Personal Web Archive: Webページのバージョン閲覧支援システム

若松 亮太 † **志築 文太郎** † **田中 二郎** †

† 筑波大学大学院 システム情報工学研究科 コンピュータサイエンス専攻

E-mail: † waka@iplab.cs.tsukuba.ac.jp, {shizuki,jiro}@cs.tsukuba.ac.jp

概要

過去に閲覧したWebページで見た情報を再び見ようとしても，Webページから見た情報が削除されていること，あるいはWebページそのものが削除されていることがしばしばある．Webアーカイブに削除された情報が存在する可能性はあるが，利用者が閲覧したバージョンが保存されているとは限らず，逆に閲覧していないバージョンが存在するため，それらのバージョンを比較しながら目的の情報を探すのは手間がかかる．本論文では，Webブラウジングの過程で個人的なWebアーカイブを作成し，保存した複数のバージョン間の差分を提示することによって，閲覧経験のある知識の再発見を支援するシステム **Personal Web Archive** について述べる．

Personal Web Archive: Support System for Browsing the Versions of Web Page

Ryota WAKAMATSU † Buntarou SHIZUKI † Jiro TANAKA †

† Graduate School of Systems and Information Engineering
University of Tsukuba

E-mail: † waka@iplab.cs.tsukuba.ac.jp, {shizuki,jiro}@cs.tsukuba.ac.jp

Abstract

We present **Personal Web Archive** which supports rediscovering the knowledge with a browseing history. We often revisit the web pages which we browsed in the past to look for the information again. However, sometimes the information or the web page itself has been deleted. Personal Web Archive is a plug-in of a web browser. It archives all the web pages during web browsing into users' personal archive. It also shows differences between multiple versions of the revisited web page in the archive to help users find required information from those versions.

1 はじめに

近年，World Wide Web（以下Web）上には無数のWebページが存在し，それらのWebページは頻繁に更新を繰り返している．このため，過去に閲覧したWebページで見た情報を再び見ようとしても，そのWebページから見た情報が削除されていること，あるいはWebページそのものが削除されていることがしばしばある．あらかじめWebページを保存しておけば元の情報を見ることは可能だが，閲覧中に必要だと思わなかった情報が後に必要になることがあるため，保存しておくべきWebページを見極めるのは困難である．

Internet Archive[1]（Wayback Machine[2]）に代表される典型的なWebアーカイブ（例えば[3, 4]）や，個々のWebサイトによるWebサイト自身の

アーカイブ（例えば，ウェブログでは投稿された記事がアーカイブとして月毎に纏められている）に削除された情報が存在する可能性はある．しかし，利用者が閲覧したバージョンが保存されているとは限らず，逆に閲覧していないバージョンが存在するため，それらのバージョンの中から目的の情報を探すのは手間がかかる．また，Greenberg らの研究によると Web 閲覧の大部分は，ブラウザのバック・フォワード，ブックマークなどを用いた同じ URL を持つ Web ページへの再訪問である [5, 6]．したがって，閲覧者は同じ URL を持つ Web ページの一部だけが異なる別のバージョンを複数閲覧していることになるが，それらの類似するバージョンを比較して，どのバージョンのどこに過去に閲覧した Web ページで見た情報が存在するかを判断するのは困難である．これらの既存のシステムの問題点を考慮すると，過去に閲覧した Web ページで見た情報を再び見るためには，以下のような特徴を持つシステムがあればよいと考えられる．

- 利用者が閲覧した Web ページのみを収集した Web アーカイブを作成する
- 多くの類似点を持つ Web ページのバージョン間の情報の比較を支援する

我々は，これらの特徴を持つシステム **Personal Web Archive** の開発を行った．図 1 にシステムの概観を示す．Personal Web Archive は，Web ページの閲覧と同時に保存することによって作成する個人的な Web アーカイブの視覚化を行う．また，Web ページのバージョン間の差分を同一のビューで提示し，類似する Web ページの比較を支援する．

