的に不安定であるため，WWW 上の情報を定期的に収集，保存するウェブアーカイビングの必要性が指摘されいる．しかし，ウェブアーカイビングは，同じ URL（Uniform Resource Locator）の情報を継続的に取得し，保存を行うため，保存スペースが大容量になるという問題点がある．そこで，WWW 上の情報の類似性に着目し，ウェブページの空間的類似性と時間的類似性を利用し，ファイル単位，分割ファイル単位それぞれでの差分格納を行うことにより，ウェブアーカイビングに適した記憶データの圧縮方式を提案する．

Web archiving with HTML script blocks and differential data storage

Masashi Fukui[2], Hirohide Endo[2]

While the number of web pages are still increasing, spacial and time consistency of these pages is unstable. To solve this problem web archiving has been promoted by the national libraries. However, in addition to archiving at the national level, web archiving at the organizational or personal level will also be needed in the future. For web archiving a large storage capacity is necessary, which translates to a high cost.In order to reduce the storage requirement, a data compression method suited to web archiving is proposed. Renew parts of data and differential parts of data between wab pages are picked up and stored.

1. はじめに

近年，インターネットの急速な普及により，WWW（World Wide Web : 以下「ウェブ」とする）において得られる情報量は数年前に比べ飛躍的に増大し, 多くの人がウェブ上の情報を参照するようになっている．しかし，ウェブ上の情報は空間的･時間的に不安定であるという問題点がある．そこで，ウェブ上の情報の文化的・社会的価値を残そうとするウェブアーカイビングの必要性が指摘されおり, 各国で取り組まれている[2]. また，ウェブアーカイビングは, 企業レベルや個人レベルにおいての過去の情報を管理する時にも適用できる．さらに, データマイニングの有効な材料としてウェブアーカイビングが利用できると考えられる[4]. ウェブ上の情報は今後さらに増加すると考えられており，また，ウェブアーカイビングの特性上，同じ URL（Uniform Resource Locator）の情報を周期的に取得，保存を行うため，ウェブ･アーカイビングにおける保存スペースは, 収集対象ウェブサイト容量収集回数になり，大容量になると考えられる. アメリカのインターネット･アーカイブが行っているウェブアーカイビングでは，1996 年から 2003 年 5 月までに収集された容量は 500 テラバイト以上になると報告されている[4]．このように，ウェブアーカイビングにおける必要記憶容量の増加は, ウェブアーカイビングシステムのコストを押し上げる. 特に, 企業レベルや個人レベルでウェブアーカイビングを行う場合, コストの増加は深刻な問題である.

そこで, 本研究では保存スペースの軽減を目的としたウェブアーカイビングシステムについて研究する．データ圧縮には，ZIP 形式や LHA 形式などに代表される様々な既存の圧縮方式があるが，それぞれ得意なファイルの種類が異なる．ウェブ上の情報は様々な種類の情報から構成されるため, 1 つの圧縮方式では充分な圧縮率が得られない. 高い圧縮率を目指すためには, 圧縮方式を使い分ける必要があり, 処理やデータ管理が複雑になり, 処理時間に問題が出ると考えられる. そこで, ウェブアーカイビングに適した保存スペース軽減方式を目指して, 本研究ではウェブページが部分的に更新されること, ウェブ上の情報は

[1] 立命館大学大学院　理工学研究科
[2] Graduate School of Science and Engineering, Ritsumeikan University

ウェブサイト単位で一つのまとまりであること，ウェブサイト内にはレイアウトの似たウェブページが多く存在すること，ウェブサイト内には共通の画像ファイルなどが多く用いられていることなどに着目した．そして，ウェブサイト単位での重複をなくすことによって，ウェブアーカイビングにおける保存スペースの軽減をはかる．ウェブページの空間的類似性と時間的類似性を利用し，ファイル単位，分割したファイル単位それぞれでの差分格納を行うことにより，ウェブアーカイビングに適した記憶容量のデータ圧縮方式を提案する．

2. ウェブアーカイビングの必要性と問題点

ここでは，ウェブアーカイビングの必要性と問題点について述べる．さらに，ウェブアーカイビングのデータ圧縮に対しての既存の圧縮方式の問題点を述べる．

2.1 ウェブアーカイビングの必要性

ウェブ上の情報の多くは頻繁に更新・削除され，日々失われている．ブックマークしていたページを再び訪れると表示されないといった状況や，内容が変わってしまっているといった状況は，誰もが一度は経験しているだろう．つまり，ウェブ上の情報は空間的・時間的に不安定であるという問題点がある．ウェブ上の情報は紙など他の媒体で同等のものが存在する場合もあるが，最初からデジタル情報として作成されたウェブ上の情報は，二度と入手することができない．ウェブページのを引用して議論をしたとき，その引用データが削除されれば，根拠が示せなくなる．

