筑波大学大学院博士課程

システム情報工学研究科修士論文

個人用Webアーカイブの閲覧支援システム

若松 亮太

(コンピュータサイエンス専攻)

指導教員　田中 二郎

2009年3月

概要

World Wide Web（以下 Web）上には無数の Web ページが存在し，それら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後から再び見ようとしても，その情報が Web ページから削除されていることや，Web ページ自体が削除されていることがしばしばある．この削除された情報や Web ページを閲覧するための方法として，Web ページの閲覧中にその Web ページの複製をローカルマシン上に保存しておき，それを閲覧する方法や，Web ページの複製を収集したものを公開する Web サービスである Web アーカイブを利用して目的の情報や Web ページを閲覧する方法がある．しかし，それらの方法では，手動で保存を行う必要があったり，第三者によって保存が行われていたりするため，閲覧したい Web ページが確実に保存されているとは限らない．また，そのインタフェースについても，保存した Web ページの閲覧を積極的に支援しているとはいえない．

本論文では，これらの問題を解決するために作成したシステム Personal Web Archive について述べる．Personal Web Archive は，閲覧者が Web 閲覧を行う過程で，閲覧した Web ページの複製を収集した個人用 Web アーカイブの作成を自動的に行う．さらに，作成した個人用 Web アーカイブ内に存在する，保存時刻が異なるが同一の URL を持つ Web ページ群に対し，その中の複数の Web ページ間の差分を同一画面内に提示することによって，それらの Web ページの比較，閲覧の支援を行う．また，本システムを利用することによって，どのように Web 閲覧が支援されるかの確認を行った. その結果と既存のシステムを用いた場合の結果の比較による本システムの有効性の検証について述べる. 最後に，本システムについての考察と今後の課題について述べる．

目 次

第 1 章　はじめに **1**
1.1　研究の背景 1
1.2　本研究の目的 2
1.3　本論文の構成 2

第 2 章　再訪問・Web アーカイブの現状と問題点 **3**
2.1　再訪問の目的と既存のシステム・機能 3
2.2　Web アーカイブ 10
2.2.1　Web ページの収集，保存の問題 10
2.3　再訪問のインタフェースの問題 13
2.3.1　Web ページの差分に関する問題 14
2.4　問題解決に必要な機能 15

第 3 章　関連研究 **16**
3.1　再訪問に関する研究 16
3.2　Web アーカイブに関する研究 16
3.2.1　作成 16
3.2.2　検索 17
3.2.3　閲覧 17
3.3　その他 18

第 4 章　Personal Web Archive **20**
4.1　Web アーカイブの作成 21
4.2　Web アーカイブの可視化 21
4.3　バージョン間の差分の提示 22
4.3.1　単数バージョンの閲覧 22
4.3.2　複数バージョンの閲覧 24
4.4　システムの詳細 26
4.4.1　単数バージョンの閲覧 28
4.4.2　複数バージョンの閲覧 29

第 5 章　実装 **31**
5.1　Web アーカイブ作成部 32

5.2 データ処理部 32
5.2.1 データの読み込み 33
5.2.2 差分の抽出 33
5.3 データ提示部 36

第 6 章 システム利用例 **37**
6.1 過去の記事を探す 37
6.2 記事の続きを読む 40

第 7 章 議論と今後の課題 **42**
7.1 再訪問のための検索について 42
7.2 個人用 Web アーカイブについて 42
7.3 Personal Web Archive について 44

第 8 章 おわりに **45**

謝辞 **46**

参考文献 **47**

図 目 次

2.1 再起率の調査結果 4
2.2 アドレス閲覧とコンテンツ閲覧の例 5
2.3 Wayback Machine 11
2.4 Web ページの更新時刻とクローラの保存時刻のずれ 12
2.5 Web アーカイブ閲覧の流れ 13

3.1 Past Web Browser の概観 18

4.1 Personal Web Archive の概観 20
4.2 時系列データの可視化 21
4.3 単数バージョンの閲覧 23
4.4 単数バージョン間の差分の提示 23
4.5 複数バージョンの閲覧 24
4.6 複数バージョン間の差分の提示 25
4.7 Personal Web Archive の構成 26
4.8 Web アーカイブ提示部 27
4.9 単数バージョンの閲覧における差分の強調 28
4.10 複数バージョンの閲覧における差分の強調 29

5.1 システム構成 31
5.2 単数バージョンの閲覧における差分の抽出 34
5.3 複数バージョンの閲覧における差分の抽出 35

6.1 トップページへの訪問 38
6.2 表示期間の変更 39
6.3 バージョンの選択 39
6.4 結果の閲覧 40
6.5 バージョンの選択 41

7.1 1GB 単価と容量の推移 44

表 目 次

2.1 再訪問の目的と Web ページの状況 7

5.1 保存するメタデータ 32

6.1 作成した個人用 Web アーカイブ 37

7.1 個人用 Web アーカイブのデータサイズ 43
7.2 ハードディスクの容量と価格 43

第1章　はじめに

1.1　研究の背景

World Wide Web（以下 Web）上には無数の Web ページが存在し，それらは頻繁に更新を繰り返している．このため，Web ページで過去に見た情報を後から再び見ようとしても，その情報が Web ページから削除されていることや，Web ページ自体が削除されていることがしばしばある．

この削除された情報や Web ページを閲覧するために，いくつかの方法がある．1 つ目に，閲覧者が明示的に Web ページを保存する方法がある．多くの Web ブラウザに備わっている「ページの保存」機能やウェブ魚拓 [1] などの Web サービスを利用することによって，ローカルマシン上や Web 上に Web ページの複製を保存しておくことができる．しかし，閲覧中に必要だと思わなかった情報が後から必要になることがあるため，保存しておくべき Web ページのすべてを閲覧時に見極めるのは困難である．

2 つ目に，閲覧者は事前に特別な作業を行わず，情報発信者や第三者が準備したシステムを利用して削除された Web ページを閲覧する方法がある．例えば，Internet Archive の運営する Wayback Machine[2] に代表される Web アーカイブ（例えば，[3, 4, 5]）や個々の Web サイトによる自身のコンテンツのアーカイブ（例えば，多くの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）に削除された情報や Web ページが存在する場合がある．しかし，上記のような一般的な Web アーカイブでは Web ページの保存をクローラに依存するため，Web ページの保存タイミングの設定や Robots Exclusion Protocol[6] の関係で，閲覧者が過去に閲覧したすべてのバージョンが保存されているとは限らない．したがって，閲覧者が探している情報や Web ページが見つからない場合がある．逆に，閲覧者自身が閲覧していないバージョンが保存されていることもしばしばある．これらのバージョンが混在する中には，更新によって一部の情報だけが異なる Web ページが大量に存在する．それらの大量の類似するバージョン群の中の，どのバージョンの，どの位置に Web ページで過去に見た情報が存在するかを判断するのは難しい．以上のように，Web アーカイブの中から目的の情報や Web ページを探し出すのは非常に困難である．

Greenberg らの研究 [7, 8] によると，Web 閲覧の大部分は Web ブラウザの「戻る」,「進む」,「履歴」,「ブックマーク」などの機能を用いた同じ URL への再訪問である．再訪問が頻繁に行われているにも関わらず，削除された情報や Web ページを閲覧する方法には上記のような問題がある．したがって，これらの問題を解決し，閲覧者の再訪問を支援する新たな手段が求められている．

1.2 本研究の目的

本論文では，Web 閲覧の過程で個人用の Web アーカイブを作成し，その中から情報を閲覧するためのインタフェースの開発を目的とする．目的を達成するための手段として，作成した個人用 Web アーカイブの中に存在する同じ URL を持つ Web ページのバージョン間の差分の提示を行う．

1.3 本論文の構成

本論文の構成について述べる．第 2 章では，Web 閲覧における再訪問と Web アーカイブの現状を考察し，その問題点を述べる．第 3 章では，本研究に関連する研究について述べる．第 4 章では，閲覧経験のある知識の再発見を支援するインタフェースを持つ試作システム Personal Web Archive について述べる．第 5 章では，Personal Web Archive の実装について述べる．第 6 章では，Personal Web Archive の利用例を述べ，評価を行う．第 7 章では，本研究に対する議論と今後の課題について述べる．最後に，第 8 章で本論文をまとめる．

第2章　再訪問・Webアーカイブの現状と問題点

Greenberg らは，Web ブラウザの「戻る」,「進む」,「履歴」,「ブックマーク」などの機能を用いて，閲覧経験のある Web ページを再度閲覧することを**再訪問（*revisiting*）**と定義した [7]．Web の閲覧者が Web ページを閲覧した回数の総数を $total\_visit\_count$ ，閲覧者が閲覧した URL の総数を $total\_URL\_count$ ，閲覧者が Web ページへ再訪問する確率を R とすると，R は式 2.1 のように表わされる．

$$R = 100 \times \frac{total\_visit\_count - total\_URL\_count}{total\_visit\_count} \tag{2.1}$$

彼らは，この確率 R を**再起率（*recurrence rate*）**と定義した．彼らは，この再起率についても調査しており，1995 年の調査 [7] では約 58%，1999 年 10 月から 2000 年 1 月までの調査 [8] では約 81%となることを明らかにした．このことから，彼らは Web とは再起システムであると述べている．また，その他の調査による再起率は，1994 年の Catledge らの調査 [9] では約 61%，2008 年の Weinreich らの調査 [10] では約 65%であるという結果が出ている．図 2.1 に再起率の調査結果のグラフを示す．以上のことから，再起率がいずれの調査でも高い値を示しており，Web 閲覧において再訪問が重要な意味を持っていることが分かる．

本章では，この再訪問に用いられる既存のシステム・機能とその問題点を分析し，それらの問題を解決するためのシステムに必要な特徴を考察する．

2.1　再訪問の目的と既存のシステム・機能

ユーザが過去に閲覧したページへの再訪問を意図的に行う際の目的は，以下の 2 種類に大きく分けられる．

1. 過去に閲覧した URL の現在のコンテンツを閲覧する目的
2. ある URL で過去に閲覧したコンテンツそのものを閲覧する目的

上記 1 の閲覧を以降では**アドレス閲覧**と呼ぶこととする．また，2 の閲覧を以降では**コンテンツ閲覧**と呼ぶこととする．以下に，この 2 種類の閲覧が行われる Web ページや閲覧の際に用いられるシステム・機能の例を挙げる．

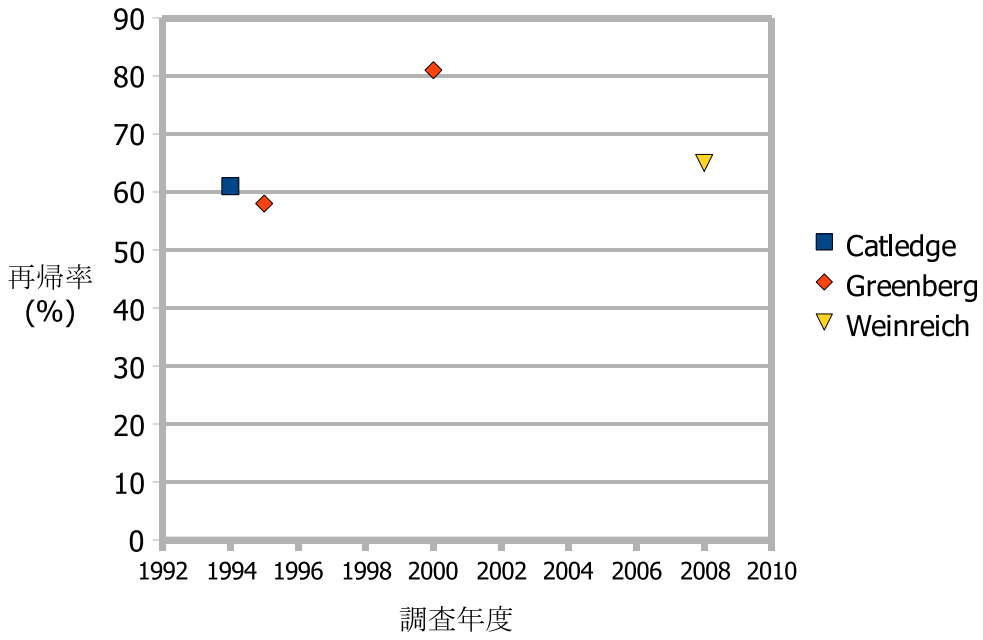

図 2.1: 再起率の調査結果

アドレス閲覧

アドレス閲覧は，ニュースサイト，ウェブログ，掲示板などの最新の情報を閲覧する際によく見られる再訪問である．ニュースサイトの閲覧者は Web ブラウザのブックマーク機能や RSS リーダ，Web ページ上のリンクなどを用いて，そのニュースサイトのトップページにアクセスする．その後，トップページに並んだ最新記事の見出し，つまりニュースサイトにおける最新の情報を閲覧する．ウェブログの場合も同様で，閲覧者はトップページにアクセスして最新記事を閲覧する．掲示板の場合は，トップページにアクセスして最新のスレッド，トピックを閲覧する他に，スレッド，トピック毎にアクセスして最新の書き込みを閲覧するなどの利用方法が見られる．

コンテンツ閲覧

コンテンツ閲覧は，プログラミング言語のリファレンスのような何度も必要となる情報を持つ Web ページ，アドレス閲覧の説明でも採り上げたニュースサイトやウェブログ，その他の多くの種類の Web ページで，過去に閲覧した情報が後から必要になった際に見られる再訪問である．過去に閲覧した情報が後から必要になる場合とは，例えば，閲覧している途中で閲覧を中断しなければならなかった Web ページを再度閲覧する場合や，他人と情報を共有するためにその情報のある Web ページを紹介する場合などが考えられる．