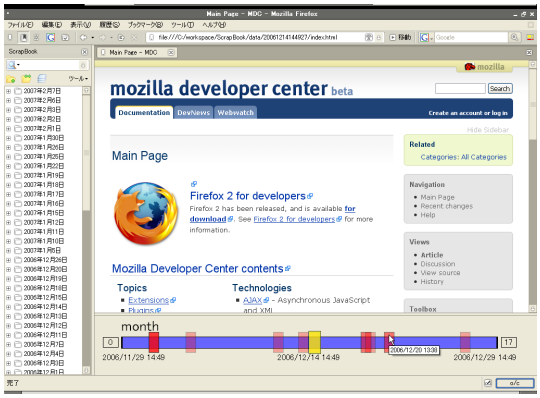

図 1: システムの概観

2 Personal Web Archive

2.1 Web アーカイブと時系列の視覚化

Personal Web Archive では，過去に閲覧した Web ページで見た情報を再び見る際の問題点を踏まえ，利用者の閲覧に即した個人的な Web アーカイブの作成を行う．利用者が Web ページを閲覧する際，その Web ページのスタイルシートや画像などを含むコンテンツをローカルディスク上に自動的に保存する．また，同時に Web ページの URL やタイトル，閲覧時刻といった閲覧に関するメタデータを記録しておく．このようにして保存された Web ページの集合を Web アーカイブとして扱う．

利用者が Web ページを閲覧する際，Web アーカイブの中から閲覧中の Web ページと同じ URL を持つ Web ページを抽出し，時系列に沿って提示する．本システムでは，図 2 に示すように，時系列を表す直線上の Web ページの時間座標に対応する位置にデータを並べることによって視覚化する．

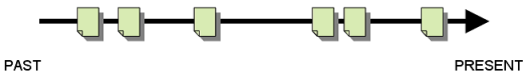

図 2: 時系列データの視覚化

このように視覚化された Web アーカイブから閲覧した Web ページで見た情報を探す際，利用者の閲覧したページのみで Web アーカイブが構成されているため，余計な情報に混乱させられることなく目的の情報を捜すことができる．また，実際の Web 閲覧履歴に基づいた Web アーカイブが閲覧時刻の間隔や位置を再現する形で視覚化されているため，利用者は目的の情報を見た時期を想起しやすくなる．

2.2 バージョン間の差分の提示

Personal Web Archive システムでは，作成した Web アーカイブの個々のバージョンを閲覧する際，直前のバージョンとの差分を強調表示し，バージョン間の変更点を閲覧者に提示する．

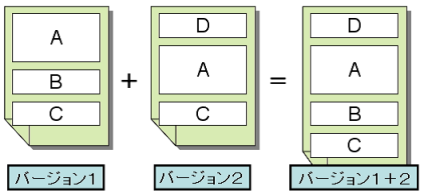

図 3: バージョン間の差分の提示

図 3 はバージョン間の差分の提示の概要を示している．バージョン 1 とバージョン 2 は連続するバージョンである．このとき，バージョン 1 の B の部分はバージョン 2 の時点では削除され，逆にバージョン 2 の D の部分が追加されている．バージョン 1+2 では，バージョン 1 及びバージョン 2 をマージし，差分を強調表示することによって，バージョン間の変更点を閲覧者に提示する．

Web アーカイブを構成する Web ページのそれぞれバージョンは，一部が異なるだけで大部分は同じ情報であることが多い．そこで，すべての情報を提示した上で変更点を強調表示することによって，利用者はその部分を追うだけで過去に閲覧した Web ページから削除されてしまった情報を探すことができる．

2.3 システムの概要

図 1 は Personal Web Archive システムを組み込んだ Web ブラウザである．一般的な Web ブラウザと同様に Web ページを提示する領域を持ち，その下部に Web アーカイブ提示領域を持つ．図 4 に Web アーカイブ提示領域を示す．まず，中央の棒状の領域は閲覧中の Web ページの時系列を表す．この領域は表示範囲を持ち，閲覧している Web ページの閲覧時刻を中心として，6 段階のスケール（decade, year, month, week, day, hour）が示す範囲を表示する．左上のテキストが現在のスケールを示している．この表示範囲の変更はマウス右クリック＋移動のマウスジェスチャで行い，上下に移動させることでスケールの変更，左右に移動させることで同一スケール内での表示範囲の前後移動を行う．棒状の領域上の小さな矩形は表示範囲内に存在する Web ページのバージョンを示し，特に，現在閲覧しているバージョンを表す矩形は異なる色で示す．また，棒状の領域の両端には表示範囲外に存在するバージョンの数を示す．それぞれのバージョンは透明度が高く設定されており，多くのバージョンが集まっている時点ほど矩形が重なるために周辺の透明度が低くなり，その時点周辺で頻繁に閲覧中の URL を閲覧していたことが視覚的に利用者に伝えられる．