そこで，ウェブ上の情報の文化的・社会的価値を残し次世代に伝えるため，各国でウェブアーカイビングを目指しての取り組みが行われている．先進的な取り組みとして，アメリカの米国議会図書館，インターネット・アーカイブ，イギリスの英国図書館，日本の国立国会図書館などが行っているプロジェクトがある．また，ユネスコの 2002 年 4 月 9 日理事会提出の報告書でも，増加する電子情報を「世界の記憶」として保存していくべきだと指摘されている[2]．

その他にも，アーカイブされたデータは，社会の現象の推移や市場調査などに利用でき，データマイニングのための有益なデータとなる．各ページが変更される度合いを調べる研究や新しい情報が出現する度合いを調べる研究においてウェブアーカイブが利用されている[4]．また，企業や個人のレベルにおいても，過去に作成された情報の管理にアーカイブデータは利用できる．

2.2 ウェブアーカイビングの問題点

ウェブアーカイビングの問題点について，日本の国立国会図書館の WARP（Web Archiving Project）[3]を例にとって述べる．WARP では，収集はロボット収集であり，保存ディレクトリ構造は，収集先のディレクトリ構造を忠実に守った形式で保存されている．そのため，この保存ディレクトリ構造では，同一 URL に対して収集回数分の保存スペースが必要である．今後，この取り組みが本格的に開始された場合，対象 URL，収集回数の増加するのに伴って必要となる保存スペースが増大すると考えられる．アメリカのインターネット・アーカイブが行っているウェブアーカイビングでは，1996 年から 2003 年 5 月までに収集された約 110 億ページで，その容量は 500 テラバイト以上になると報告されている[4]．

以上のことからウェブアーカイビングには，大量の保存スペースが必要であり，保存する装置や，管理において大きな負担になると考えられる．また，企業や個人レベルにおいてウェブアーカイビングを行う場合，大容量化によって装置や管理におけるコストの増加が問題となると考えられる．

2.3 既存の圧縮方法

データ圧縮については様々な研究が進められており，多くの圧縮形式が存在する．圧縮にはそれぞれ向き，不向きがあり，データ形式に対応して様々な圧縮方式が提案されている．例を挙げると，画像ファイル用の圧縮である JPEG（Joint Photographic Expert Group）や GIF（Graphics Interchange Format），新しいものでは XML ファイル用の Xmill や XML ZIP などである[5]．また，圧縮率だけではなく，処理速度を重視したものや，バランスを重視したものなど，数多くの圧縮方式が存在する[8]．

ウェブアーカイビングにおいて既存の圧縮を利用する場合，2 つの問題が生じる．ひとつは，構成ファイルの多様さである．ウェブ上の情報は様々な種類のファイルによって構成される．HTML（Hyper Text Markup Language）ファイルや画像ファイル，他にも PDF （Portable Document Format）ファイルや音声ファイル，実行形式ファイルなどである．これらのデータを１種類の圧縮方式で圧縮する場合，圧縮率は最適にならない．

２つ目は，処理速度の問題である．ウェブ上の様々なファイル形式に対して，対応する圧縮方式を使い分けた場合，処理が複雑化し，処理速度の面で問題が出てくると考えられる．

これらのことから，既存の圧縮方式は，ウェブアーカイビングの圧縮方式として改善すべき余

地があると考えられる．

3. ウェブアーカイビングに適したデータ圧縮方式の提案

ここでは，これまでに述べた問題点の解決策を提案する．本研究では，「ウェブアーカイビングにおける保存スペースの軽減」を目的とし，ウェブアーカイビングに適した圧縮方法を提案する．

3.1 保存スペースの軽減方法

本研究では，保存スペースの軽減を達成するため「差分格納」という方法を用いる．保存スペースの軽減方法に差分格納を提案するに至ったのは，ウェブページが一部分のみ更新されることが多いという性質を持っているためである．

ウェブ上の情報は，日々更新されており，同じ URL にアクセスしても，日付が違うと，ページが変わっているという状況は，よく経験する．そして，その変更は，一部分のみが更新されている場合が多い（図 1）．そこで，「一部分のみが更新されている場合が多い」という性質を利用し，保存スペースを軽減させるために差分格納という方法を用いる．つまり，収集するページの更新部分のみを保存し，変化のない部分においては，以前に収集したページを利用することによって，保存スペースの軽減を図る．