図 2.2: アドレス閲覧とコンテンツ閲覧の例

図 2.2 にニュースサイトとウェブログでのアドレス閲覧とコンテンツ閲覧の例を示す．図 2.2 上部のニュースサイトの左側の赤枠内には，見出しと画像と要約を持つ記事が 1 つ，見出しのみを持つ記事が 5 つある．また，右上の赤枠内には RSS フィードを配信していることを示すアイコンがある．左側の赤枠内の情報は Web ページが更新される度に変更されるため，URL の閲覧を目的とした場合，閲覧者はその情報を Web ブラウザで読んだり，RSS リーダで配信されている同じ内容のフィードを読んだりすることになる．図 2.2 下部のウェブログでも同様で，左側の赤枠内には最新の記事があり，閲覧者は Web ブラウザでその記事を読むことになる．また，右側の青枠内には「最新のブログ記事」と「月別アーカイブ」があり，コンテンツ閲覧を目的とした場合，過去に閲覧した記事を読もうとする閲覧者は，その中から目的の記事を探すことになる．

アドレス閲覧とコンテンツ閲覧を目的として再訪問を行うとき，ある URL を持つ Web ページの過去に閲覧した時点での状況と現在閲覧している状況が相違していることがしばしばある．この相違している状況には，次のようなものがあると考えられる．まず，Web ページのアドレスの状況について,「URL が存在しない」,「URL が存在する」という 2 種類の状況が考えられる．なお，Web ページが移動して URL が変更された場合も「URL が存在しない」として扱う．次に，Web ページのコンテンツの状況について,「全体が変更されている」,「一部が変更されている」,「過去と同一である」という 3 種類の状況が考えられる．一般に，アドレス閲覧を目的とした場合は Web ページの最新のコンテンツを閲覧できればよいため，コンテンツの状況の相違は目的達成の妨げにならない．一方，コンテンツ閲覧を目的とした場合は Web ページの過去と同じコンテンツを閲覧しなければならないため，コンテンツの状況の相違が目的達成の可否に大きく関わってくる．2 種類の再訪問の目的と 2 × 3 種類の過去に閲覧した時点と現在のコンテンツの相違の状況の組み合わせに対して，既存のシステム・機能がどの程度対応できているかを表 2.1 に示す.

表 2.1 の 1 行目は再訪問の目的を，2-3 行目はアドレスとコンテンツのそれぞれの状況を示している．4 行目以降は既存のシステム・機能の対応の程度を以下の 3 段階で評価している．

- 目的を達成できない（×）
- 目的をある程度達成できる（△）
- 目的を達成できる（○）

また，最終列では，既存のシステム・機能を用いて目的を達成するために，事前に何らかの準備が必要かどうかを評価している．

表 2.1: 再訪問の目的と Web ページの状況

再訪問の目的		アドレス閲覧						コンテンツ閲覧						準備
ページの状況	アドレス	存在しない			存在する			存在しない			存在する			
	コンテンツ	全体	一部	同一	全体	一部	同一	全体	一部	同一	全体	一部	同一	
既存のシステム・機能	ブックマーク	-	-	-	○	○	○	×	×	×	×	△	○	要
	Web 閲覧履歴	-	-	-	○	○	○	×	×	×	×	△	○	不
	RSS・Atom	-	-	-	○	○	○	×	×	×	×	△	○	要
	クローラ型 Web アーカイブ	-	-	-	×	×	×	△	△	△	△	△	△	要
	利用者登録型 Web アーカイブ	-	-	-	×	×	×	△	△	△	△	△	△	要
	Web ページのスクラップ	-	-	-	×	×	×	○	○	○	○	○	○	要
	Web 検索	-	-	-	△	△	△	×	×	×	×	△	△	不

表中の既存のシステム・機能の各項目について説明する.

ブックマーク

「ブックマーク」とは，Web ブラウザの標準的な機能として提供されているブックマーク機能のことである．また，Google Bookmarks[11] のような Web サービスもこれに含む.

Web 閲覧履歴

「Web 閲覧履歴」とは，Web ブラウザの標準的な機能として提供されている履歴機能のことである．また，Google Web History[12] のような Web サービスもこれに含む.

RSS・Atom

RSS とは，RSS 1.0（RDF Site Summary）[13]，および，RSS 2.0（Really Simple Syndication）[14] である．Atom とは,（Atom Syndication Format）[15] である．これらは，Web ページの見出し，要約，更新時刻などを記したフォーマットである．ここで,「RSS・Atom」とは，これらのフォーマットに沿って記述されたフィードを講読するための RSS リーダのことである．RSS リーダには Web ブラウザに組み込まれた形で存在するものもある.

クローラ型 Web アーカイブ

「クローラ型 Web アーカイブ」とは，クローラと呼ばれる Web 上の文書や画像などのコンテンツを収集するプログラムによって保存された Web 全体のアーカイブの閲覧を提供するサービスである．代表的なクローラ型 Web アーカイブとして，Internet Archive による Wayback Machine[2] がある．また，Google などの Web 検索エンジンのキャッシュについても，過去のバージョンが閲覧できる機能を持っている点において，一種のクローラ型 Web アーカイブといえる．

利用者登録型 Web アーカイブ

「利用者登録型 Web アーカイブ」とは，利用者に指定された Web ページのコンテンツを保存し，その閲覧を提供するサービスである．このようなサービスが一般的に Web アーカイブと呼ばれることはないが，サービスの性質上ここでは利用者登録型 Web アーカイブと呼ぶ．例として，hanzo:web[16]，ウェブ魚拓 [1] などがある．

Web ページのスクラップ

「Web ページのスクラップ」とは，閲覧中の Web ページを保存する機能のことである．これには，Web ブラウザの標準的な機能として提供されているものや，五味渕らによる Mozilla Firefox の拡張機能 ScrapBook[17]，Microsoft Internet Explorer 5.0 Macintosh Edition の Scrapbook 機能などがある．また，五味渕らの ScrapBook には，閲覧した Web ページを自動的に保存する機能もある．

Web 検索

「Web 検索」とは，Google や Yahoo!などの Web 検索エンジンである．ただし，そのキャッシュについてはクローラ型 Web アーカイブに含めるため，Web 検索としては扱わないこととする．

既存のシステム・機能についての評価の詳細を説明する前に，評価の項目の左から 1 列目から 3 列目までの「-」としている部分について説明する．この部分は,「アドレス閲覧を目的としたときに，URL が存在しない」場合である．この場合，目的の達成が不可能なことは明らかなため，表中のすべての既存のシステム・機能において評価から除外している．

次に，既存のシステム・機能の評価について述べる．

ブックマーク，Web 閲覧履歴，RSS・Atom

この 3 種類は，システム・機能を使う際に URL があらかじめ分かっていることが特徴である．URL さえ分かっていれば,「アドレス閲覧を目的としたときに，URL が存在する」場合，コンテンツがどのように変更されていたとしても目的を達成することができる．したがって，評価の 4 列目から 6 列目までが○となる．次に,「コンテンツ閲覧を目的としたときに，URL が存在しない」場合，この 3 種類は URL しか手がかりとして持たないために，URL が存在しないとコンテンツを閲覧することができない．したがっ

て，評価の 7 列目から 9 列目までが×となる．次に,「コンテンツ閲覧を目的としたときに，URL が存在する」場合，コンテンツの全体が変更されていると目的のコンテンツがないため×，コンテンツの一部が変更されていると目的のコンテンツの有無が不明なため△，コンテンツが同一であると目的のコンテンツがあるため○となる．最後に,「ブックマーク」と「RSS・Atom」は事前に URL の登録が必要なため，事前に準備が必要である．

クローラ型 Web アーカイブ

まず,「アドレス閲覧を目的としたときに，URL が存在する」場合，URL を入力して保存されている最新のバージョンを閲覧したとしても，それが元の Web ページの最新のコンテンツと一致するかどうか不明なため，評価の 4 列目から 6 列目までが×となる．次に,「コンテンツ閲覧を目的とした」場合，クローラ型 Web アーカイブでは，現在の Web ページの状況に関わらず過去のバージョンのコンテンツを閲覧することができる．ただし，クローラ型 Web アーカイブには目的のコンテンツを持つバージョンが保存されていない場合がある．詳細については 2.2.1 節で述べる．したがって，評価の 7 列目から 12 列目までが△となる．最後に，クローラ型 Web アーカイブを利用するためには，入力に用いる URL を保持しておく必要があるため，事前に準備が必要である．

利用者登録型 Web アーカイブ

利用者登録型 Web アーカイブは，Web ページの収集方法以外ではクローラ型 Web アーカイブと同様の性質を持つため，その評価も等しくなる．ただし，利用者登録型 Web アーカイブにも目的のコンテンツを持つバージョンが保存されていない場合があるが，その理由については若干異なる．この詳細についても 2.2.1 節で述べる．また，利用者登録型 Web アーカイブを利用するためには，過去に閲覧した時点で Web ページを登録しておく必要があるため，事前に準備が必要である．

Web ページのスクラップ

まず,「アドレス閲覧を目的としたときに，URL が存在する」場合，この種類のシステム・機能においても，前述の 2 種類と同様の理由で評価の 4 列目から 6 列目までが×となる．一方,「コンテンツ閲覧を目的とした」場合，過去に閲覧した時点で Web ページを保存しておけば，目的のコンテンツを閲覧することができる．したがって，評価の 7 列目から 12 列目までが○となる．最後に，過去に閲覧した時点で Web ページを登録しておく必要があるため，事前に準備が必要である．

Web 検索

まず,「アドレス閲覧を目的としたときに，URL が存在する」場合，コンテンツがどのように変更されていたとしても目的を達成することができる．ただし，その URL に訪問するためには，検索クエリとして用いるキーワードを上手く設定する，過去に検索して閲覧したときの検索クエリとして用いたキーワードを記憶から想起する，または，検索クエリの履歴から選択する，などの行動が必要となる．したがって，評価の 4 列目か

ら 6 列目までが△となる．次に,「コンテンツ閲覧を目的としたときに，URL が存在しない」場合，URL が存在しないと検索結果のリンク先が見つからない，または，検索結果にその URL が現れないため，コンテンツを閲覧することができない．したがって，評価の 7 列目から 9 列目までが×となる．次に,「コンテンツ閲覧を目的としたときに，URL が存在する」場合，前述と同じく必要な行動があることを前提として，コンテンツの全体が変更されていると目的のコンテンツがないため×，コンテンツの一部が変更されていると目的のコンテンツの有無が不明なため△，コンテンツが同一であると目的のコンテンツがあるため△となる．ここで，Web 検索については，アドレスやコンテンツは異なるが必要な情報が存在する Web ページが検索結果に現れることがあるため，再訪問以外の方法でも情報を発見できる．

以上の評価結果より，アドレス閲覧において,「ブックマーク」,「Web 閲覧履歴」,「RSS・Atom」が良い評価を得ていることが分かる．特に,「Web 閲覧履歴」は利用のために閲覧以外の作業を必要としない点において優れているといえる．一方，コンテンツ閲覧においては,「Web ページのスクラップ」が優れた評価を得ている．ただし，過去に閲覧した時点で保存しなければ利用できないという点は，過去に閲覧した情報が後から必要になった際によく見られる再訪問であるコンテンツ閲覧において，大きな問題である．また,「クローラ型 Web アーカイブ」と「利用者登録型 Web アーカイブ」も多少の評価を得ているが，この 2 種類についても Web ページの収集，保存の不確実さと事前準備が必要な点において問題を抱えている．

2.2 Web アーカイブ

Web アーカイブとは，これまで述べてきたように，Web 上の文書や画像などのコンテンツを収集，保存し，Web 全体のアーカイブとして公開している Web サービスである．クローラ型 Web アーカイブの代表的な例として，Wayback Machine のインタフェースを図 2.3 に示す．図上部は，Web アーカイブの検索インタフェースである．URL と検索する期間の年月日，その他のオプションとして検索するファイルタイプなどを設定するフォームを持つ．図下部は検索インタフェースにおいて，`http://tsukuba.ac.jp/` を検索した結果である．まず，検索結果を年毎に列に纏められている．その中で各年毎の結果の数を表記し，日付毎に順に並べている．また，更新があったバージョンについては，日付の隣に「*」が記されている．閲覧者は，各日付のアンカーテキストを持つリンクをクリックすることで，その日付における Web ページのバージョンを閲覧することができる．利用者登録型 Web アーカイブについては，これに URL 登録用のフォームが加わる程度のインタフェースを持つ．

2.2.1 Web ページの収集，保存の問題

クローラ型 Web アーカイブ，および，利用者登録型 Web アーカイブの問題点として，目的のコンテンツを持つバージョンが保存されていない場合があると 2.1 節で述べた．この原因としては以下のようなものがある．

INTERNET ARCHIVE
WaybackMachine

Advanced Search

find this **URL** http://www.tsukuba.ac.jp/

between these **dates** (optional) Month Day Year — Month Day Year

Go Wayback

Other Advanced Search Options

URL Matching	○ Retrieve page that most closely matches search criteria ◉ List all pages that match search criteria
Aliases	◉ Merge aliases (search results for yahoo.com, www.yahoo.com and yahoo.com/index.html will be merged together) ○ Show aliases separately (a search for yahoo.com will list www.yahoo.com separately) ○ Don't show aliases (a search for yahoo.com will not show www.yahoo.com)
Redirects	◉ Hide redirects (on the search results, we will not display pages that redirect to other pages) ○ Flag redirects (on the search results, we will mark all pages that redirect to another page with an 'r') ○ Show redirects (on the search results, we will display pages that redirect)
File Types	All types — Will only display files of the type you specify
Duplicates	☐ Show duplicates (if we have 20 identical versions of a page on the same day, we will show them all)
Comparison	☐ Show checkboxes to allow comparison of 2 versions of a page. Comparison technology provided by Docucomp.
Convert to PDF	☐ **(BETA)** Provide links to a service that will convert a version of a web page to PDF format. Conversion technology provided by 2Convert.

