データベースシステム 129－19
（２００３．１．２４）

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

安川　美智子[†]　山田　篤[†,††]　星野　寛[†,††]
大瀬戸　豪志[†,†††]　上林　彌彦[†]

Web 閲覧の機会が増える中で，閲覧した Web コンテンツを後で再閲覧，再利用する必要性が高まっている．そのためには，後で利用したいコンテンツを含む，有用な Web ページを保存（アーカイブ）しておくこと，そして，以前に保存しておいた Web ページに，必要に応じて，効率よくアクセスできることが重要である．また，Web コンテンツを効率よく収集，再利用するためには，他者との間でコンテンツやコンテンツのメタデータを流通させることが必要になってくる．このため我々は，個人用の Web アーカイブにおけるデータ管理と利用の手法を提案し，個人用 Web アーカイブシステムのプロトタイプを開発した．提案システムでは，キャッシングプロキシの仕組みを利用した個人用のプロキシを用いて，ユーザが閲覧した Web ペー ジ，及び，ユーザの Web 閲覧履歴を蓄積し，ユーザが閲覧した Web ページ中のテキスト情報と，ユーザが用いた Web 検索エンジンに対する検索語をもとに，Web ページの自動分類を行う．これにより，ユーザは，過去に閲覧した Web ページや他者が閲覧した有用な Web ページに含まれる Web コンテンツを効率よく閲覧することができ，Web コンテンツの収集や再利用を行いやすくなる．

A Personal Archiving System for Collection and Reuse of Web Contents

Michiko Yasukawa,[†] Atsushi Yamada,[†,††] Hiroshi Hoshino,[†,††]
Takashi Oseto[†,†††] and Yahiko Kambayashi [†]

Recently, the opportunities of web browsing are increasing. It intensifies the need for re-access and reuse of web contents. Therefore an archiving of worth web pages that include desirable web contents for reuse, and a efficient re-accessing of restored web pages and web contents in the archives are required. Furthermore, it has become necessary that web content and its meta-data are exchanged among users to collect and reuse the web contents effectively. In this paper, we propose a method of the archiving of web pages and a prototype of the personal archiving system is explained. Our system utilized the mechanism of caching proxy and collect web pages, contents and user's profile data, such as a browsing history of the user. Web pages in archives are automatically categorized based on the text data in browsed web pages and search words for web search engines. With this categorization, a user of the archive system can efficiently re-access worth web pages and contents that were browsed by the user or by other users. It is supposed to help the collection and reuse of web contents effectively.

1. はじめに

Web 上で提供され，自由に閲覧できるコンテンツの中には，素材として利用価値の高いものが多数ある．このような Web 上の有用なコンテンツを一度閲覧や視聴するだけにとどまらず，必要に応じて後で再び閲覧や視聴することや，Web コンテンツを素材として利用して，二次的なコンテンツを作成できることが望ましい．Web 上では，膨大な数の Web ページ，Web ページ中に含まれるコンテンツが提供されており，少数の有用な情報を探し出すことは難しく，時間がかかる．このため，一度アクセスした情報に後で再びアクセスする際に，同じ手間をかけないことを目的として，ユーザのアクセス履歴やブックマークを使ったユーザ支援がこれまでに多数，提案されてきている．また，Web の普及にともない，再アクセスを支援する研究や，検索機能の向上に関する研究がこれまでに多数行われている．現在では，Web サーチエンジンを利用した検索は，以前と比較して，幅広いユーザの要求を満たすものになってきている．また，数年前までは，

† 京都大学大学院情報学研究科社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University
†† 財団法人 京都高度技術研究所
ASTEM RI.
††† 立命館大学大学院法学研究科
Graduate School of Law, Ritsumeikan University

現在と比較して一般ユーザが利用可能なコンピュータの性能はあまり高くなく，ハードディスクも小容量であったのに対して，現在は，ノート PC など小型のものも含め，コンピュータの CPU，メモリなど性能は向上しており，HDD も大容量化，小型化が進んでいる．コンテンツの視聴や閲覧など一次利用についてのユーザ環境は，Web の出現当初と比較してかなり向上したといえる．しかし，個々のユーザのコンテンツ閲覧やコントロール，コンテンツの二次利用については，まだ十分に検討されているとは言えない．