まず，HTML ファイルをタグを基に解析し，分割する．2 回目以降の収集時には以前に保存した分割ファイルと比較を行い，必要な分割ファイルのみを保存する．閲覧時には，必要な分割ファイルを再構築し，指定された日時のウェブページを復元する（図 2）．

図 1.ウェブページの変更例

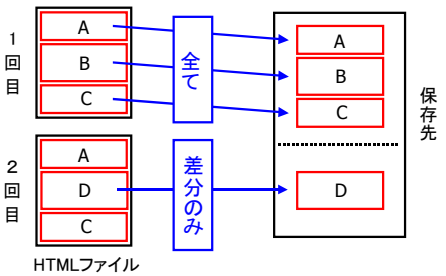

図 2. 差分格納例

3.2 差分格納の適用範囲

つぎに，差分格納をどのように適用すればウェブアーカイビングにおいて効果が発揮するかを検討する．3.1 での「ウェブ上の情報は一部分のみが更新されることが多い」という「時間的類似性」の他に，ウェブ上の情報には「同一ウェブサイトには似たレイアウトのウェブページが多く存在する」という「空間的類似性」がある．このことからウェブサイト内のページとも比較し，差分格納を行うことで，更なる保存スペース軽減が期待できる．

つぎに，ウェブ上の情報の構成要素について考える．これまでの差分格納は，HTML ファイルのみについて述べてきた．しかし，ウェブサイトを構成する要素としては HTML ファイル以外に，画像ファイルや PDF ファイルなどさまざまなファイルが挙げられる．過去のウェブ上の情報を復元するにはこれらのファイルも必要となってくるので，これらのファイルも収集しなければならない．また，HTML ファイル以外のファイルにおいても，ここまで述べてきた「時間的類似性」と「空間的類似性」がある．2 つの HTML 分割ファイルが一致している場合，そこに含まれている画像ファイルは同一のものである．そのため，差分格納を HTML ファイル以外のファイルにも適用することで，保存スペースの軽減がさらに図れると考えられる．しかし，HTML ファイルと異なり，画像ファイルや PDF ファイルは，分割することが難しい．これらのファイルでは，一部分だけの更新は HTML ファイルのように多くはないと考えられるので，HTML ファイル以外のファイルについては，ファイルの分割は行わず，ファイル自体を一つのブロックとみなし，差分格納を適用することで，保存スペースの軽減の軽減効果があると考えられる．

3.3 収集単位

ここでは，ウェブアーカイビングにおける収集単位について述べる．結論から述べると「ウェブサイト単位」が妥当であると考えられる．これには，差分格納からの観点とウェブアーカイビング本来の目的からの観点からの 2 つの理由がある．

まず，差分格納からの観点であるが，3.2 での空間的類似性についての調査により，ウェブ上の情報には空間的な類似性がみつがったが，レイアウトの似たページは，同一ウェブサイトに多い．これは，ウェブサイトがある一定の目的の基に作成されるためであり，作成者が同じであったり，利用者にとってウェブサイト内は同じレイアウトであった方が操作性が良いなどの理由からである．

つぎに, ウェブアーカイビング本来の目的からであるが, 現在の図書館では, 紙媒体での情報は書誌という単位で管理されている. これは, 情報が 1 ページだけでなくある程度のまとまりをもって一つの情報を発信しているためである. ウェブ上の情報もこれと同様で, 1 つのウェブページでは十分な情報量とは言えず, ウェブサイト単位で情報を発信しているため, ウェブアーカイビングにおける収集単位もウェブサイト単位が妥当であると考えられる. 実際に, 国立国会図書館をはじめ, 多くの組織がウェブサイト単位でのウェブアーカイビングを行っている.

4. システム構成

本研究では 3 章で述べた提案を検証するため, ウェブアーカイビングシステムを構築した.

4.1 システム構成の概要

本システムは大きく分けて二つの部分に分かれる（図 3）.

（1） ウェブサイトの情報を収集し，保存する部分

（2） 収集するウェブサイトの設定や保存した情報を表示する部分

（1）は，ユーザが設定したウェブサイトの収集条件をもとに, 情報を取得し, 保存スペース軽減処理を行い，保存する部分である（以下,「収集・保存部」と言う）. （2）は，管理者や利用者などのユーザのインターフェースとなる部分であり, 収集するウェブサイトの追加や変更, 保存されている情報を再構築し, 表示する部分である（以下,「インターフェース部」と言う）. 本研究では,（1）の部分に主眼を置いているため,（2）の部分においては, 最低限必要な機能のみを実装した.