INTERNET ARCHIVE
WaybackMachine

Enter Web Address: http:// All Take Me Back Adv. Search Compare Archive Pages

Searched for http://www.tsukuba.ac.jp/ **605** Results

Note some duplicates are not shown. See all.
* denotes when site was updated.
Material typically becomes available here 6 months after collection. See FAQ.

Search Results for Jan 01, 1996 - Jun 18, 2008

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
0 pages	2 pages	2 pages	10 pages	20 pages	16 pages	30 pages	12 pages	108 pages	220 pages	91 pages	62 pages	1 pages
	Oct 08, 1997 * Nov 08, 1997	Jan 22, 1998 * Dec 12, 1998 *	Jan 25, 1999 * Feb 03, 1999 * Feb 08, 1999 * Feb 10, 1999 Apr 22, 1999 * Apr 28, 1999 * Apr 29, 1999 May 01, 1999 Oct 10, 1999 * Nov 16, 1999 *	Feb 26, 2000 * Mar 01, 2000 Mar 02, 2000 May 10, 2000 * May 11, 2000 May 11, 2000 * May 19, 2000 May 20, 2000 Jun 07, 2000 * Jun 12, 2000 * Jun 19, 2000 * Jun 21, 2000 Jul 06, 2000 Aug 15, 2000 * Aug 31, 2000 * Oct 07, 2000 * Oct 13, 2000 * Oct 19, 2000 * Dec 02, 2000 * Dec 07, 2000 *	Feb 02, 2001 * Feb 16, 2001 * Feb 26, 2001 * Mar 01, 2001 * Mar 02, 2001 Mar 31, 2001 * Apr 04, 2001 Apr 05, 2001 Apr 18, 2001 Apr 20, 2001 May 08, 2001 May 30, 2001 * Jun 03, 2001 Jul 21, 2001 * Aug 15, 2001 * Sep 23, 2001 *	Apr 02, 2002 * Apr 08, 2002 May 26, 2002 * Jun 03, 2002 Jul 20, 2002 * Sep 13, 2002 * Sep 14, 2002 Sep 16, 2002 Sep 17, 2002 Sep 20, 2002 Sep 21, 2002 Sep 22, 2002 Sep 24, 2002 Sep 25, 2002 Sep 26, 2002 Sep 28, 2002 Sep 29, 2002 Oct 01, 2002 * Oct 04, 2002 * Oct 07, 2002 Oct 14, 2002 * Oct 21, 2002 * Oct 28, 2002 * Nov 04, 2002 Nov 11, 2002 * Nov 21, 2002 * Nov 27, 2002 Nov 29, 2002 Dec 02, 2002 * Dec 04, 2002	Feb 02, 2003 * Feb 08, 2003 Feb 18, 2003 Mar 19, 2003 * Mar 20, 2003 Apr 01, 2003 * Apr 20, 2003 * May 22, 2003 * Jun 21, 2003 * Jul 26, 2003 * Aug 10, 2003 * Dec 28, 2003 *	Feb 13, 2004 * Apr 02, 2004 * Apr 03, 2004 * May 25, 2004 * Jun 02, 2004 * Jun 06, 2004 * Jun 11, 2004 * Jun 12, 2004 Jun 16, 2004 * Jun 18, 2004 * Jun 19, 2004 Jun 22, 2004 * Jun 23, 2004 Jun 24, 2004 * Jun 25, 2004 * Jun 26, 2004 Jun 28, 2004 Jun 29, 2004 Jun 30, 2004 * Jul 01, 2004 * Jul 02, 2004 * Jul 03, 2004 Jul 04, 2004 Jul 05, 2004 Jul 06, 2004 * Jul 07, 2004 * Jul 10, 2004 * Jul 11, 2004 Jul 12, 2004 Jul 14, 2004 * Jul 15, 2004 * Jul 16, 2004 *	Jan 07, 2005 * Jan 09, 2005 Jan 10, 2005 Jan 13, 2005 * Jan 16, 2005 * Jan 17, 2005 * Jan 17, 2005 * Jan 20, 2005 * Jan 21, 2005 * Jan 27, 2005 * Feb 03, 2005 * Feb 03, 2005 * Feb 04, 2005 * Feb 05, 2005 * Feb 07, 2005 * Feb 15, 2005 * Feb 17, 2005 * Feb 18, 2005 * Feb 19, 2005 * Mar 01, 2005 * Mar 01, 2005 * Mar 02, 2005 * Mar 04, 2005 * Mar 08, 2005 * Mar 09, 2005 * Mar 15, 2005 * Mar 18, 2005 * Mar 19, 2005 * Mar 21, 2005 * Mar 22, 2005 * Mar 23, 2005 * Apr 01, 2005 *	Jan 01, 2006 * Jan 01, 2006 * Jan 01, 2006 * Jan 01, 2006 * Jan 01, 2006 * Jan 02, 2006 * Jan 02, 2006 * Jan 02, 2006 * Jan 02, 2006 * Jan 02, 2006 * Jan 02, 2006 * Jan 02, 2006 * Jan 05, 2006 * Jan 06, 2006 * Jan 06, 2006 * Jan 06, 2006 * Jan 07, 2006 * Jan 07, 2006 * Jan 07, 2006 * Jan 07, 2006 * Jan 07, 2006 * Jan 07, 2006 * Jan 07, 2006 * Jan 08, 2006 * Jan 08, 2006 * Jan 10, 2006 * Jan 10, 2006 * Jan 10, 2006 * Jan 12, 2006 * Jan 12, 2006 *	Jan 03, 2007 * Jan 04, 2007 * Jan 05, 2007 * Jan 09, 2007 * Jan 10, 2007 * Jan 13, 2007 * Jan 15, 2007 * Jan 18, 2007 * Jan 20, 2007 * Jan 24, 2007 * Jan 25, 2007 * Feb 02, 2007 * Feb 02, 2007 * Feb 06, 2007 * Feb 09, 2007 * Feb 10, 2007 * Feb 16, 2007 * Feb 24, 2007 * Mar 04, 2007 * Mar 12, 2007 * Mar 21, 2007 * Apr 05, 2007 * Apr 06, 2007 * Apr 07, 2007 Apr 26, 2007 * May 04, 2007 * May 09, 2007 * May 26, 2007 * Jun 05, 2007 * Jun 09, 2007 * Jun 10, 2007 * Jun 13, 2007 *	Jan 24, 2008 *

図 2.3: Wayback Machine

1. Robots Exclusion Protocol[6] によりクローラのアクセスが拒否されることがある
2. Web ページへのリンクが張られていない場合がある
3. Web ページを収集するタイミングをクローラに依存する
4. Web ページの権利者によって削除されることがある

1，2，3 はクローラ型 Web アーカイブに特有の原因である．1 について，Robots Exclusion Protocol とは，クローラの行動を制御するための規約である．Web サイトのルートにクローラのアクセスを拒否するよう記述した robots.txt ファイルが存在すると，クローラはそのファイルで指定されたファイル，フォルダを収集，保存することができなくなる．2 について，クローラはリンクを辿りながら Web ページを巡回するため，リンクが張られていない Web ページは収集，保存することができない．3 について，クローラは独自のタイミングで Web を巡回しており，そのタイミングは Web ページの更新と連動している訳ではない．したがって，Web ページの更新後，クローラによる収集が行われる前に次の更新が起こることがある．図 2.4 に Web ページの更新とクローラの保存の例を示す．図中では，点線の時刻にクローラによる保存が行われているが，2 回目の保存と 3 回目の保存の間，および，4 回目の保存と 5 回目の保存の間にそれぞれ 2 回ずつ Web ページの更新が行われている．青色のシンボルで表したそれらの更新の 1 回目によるバージョンは，Web アーカイブに保存されない．また，赤色のシンボルで表したバージョンは，Wayback Machine において「*」が記されていたバージョンで，直前に保存したバージョンからコンテンツの更新があったことを示している．

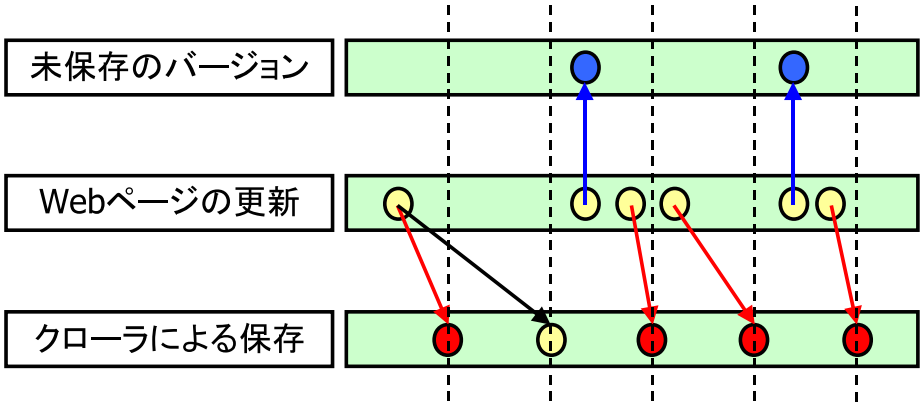

図 2.4: Web ページの更新時刻とクローラの保存時刻のずれ

4 は 2 種類の Web アーカイブに共通する原因である．Web ページは一般的に著作権付きの情報であり，Web アーカイブはそれを公開する法的権限を保持していない．したがって，権利者からの要請があった場合，Web アーカイブはその Web ページのアーカイブを削除するのが一般的である．

2.3　再訪問のインタフェースの問題

前節まででは，再訪問に利用できる既存のシステム・機能について述べてきたが，それらの閲覧のためのインタフェースについては論じてこなかった．ここでは，再訪問のインタフェース，特にコンテンツ閲覧におけるインタフェースの問題について述べる．図 2.3 に示した Wayback Machine のインタフェースに注目する．Wayback Machine では，各日付のリンクをクリックすることによりその日付のバージョンを閲覧する．ここで閲覧するのは，元の Web ページの複製である．これは，閲覧したいコンテンツを過去に閲覧した時期と Web ページが収集された時期が一致し，どのバージョンがそのコンテンツを保持しているかはっきり分かっている場合であれば問題ない．しかし，目的のバージョンが曖昧で，複数の候補の中からそのバージョンを絞り込む場合には困難を伴う．閲覧者は図 2.5 のようなフローチャートに沿って行動すると想定される．

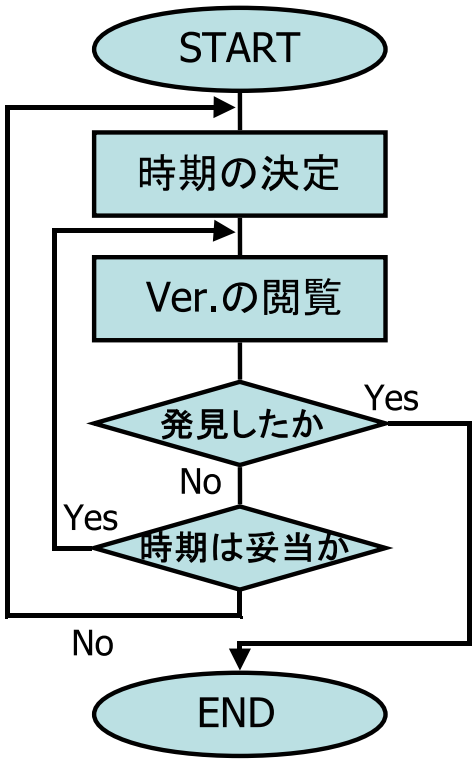

図 2.5: Web アーカイブ閲覧の流れ

それぞれのノードは以下の処理，判断である．

時期の決定

閲覧したいコンテンツを持つある程度の時期を予想する．この時期の中には複数のバージョンが存在する場合がある．

Ver. の閲覧

予想した時期に 1 つのバージョンのみ存在する場合，そのバージョンを閲覧する．複数のバージョンが存在する場合，その中から 1 つのバージョンを選択し，閲覧する．

発見したか

閲覧したバージョンに目的のコンテンツが存在し，それを発見できたかどうか．発見できた場合は Yes であり，そうでない場合は No である．

時期は妥当か

「時期の決定」で予想した予想したある程度の時期は正しいかどうか.「Ver. の閲覧」で選択しなかったバージョンに目的のコンテンツがありそうな場合は Yes であり，予想した時期から別のバージョンを選択し，閲覧する．一方，選択したバージョンを実際に閲覧した結果，目的のコンテンツを持つバージョンがありそうな時期が異なりそうな場合は No であり，別の時期から探し直すことになる．

ここで，閲覧者は目的のコンテンツを発見しない限り複数のバージョンの閲覧を繰り返すことになるが，それらのバージョンは一部のみが異なる類似した Web ページである．したがって，異なるバージョンを閲覧する度に，一度確認したコンテンツを再び確認することになったり，どこが変更されているのかを読んで確認したりしなければならない．これには多大な労力を要する．このため，目的のコンテンツを持つバージョンが見つかる前に諦めなければならないこともある．また，Web アーカイブ内に目的のコンテンツを持つバージョンが存在しなかった場合も，時間を浪費するだけである．

2.3.1 Web ページの差分に関する問題

一部のみが異なる類似した複数のバージョンの閲覧，比較を行うためには，Web 上のコンテンツに限らず，ドキュメントファイルやプログラムのソースファイルの比較などでも頻繁に用いられるファイル間の差分を抽出して提示する方法が有効であると考えられる．ここで，更新によって作られる古いバージョンと新しいバージョンに存在する情報は，ファイル間の差分に着目することにより以下のように分けることができると考えられる．