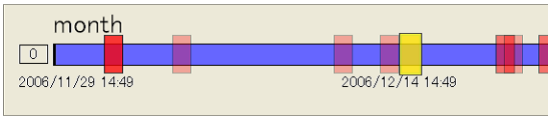

図 4: Web アーカイブ提示領域

単一バージョンの閲覧

利用者が時系列上においてバージョンを示す矩形をクリックするとき，そのバージョンが選択される．システムは選択されたバージョンとその直前のバージョンと比較し，追加された情報や削除された情報といった差分を強調表示する．図 5 に単一バージョンの差分の強調表示の例を示す．追加された情報は背景を赤色に強調表示する．また，削除された情報は元の位置に復元し，背景を青色に強調表示をする．

利用者は強調表示された部分を確認しながらバージョンを順に閲覧していくことにより，目的の情報を見つけることができる．例えば，利用者が探している情報が現在閲覧しているバージョンに追加された情報より後に追加された情報だと感じたならば，閲覧するバージョンを新しくすればよい．

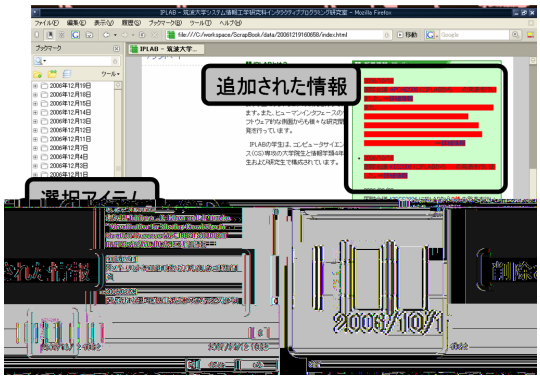

図 5: 更新差分の提示

複数バージョンの閲覧

利用者が時系列上においてドラッグ&ドロップを行い，その範囲に複数のバージョンを示す矩形が存在するとき，それらバージョンがすべて選択される．システムは選択されたすべてのバージョンを比較し，それらの差分をマージし，単一バージョンの閲覧を行う場合と同様に強調表示する．図 6 に複数バージョンの差分の強調表示の例を示す．単一のバージョンが選択された場合の提示方法に加え，追加や削除の鮮度によって背景色の濃度を調節し，新しく追加された情報（古く削除された情報）ほど濃度を高くする．

利用者は，記憶から想起するなどによってある程度の時期を推定して範囲選択を行い，その結果から目的の情報を探していく．現在強調表示されている最も新しい追加情報された情報（古い削除された情報）より未来（過去）に目的の情報が追加された（削除された）と感じた場合，選択範囲を未来方向（過去方向）に拡げて探すことにより，目的の情報を見つけることができる．また，選択範囲のバージョン内に探している情報が存在すると感じた場合，選択範囲を絞っていけばよい．

図 6: 複数バージョン間の差分

3 実装

Personal Web Archive システムは Web ブラウザ Firefox の拡張機能として開発を行った．これは JavaScript，XUL（XML-based User-interface Language）などの言語により記述される．Web ブラウザに組み込む形となるため，普段の Web ブラウジングの環境に近い形でデータ収集や評価を行うことができる．

3.1 閲覧と同期したアーカイブの作成

Web アーカイブの作成には，Firefox の拡張機能として公開されている ScrapBook[7] とそのアドオン機能である AutoSave を用いた．Web ページが読み込まれると同時に，Web ページのコンテンツの保存と Web ページが持つメタデータ（URL，タイトルなど）が RDF 形式で保存される．本システムでは，この ScrapBook によって収集された Web ページの集合を Web アーカイブとして扱う．

3.2 バージョン間の差分の提示

単一のバージョンが選択されている場合について述べる．選択されているバージョンと直前のバージョンの DOM（Document Object Model）ツリーを走査し，それぞれの DOM ツリーを比較し，追加された HTML タグ・削除された HTML タグを検出する．削除された HTML タグが検出されているとき，閲覧中のバージョンの DOM ツリーにその HTML タグを復元する．さらに，復元された HTML タグと検出された追加された HTML タグに対し，id 属性と class 属性を加え，それぞれのスタイルシートに合わせたデザインで閲覧者に提示する．単一のバージョンが選択されている場合の DOM ツリーへの操作例を図 7 に示す．

複数のバージョンが選択されている場合，最新のバージョンと直前のバージョンに対して単一のバージョンが選択されている場合と同じ操作を行い，変更された DOM ツリーとさらに前のバージョンに対してさらに同じ操作を繰り返す．複数のバージョンが選択されている場合の DOM ツリーへの操作例を図 8 に示す．

```
<body>
  <div>記事 C</div>
  <div>記事 B</div>
  <div>記事 A</div>
</body>
```