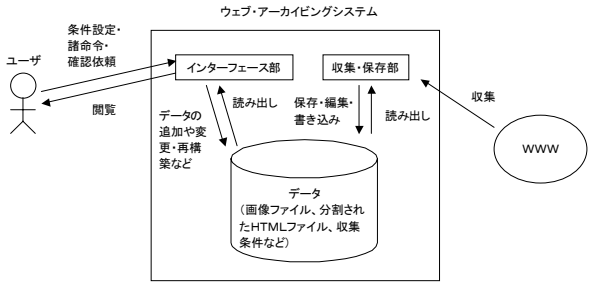

図 3.システムの概要

4.2 本システムで必要となるデータ

本システムは, 収集するウェブサイトの情報をデータベースで管理する. この情報を基に収集や再構築を行う. データベースはウェブサイト管理テーブルと収集管理テーブルの 2 つのテーブルで構成される.

ウェブサイト管理テーブルは, ウェブサイトの収集条件を格納するテーブルで, ウェブサイト番号, ウェブサイトのドメイン, トップページのファイル名, 収集する深さ, 収集する周期, 前回収集した日付の 6 つのフィールドから構成される. 収集管理テーブルは, いつ, どのウェブサイトを収集したかを記録するためのテーブルで, ログ番号, ウェブサイト番号, 日付の 3 つのフィールドから構成される.

本システムは, データベースの他に, ウェブサイトを構成するデータが必要となる. つまり, 収集したウェブサイトを構成するファイルや差分格納によって分割されたファイル, また, それらを再構築するためのインデックスファイルなどである（以下,「アーカイブデータ」と言う）. インデックスファイルは, ウェブサイトの構成をファイル単位で管理するものと, HTML ファイルの分割ファイルを管理するものの 2 種類存在する（以下, 前者を「リストファイル」, 後者を「インデックスファイル」と言う）. アーカイブデータは, ウェブサイトごとに保存される. 各ウェブサイトディレクトリの下には, 収集した日付のディレクトリがあり, その下にアーカイブデータが保存される（図 4）. アーカイブデータは, ウェブ上のファイル構成を忠実に再現した形で保存される. 実際のウェブ上のファイル構成と異なるのは,

（1） 差分格納によって重複が見つかり, 削除されたデータが存在しないこと

（2） HTML ファイルは, 分割され, 差分を取り, 保存すべき差分分割ファイルとインデックスファイルのみが, HTML ファイルと同名のディレクトリの下に保存されること

（3） 日付ディレクトリの下に, HTML ファイル用とその他のファイル用の 2 つのリストファイルが存在すること

である.

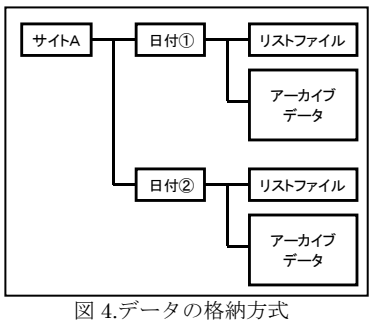

図 4.データの格納方式

4.3 ウェブサイトの情報の収集・保存

収集・保存部は, Java で開発されたプログラ

ムである．収集・保存部は，一度起動すると，エラーなどの例外を除き，終了することはなく，永久的に処理を続ける（図 5）．収集・保存部は，起動されると日付が変わるまで待機し，日付が変わると収集作業を開始する．収集作業が終わると，日付が変わるまで待機する．

収集処理は，まず，データベースのウェブサイト管理テーブルからデータを読み込み，登録されているウェブサイトを先頭からチェックしていく．本日の日付と収集対象ウェブサイトの周期と前回の収集日から，そのウェブサイトを収集すべきか判断する．収集するべきと判断されると，収集対象ウェブサイトのドメイン，トップページ，収集する深さのデータを基に，そのウェブサイトにアクセスし，ワークディレクトリにミラーリングする．次に，ミラーリングされたファイルを基に，リストファイルを作成する．リストファイルは，HTML ファイル用とそれ以外のファイル用の 2 種類作成する．2 つのリストファイルは，共に書式は同じであり，ウェブサイトを構成するファイル一つにつき，ファイル名，ファイルのパス，最終更新日，ファイルサイズ，リンク用パスを出力したものである．リンク用パスとは，差分格納処理を行った際に重複が見つかり，そのファイルを削除した場合に，代わりに利用するファイルのパスのことである．リストファイルが作成された時点では，リンク用パスは全て「no」という値になっている．