1. 新しいバージョンのみに存在する，更新によって追加された情報
2. 古いバージョンのみに存在する，更新によって削除された情報
3. 両方のバージョンに存在する情報

上記 1 の情報を以降では**追加情報**と呼ぶこととする．追加情報は，ウェブログやニュースサイトのトップページ，電子掲示板のスレッドなど更新頻度の高い Web ページでよく見られる．また，2 の情報を以降では**削除情報**と呼ぶこととする．削除情報は，追加情報と同様にウェブログやニュースサイトのトップページなど更新頻度の高い Web ページでよく見られる．一方，

電子掲示板のスレッドのような記事の削除が少ない Web ページではあまり見られない．なお，古いバージョンに存在した情報が修正されて別の情報に変化した場合は，修正前の情報を削除情報，修正後の情報を追加情報と考える．一部のみが異なる類似した複数のバージョンの閲覧，比較を行いながら情報を探すとき，この追加情報，削除情報を元に探している情報に近づいているかどうかを判断することになる．例として，以下のような判断が考えられる．

- 探しているコンテンツは追加情報よりも新しい情報である
- 探しているコンテンツは削除情報よりも古い情報である
- 探しているコンテンツと一緒に削除情報を閲覧したことがある

しかし，Wayback Machine などのインタフェースなどでは，これらの情報を把握するのは非常に困難である．Wayback Machine では，Web ブラウザのウィンドウやタブに異なるバージョンを提示し，それらを順に，あるいは並べて閲覧することになる．ここで，比較したバージョン間に追加情報が存在するかどうかを確かめるには，Web ページの日付や内容などを確認する必要がある．しかし，実際に追加情報が存在するかどうか分かりにくいことや，存在する場合もどの程度の量が追加情報であるか分かりにくいことが判断を難しくする．削除情報の場合についても，まったく同様である．

2.4 問題解決に必要な機能

本論文では，以上の既存のシステム・機能の問題点を考慮し，閲覧経験のある Web ページを再度閲覧するのを支援するためには，以下のような機能を持つシステムがあればよいのではないかと考えた．

1. 閲覧した Web ページを確実に，かつ自動的に収集する
2. 複数の類似した Web ページの比較，閲覧を支援する

1 の機能は，閲覧経験のある Web ページを後から再訪問しなければならなくなったときに，目的のコンテンツを確実に閲覧できるようにしておくために必要な機能である．2.1 節では，既存のシステム・機能において，そのシステム・機能を利用するために，過去に閲覧した時点で何らかの作業が必要である点，特に Web ページを保存しておかなければならないという点が問題となった．また，2.2 節では，目的のコンテンツを持つバージョンが保存されていない場合があることを問題として取り上げた．したがって，これらの問題を解決するために，閲覧した Web ページ，つまり，後から再訪問する可能性のある Web ページを自動的かつ確実に保存しておく機能が必要と考えられる．

2 の機能は，再訪問により情報を探す際のインタフェースを改善するための機能である．2.3 節では，Web ページへの再訪問によって情報を探す際，既存のシステム・機能のインタフェースでは，類似した各々のバージョンの比較，閲覧が非常に難しいことを問題とした．したがって，複数の類似した Web ページを比較，閲覧するための機能が必要と考えられる．

第3章　関連研究

3.1　再訪問に関する研究

これまで，Web 閲覧における利用者の行動調査が何度も行われている．例えば，1994 年の Catledge らの調査 [9]，1995 年の Tauscher らの調査 [7]，2000 年の Cockburn らの調査 [8] などがあった．近年の調査では，2008 年の Weinreich らの調査 [10] などがある．これらの多くは，サーバマシン上のプロキシを経由した被験者の行動を観測し，その結果を分析している．これらの調査結果として，Web 閲覧における利用者の行動や Web ページ，Web サイト，ひいては Web 全体の構造に関する問題などが明らかになってきている．本研究では，これらのいずれの調査でも取り上げられている再起システムとしての Web に着目し，再訪問した際の Web ページを閲覧するためのインタフェースを改善したシステムの開発を行った．

3.2　Web アーカイブに関する研究

これまで，多くの Web アーカイブに関する研究が行われている．それらの研究の中では，Web アーカイブを作成する方法，作成した Web アーカイブから Web ページを検索する方法，検索した Web ページを閲覧する方法など，主として扱っている点も異なっている．ここでは，上の 3 つの点について本研究と関連する研究を説明する．

3.2.1　作成

Web ページを収集して Web アーカイブの作成する方法に関する研究がいくつも行われている．また，その中には個人用の Web アーカイブの作成についての研究もいくつかある．

Rao らの Proxy-Based Personal Web Archiving System[18] では，閲覧者がサーバマシン上のプロキシ経由で閲覧した Web ページをサーバマシン上に保存する．このシステムを利用しての閲覧を繰り返すことにより，その閲覧と等しい数の Web ページが収集される．彼らは，その Web ページ群を個人用 Web アーカイブとして扱っている．本研究では，彼らと同様に個人用 Web アーカイブを対象として研究を行った．

安川らの Personal Archive Proxy[19] では，サーバマシン上のプロキシ経由で閲覧した Web ページを保存するが，サーバマシンでは管理のために Web ページの分類を行うに止め，Web ページはクライアントマシン上に保存している．彼らは，このクライアントマシンへの保存を，Web アーカイブにおける通信コストの問題，サーバマシン上に閲覧経験に基づいた Web

アーカイブを保存することによるプライバシ上の問題，著作権上の問題に対する解決策とした．本研究では，この点において同じ立場をとる．

その他として，クローラ型 Web アーカイブにおいて，クローラによる Web ページ収集の精度と効率を改善するために，Web ページ毎の更新頻度によって調整された保存頻度で Web ページの収集を行うクローラの研究が田村らによって行われた [20]．また，Web ページを保存する容量を抑えるために，Web ページの差分のみを収集する Web アーカイブの研究が福井らに行われた [21]．また，第三者ではなく情報発信者が主導して Web アーカイブを作成することにより，Web ページ収集の精度と効率を改善するシステムの開発が柊らによって行われた [22].

これらの研究では，Web ページを収集して Web アーカイブを作成するための方法については詳しく述べられている．一方で，作成した Web アーカイブを閲覧する方法についてはあまり触れられていない．本研究では，この Web アーカイブの閲覧手法に着目して研究を行った．

3.2.2 検索

Web アーカイブの作成に加えて，保存した個々の Web ページを再訪問するための検索に着目した研究が行われている．

例として，角谷らの研究を採り上げる．彼らは，個々の Web ページに出現するキーワードから時期毎のトピックを抽出する研究 [23] や，そのトピックと閲覧者による検索クエリとしてのキーワードとの関係から閲覧者の質問意図を抽出する研究 [24] を行った．

閲覧の繰り返しによってローカルマシン上に個人用 Web アーカイブを作成し，閲覧中の Web ページに関連のある Web ページをその個人用 Web アーカイブから抽出して提示する History-Centric Browsing システム [25, 26, 27] の開発が白井らに行われた．ここで，関連のある Web ページの提示とは，閲覧時間の近い Web ページ，同一 URL を持つ Web ページ，内容の類似する Web ページのサムネイルを数枚ずつ提示することによって行われている．

これらの研究では，検索によって保存した個々の Web ページを再訪問するまでの方法について詳しく述べられている．一方で，保存した個々の Web ページを再訪問した後，実際に Web ページを見るためのインタフェースについてはほとんど触れられていない．本研究では，この再訪問した Web ページを閲覧するためのインタフェースに着目して研究を行った．

3.2.3 閲覧

再訪問した Web ページを閲覧するためのインタフェースの研究として，Jatowt らの Past Web Browser[28, 29] がある．彼らは，複数のクローラ型 Web アーカイブのデータをマージし，それを時系列データとして可視化した．さらに，その時系列データから個々の Web ページを提示する際，直前のバージョンとの差分の提示を行う．図 3.1 に Past Web Browser の概観を示す．

図中の直線上に提示された赤いシンボルが Web ページのバージョンを可視化したものである．また，閲覧中の Web ページ内に黄色と青色で強調された部分があるが，黄色の部分が追

図 3.1: Past Web Browser の概観

加情報，青色の部分が削除情報を表している．削除情報については表示後に少しの時間が経過すると，点滅後表示されなくなる．

本研究では，クローラ型 Web アーカイブに収集されているデータではなく，閲覧者によって収集された個人用 Web アーカイブの可視化を行う点について彼らの研究と異なる．また，彼らは隣接する 2 つのバージョン間の差分の提示を行っているが，本研究では複数のバージョン間の差分を同一画面上に提示し，比較することによって個人用 Web アーカイブの閲覧を支援するインタフェースの開発を行った．

3.3 その他

本研究で扱う個人用 Web アーカイブのように個人の視点で蓄積されたデータを扱う研究が盛んに行われている．例えば，Dumais らの Stuff I’ve Seen[30] や佐藤の dripdrop[31] では，電子メールや Web ページ，その他の文書などのファイルを，時刻や文書の作者などのコンテキスト情報を元に検索するためのシステムの開発を行っている．また，Google による Google Desktop[32] などがすでに一般に公開され利用されている．

本研究では，複数のバージョン間の差分を同一の Web ページ内にまとめて提示する．それらの差分には情報の現れる時期によって鮮度に差ができる．Web ページ上の情報の鮮度を可視化した研究として塚田らの Dying Link[33, 34, 35] がある．彼らは，Web ページ上のリンク

のアンカーテキストに，更新時刻に従って「掠れていく」ような視覚効果を持たせることで，Web ページ上の情報の鮮度を表現している．その他，一般的な情報の鮮度の可視化手法としては，時間的要素を色や透明度で表すものが多い．

第4章　Personal Web Archive

本研究では，2.4 節で述べた機能を持つシステム Personal Web Archive の開発を行った．図 4.1 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した個人用 Web アーカイブに対し，可視化を行う．また，その個人用 Web アーカイブの可視化にあたり，複数の 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 4.1: Personal Web Archive の概観

4.1 Web アーカイブの作成

Personal Web Archive では，閲覧者が Web ページを閲覧する際，その Web ページをローカルディスク上に自動的に保存する．この方法によって，閲覧したページが確実かつ自動的に保存され，閲覧を繰り返すことによって閲覧経験のある Web ページの集合が作成される．本研究では，この作成された Web ページの集合を個人用 Web アーカイブとして扱う．

ここで，Web ページを保存する際に保存するものは，文書や画像などの実体をもつファイルとしてコンテンツそのものと，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である．具体的には以下のようなものを保存する．

- コンテンツ
  - 文書ファイル（html，htm など）
  - 画像ファイル（bmp，jpg など）
  - 音声ファイル（mp3，wma など）
  - 動画ファイル（avi，wmv など）
  - その他のファイル（pdf，doc など）
- メタデータ
  - 管理用 ID
  - 閲覧時刻
  - タイトル
  - URL

4.2 Web アーカイブの可視化

閲覧者が Web ページを閲覧する際，個人用 Web アーカイブの中から閲覧中の Web ページと同一 URL を持つ Web ページを抽出し，時系列順に並べて提示する．Personal Web Archive では，図 4.2 に示すような時系列を表す直線上の，個々の Web ページの閲覧時刻に対応する位置にデータを並べることによって可視化する．

図 4.2: 時系列データの可視化

このように可視化された個人用 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，閲覧者が過去に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 閲覧の経験に基づいた Web アーカイブが閲覧時刻の間隔や位置を再現する形で可視化されているため，閲覧者は目的の情報を見た時期を想起しやすくなる．

4.3 バージョン間の差分の提示

Web アーカイブの同一 URL を持つ Web ページ群の個々の Web ページ，つまり，個々のバージョンは，Web ページ内の一部の情報が異なるのみで，大部分は以前と同じ情報であることが多い．したがって，バージョン間の差分を提示することによって，閲覧者が複数のバージョンを連続して閲覧する際，その差分に注目するだけでそれぞれのバージョンを比較しながら閲覧できるようになると考えられる．

Personal Web Archive では，個々のバージョンを閲覧する際，バージョン間の差分を閲覧者に提示する．ここで，1 つのバージョンを閲覧する場合と複数のバージョンを閲覧する場合に対して，異なる方法で差分を強調して提示する．以降，この 2 種類の場合に扱うバージョンのそれぞれを**単数バージョン**，および，**複数バージョン**と呼ぶ．

4.3.1 単数バージョンの閲覧

図 4.3 に，単数バージョンを閲覧する際の動作を示す．まず，前述のように同一 URL を持つ Web ページが抽出され，時系列データとして可視化されている．この時系列データ上のある連続したバージョンを，それぞれ①と②とする．ただし，①と②の間には，削除情報と追加情報があるものとする．この時系列データ上の②を閲覧する．従来のインタフェースでは，図 4.3 上部のように②がそのまま提示される．一方，Personal Web Archive では，図 4.3 下部のように，隣接する直前のバージョンである①と②をマージしたもの提示し，さらに，①と②の差分である①に存在する削除情報と②に存在する追加情報を強調して提示する．

図 4.4 において，①と②の差分の提示についての詳細を説明する．①には，情報 A，B，C が存在する．②には，情報 A，B，D が存在する．①＋②がマージされた結果である．①と②の間では，情報 C が削除されていることが分かる．つまり，情報 C は削除情報である．一方，情報 D が追加されている．つまり，情報 D は追加情報である．これらの差分は，①＋②に示したように，削除情報は青い背景色，追加情報は赤い背景色を持つ情報として閲覧者に提示される．