バージョン 1

```
<body>
  <div>記事 D</div>
  <div>記事 C</div>
  <div>記事 B</div>
</body>
```

バージョン 2

```
<body>
  <div class="add" id="add1">
       記事 D
  </div>
  <div>記事 C</div>
  <div>記事 B</div>
  <div class="del" id="del1">
       記事 A
  </div>
</body>
```

バージョン 1+2

図 7: DOM ツリーへの操作（単一）

4 関連研究

杉浦，古関らの Internet Scrapbook[8] では，Web ページ内の一部または全部の情報を切り取り，閲覧者が必要な情報を集めた仮想的な Web ページを作成する．Web ページ内の情報を保存し，蓄積する点において我々の研究と同様であるが，保存に明示的な操作が必要なため，保存しておくべき Web ページ内の情報を確実に収集することは困難である．

白井らは閲覧した Web ページを保存し，閲覧中の Web ページに類似する Web ページ，関連のある Web ページを Web アーカイブ内から提示するシステムの構築を行った [9]．彼らのシステムでは，それぞれのページへのインデックスを提示し，実際の閲覧は従来のインタフェースで行った．一方，我々は同じ URL を持つ Web ページ，つまり，別のバージョンの Web ページを同一のビューで閲覧，比較することに重点を置いた．

```
<body>
  <div>記事 B</div>
  <div>記事 A</div>
  <div>記事 Z</div>
</body>
```

バージョン 0

```
<body>
  <div class="add" id="add1">
       記事 D
  </div>
  <div>記事 C</div>
  <div>記事 B</div>
  <div class="del" id="del1">
       記事 A
  </div>
</body>
```

バージョン 1+2

```
<body>
  <div class="add" id="add1">
       記事 D
  </div>
  <div class="add" id="add2">
       記事 C
  </div>
  <div>記事 B</div>
  <div class="del" id="del1">
       記事 A
  </div>
  <div class="del" id="del2">
       記事 Z
  </div>
</body>
```

バージョン 0+(1+2)

図 8: DOM ツリーへの操作（複数）

また，A. Jatowt らの Past Web Browser[10] は，Internet Archive[1] のような Web アーカイブを時系列データとして視覚化し，その時系列データとしての Web ページを提示する際に直前のバージョンとの差分の提示を行う．我々のシステムは，利用者の Web 閲覧によって収集された利用者に最適化された Web アーカイブの視覚化を行い，複数のバージョンを同一のビューで閲覧，比較する点において異なると言える．

5 議論

本論文では，インタフェースの比較対象として Internet Archive[1]（Wayback Machine[2]）に代表される典型的な Web アーカイブを取り上げた．典型的な Web アーカイブでは，過去に閲覧した Web ページで見た情報を再び見るために，目的の情報を持つ Web ページの URL が分かっていなければならない．同様に本システムでは，目的とする Web ページの URL が分かっている状況，つまり，Web ブラウザのブックマークや Web 閲覧履歴，アドレスバー，Web ページに張られているリンクなどから目的の情報を持つ Web ページを訪問できる状況，または，既にその Web ページを閲覧中である状況を利用場面として想定する．過去に閲覧した Web ページへの再訪問およびその手段については，第一章で述べた Greenberg らによって研究が行われている [6, 11]．一方，我々は再訪問後の Web ページの閲覧に焦点を当てて研究を行っている．

Web ブラウザの Web 閲覧履歴や更新日時でソートしたファイルの提示などに用いられている時系列表示は，図 9 のような単純な一次元リストで行われることが多い．しかし，このような一次元リストではデータ間の時間間隔や多くのデータが集まっている時間座標を想起することは難しい．

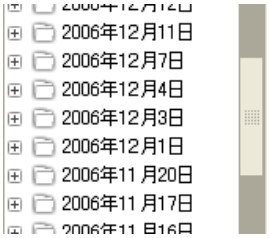

図 9: 単純な一次元リストによる時系列表示

本システムでは，閲覧中の Web ページの時系列を表す矩形上にバージョンを配置することによって Web アーカイブの提示を行ったが，この方法以外にも時系列を持つデータの時系列の提示方法は考えられる．例えば，佐藤による dripdrop[12] では，時系列を表す矩形に対し，周期とスケールを持つリングを用いて操作するインタフェースを提示している．その他の時系列の提示・操作インタフェースについても議論する必要がある．