ウェブサイトが初めての収集でなければ，「ファイル単位での時間的類似性を利用した差分格納」を行い，前回収集したファイルとの重複を検査する．重複があったファイルは，削除対象となり，リストファイルのリンク用パスを書き換える．「ファイル単位での時間的類似性を利用した差分格納」は，初めて収集する場合は前回の収集ファイルが存在しないので適用されない．次に，「ファイル単位での空間的類似性を利用した差分格納」を行い，今回収集したウェブサイト内でのファイルの重複を検査する．この後，HTML は，

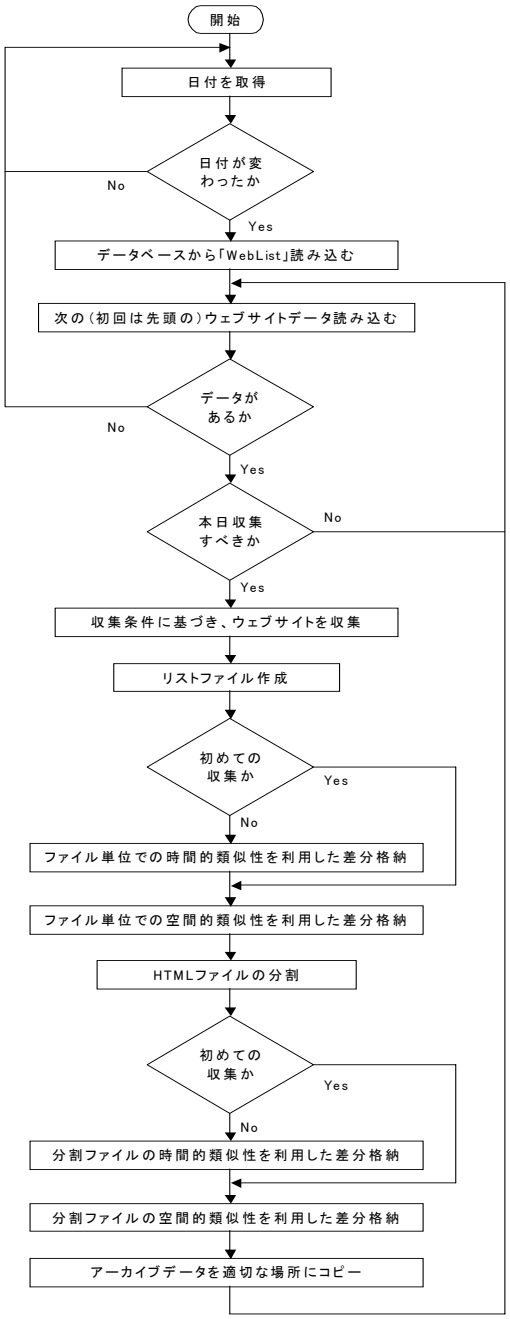

図 5.収集･保存部の処理

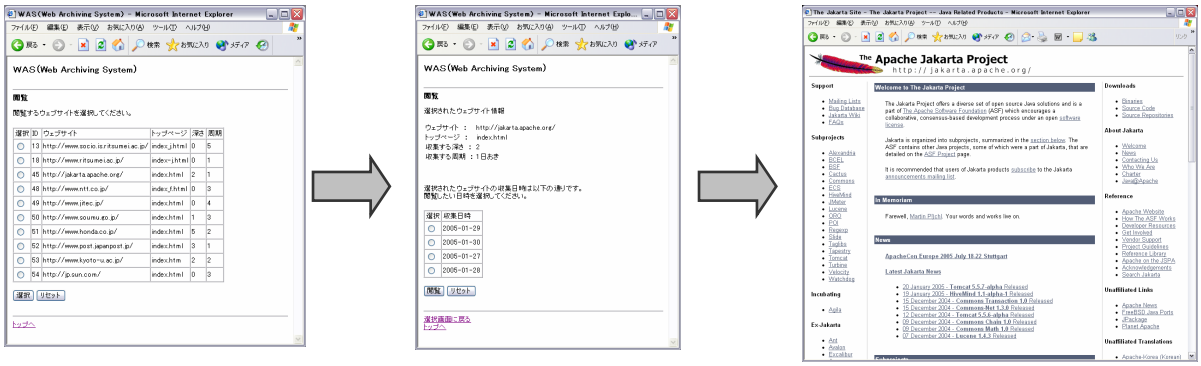

図 6.収集した情報の再構築手順

内部のタグを基に分割され,インデックスファイルと分割ファイルに変換される.分割ファイルは,HTML ファイルを分割したソースコードファイルで，インデックスファイルは，元の HTML ファイルがどのような分割ファイルから構成されるかを記述する.分割された HTML ファイルは,次に「分割ファイルの時間的類似性を利用した差分格納」を行い，前回収集された HTML ファイルとの分割ファイル単位での重複を検査する.「分割ファイルの時間的類似性を利用した差分格納」も，「ファイル単位での時間的類似性を利用した差分格納」と同様に，初めて収集する場合は適用されない．最後に，「分割ファイルの空間的類似性を利用した差分格納」により，今回収集された HTML ファイル同士の分割ファイル単位での重複検査を行う．そして，これらの検査で発見した重複を除去し,残ったファイル及びリストファイルを適切な場所に格納する.