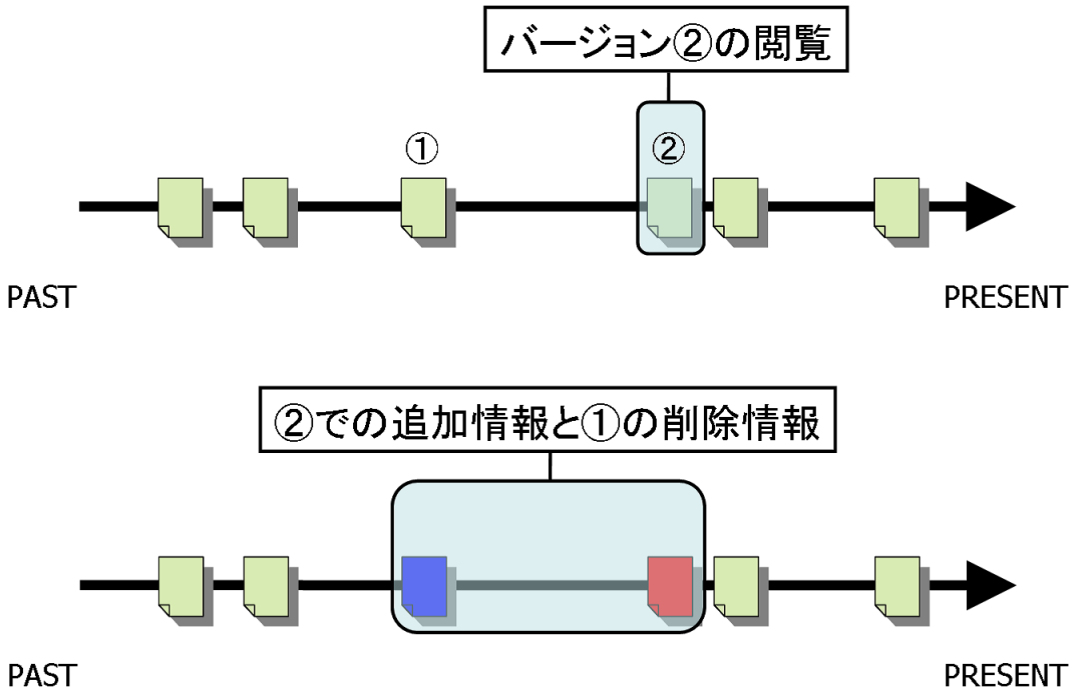

図 4.3: 単数バージョンの閲覧

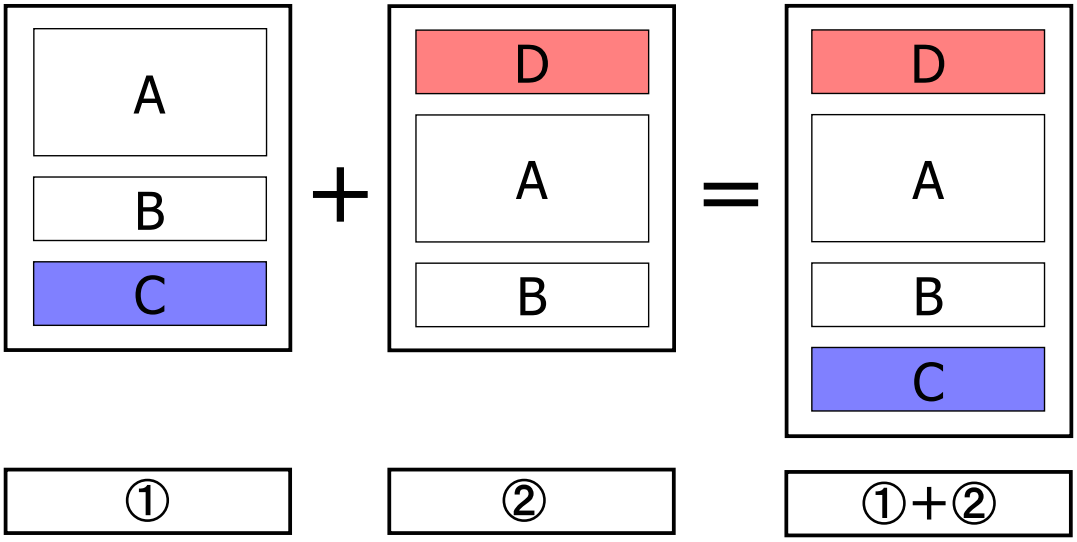

図 4.4: 単数バージョン間の差分の提示

4.3.2　複数バージョンの閲覧

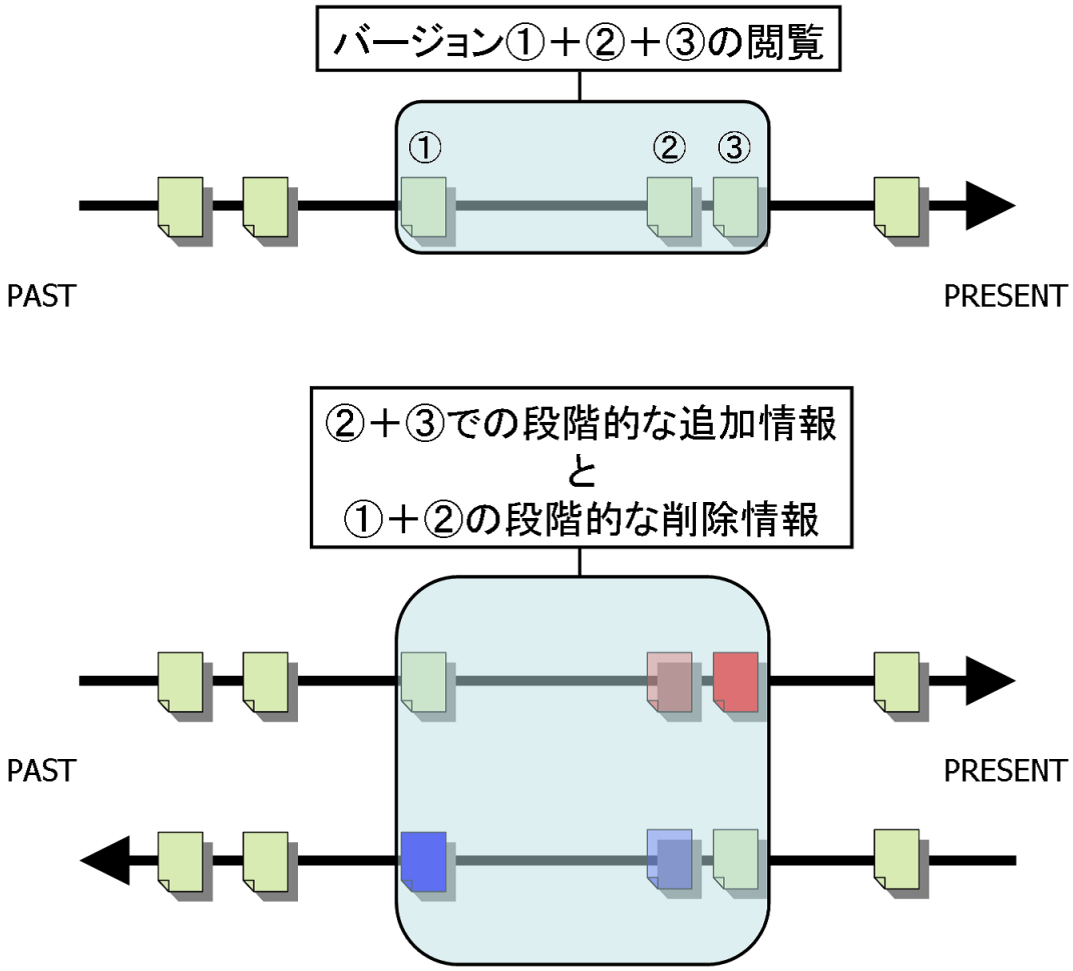

図 4.5: 複数バージョンの閲覧

図 4.5 に，複数バージョン（ここでは，3 つのバージョン）を閲覧する際の動作を示す．まず，単数バージョンの閲覧の際と同様に，同一 URL を持つ Web ページが抽出され，時系列データとして可視化されている．この時系列データ上のある連続した 3 つのバージョンを，それぞれ①，②，③とする．ただし，①と②の間，②と③の間には，それぞれ削除情報と追加情報があるものとする．この時系列データ上の①，②，③を同時に閲覧する．従来のインタフェースでは，それぞれのバージョンが Web ブラウザの異なるウィンドウやタブにそのまま提示される．一方，Personal Web Archive では，図 4.5 下部のように，①，②，③をマージしたもの提示する．さらに，単数バージョンの閲覧の際と同様に，①と②の間の差分，②と③の間の差分を提示することになるが，この提示する方法を異なるものとなる．まず，追加情報には③で追

加された最新の情報と②で追加された 1 段階古い情報が存在する．つまり，情報の鮮度に差があることになる．したがって，その鮮度の差を示すために，図 4.5 下部の右向き矢印上で示したように，追加情報を段階的に強調して提示する．また，削除情報についても同様で，①で削除された最古の情報と②で削除された 1 段階新しい情報が存在する．したがって，削除情報も段階的に強調して提示する．ここで，鮮度を表すための段階的な強調の方法は，関連研究で触れたように文字を掠れさせる，色を変える，不透明度を変えるなどの方法が考えられるが，可読性の問題などを考慮して，本システムでは背景色の不透明度を変更する方法を採用する．

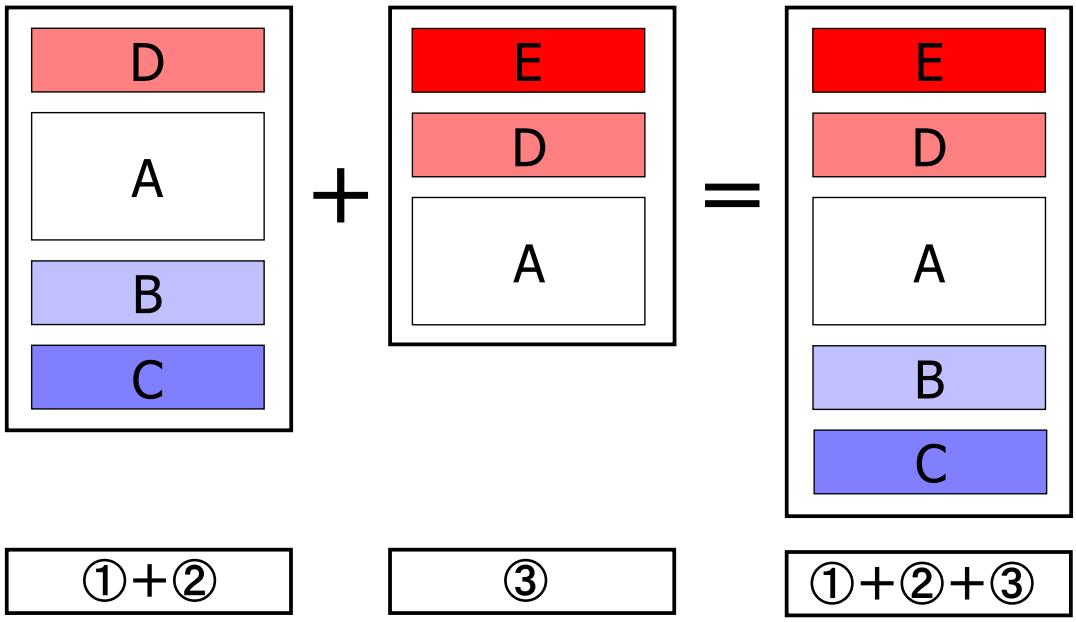

図 4.6: 複数バージョン間の差分の提示

図 4.6 において，①，②，③の差分の提示についての詳細を説明する．①と②については 1 つのバージョンの閲覧の際と同様に，①には，情報 A，B，C が存在する．②には，情報 A，B，D が存在する．①＋②がマージされた結果であり，情報 D が追加情報，情報 C が削除情報である．③には，情報 A，D，E が存在する．①＋②＋③がマージされた結果である．このマージは，①と②のマージの結果に，さらに③を加える形で行っている．①＋②と③の間では，情報 E が追加されていることが分かる．つまり，情報 E は追加情報である．一方，情報 B が削除されている．つまり，情報 B は削除情報である．ここで，複数バージョンの閲覧の場合，前述のように差分の情報の鮮度が異なる場合がある．まず，追加情報について，情報 E が最新の情報，情報 D が 1 段階古い情報である．したがって，①＋②＋③では，情報 E は不透明度の高い赤い背景色，情報 D は不透明度の一段階低い赤い背景色として提示される．削除情報について，情報 C が最古の情報，情報 B が 1 段階新しい情報である．したがって，情報 C は不透明度の高い青い背景色，情報 B は不透明度の一段階低い青い背景色として提示される．

4.4　システムの詳細

図 4.7: Personal Web Archive の構成

図 4.7 は Personal Web Archive を組み込んだ Web ブラウザである．まず，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つ．加えて，その下部に個人用 Web アーカイブを提示する領域を持つ．以降，それぞれの領域を Web ページ提示部，Web アーカイブ提示部と呼ぶ．

図 4.8 に Web アーカイブ提示部の詳細を示す．まず，中央の棒状の青い領域は時系列を表す．この時系列上に，現在閲覧している Web ページと同じ URL を持つ Web ページ，つまり，その URL に対する異なるバージョンの情報を個人用 Web アーカイブから抽出して提示する．現在閲覧している Web ページのバージョンの閲覧時刻を中心として，決められた期間についての情報を提示する．決められた期間とは，decade，year，month，week，day，hour の 6 種類であり，全体として 1 期間，すなわち，中心から 1/2 期間ずつの時期を提示する．例えば，期間が year の場合，前後半年分の情報を提示することになる．図 4.8 中の左上の week というテキストが現在の期間を示している．棒状の青い領域上の小さな赤い矩形は，表示している時期に存在する Web ページのバージョンを表しており，特に，現在閲覧しているバージョンを表す矩形は黄色く強調表示している．また，青い領域の両端の数字は，表示している時期の前後に存在するバージョンの数を表している．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期ほど矩形が多く重なるために周辺の透明度が低くなり，その時期に頻繁に閲覧中の URL を閲覧していたことが可視化されて閲覧者に伝えられる．