現在，あらゆる情報が Web 上で提供されるようになってきていることから，Web を閲覧する機会が増えており，閲覧した Web ページやページ中のコンテンツを効率よく管理したいという要求が高まっている．Web 閲覧などのユーザ支援を目的として Web サーバとユーザとの間に入ってユーザを支援する仕組みが提案されている[1]．またユーザ支援の仕組みを容易に実現できるミドルウェアとして，個人用のアーカイビングプロキシが提案されている[2)3)]．

しかし，これらの提案では，Web 上のデータをどのように保管，管理，利用するか，ということについて十分な検討がなされていない．Web のアーカイブについては，現在，その重要性が取り上げられ，アーカイブにおける課題や，どのようなデータを，どのような方法で保存すべきかなどアーカイブの方法が，検討され始めている[4)5)6)]．

以上のことを背景に，本論文では，Web コンテンツを効率よく収集，再利用するための個人用のアーカイブシステムを提案した．我々のシステムは，従来の Web 閲覧支援で実用上，本質的な問題となる通信上の問題，プライバシ上の問題，著作権上の問題を，ユーザのマシン上で個人的に運用するアーカイブシステムにより，解決している．また，従来のプロキシキャッシュの仕組みを単純に適用するだけでは保存できなかったデータについてもファイル名の付け方を工夫することで保存できるようにした．さらに，アーカイブに保存されているデータの効率の良い管理と閲覧のためのメタデータとテキスト自動分類についても検討した．

以下，本論文では，2 章で，Web コンテンツの収集と再利用における課題について説明し，3 章で，その課題への取り組みとして我々が提案する個人用アーカイブシステムについて述べる．また，4 章では，アーカイブに保管したページやコンテンツを効率よく参照できるようにするために提案システムに適用した，テキスト自動分類の手法について述べる．5 章では，開発したプロトタイプについて述べ，6 章でまとめと今後の課題について述べる．

2. 研究の背景と課題

インターネット上の膨大な情報の中から，少数の有用な情報を探し出すことは難しく，時間がかかるため，一度アクセスした情報に再びアクセスする際に，同じ手間をかけないことが望ましい．そこで，一度アクセスした情報への再アクセスを効率化することを目的として，これまでに，ユーザのアクセス履歴やブックマークを使ったユーザ支援が，これまでに多数提案されている．

文献[7)] では，ユーザが作成したブックマークからタイトルや著者名などのメタデータを抽出して，未分類の URL に対して，索引付けを行い，分類するシステムが提案されている．また，ユーザが入力した好みに合うもの，ユーザの頻繁にアクセスするものをホットリストとしてユーザに提示する，興味のないものを自動削除する，更新があったページをユーザに知らせる，などの機能がある．

文献[8)] では，プロキシを使って，ユーザが閲覧するページの上部に CGI のフォームを挿入し，ユーザにページを分類させて，ブックマークの格納と索引付けを行う手法が提案されている．

また，自然言語の方法とタキソノミ（分類学）の方法を使ってブックマークの管理支援を行う[9)] や，ユーザが閲覧したページのキャプチャをサムネイルとして使ってブックマークのインタフェースを改良する研究[10)]，CBR（事例ベース推論）に基づき，エージェントがユーザのブックマーク分類に関する戦略を学習して，他のエージェントからブックマークを受け取り，ユーザのリポジトリに追加する[11)] が提案されている．また，URL のリストを集めてユーザに推薦をするシステム[12)] や，ページのタグの解析をしてページの特徴抽出をしてユーザにブックマークを提示する[13)] などもある．また，[14)] は，ユーザがなぜそのページを保存したか，あとで思い出しやすくするために，ページの断片にコメントを付けて保存しておき，その保存された部分とコメントの索引付けを行う手法を提案している．しかし，従来提案されたユーザ支援には以下のような問題点と解決すべき課題がある．

課題 (1)

ユーザと Web サーバとの間で，第三者の立場で行うユーザ支援は，通信効率化，プライバシ保護，著作権侵害の点から，実用化の点で問題がある．まず，通信上の問題として，ブックマーク管理などのサーバを介することは，ユーザは，目的の Web ページをもつ

サーバへアクセスする前に，別途，ユーザ支援のためのサーバへアクセスすることが必要になる．ユーザがブックマークにアクセスするという行為自体は簡単で短時間に行われるものであるにもかかわらず，間に入るサーバアクセスのために，時間的，通信的にコストの高いものになってしまう．