以上の作業を,データベースに登録されている全てのウェブサイトを対象に行い，再び，日付が変わるまで待機する.

4.4 ユーザインターフェース

本システムは，4.3 で述べた収集・保存部の収集条件の設定や保存したウェブサイトの閲覧を行うためのユーザインターフェースを実装している．インターフェース部は，HTML ファイルと JSP ファイルで構成され，ブラウザで動作する.機能としては,収集ウェブサイトの追加機能,収集条件の変更機能,収集ウェブサイトの削除機能，保存したウェブサイトの閲覧機能，ログファイルの閲覧機能である.

情報の再構築は，閲覧機能を用いる．ユーザは，表形式で表示されたウェブサイト，及び，閲覧したい日付を選択する．システムは，渡されたウェブサイトの識別データと収集日データを基に，ウェブサイトを再構築する．まず，ウェブサイトの識別データと収集日データから該当するリストファイルを検出する.リストファイルを先頭から順番に読み取り,ウェブサイトを構成しているファイルを再構築用ディレクトリに作成していく．ファイルが HTML ファイルの場合は,ファイル名と同名のディレクトリを探し,その中に含まれるインデックスファイルを読み込み,必要とする分割ファイルをつなぎ合わせ，元の HTML ファイルを再現する．ウェブサイトの再構築の終了後,そのウェブサイトのトップページを表示する（図 6）.

5. 評価

本研究での提案を 4 章で述べたシステムを用いて評価を行った．実験結果と考察を述べる.

5.1 圧縮率

ここでは,収集したデータの圧縮率について評価する．収集は，本システム，ミラーリングソフトの「wget[9]」の 2 種類で行い，比較する．圧縮率は，wget を基準にしている．wget はミラーリングソフトであり,ウェブ上の情報を忠実に再現するため,ウェブ上のデータと同一の値になる.

ニュースサイトが 2 サイト，企業・団体サイトが 9 サイト，個人サイトが 4 サイトの合計 15 サイトをそれぞれ 5 回ずつ収集した結果,HTML ファイルで 38.4%,その他のファイルで 20.6%,

表 1.収集結果

		wget	本システム
ファイル容量（バイト）	全体	769,451,591	183,685,241
	HTMLファイル	139,822,133	53,702,638
	その他のファイル	629,629,458	129,982,603
圧縮率	全体	100.0%	23.9%
	HTMLファイル	100.0%	38.4%
	その他のファイル	100.0%	20.6%