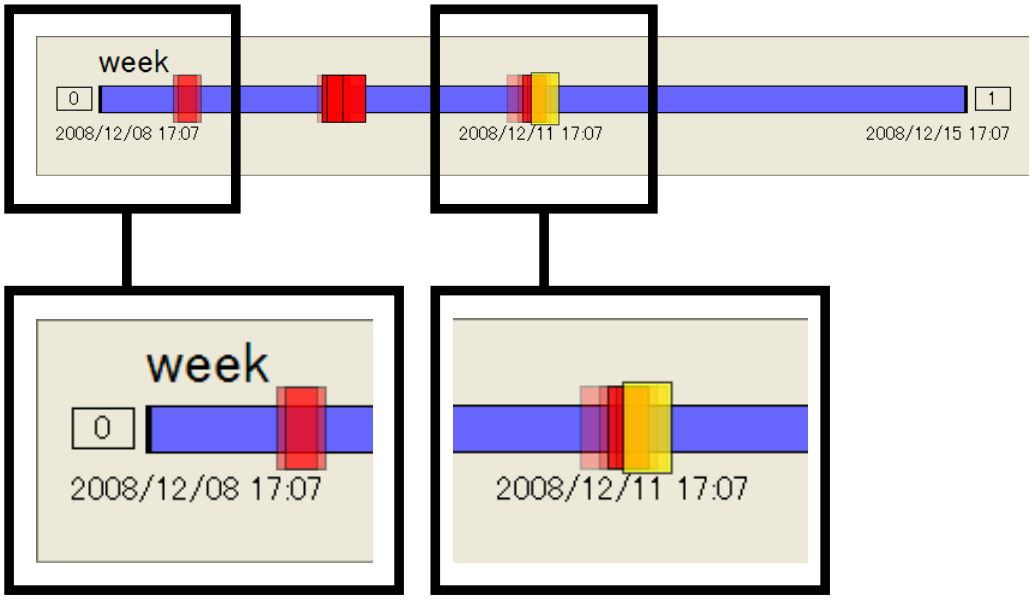

図 4.8: Web アーカイブ提示部

Web アーカイブ提示部上では，以下のような操作を行うことができる．

単数バージョンの選択
: 表示されているバージョンを表す矩形上でマウスのクリックを行うことによって，そのバージョンが選択される．

複数バージョンの選択
: マウスのドラッグによって範囲調節し，マウスのリリースによって範囲を決定する．決定した範囲にバージョンを表す矩形が存在すれば，それらのすべてのバージョンが選択される．

バージョン閲覧時刻の提示
: 表示されているバージョンを表す矩形上でマウスをホバーさせると，そのバージョンの閲覧時刻がポップアップにより提示される．

表示期間，時期の変更
: 表示期間，時期の変更をマウスの右ボタンのドラッグと移動の組み合わせによるジェスチャで行う．マウスを上下に移動させることで表示期間の変更を行う．上に移動させることで表示期間を 1 段階大きくし，下に移動させることで 1 段階小さくする．また，マウスを左右に移動させることで表示時期の変更を行う．左に移動させることで 1 期間過去の時期を提示し，右に移動させることで 1 期間未来の時期を提示する．

4.4.1 単数バージョンの閲覧

本システムは，Web アーカイブ提示部で単数バージョンの選択が行われたとき，選択されたバージョンとその直前のバージョンをマージし，Web ページ提示部に提示する．さらに，それらのバージョンの差分を抽出し，追加情報と削除情報を強調して提示する．バージョン間の差分を強調された Web ページの例を図 4.9 に示す．追加情報は背景色を赤色に強調されて提示される．また，削除情報は直前のバージョンに存在した位置に復元され，さらに，背景色を青色に強調されて提示される．

閲覧者は，強調された差分を確認しながら Web ページ提示部に提示されたバージョンを順に閲覧していくことにより，過去に閲覧したバージョンから目的の情報を探すことになる．作業の流れを説明する．まず，閲覧者は追加情報を確認する．記憶の中の目的の情報と追加情報を比較し，それらの情報が現れた時刻の前後関係を記憶から想起する．その結果，追加情報より後の情報だと判断したならば，より新しいバージョンの追加情報を閲覧する．前の情報だと判断したならば，今度は，削除情報と比較する．その結果，削除情報より前の情報だと判断したならば，より古いバージョンの削除情報を閲覧する．後の情報だと判断したならば，目的の情報は追加情報と削除情報の間，つまり現在閲覧中のバージョンに存在している可能性が高いと分かる．

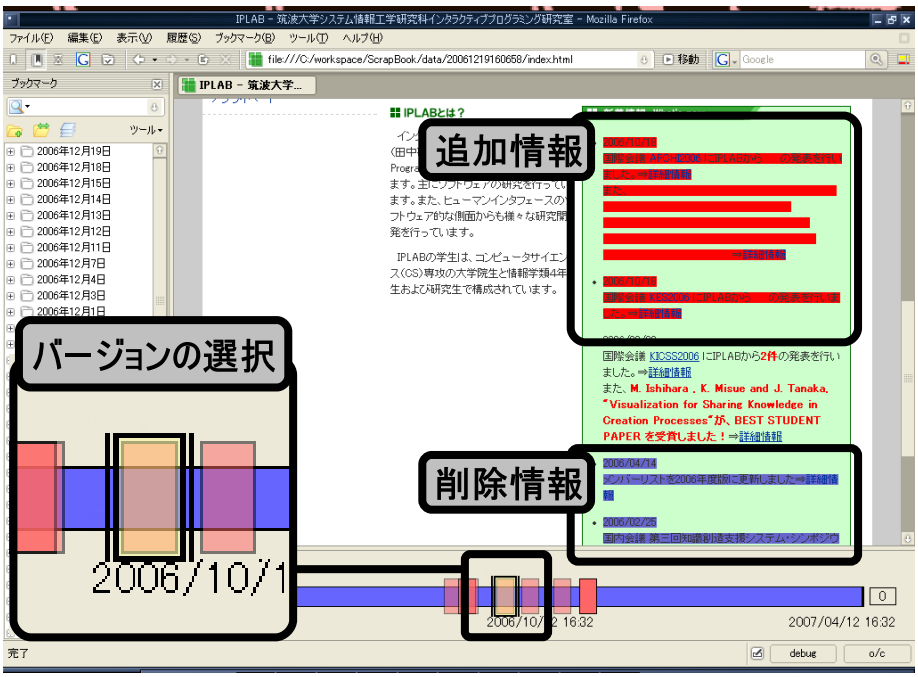

図 4.9: 単数バージョンの閲覧における差分の強調

4.4.2　複数バージョンの閲覧

本システムは，Web アーカイブ提示部で複数バージョンの選択が行われたとき，選択されたすべてのバージョンをマージし，Web ページ提示部に提示する．さらに，単数バージョンの閲覧の場合と同様に，追加情報と削除情報を強調して提示する．さらに，追加，削除の鮮度を基に計算された不透明度をそれらの追加情報，削除情報に与える．新しい追加情報であるほど，また，古い削除情報であるほど不透明度を高く設定する．複数バージョン間の差分を強調された Web ページの例を図 4.10 に示す．

図 4.10: 複数バージョンの閲覧における差分の強調

閲覧者は，単数バージョンの閲覧の場合と同様に，強調された差分を確認しながら Web ページ提示部に提示されたバージョンを順に閲覧していくことにより，過去に閲覧したバージョンから目的の情報を探すことになる．作業の流れを説明する．まず，閲覧者は最新の追加情報を確認する．記憶の中の目的の情報と最新の追加情報を比較し，それらの情報が現れた時刻の前後関係を記憶から想起する．その結果，追加情報より後の情報だと判断したならば，より新しいバージョンの追加情報を閲覧する．前の情報だと判断したならば，今度は 2 番目に新しい追加情報と比較する．以下，判断と次の追加情報との比較を繰り返す．目的の情報が追加情報として存在していた場合，この繰り返しの過程で閲覧者に発見されることとなる．目的の情報が追加情報として存在していなかった場合，今度は，最古の削除情報と比較する．その結果，削除情報より前の情報だと判断したならば，より古いバージョンの削除情報を閲覧する．後の情報だと判断したならば，今度は 2 番目に古い削除情報と比較する．以下，判断

と次の削除情報との比較を繰り返す．目的の情報が削除情報として存在していた場合，この繰り返しの過程で閲覧者に発見されることとなる．目的の情報が削除情報として存在していなかった場合，目的の情報は追加情報と削除情報の間，つまり現在閲覧中のバージョンの追加情報でも削除情報でもない情報として存在している可能性が高いと分かる．

複数バージョンの閲覧は，ある期間の情報を内へと絞り込んでいく形で用いられるため，その期間内に目的の情報があると分かっている場合，高い効果が得られると考えられる．

第5章　実装

Personal Web Archive は，Mozilla Foundation による Web ブラウザ Mozilla Firefox のバージョン 1.5 以降で動作する拡張機能として作成を行った．拡張機能とは，インストールすることで Web ブラウザ本体に様々な機能を提供するプログラムのことである．拡張機能の開発には，インタフェースの記述のために XUL（XML-based User-interface Language）という Mozilla Foundation のアプリケーションの開発専用に用いられる言語を，処理の記述のために JavaScript を主として用いる．また，XPCOM（Cross Platform Component Object Model）という C++で記述されたクロスプラットフォームで動作するコンポーネント技術を利用することもでき，その API も公開されている．Mozilla Firefox の拡張機能として開発を行った理由として，以下の理由がある．

- クロスプラットフォームで動作すること
- 本体の Web ブラウザの機能をそのまま利用できること

このような理由から，本システムでは，閲覧者の普段通りの Web 閲覧の環境にできるだけ近い形，および，幅広い環境でのデータ収集や評価を行うことができると想定している．

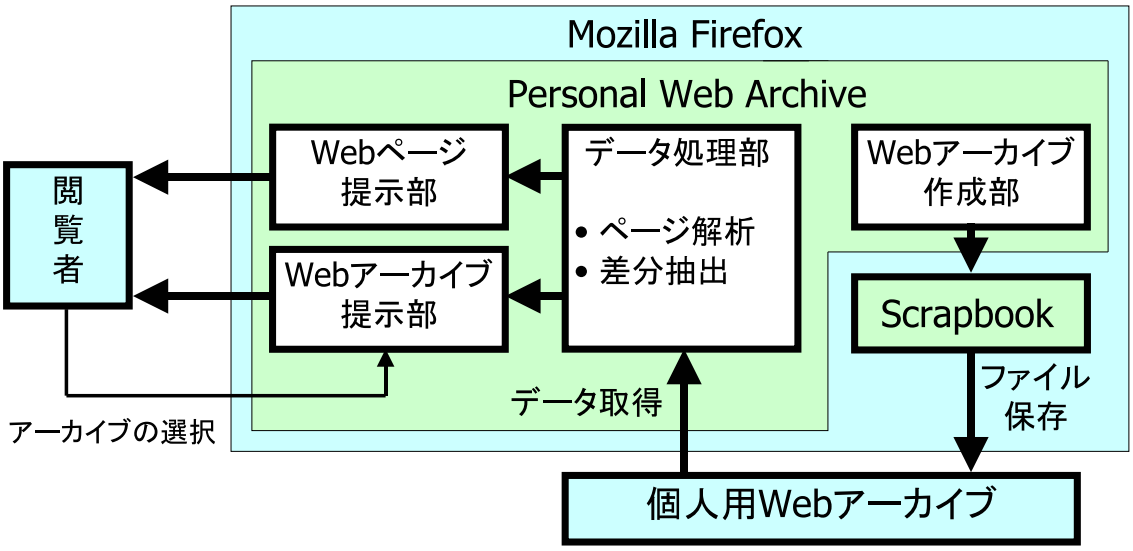

図 5.1: システム構成

図 5.1 にシステム構成を示す．本システムは，Web アーカイブ作成部，データ処理部，データ提示部（Web アーカイブ提示部，Web ページ提示部）に分けられる．

5.1 Web アーカイブ作成部

個人用 Web アーカイブの作成には，Mozilla Firefox の拡張機能として公開されている ScrapBook [17] で実装されている Web ページの保存を行う関数を利用した．この関数は Mozilla Firefox に ScrapBook をインストールすることで他の拡張機能からも利用可能となる．本システムは，Web ページが読み込また際，この関数を呼び出し，4.1 節で述べた Web ページのコンテンツを保存する．それぞれの Web ページは，ScrapBook の仕様に従い，指定されたフォルダ内に閲覧時刻を基に命名されたフォルダに保存される. 例えば，2009 年 1 月 1 日 1 時 1 分 1 秒に閲覧した Web ページなら 20090101010101 というフォルダ名となる．また，リンク先の再帰回収は行わず，表示されている画像などは相対リンクの形で同フォルダに保存される．

Web ページの保存と同時に，Web ページの持つメタデータの保存も行う．ここで，メタデータは RDF（Resource Description Framework）という枠組みを用いて保存する．RDF は，Web 上のリソースに対してメタデータを記述する用途を意図して設計された XML ベースの枠組みであり，RSS などでも利用されている．本システムでは，Web ページの集合である Web アーカイブの管理にも RDF が有効であると考え，これを用いている．1 つの RDF ファイル index.rdf の中に，保存した Web ページのそれぞれを別に表すノードを作り，さらに，そのノードの中に属性として以下の表 5.1 に示したものを記述する．