複数のユーザがアクセス可能な状態になっているプロキシサーバでは，プライバシの問題と著作権の問題が生じる．ブックマークやページに対するコメント，ページ閲覧における興味や嗜好などの情報はユーザにとって他人に公開したくないプライバシデータである．Web サーバから個人情報が流出するなどの事故がおきており，再アクセス支援の目的で蓄積するブックマーク等のプライバシデータも第三者に保管させるとこうした危険にさらされる可能性がある．また，ユーザの個人のコンピュータ（クライアント）と，WWW サーバとの間に入るプロキシは，ユーザと WWW サーバのコンテンツの著作者と関係において，第三者となり，この第三者が，ページのコピーや，ユーザがコメントを付加したコンテンツを保持し，そのユーザや他のユーザに参照させることは著作権上の問題を生じる．

課題 (2)

従来のブックマークの研究は，前提として，利用可能な検索エンジンがあまり有用ではなく，ユーザは検索エンジンを駆使し，ネットサーフィンを行うことを想定しており，ユーザが苦労して集めた少数の有用なサイトの所在情報を有効活用することが主眼であったが，現在の検索エンジン（たとえば Google[15] や Yahoo![16] など）は，かなり改良されている．また，ユーザがブックマークに追加したい有用なサイトは多数あることから，一時利用するためのサイトをブックマークで管理しようとするとブックマークは増加する一方であり，使い勝手が悪くなる．固有名詞を含むなど特定性の高い情報はブックマークをたどるより，検索エンジンを用いたほうが早い場合もある．一方で，ユーザによる，コンテンツの収集や二次利用を目的とした，個別化やユーザ支援については，現在のところ十分に検討がされているとはいえない．

ユーザが Web コンテンツの収集と再利用行う際に，求められるユーザ支援は，次のようなものである．

- 有用なページとその関連ページを一緒に見たい
- 未分類のページを分類してほしい
- あるページに類似のページを見たい
- 他のユーザに，このページに似ているものを知っていたら教えてほしいと問い合わせたり，他のユーザからの同様の問い合わせに答えたりできるようにしたい

未分類ページの分類の仕方としては，システムが独自にカテゴリを決めて分類する方法と，ユーザがあらかじめ決めたカテゴリにシステムが対応付けをする方法が考えられる．ユーザがあらかじめ何らかのカテゴリ情報をもっていれば，システムがユーザに合わせる方法のほうがユーザにとっては使いやすいと考えられる．ユーザは，一般的な情報の中から，自分の視点・見方で判断した，自分にとって有用な情報を見つけ出している．このため，分類の際には，一般的な関連付けではなく，そのユーザが持っている視点・見方を反映することが必要である．

Web コンテンツの収集と再利用を支援するためには，上に述べた従来手法の問題，残された課題を解決し，有効でかつ効率の良いコンテンツ保管，利用の仕組みを実現することが課題となる．

3. 個人用アーカイブシステム

上で述べた，Web コンテンツの収集と再利用における課題 (1) を解決するため，我々は個人用のアーカイブシステムを提案する．まず 3.1 では，研究の前提となる事項をまとめ，個人によるアーカイブ構築について述べる．つぎに，アーカイビングの手法と，保管されたデータの管理・利用のためのメタデータについて説明する．

3.1 個人によるアーカイブ構築

ユーザがコンテンツを再び閲覧する際に，サーバが別のところにあると，ユーザ支援のためのアクセスに通信の負荷がかかってしまうことから，ユーザ支援はできるだけ，ユーザの近いところで行うのがよいと考えられる．また，我々は各ユーザのプライバシ情報を，ユーザのコントロールの及ばないところに保管するのは望ましくないから，プライバシ情報はユーザが管理できるマシン上に保管するべきである．第三者がアーカイブデータを保管することは著作権上の問題が生じる可能性があり，また，複数のユーザ間で知識を交換する際にコピーを流通させると著作権上の問題が生じる．著作権上の問題を生じさせず，また，ユーザの利便性を損なわないシステムを実現することが必要である．本研究では，以下のことを前提とする．

- 私的な複製は自由に行える
- メタデータの交換は自由に行える

ユーザが自分で利用するために，コピーを自分のマシン上に保管することについては，私的複製であり，著作権法上の問題を生じない．ただし，コピープロ

テクトのついているものは，私的な複製であっても自由に行えないことから，本システムでは，音楽・映像など特殊なフォーマットで流通し，コピープロテクトがかけられているものはアーカイブの対象としない．Web 上で自由に閲覧され，ユーザ側で印刷，ファイルに保存，媒体に保存できる Web 上のコンテンツを対象とする．従来は，個人で使えるマシンの性能が低かったが，現在では，各個人が，アーカイブを構築するのに十分な性能のマシンが一般に利用可能になっている．複数のユーザ間で知識を交換する際に，データの実体を流通させると著作権侵害の問題が生じるが，メタデータ（書誌情報と呼ばれる）は自由に流通させることができる．