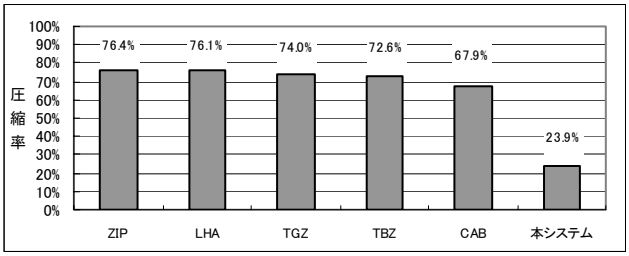

図 7.主な圧縮形式との比較

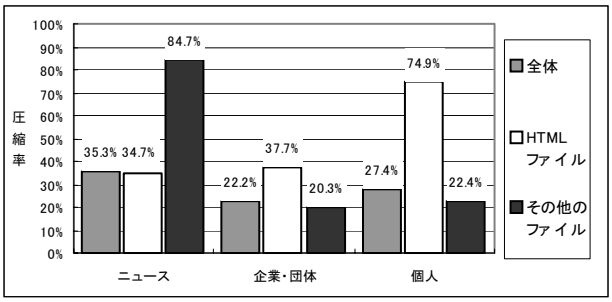

図 9.ジャンル別圧縮率

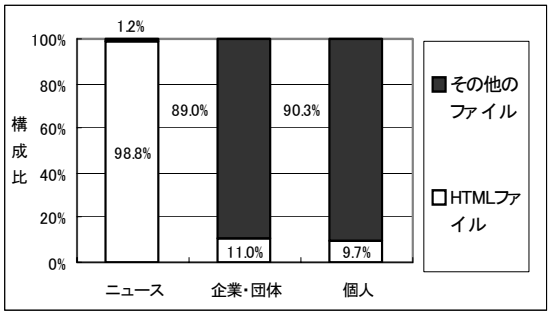

図 8.ジャンル別構成比

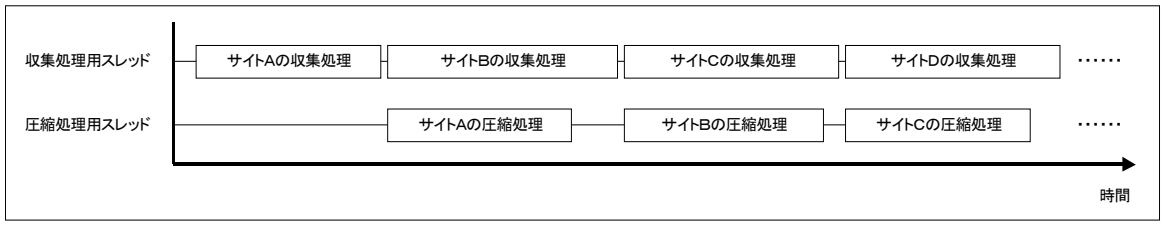

図 10.スレッド処理例

表 2.処理時間と割合

	収集時間	圧縮時間	その他	全体
時間(ミリ秒)	1259758	877980	29809	2167547
割合(%)	58.1%	40.5%	1.4%	100.0%

全体で 23.9%となった（表 1）．これは，一般的に使用されている圧縮形式よりも高い圧縮率であった（図 7）．圧縮率は，wget で収集したデータを，フリーウェアの「+Lhaca（Version 1.20）[10]」を用いて，ZIP 形式，LHA 形式，CAB 形式，TGZ 形式，TBZ 形式で圧縮した結果から求めた．この結果からも，本提案の時間的類似性，空間的類似性を利用した差分格納は，ウェブアーカイビングに適していると考えられる．

次に，収集したウェブサイトをジャンル別に考察する．実験に用いた 15 サイトを「ニュース」サイト，「企業・団体」サイト，「個人」サイトに分け，それぞれのファイル容量の構成比，圧縮率を求めた（図 8，図 9）．

全体の圧縮率で最も圧縮できなかったのはニュースサイト（35.3%）であった．ニュースサイトは，3 つのジャンル中，一回の更新で行われる変更が最も多い．このため，全体の圧縮率が低くなったと考えられる．しかし，ニュースサイトの HTML ファイルの圧縮率は，全ジャンル中，最も高い（34.7%）．これは，ニュースサイトには，同じレイアウトのページが数多く存在し，差分格納を適用しやすかったからであると考えられる．逆に，その他のファイルは，84.7%と他の 2 ジャンルに比べて，大幅に低い圧縮率となった．ニュースサイトでは，画像などの HTML ファイル以外のファイルは，レイアウト用や記事の写真などに使われる．その中でも，記事に使われる画像が多く，その画像は記事ごとに違うものが使われるので，重複が見つかりにくく，低い圧縮率となっている．しかし，構成比からもわかるように，HTML ファイルがウェブサイトのほとんどを占める（98.8%）ため，その他のファイルの圧縮率の低さは，全体の圧縮率においてそれほど影響は出ていない．

他に，圧縮率において他の 2 ジャンルに比べて大幅な違いを見せたのは，個人サイトにおける HTML ファイルである（74.9%）．個人サイトの HTML ファイルが大幅な低い圧縮率となったのは，ページ間におけるレイアウトの統一の無さである．また，表示におけるレイアウトが似ていても，使用しているタグが違うことなどにより，タグによって HTML ファイルを分割する本システムでは重複を見つけられなかったためである．これは，企業や団体のサイト，ニュースサイトは規模が大きく，HTML ファイル作成支援ソフトやテンプレートなどを使用し，作成しているのに対し，個人ページでは管理者の手作業による割合が高いためであると考えられる．

5.2 処理時間

本研究の一番の目的は，ウェブ上の情報の圧縮であり，処理時間は主眼ではないが，大きな問題とならないかの評価と，今後の改善点を見つける材料のため，処理時間においても計測した．5.1 で行った実験において，収集にかかった時間，及び，収集した情報の閲覧時にかかる時間を計測した．本実験に使ったコンピュータのスペックは，OS が Windows XP，CPU（Central Processing Unit）が Pentium4 の 2.8GHz，メインメモリが 1G バイトである．