表 5.1: 保存するメタデータ

プロパティ名	説明
id	閲覧時刻から作成した ID 西暦から秒までを並べたもの
title	Web ページのタイトル
url	Web ページの URL

5.2 データ処理部

データ処理部では，個人用 Web アーカイブからのデータの読み込みやバージョン間の差分の抽出などの処理を行う．

5.2.1 データの読み込み

Web ページを閲覧する際，Web アーカイブ作成部で作成した RDF ファイル index.rdf を XML ファイルとして読み込む．次に，その XML ファイルの DOM（Document Object Model）[36] ツリーにアクセスを行い，属性 url の値として閲覧している Web ページの URL と同じ URL を持つ Web ページのノード，および，そのノードの属性が持つ値をメタデータをすべて抽出しておく．

5.2.2 差分の抽出

バージョン間の比較を行う際，その差分の抽出を行う．比較が行われるそれぞれのバージョンの HTML の DOM ツリーを走査し，削除された HTML タグと追加された 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，新しい方のバージョンの HTML 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 タグに鮮度によって決められた class 属性を与え，style 属性を用いて class 属性毎に異なる不透明度を設定する．追加された HTML タグが検出されているときも同様で，その HTML タグに鮮度によって決められた class 属性を与え，style 属性を用いて class 属性毎に異なる不透明度を設定する．

鮮度と不透明度の設定について述べる．不透明度は Mozilla Firefox 独自の style 属性のプロパティ-moz-opacity により設定する．-moz-opacity は 0.00〜1.00 の値を取り，0.00 が完全な透明，1.00 が完全な不透明となる．変数 $number\_of\_version$ と $freshness\_of\_information$ を以下のように定義する．

number_of_version
: 比較されるバージョンの数.

freshness_of_information
: 情報の鮮度．追加された HTML タグについては，新しいバージョンに存在するものから順に 1，2，…，$number\_of\_version$ - 1 となる．一方，削除された HTML タグについては，古いバージョンに存在するものから順に 1，2，…，$number\_of\_version$ - 1 となる．

このとき，-moz-opacity の値は以下の式 5.1 で求められる．

$$\text{-}moz\text{-}opacity = \frac{number\_of\_version - freshness\_of\_information}{number\_of\_version - 1} \tag{5.1}$$

例として，$number\_of\_version$ が 5 だった場合，追加された HTML タグの-moz-opacity は，古いバージョンから順に，なし（追加自体がない），0.25，0.50，0.75，1.00（不透明）となる．一方，削除された HTML タグの-moz-opacity は，古いバージョンから順に，1.0（不透明），0.75，0.50，0.25，なし（削除自体がない）となる．

図 5.2 に単数バージョンの閲覧におけるバージョンの比較の際に行われる Web ページへの処理を示す．これは，4.3.1 節の図 4.4 で示したものを HTML 構造で表した図である．まず，バージョン②に削除情報 C の HTML タグが挿入される．さらに，その HTML タグには，del1 の値を持つ class プロパティが与えられる．一方，追加情報 D の HTML タグには，add1 の値を持つ class プロパティが与えられる．

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

バージョン①

```
<body>
  <div>D</div>
  <div>A</div>
  <div>B</div>
</body>
```

バージョン②

```
<body>
  <div class="add1">
       D
  </div>
  <div>A</div>
  <div>B</div>
  <div class="del1">
       C
  </div>
</body>
```

バージョン①＋②

図 5.2: 単数バージョンの閲覧における差分の抽出

複数バージョンに閲覧においても，単数バージョンの閲覧と共通して前述の処理を行う．ただし，複数バージョンの閲覧の場合はこの処理を古いバージョンから順に繰り返して行う．

図 5.3 に複数バージョンの閲覧におけるバージョンの比較の際に行われる Web ページへの処理を示す．これは，4.3.2 節の図 4.6 で示したものを HTML 構造で表した図である．バージョン①＋②について，比較の始めに行われる鮮度と不透明度の計算が異なるため，単数バージョンの場合とは結果が異なっている．追加情報 D の $freshness\_of\_information$ は 2 であるため，その HTML タグには，add2 の値を持つ class プロパティが与えられる．他の情報についても鮮度と不透明度の計算が行われ，E は 1 の追加情報，C は 1，B は 2 の削除情報となる．この計算結果に基づいてそれぞれの class プロパティに値が与えられる．

```
<body>
  <div class="add2">
       D
  </div>
  <div>A</div>
  <div>B</div>
  <div class="del1">
       C
  </div>
</body>
```

バージョン①＋②

```
<body>
  <div>E</div>
  <div>D</div>
  <div>A</div>
</body>
```

バージョン③

```
<body>
  <div class="add1">
       E
  </div>
  <div class="add2">
       D
  </div>
  <div>A</div>
  <div class="del2">
       B
  </div>
  <div class="del1">
       C
  </div>
</body>
```

バージョン①＋②＋③

図 5.3: 複数バージョンの閲覧における差分の抽出

5.3 データ提示部

データ提示部では，データ処理部での処理結果を出力する．

Web アーカイブ提示部では，index.rdf から抽出したメタデータの閲覧時刻を元に，Mozilla Firefox のインタフェース記述のための XUL 上に，それぞれのバージョンを表すノードを作成する．さらに，ノード毎に閲覧時刻とローカルマシン上での URI を属性として与える．このノードを時系列上の閲覧時刻に対応する位置に提示する．また，その提示したノードを閲覧，選択するためのイベントハンドラを与えておく．

Web ページ提示部では，前節で述べた差分の強調された Web ページを提示する．Web ページ提示部は元の Web ブラウザの機能として備わっている部分であるので，Web ブラウザの基本的な機能はそのまま利用できる．

第6章　システム利用例

ここでは，本システムの利用例とその評価について述べる．本システムの利用にあたり，Personal Web Archive と ScrapBook をインストールした Mozilla Firefox を用いて Web 閲覧を行い，個人用 Web アーカイブの作成を行った．個人用 Web アーカイブの作成は，主に 3 つの期間に行われた．以下の表 6.1 にその詳細を示す.

表 6.1: 作成した個人用 Web アーカイブ

	期間 1	期間 2	期間 3	合計
期間	2006 年 9 月 ～ 2007 年 10 月	2008 年 2 月 ～ 2008 年 6 月	2008 年 12 月 ～ 2009 年 1 月	-
日数	約 400 日間	約 100 日間	約 40 日間	約 540 日間
Web ページ数	14,669 個	2,655 個	1,401 個	18725 個
データサイズ	1741.23MB	400.96MB	422.81MB	2565MB

この個人用 Web アーカイブに対し，上記の Web ブラウザで Web 閲覧を行った際に見られた 2 種類の例を以下に示す.

1. ニュースサイトのトップページで過去の記事を探す例
2. ウェブログで以前の閲覧した記事の次の記事を読む例

6.1　過去の記事を探す

1 の例について述べる．閲覧者は，頻繁に閲覧しているニュースサイトで一ヶ月前に閲覧した記事を探しているとする．まず，閲覧者はブックマークからニュースサイトのトップページにアクセスする．すると，Web アーカイブ提示部に個人用 Web アーカイブに存在する過去の Web ページのバージョンが提示される．また，どの時刻からどの時刻までの時期のバージョンが表示されているかと，その範囲以外に過去のバージョンが存在する数が両端に提示される．中央の時刻は現在の時刻である（図 6.1）.

ここで，現在の表示期間は month なので，前後半月分のデータが提示されていることになり，一ヶ月前のバージョンは提示されていない．そこで，Web アーカイブ提示部上で 4.4 節で

図 6.1: トップページへの訪問

述べたジェスチャを使い，表示期間を year に変更する．すると，一年分のバージョンが Web アーカイブ提示部に現れる（図 6.2）．なお，現在閲覧しているバージョンについては，次回の閲覧，または，他のバージョンの閲覧を行うと提示され始める．

ここで，中央から少し左にいくつかのバージョンが存在することが分かる．表示期間が year であることを考慮すると，それらのバージョンが約一ヶ月前に閲覧したバージョン群であり，それらの中のいずれかに探している記事が存在していると予想できるため，それらのバージョンを纏めて選択する．すると，選択されたバージョン群がマージされ，差分が強調された Web ページが提示される（図 6.3）．

提示された Web ページをスクロールし，強調されている差分に注目しながら内容を確認する（図 6.4）．トップニュースがいくつか鮮度の順に提示された後，他のニュースの見出しが並んでいるのが分かる．探しているニュースがトップニュースであったか，それ以外であったかなどを手がかりにして閲覧を続け，そのニュースが見つかればその詳細の閲覧を行う．以上の作業により目的が達成される．

一方，一般的な Web ブラウザでこの作業を行った場合について述べる．新聞社が運営するニュースサイトのいくつかで過去の記事を探そうとした場合，過去の記事へのリンクが見つけられず，代わりに有償の記事データベースを公開している Web サイトがほとんどであった．ただし，調査したすべての Web サイトにはキーワードによる記事検索用のフォームが用意されていた．したがって，これらの Web サイトから過去に閲覧した記事を探す場合，上記のフォームからキーワード検索により記事を探す方法，Web 履歴から記事を探す方法を用いた．しか

図 6.2: 表示期間の変更

図 6.3: バージョンの選択

図 6.4: 結果の閲覧

し，これまで述べてきたように，それらの方法は確実性やインタフェースに問題があり，必ずしも目的の記事を発見できるとは限らず，また，発見できた場合でも多くの時間を浪費することが多くあった．

6.2 記事の続きを読む

2 の例について述べる．更新毎に閲覧するわけではないが，機会があれば未読の情報は読むという閲覧方法をとる Web ページがいくつかあり，そのような Web ページが主にウェブログであった．この種類の Web ページを閲覧する際，本システムを Web ページの「栞」のように用いることで，閲覧の効率が上がることがあった．

閲覧者は，ウェブログを閲覧する際，まずそのトップページにアクセスする．すると，1 の例の場合と同様に，Web アーカイブ提示部に Web ページの過去のバージョンが提示される．次に，提示された中で最近のバージョンの閲覧時刻の確認を行う（図 6.5）．

最近の閲覧時刻が分かると，多くのウェブログに用意されている月毎に纏められた記事のアーカイブや記事へのリンクを持つカレンダーを利用して，最近の閲覧時刻以降の記事のうち最も古い記事の閲覧を行う．以上の作業により目的が達成される．

一方，一般的な Web ブラウザでこの作業を行った場合について述べる．まず，閲覧しているウェブログにおける最近の閲覧時刻の想起を行った．次に，想起した時期の月の記事のアーカイブやカレンダーの日付の選択を行った．しかし，記事のアーカイブの場合，一ヶ月分並

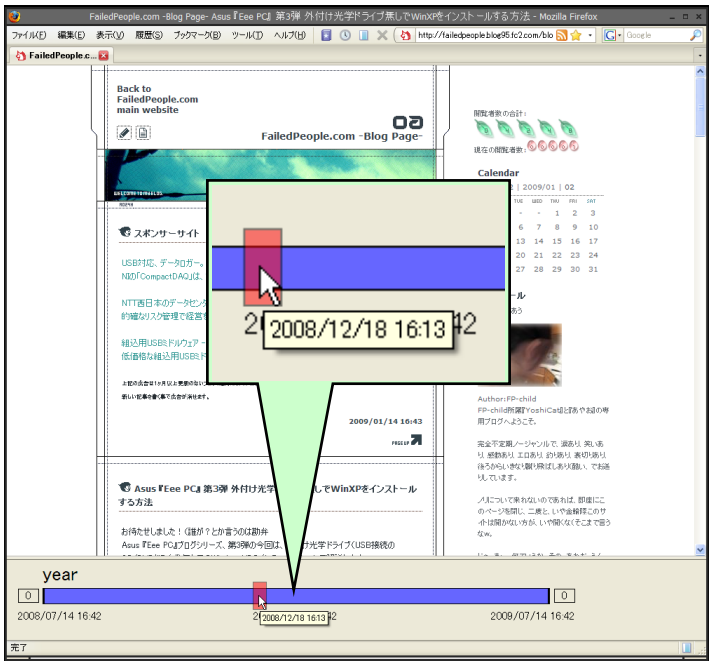

図 6.5: バージョンの選択

んだ記事群の中から，記事の日付だけでなく内容を確認しながら目的の記事を探さなければならなかった．また，カレンダーの場合，選択した日付の記事の日付と内容と確認し，閲覧していた場合は次の記事を，閲覧していなかった場合は前の記事を確認するという作業を繰り返す必要があった．

この利用方法は，前述の記憶からの想起を出発点とする既存の方法よりも時間効率が良く，また，発見した記事の正確さも高いものと思われた．また，最初の時点でどの程度の期間の記事を読めばよいかも分かるため，閲覧作業に必要な時間も予測しやすくなった．以上から，本システムを Web ページの「栞」のように扱う方法も有効であると考えられる．

第7章　議論と今後の課題

7.1　再訪問のための検索について

本論文では，インタフェースの比較対象として Wayback Machine[2] に代表されるクローラ型 Web アーカイブを取り上げた．クローラ型 Web アーカイブでは，再訪問のために目的の Web ページの URL が分かっていなければならない．本システムでも，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履歴，Web ページ上のリンクなどから目的の Web ページを訪問できる状況，または現在その Web ページを閲覧中である状況を利用場面として想定して研究を行った．

Web ページへの再訪問の手段については，関連研究で述べた Web アーカイブの検索に関する研究に加え，Greenberg らの研究 [37, 38] など，多くの研究が行われている．本研究でも，Personal Web Archive をインストールした Web ブラウザ全体としては，作成した Web アーカイブからのキーワード検索の結果を並べる程度が可能となっているが，システムを実用するにあたっては，本システムに適した検索インタフェースを開発する必要があると思われる．