複数のバージョンを閲覧するとき，表示範囲内で追加された情報が削除される場合がある．現在の実装ではこのような情報は，最終的に削除されているため削除された情報として扱っている．しかし，閲覧した情報を再現するためには，追加・削除の両方として扱う方がよい場合もあるため，本システムでは，両方を表示することも可能である．

ローカル PC 上に Web アーカイブを作成することによって，ハードディスクの容量が消費される問題について述べる．2001 年の矢野 [13] の調査によると，Web ページを一年間に閲覧する量は画像ファイルを含めて 1.6GByte であった．一方，2007 年の我々の調査では 2.1GByte であり，約 1.3 倍に増加している．1999 年から隔年のハードディスクの中で，容量 1GByte あたりの価格が安い商品を調査した結果を表 1 に示す（Impress Watch[14] 調べ）．表中の「容量」はハードディスクの容量（GByte），「GB 単価」は 1GByte あたりの価格（円），「平均価格」はその平均価格（円）を示す．

表 1: ハードディスクの容量と価格

調査年度	容量	GB 単価	平均価格
1999	10	2196	21,959
2001	80	348	27,866
2003	120	120	14,378
2005	160	51	8,092
2007	250	32	7,886

表 1 の 2001 年と 2007 年を比較する．2001 年の GB 単価 348 円に対し，2007 年では 32 円となり，約 1/11 倍まで低下している．すなわち，閲覧する量が約 1.3 倍しか増加していないのに対し，等価で手に入るハードディスクは約 11 倍まで増加している．また，ハードディスクの容量自体も 3 倍以上に増加している．したがって，ハードディスクの容量の消費に関しては考慮しなくても問題ないといえる．

また，システムの使用感に対する知見を得るためにユーザテストを行うことが今後の課題である．

6 まとめ

本研究では，閲覧経験のある Web ページに存在する情報の再発見を支援するシステム **Personal Web Archive** を実装を行った．本システムは，Web ページの閲覧と同時に保存することによって作成する個人的な Web アーカイブの視覚化を行う．また，Web ページのバージョン間の差分を同一のビューで提示し，類似する Web ページの比較を支援する．このシステムを用いることによって，利用者は Web 上で過去に得た知識の想起を容易に行うことができると考えられる．

参考文献

[1] Internet archive. `http://www.archive.org/`.

[2] Wayback machine. `http://www.archive.org/web/web.php`.

[3] Election 2002 web archive browse. `http://lcweb4.loc.gov/elect2002/`.

[4] The september 11 web archive. `http://september11.archive.org/`.

[5] Linda Tauscher and Saul Greenberg. Revisitation patterns in world wide web navigation. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 399–406. ACM Press, 1997.

[6] Andy Cockburn, Saul Greenberg, Steve Jones, Bruce McKenzie, and Michael Moyle. Improving web page revisitation: Analysis, design and evaluation. *IT & Society*, Vol. 1, 3, Winter 2003, pp. 159–183, 2003.

[7] Scrapbook. `http://amb.vis.ne.jp/mozilla/scrapbook/`.

[8] Atsushi Sugiura and Yoshiyuki Koseki. Internet scrapbook: automating web browsing tasks by demonstration. In *UIST '98: Proceedings of the 11th annual ACM symposium on User interface software and technology*, pp. 9–18. ACM Press, 1998.

[9] 白井良成, 中小路久美代, 山本恭裕. インタラクションヒストリによる web ブラウジング拡張. インタラクション 2006, 情報処理学会, pp. 223–224, 2006.

[10] Adam Jatowt, Yukiko Kawai, Satoshi Nakamura, Yutaka Kidawara, and Katsumi Tanaka. A browser for browsing the past web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pp. 877–878. ACM Press, 2006.

[11] Shaun Kaasten and Saul Greenberg. Integrating back, history and bookmarks in web browsers. In *CHI '01: CHI '01 extended abstracts on Human factors in computing systems*, pp. 379–380. ACM Press, 2001.

[12] 佐藤省吾. メタデータに基づく検索/閲覧インタフェース. Master's thesis, 筑波大学大学院理工学研究科, 2006.

[13] 矢野慎一郎. Web ブラウザにおける時間情報を考慮した履歴機能の検討と実装. 卒業論文, 筑波大学第三学群工学システム学類, 2001.

[14] Impress watch. `http://www.watch.impress.co.jp/`.