以上のことから，我々は，ユーザのコンテンツ収集と再利用を支援する上でもっとも適した環境は，個人用のアーカイブであると考えた．我々が提案する個人用の Web アーカイブシステムのシステム構成を図 1 に示す．

アーカイビングプロキシは，ユーザのマシン上で動作し，ユーザと上位のサーバとの間で送受信される全てのデータを保存する．

3.2 アーカイビングの方法

ユーザが Web で参照したデータを保管するツールやシステムには，それ自身が単独で動作するアプリケーション型のものと，ブラウザと WWW サーバとの間で動作するプロキシ型のもの（たとえば Squid[17]，WWWOFFLE[18]）がある．アプリケーション型のものは，ファイルの一括ダウンロードや Web ページの巡回など，有用な情報を選択的にアーカイブできるが，アーカイブ作業をユーザが細かく指定しなければならない．一方，プロキシ型のものは，ユーザがアクセスしたものをキャッシングプロキシの原理で自動的に保管するため，アーカイブ作業においてユーザの手間がかからないというメリットがある．

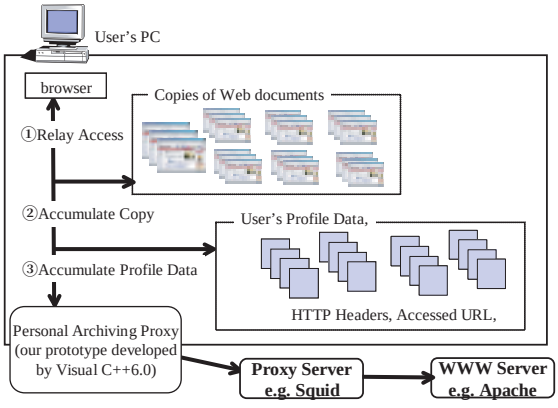

図 1 個人用 Web アーカイブシステム
Fig. 1 Personal Web Archiving System.

キャッシュとアーカイブは，データを保存するという点では同じであるが，その目的が異なる．通信の効率化を目的とする場合は，プロキシキャッシュが有効であるが，アーカイブを目的とする場合に，プロキシキャッシュを使うと問題が起きる．なぜなら，キャッシュは，有効なキャッシュデータのみ保存し，無効なデータ（古いデータ）は捨てられるからである．アーカイブにおいては，古いバージョンを保存したい場合もある．

また，動的に生成されるページなど，no-cache（レスポンスを保存しないように）と指示されたデータは，キャッシュデータとしては保存されない．no-cache と指示されたデータや動的に生成されるページは，通常はキャッシュの対象外とされる．動的に生成されるページのキャッシュに関する研究として文献[19)20)] などがあるが，これは通信効率化のためのキャッシュを行うものであり，古いデータは捨てられる．

アーカイブのためには，キャッシュとは異なるデータ保存の仕組みを検討しなければならない．まず，どのようにファイルを保存するかが問題になる．ファイル名の付け方として 2 つある．

(1) もとの URL をそのままファイル名として使う
(2) URL を別のファイルに置き換える

静的なページだけをアーカイブの対象とする場合は (1) が，もっとも簡単で，高速である．しかし，動的に生成されるページを (1) の方法で保管しようとすると問題が生じる．動的に生成されるページの URL は引数部分が長く，URL をファイル名として保存しようとすると，ファイル名が長すぎるために保存できない場合がでてくる．また，ファイル名としては使用できない文字が URL に含まれる場合もある．

動的なページや長い URL に対応するためには，(2) の方法が適している．我々のシステムでは動的なページも含めて，ユーザがアクセスしたページを全て保存したいので，(2) を採用する．

URL を別のファイル名に置き換える際には，以下の点を考慮しなければならない．

- ファイル名の重複を避ける
- 読み出しを効率よく行えるようにする

(2) のもっとも単純な方法は連番にすることである．しかし，各マシンで 0 や 1 から始まる番号をふると，あるマシンのアーカイブと別のマシンのアーカイブの間でデータの移動やコピーをする際にファイル名が重