7.2　個人用 Web アーカイブについて

作成した個人用 Web アーカイブについての考察を行う．まず，そのデータサイズについて，ローカルマシン上に保存を行うことからディスクの容量を圧迫するという懸念があった．2001 年に行われた本研究室内での矢野の調査 [39] では，画像ファイルを含めた Web ページを一年間に閲覧する容量は約 1,600MByte という結果がでている．また，2006 年，および 2007 年に行われた本研究室内での小澤の調査 [40] でも 2 名の Web 閲覧による Web ページの収集が行われており，約 365 日間で 6884 個，約 112 日間で 7147 個の Web ページが収集されている．前章の表 6.1 に示したように，合計約 540 日の間に収集した 17842 個の Web ページのデータサイズは約 2,565MByte であり，Web ページ 1 個あたりの平均データサイズは約 143.76KByte となる．これが小澤の調査にも適用できるとすると，2 名の結果はそれぞれ，365 日間で約 990MByte，112 日間で約 1,027MByte となる．以上をまとめたものを表 7.1 に示す．

表中の平均は，1 日の Web 閲覧を行う平均データサイズを示しているが，2001 年から 2008 年までの間に大きく変化した傾向はない．一方，過去に 10 年間に販売されたハードディスクの中で，容量 1GByte あたりの価格が最も安い製品を調査した結果を表 7.2 に示す（Impress Watch[1]調べ）．表中の「容量」はハードディスクの容量（GByte）,「1GB 単価」は 1GByte あ

[1] http://www.watch.impress.co.jp/

たりの価格（円）,「平均価格」はその製品の平均価格（円）である．

表 7.1: 個人用 Web アーカイブのデータサイズ

	期間	日数	データサイズ	平均
矢野	2001 年	365 日	1,600MB	4.38MB
小澤 1	2006 年 6 月～2006 年 12 月, 2007 年 6 月～2007 年 12 月	365 日	990MB	2.71MB
小澤 2	2006 年 6 月～2006 年 10 月	112 日	1,027MB	9.17MB
本研究 1	2006 年 9 月～2007 年 10 月	400 日	1,741MB	4.35MB
本研究 2	2008 年 2 月～2008 年 6 月	100 日	401MB	4.01MB
本研究 3	2008 年 12 月～2009 年 1 月	40 日	423MB	10.58MB

表 7.2: ハードディスクの容量と価格

調査年度	容量	1GB 単価	平均価格
2000	15	751.6	13,356
2001	80	332.1	28,370
2002	80	197.3	15,864
2003	120	120.7	14,974
2004	160	65.5	10,936
2005	160	48.6	8,187
2006	250	42.0	10,884
2007	250	29.9	7,878
2008	500	10.0	10,649
2009	1000	7.3	7,911

表 7.2 において，前述の調査が行われた 2001 年，2006 年，2007 年，2008 年を比較する．2001 年の 1GB 単価 332.1 円に対し，2006 年は 42 円であり，約 1/8 になっている．さらに，2006 年と 2007 年，2007 年と 2008 年の比較では，それぞれ約 3/4，約 1/3 と 1GB 単価は徐々に下がっている．また，表中の 1GB 単価と容量について，その推移を図 7.1 に示す．グラフより，1GB 単価が下がっているのと同時に，ハードディスクの大容量化が進んでいることが分かる．このように，Web 閲覧により保存するデータサイズに変化がないのに対し，ハードディスクの低価格化，大容量化が進んでおり，ハードディスクの容量の圧迫に関しては懸念する必要はないと考えられる．ただし，本研究では動画の保存を行ったと述べたが，Flash 内から呼び出される動画ファイルについては保存していない．多くの動画共有サイトでは，この方法で動画が配信されており，それらの Web サイトを対象として動画ファイルの保存を行う場合は，新たな調査が必要となると考えられる．

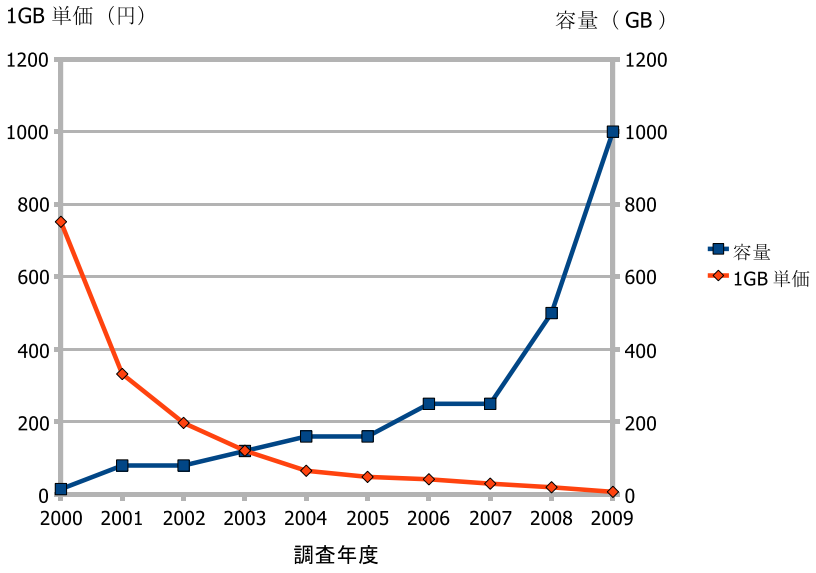

図 7.1: 1GB 単価と容量の推移

7.3 Personal Web Archive について

差分の抽出の精度について述べる．本研究では，差分の抽出に Web ページの HTML の DOM ツリーの比較という方法をとった．この方法では，HTML の仕様に従って記述されていない Web ページや，スタイルシートを多用して情報の位置を記述している Web ページに対して差分の抽出の精度が悪化する．また，Ajax などにより動的に構成が変化する Web ページでは，変化後の状態は保存されない．以上のような Web ページに対しても差分を抽出する方法が求められる．

時系列データの提示方法について述べる．Web ブラウザの履歴や更新日時によるファイルのソートなどの時系列データの提示は，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による時系列データの提示では，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．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閲覧時刻を考慮してバージョンの提示を行った．この方法は，一次元リストによる時系列データの提示と比較すると優れているといえるが，十分な設計が行われているとはいえない．したがって，本研究の目的に合致した新しいインタフェースを設計することが今後の課題である．

第8章　おわりに

本論文では，過去に閲覧したWebページへの再訪問を支援するシステムPersonal Web Archiveについて述べた．Personal Web Archiveは，閲覧者はWeb閲覧を行う過程で，閲覧したWebページの複製を収集した個人用Webアーカイブの作成を自動的に行う．さらに，作成した個人用Webアーカイブ内に存在する，保存時刻が異なるが同一のURLを持つWebページ群に対し，その中の複数のWebページ間の差分を同一画面内に提示することによって，それらのWebページの比較，閲覧の支援を行う．また，本システムを利用することによって，どのようにWeb閲覧が支援されるかの確認を行った. さらに，その結果と既存のシステムを用いた場合の結果の比較による本システムの有効性の検証について述べた. 最後に，本システムについての考察と今後の課題について述べた.

謝辞

本論文の執筆にあたり，指導教員として丁寧な御助言と御指導を頂いた田中二郎先生に心より感謝いたします．先生には貴重な研究資料，快適な研究環境など様々な点において御助力を頂きました．厚く御礼申し上げます．志築文太郎先生には，日常のゼミ活動やミーティングなどを通し，研究の着手から論文の執筆まで，研究全般に対する丁寧な御助言と御指導を頂きました．心より感謝いたします. 三末和男先生，高橋伸先生にはゼミ活動などの機会に有益な議論の機会を与えていただきました. 心より感謝いたします. 最後に，田中研究室の皆様にも大変お世話になりました．とりわけ，WAVE チームの皆様には研究の全般にわたり貴重な御意見を頂きました．ここに深く御礼申し上げます.

参考文献

[1] 株式会社アフィリティー. ウェブ魚拓. http://megalodon.jp/.

[2] Internet Archive. Wayback Machine. http://www.archive.org/web/web.php.

[3] The Singapore Internet Research Centre. Asian Tsunami Web Archive. http://september11.archive.org/.

[4] september11.archive.org. The September 11 Web Archive. http://september11.archive.org/.

[5] The Library of Congress. United States Election 2002 Web Archive. http://lcweb4.loc.gov/elect2002/.

[6] robotstxt.org. The Robots Exclusion Protocol. http://www.robotstxt.org/.

[7] 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.

[8] 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.

[9] Lara D. Catledge and James E.. Pitkow. Characterizing browsing strategies in the World-Wide Web. In *Proceedings of the Third International World-Wide Web conference on Technology, tools and applications*, pp. 1065–1073, New York, NY, USA, 1995. Elsevier North-Holland, Inc.

[10] Harald Weinreich, Hartmut Obendorf, Eelco Herder, and Matthias Mayer. Not Quite the Average: An Empirical Study of Web Use. *ACM Trans. Web*, Vol. 2, No. 1, pp. 1–31, 2008.

[11] Google, Inc. Google Bookmarks. http://www.google.com/bookmarks/.

[12] Google, Inc. Google Web History. http://www.google.com/history/.

[13] web.resource.org. RDF Site Summary (RSS) 1.0. http://web.resource.org/rss/1.0/.

[14] RSS Advisory Board. RSS 2.0 Specification (version 2.0.10). http://www.rssboard.org/rss-specification.

[15] The Internet Engineering Task Force. Request for Comments: 4287. http://www.ietf.org/rfc/rfc4287.txt.

[16] Hanzo Archives Limited. Hanzo:web. http://hanzoweb.com/.

[17] 五味渕大賀. ScrapBook. http://amb.vis.ne.jp/mozilla/scrapbook/.

[18] Herman Chung-Hwa Rao, Yih-Farn Chen, and Ming-Feng Chen. A Proxy-Based Personal Web Archiving Service. *SIGOPS Oper. Syst. Rev.*, Vol. 35, No. 1, pp. 61–72, 2001.

[19] 安川美智子, 山田篤, 星野寛, 大瀬戸豪志, 上林彌彦. Web コンテンツの収集と再利用を支援する個人用アーカイブシステム. 情報処理学会研究報告, Vol. 2002-DBS-129, pp. 139–146, 2003.

[20] 田村孝之, 喜連川優. 大規模 web アーカイブのための更新クローラの設計と実装. 電子情報通信学会論文誌 D, Vol. J91-D, pp. .551–559, 2008.

[21] 福井雅士, 遠藤裕英. ウェブアーカイブを目的とした html スクリプトブロック化と差分格納方式. 情報処理学会研究報告, Vol. 2005-FI-78, 2005-DD-49, pp. 33–40, 2005.

[22] 柊和祐, 阪口哲男, 杉本重雄, 田畑孝一. 情報発信組織主導の web アーカイブシステム. 情報処理学会研究報告, Vol. 2003-FI-73, pp. 77–84, 2003.

[23] 角谷和俊, 田中克己. Web アーカイブのための時間情報管理とその応用. 情報処理学会研究報告, Vol. 2003-DBS-131, pp. 109–116, 2003.

[24] 賀家智代, 角谷和俊. Web アーカイブのための質問キーワードの順序依存を考慮した時系列ページ検索. 情報処理学会研究報告, Vol. 2005-DBS-137, pp. 91–97, 2005.

[25] Yoshinari Shirai, Yasuhiro Yamamoto, and Kumiyo Nakakoji. A History-Centric Approach for Enhancing Web Browsing Experiences. In *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*, pp. 1319–1324, New York, NY, USA, 2006. ACM Press.

[26] 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 論文集, 情報処理学会, pp. 223–224, March 2006.

[27] 白井良成, 中小路久美代, 山本恭裕, 平田圭二. インタラクションヒストリを顧みるための表現系と操作系の試作. 情報処理学会研究報告, Vol. 2006-HI-121, pp. 9–16, 2006.

[28] Adam Jatowt, Yukiko Kawai, Satoshi Nakamura, Yutaka Kidawara, and Katsumi Tanaka. Journey to the Past: Proposal of a Framework for Past Web Browser. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pp. 135–144, New York, NY, USA, 2006. ACM.

[29] 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.

[30] Susan Dumais, Edward Cutrell, JJ Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins. Stuff I've Seen: A System for Personal Information Retrieval and Re-Use. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 72–79, New York, NY, USA, 2003. ACM.

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

[32] Google, Inc. Google Desktop. `http://desktop.google.com/`.

[33] Koji Tsukada, Satoru takebayashi, and Toshiyuki Masui. Dying Link. In *Proceedings of the 10th International Conference on Human-Computer Interaction (HCI 2003)*, Vol. 3 (Human-Central Computing), pp. 1353–1357, June 2003.

[34] 塚田浩二, 高林哲, 増井俊之. 廃れるリンク. 情報処理学会論文誌, Vol. 43, No. 12, pp. 3718–3721, December 2002.

[35] 塚田浩二, 高林哲. 廃れるリンク. インタラクション 2002 論文集, 情報処理学会, pp. 73–74, March 2002.

[36] W3C. Document Object Model (DOM) Specifications. `http://www.w3.org/DOM/DOMTR`.

[37] Andy Cockburn, Saul Greenberg, Bruce Mckenzie, Michael Jasonsmith, and Shaun Kaasten. WebView: A Graphical Aid for Revisiting Web Pages. In *Proceedings of OZCHI'99, Australian Conference on Human Computer Interaction*, pp. 15–22, 1999.

[38] 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.

[39] 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.

[40] 小澤崇記. スレッドに基づく Web 閲覧履歴検索インタフェース. 修士論文, 筑波大学大学院システム情報工学研究科, 2008.