収集・保存時の処理時間を収集時間，圧縮時間，その他に分けて計測した．収集時間は，ウェブ上から情報を収集する時間であり，圧縮時間は，収集したファイルに差分格納を適用するのにかかった時間である．それぞれのウェブサイトを 1 回収集するのにかかった時間の平均の合計と割合を表 2 に示す．本システムにおける処理は，収集時間と圧縮時間が全体のほとんどを占め（98.6%），その他の時間は無視できる範囲（1.4%）であった．収集時間と圧縮時間は，それぞれ全体の 58.1%と 40.5%であり，収集時間の方が長いことがわかった．収集時間は，主にウェブ上からのダウンロードにかかった時間であり，通信回線に依存し，CPU にはあまり負担をかけない．逆に圧縮時間は，CPU の処理能力に依存し，通信環境は無関係である．そのため，スレッド処理などを用いて，収集処理と圧縮処理を別々に処理した場合，圧縮時間の方が収集時間より短ければ，全体の処理時間において，圧縮時間はそれほど問題とならない（図 10）．実験の結果の収集時間 58.1%，圧縮時間 40.9%は，収集時

間より圧縮時間が短いということなので，収集・保存において，本提案の圧縮は大きな問題とならないことがわかる．

次に，収集したデータの復元にかかる時間について述べる．収集したデータを復元する時間を計測した．収集したウェブサイトの情報に対して，ウェブサイト，日付を指定した収集一回あたりの情報の復元にかかる時間を計測し，その合計，及び，平均を計算した．今回の実験では，一回の収集分の復元に最も時間のかかったウェブサイトは，約 16.6 秒，最速は約 0.5 秒，平均は，約 5.5 秒であった．本システムでは，ウェブサイト全体を復元し，トップページを表示する．そこで次に，ウェブページ単位で復元する場合を考える．今回の収集全ての情報を復元するのにかかった時間は，411256 ミリ秒であった．また，その際，復元されるデータの総容量は表 1 より 769451591 バイトである．さらに，ウェブページ 50 ページを調べたところ，一つのウェブページを構成するのに必要なファイルは 20.78 ファイル，容量は 89485.28 バイトであった．よって，ウェブページ 1 ページを復元するのに必要な時間は，

89485.28/(769451591/411256)=約 47.8 ミリ秒

である．一つのウェブページの復元に約 0.05 秒という値は，ウェブ上の閲覧においてほぼ問題ないと思われる．

6. おわりに

本研究では，現在各国の国立図書館などで取り組まれているウェブアーカイビングにおける保存スペースの軽減のための，データ圧縮方法を開発することを目的とした．ウェブ上の情報のデータ圧縮を実現するため，既存の圧縮形式を用いるのではなく，ウェブ上の情報の空間的類似性と時間的類似性に着目し，ウェブサイト内の重複を無くすことにより，ウェブアーカイビングに適した差分格納方式を提案した．

本提案を実装したシステムを構築し，ウェブ上の情報に見立てたミラーリングソフトの容量と比較することで，本提案のデータ圧縮方式がウェブアーカイビングに適した圧縮方式であることついて考察した．

今後は，さらに大規模な収集を行い，調査していく．そして，圧縮率の向上，圧縮時間の短縮を目指す．

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，情報学基礎 No.071-012 (2003)

[2] 富岡麻理：『ウェブ・アーカイビングの現状』，慶應義塾大学，http://www.slis.keio.ac.jp/gsec2001.pdf (2001)

[3] 国立国会図書館インターネット資源選択的蓄積実験事業(WARP)，国立国会図書館，http://warp.ndl.go.jp/

[4] 豊田正史，喜連川優：大規模 Web アーカイブからのデータマイニング，情報処理学会 会誌，情報処理 Vol.46 No.1 pp46-51 (2005)

[5] 大塚信吾，宮崎収兄：二段階圧縮法の XML への適用，情報処理学会 研究報告，データベースシステム No125-53 (2001)

[6] 遠藤裕英：スクリプト細分割記憶による Web アーカイブ方式，情報処理学会 研究報告，情報学基礎 No.73-8 (2003)

[7] 中山雅照：更新履歴を利用した Web ページの送信データ量削減方式，立命館大学大学院理工学研究科修士論文 (2002)

[8] アーカイブ形式解説，http://www.nemu.to/Beta/

[9] wget，The GNU Project，http://wget.sunsite.dk/

[10] 村山富男：+Lhaca，http://park8.wakwak.com/~app/Lhaca/