複する．ある一人のユーザのアーカイブデータは異なるマシン間であっても，重複しないことが必要である．

ファイル名の重複を避け，書き込みと読み出しを効率よく行うため，URL をキーとする一定長さのハッシュ文字列を生成し，その文字列をファイル名として使用する方法が考えられる．同じ URL に対して，異なるバージョンのページが生成される可能性があることから，いつのバージョンかを示す時間を表す文字列を加えた，“URL” + “参照日時” の文字列をキーにして，ハッシュ文字列でファイル名を割り当てる方法を考えた．しかし，この方法では，変更されていない静的なページで同じデータが重複して保存されることになる．そこで，我々のシステムでは，HTTP レスポンスヘッダの「Last-Modified（サーバがリソースを最後に更新した時刻）」エンティティの設定されているページに対しては，“URL” + “Last-Modified” をキーとするハッシュ文字列で，そして,「Last-Modified」の設定されていないページに対しては “URL” + “参照日時” をキーとするハッシュ文字列でファイル名を生成することとした．

参照日時については，HTTP ヘッダのサーバからのレスポンスヘッダに「Date（メッセージが作成された日時を表すためにサーバがすべてのレスポンスにおいて生成する」エンティティがあるが，これは，異なるサーバ間で時間が一致しない場合があるため，アーカイブデータの参照時間を表すデータとして適さない．そこで，参照日時を表す文字列は，ユーザのマシンの時間をもとに生成することとした．

上記により，データ書き込みの際の，ファイル名重複の問題は解決される．一方，読み出しについては，URL のハッシュ文字列がそのままファイル名に対応している場合は，読み出しのとき，URL から読み出すべきファイル名がわかる．しかし，ハッシュ文字列＋時間のファイル名にした場合，もとの URL からだけでは，読み出すべきファイル名がわからないので，対応付けを別途保存しておく必要がある．

そこで，我々のシステムでは，どのバージョン URL がどのローカルファイルに対応するかを URL のメタ情報ファイル（URL に一意に対応するファイル）に記述しておくこととした．メタ情報ファイルに記述するメタデータの詳細については次節で述べる．

我々のシステムの方針は，ユーザとサーバの間でやりとりされた全ての情報を保存することである．これは，無駄な情報を保存する可能性があるが，何を保存するべきかはあらかじめわからないので，通信時に，どれを保存するかを判断すると，オーバヘッドが大きくなるので，全て保存して，後で不要なものを削除するという手法がよいと考えたからである．

次に問題になるのは，ユーザとサーバの間でやりとりされる情報を，どこで区切って一つのファイルにするか，ということである．HTTP の通信は，クライアントからのリクエスト，サーバからのレスポンスにより行われるため，Web アーカイブの対象となるデータは時系列的なデータである．

全てのリクエスト，レスポンスを一つのファイルに収めることも可能だが，後で，言語処理，画像処理など，データごとに適した処理を効率よく行うためにはデータの性質を考慮してファイルを分けたほうがよいと考えた．我々のシステムでは，HTML に対してのインデクシングを行う (これについては 4 章で述べる) ため，HTML データを分けて保存することにした．

3.3 メタデータ

静的なメタデータ

プロキシは，ユーザとサーバの間の通信を中継する際にデータのコピーを保管し，ユーザからの要求に応じてアーカイブデータを読み出してユーザに渡す必要がる．そのためには，アーカイブにデータが保管されているか，されている場合，どのデータが必要とする URL に対応しているか，また，複数のバージョンがある場合，各バージョンの参照日時はいつか，最新のものはどれか，などの情報を静的なメタデータとしてメタ情報ファイルに保存する．我々のメタデータでは，XML 風に改良した HTTP ヘッダのフォーマットを用いている．これは，通信時のオーバヘッドをできるだけ少なくするため HTTP ヘッダの文字列をそのままの形式で保存する（エスケープ処理を行わない）こと，異なる改行コードでメタ情報ファイルを移動したときエンティティの区切りが失われないようにする（開始タグを付ける）という 2 つの点を考慮したためである．

動的なメタデータ

動的なメタデータは，アーカイブ内に保存されている HTML データ，メタデータ，HTTP ヘッダから必要に応じて生成されユーザによるアーカイブ参照や，他のユーザとの情報交換において利用される．たとえば以下のものがある．

- ページのタイトル
- 先頭文字列
- キーワード
- アーカイブデータのハッシュ値

URL や参照日時だけでは，異なるユーザ間で，データの同一性を保障できないため，同じページに興味が

あるなどの情報のやりとりを正確におこなえない．しかし，ページのデータの実体をやりとりすると著作権の問題がおきるので，ファイルのハッシュ値を用いる．ファイルのハッシュ値は，アーカイブに保管されているデータ実体のファイルをキーとして計算される．

4. テキスト自動分類手法の適用

2 章で述べた，Web コンテンツの収集と再利用における課題 (2) を解決するため，我々は個人用のアーカイブシステムにおけるユーザ支援に，テキスト自動分類の手法を適用する．我々の目的は，ユーザの既定のカテゴリ（たとえばブックマーク）に，ユーザが参照した Web コンテンツや他のユーザがアーカイブに保存している Web コンテンツの所在情報（URL）を自動分類することである．ユーザ支援のためには，自動分類に，個々のユーザが持っている視点や見方を反映しなければならない．我々は，自動分類の前処理である索引語と辞書の作成で，ユーザの視点や見方が反映されるように工夫した．

Web コンテンツはマルチメディア情報であるが，現在はテキストコンテンツのみを索引付けの対象とし，分類対象は URL に対応付けられている Web ページとする．

まず 4.1 と 4.2 では，自動分類のための前処理である索引語の選択と辞書の構築について述べる．そして，4.3 では，構築した辞書をもとに，文書の類似度を計算し，テキスト自動分類を行う手法について述べる．

4.1 索 引 語

索引語とは文書の内容を表す要素であり，特徴語と呼ばれる場合や，単に語と呼ばれる場合もあるが，ここでは索引語という呼び方で統一する．文書から索引語を抽出する処理は，索引付けと呼ばれる．索引付けの目的は，文書中から，その文書を特徴付ける語を漏れなく抽出することである．何度も言及される語は重要であると考えられることから，文書中の語の出現頻度に基づく索引付けが用いられることが多い．

しかし，文書中の語の出現頻度に基づく索引付けでは，個々のユーザが持つ「見方 (viewpoint)」が反映されないという問題がある．このため，あるユーザにとっては，文書の内容を特徴づける上で重要な語であっても，出現頻度が低ければ，より頻度の高い他の語が重要視されてしまう場合がある．たとえば，ユーザの参照した文書が,『懐石料理』『京料理』『祇園祭』『天神祭』に関するものである場合を考える．ユーザが『懐石料理』と『京料理』の関連性が高く,『祇園祭』と『天神祭』の関連性が高いと考えていても,『懐石料理』と『天神祭』に関する文書で,「大阪」という語が頻出であり,『京料理』と『祇園祭』に関する文書で,「京都」という語が頻出であれば,『懐石料理』と『天神祭』,『京料理』と『祇園祭』が関連付けられてしまう．

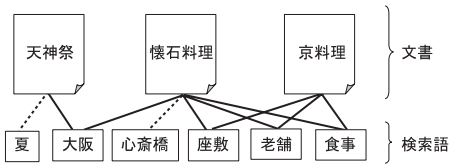

図 2 文書と検索語の関係
Fig. 2 Document and Search Word.

このような問題を解決するために，我々は，文書中の出現頻度の高い語ではなく，ユーザが検索エンジンに入力した検索語を索引語として，文書の関連付けを行う手法を提案する．ユーザが検索エンジンに対して入力する検索語は，ユーザが検索しようとする文書に対して持つ「見方」を表していると考えられることから，検索語を索引語とすることにより，ユーザの持つ視点を反映し，文書中に含まれる重要でない他の語の影響を受けない文書の関連付けが行えると考えられる．

たとえば，あるユーザが『懐石料理』に関する文書を「大阪」「心斎橋」「座敷」「老舗」「食事」,『京料理』に関する文書を「京都」「鴨川」「座敷」「老舗」「食事」,『祇園祭』に関する文書を「京都」「夏」「山鉾」「厄除け」「浴衣」,『天神祭』に関する文書を「大阪」「夏」「風物詩」「みこし」「浴衣」で検索した場合，たとえば,『懐石料理』と『京料理』では共通の検索語が 3 つあるのに対し,『懐石料理』と『天神祭』では共通の検索語が 1 つしかないため,『懐石料理』と『京料理』が関連付けられることとなる．このように，文書中の語の出現頻度ではなく，検索語に注目した関連付けを行うことで，このユーザに対しては,『懐石料理』と『京料理』，及び『天神祭』と『祇園祭』の関連付けが行われることとなる．

Web ページが，どのような検索語により検索されたかは，HTTP リクエストヘッダの「Referer」をたどり，検索結果ページを特定し，アーカイブデータから検索結果ページを取得して，検索結果ページの中に記述されている，キーワードを抽出することで得られる．

4.2 辞書の構築

ページ中に出現する語のなかには，ページを特徴付ける上で役に立たない不要語も多数あり，これらの語を省くことが必要である．大規模テキストデータを対象としたテキスト自動分類では，文書中の語をもとに，様々な辞書の構築が試みられている[21)22)23)]．たとえば以下のようなものがある．

- 一定回数に満たない出現頻度の語は重要ではないとして除外（たとえば5回未満の出現頻度であれば除外）.
- あらかじめストップワードリストを作成しておき，不要語を除外する.
- 機能語は除外する.
- 固有名詞や数字の一部は除外する.
- 全体を通じて，極めて稀にしか出現しない語（たとえば1件の文書でだけ出現する語）は除外する
- 品詞情報に基づき語の選択をする.

個人のアーカイブにおいては，あらかじめストップワードとしてどの語を登録しておくべきかが明確でないのでストップワードリストは用いない．また，文書数が少ないため，固有名詞や出現頻度の少ない語を除外すると，重要な特徴が失われる可能性があるため除外しない．数字や機能語については，文書を特徴付ける語としてはあまり役立たないことが多いことから，除外することとした．出現数の少ない語が，非重要語とは限らないので，除外しない．また，品詞情報を得るために形態素解析を利用する際に，辞書に登録されていない語は未知語として扱われる．しかし，未知語は重要な特徴語である可能性が高いことから索引語に含めることとした.

形態素解析には茶筌[24]を使用し，文書中の語のうち，以下の品詞が割り当てられた語を索引語として選んで，辞書に登録する.

- 名詞-一般
- 名詞-固有名詞
- 名詞-サ変接続
- 名詞-形容動詞語幹
- 未知語

以上により，ユーザがあらかじめ作成したカテゴリごとに，辞書を以下の2種類の辞書を構築する.

(1) カテゴリに分類されたページを検索する際に使用された語（検索語）の辞書
(2) カテゴリに分類されたページ中に出現する語（文書中の語）の辞書

ただし，ユーザが検索エンジンを使用せずに，Web閲覧を行った場合は，ページに対応する検索語はないため，検索語辞書は作成されない場合もある.

4.3 類似度計算と自動分類

分類対象となるページ（ターゲットページと呼ぶ）についても，4.2で述べた方法に従い，検索語の辞書と，文書中の語の辞書を構築する.

カテゴリ，ターゲットページのそれぞれについて構築した検索語の辞書，文書中の語の辞書をもとに，次のような辞書の使い分けと類似度計算を行う.

(1) カテゴリ，ターゲットページの両方で，検索語の辞書を用いた類似度計算
(2) カテゴリには文書中の語，ターゲットページには検索語の辞書を用いた類似度計算
(3) カテゴリには検索語辞書を，ターゲットページには文書中の語の辞書を用いた類似度計算
(4) カテゴリ，ターゲットページの両方で，文書中の語の辞書を用いた類似度計算（検索語辞書を用いない）

(1) は，検索語が，ユーザのもつ視点を表す信頼性の高い索引語であることから，ユーザのもつ視点を反映した最も精度のよい分類手法である．しかし，(1) の類似度計算を行うためには，ターゲットページとカテゴリの両方に検索語辞書が構築されている必要がある．カテゴリやターゲットページの検索語辞書が構築できない場合の代替案として，(2)(3) や (4) を利用する．類似度計算では，参照する辞書の語のページ中の出現回数をカウントし，出現頻度を重みとして索引語の重み付けを行い，索引語の異なり数を次元とするベクトルを生成して，余弦 (Cosine) を計算する.

5. プロトタイプ

我々は，提案手法に基づき個人用アーカイブシステムのプロトタイプを開発した．プロトタイプの開発には，VC++6.0 を使用した.

通信のプロトコルは，HTTP/1.0[25] に準拠しており，一般的なブラウザとサーバの間で基本的な Web サービスの中継を行えるようになっている．プロキシのコントロール（起動と停止）は図3のようなユーザインタフェースを用いて行う.

6. おわりに

本論文では，Web コンテンツを効率よく収集，再利用するための個人用のアーカイブシステムを提案した.

著作権上の問題を生じさせないため，提案システムでは他者との知識共有にコンテンツの所在情報などのメタデータを使用している．アーカイブに保存されている Web コンテンツのデータ実体を他のユーザと共有するためには，コンテンツの著作権処理が必要になる．我々は，これまでに Web コンテンツ再利用のためのライセンス処理システムを文献[26] において提案し

図 3 コントロールパネル
Fig. 3 Control Panel

ており，今後，個人用のアーカイブシステムに著作権処理のためのライセンスの仕組みを取り入れる予定である．ライセンス処理の仕組みを，アーカイブシステムに取り入れることで，さらに効果的，効率的な Web コンテンツの収集や再利用が可能になり，また，二次的な Web コンテンツの作成や流通も可能になる．

参 考 文 献

1) Maglio, P. and Barrett, R.: Intermediaries personalize information streams, *Communications of the ACM*, Vol. 43, No. 8, pp. 96–101 (2000).
2) Rao, H.: A Proxy-Based Personal Web Archiving Service.
3) Rao, H.: A Proxy-Based Personal Web Archiving Service, *Operating Systems Review*, Vol. 35, No. 1, pp. 61–72 (2001).
4) Digital Preservation Coalition: DPC Forum Web-archiving – managing and archiving online documents and records (2002).
5) Redfern, C.: Web-archiving: an introduction to the issues (2002).
6) Union College: Archiving WWW Sites WORKSHOP OBJECTIVES (2002).
7) Li, W. et al.: PowerBookmarks: a system for personalizable Web information organization, sharing, and management, *Proc. WWW8*, pp. 297–311 (1999).
8) Keller, R. et al.: A Bookmarking Service for Organizing and Sharing URLs, *Proc. WWW6*, pp. 1103–1114 (1997).
9) Torrance, M. C.: Active Notebook: A Personal and Group Productivity Tool for Managing Information, *Proc. AAAI Fall Symposium*, pp. 131–135 (1995).
10) Kaasten, S. and Greenberg, S.: Integrating Back, History and Bookmarks in Web Browsers, *Proc. CHI'01*, pp. 379–380 (2001).
11) Malek, M. and Kanawati, R.: A Cooperating Hybride Neural-CBR Classifiers for Building On-line Communities (2001).
12) Chen, L. and K., S.: WebMate : A Personal Agent for Browsing and Searching, *Proc. Agents'98*, pp. 132–139 (1998).
13) 森幹彦, 山田誠二: ブックマークエージェント：ブックマークの共有による情報検索の支援, 電子情報通信学会論文誌, Vol. J83-D-I, No. 5, pp. 487–494 (2000).
14) Denoue, L. and Vignollet, L.: New ways of using Web annotations (2000).
15) Google: http://www.google.com/intl/ja/.
16) Yahoo! JAPAN: http://www.yahoo.co.jp/.
17) Cache, S. W.P.: http://www.squid-cache.org/.
18) The WWWOFFLE Homepage: http://www.gedanken.demon.co.uk/wwwoffle/.
19) Datta, A. et al.: Dynamic content acceleration: A caching solution to enable scalable dynamic web page generation (2001).
20) Datta, A. et al.: Proxy-Based Acceleration of Dynamically Generated Content on the World Wide Web: An Approach and Implementation (2002).
21) Apte, C., Damerau, F. and Weiss, S. M.: Automated Learning of Decision Rules for Text Categorization, *Information Systems*, Vol. 12, No. 3, pp. 233–251 (1994).
22) 平博順, 春野雅彦: Support Vector Machine によるテキスト分類における属性選択, 情報処理学会論文誌, Vol. 41, No. 4, pp. 1113–1123 (2000).
23) 相澤彰子: Naive 手法による大規模テキスト分類問題へのアプローチ (2002).
24) 松本裕治ほか: 日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書 (2000).
25) Hypertext Transfer Protocol – HTTP/1.0: http://www.ietf.org/rfc/rfc1945.txt.
26) Yasukawa, M., Yamada, A., Hoshino, H., Oseto, T., Iwaihara, M. and Kambayashi, Y.: A Method for Making Dynamic License Agreements in Reuse of Web contents, 情報処理学会論文誌データベース, Vol. 43, No. SIG2(TOD13), pp. 179–191 (2002).