履歴分類の提示とアノテーションによるリファインディング支援

飯野 亜耶[†] 奥野 拓[††]

既読情報を再検索する行為をリファインディングと呼ぶ．本研究では Web 情報のリファインディングを対象に，分類された閲覧履歴の提示とマーカーを用いた支援手法を提案する．Web ページを閲覧している際，異なる目的の情報探索が同時に進行している場合があり，単純に見たページを時系列順で羅列しても効果的に記憶を想起できないと考えられる．そこでまず，内容の関連度をもとに閲覧履歴を分類した結果をユーザへ提示する手法を提案した．またページ内のテキストに直接マーカーを引き，ブックマークの代わりに用いることを提案した．本論文では分類手法を実際の履歴に適用し，閲覧目的ごとに分類できるか，また分類結果がリファインディングの情報提示に有効であるかを考察した．

Refinding Support by Using Clustered Web Browsing History and Annotation

Aya Iino[†] and Taku Okuno[††]

Refinding is searching information that has seen before. The objective of this research is refinding information on the Web. We propose a system for refinding support using clustered Web browsing history and markers. When users watching Web pages, they often browse for some different purposes in parallel. Therefore, even if another support system displays browsing history in chronological order, they may not been retrieved memory effectively. For this reason, the system shows browsing history clustered based on content association between pages. Additionally, it provides a function to mark texts in pages as to substitute for bookmark. This paper describes the result of evaluation of accuracy of clustering and how clustered history effective for information presentation in refinding.

1. はじめに

既読情報を再検索する行為をリファインディングと呼ぶ．本研究では，既読の Web ページをリファインディングできない，またはリファインディングに膨大な時間と手間がかかるという問題に取り組んでいる．研究の目的は，従来のリファインディング手段の問題点を改善し，コストの低い情報保存の手法やリファインディング手法の実現させることである．

2. 研究の位置づけ

2.1 従来のリファインディング支援手法と課題

リファインディング支援では，新規の情報を求めるための支援とは別に，専用の支援が必要であると Capla ら[1] が述べている．現在提案されているアプローチは，大きく二つの方向性に分類できる．

一つ目は Keeping Found Things Found[2] と呼ばれる方向性である．これは，ブックマークなど情報保存・整理手法を改善し，情報整理の簡単化やアクセス効率の向上を目指すものである．Web ページの全体または一部を保存し，全文検索を行えるツール ScrapBook[3] 等がこれにあたる．二つ目はユーザの履歴情報を解析し，目的の Web ページを推定するものである．

前者の手法では，ユーザが保存しなかったページのリファインディングには対応できない．後者の場合は，ユーザが意図的に保存したい情報も，一様に履歴中の一つの情報として処理され，差別化したいと考えたページの情報が埋もれてしまう．この二つの問題はトレードオフの関係になっており，どちらか一方に偏った手法では，ユーザが満足のいく支援を行うことは難しい．そこで本研究では，保存手法と履歴の解析の両面からアプローチし，履歴と保存情報を横断的に検索する手法を構築する．

2.2 横断的検索の事例

履歴やユーザの手で保存した情報を横断的に検索す

† 公立はこだて未来大学 大学院
Graduate School of Systems Information Science Future University Hakodate
†† 公立はこだて未来大学
Future University Hakodate

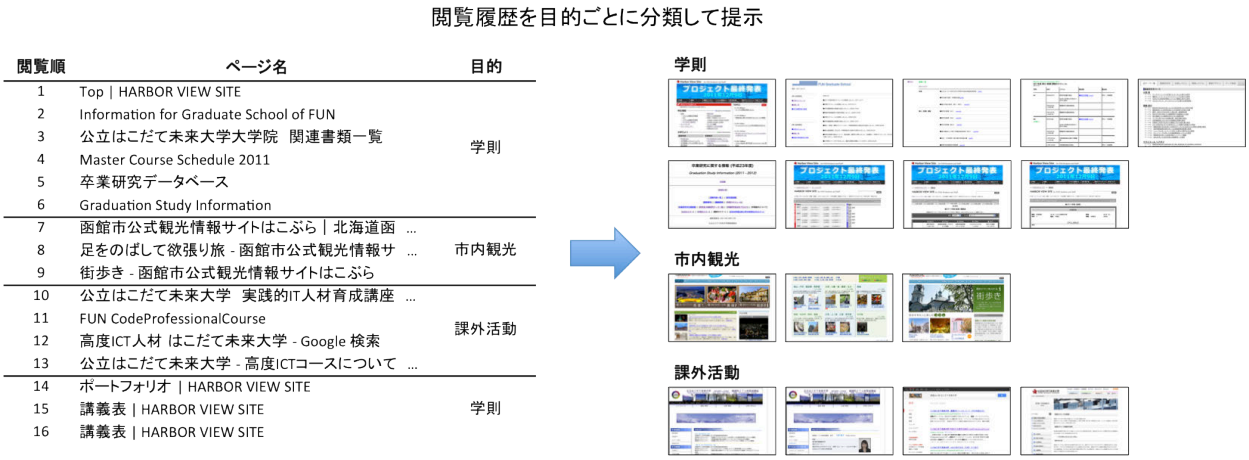

図 1 異なる閲覧目的と分類履歴の提示
Fig. 1 Different purposes of browsing and clustered browsing history

る支援の例と考察を述べ，本研究の位置づけを定める．

馬ら[4] は，ScrapBook[3] と履歴とブックマークを横断的に検索する仕組みを提案している．この横断検索を用いれば，それぞれに保存された情報を一括で探すことができる．そのため検索機能を使い分ける必要がなく，シームレスなリファインディングが実現できる．しかし，検索キーワードの記憶違いなどで，求めるページが検索結果に現れない場合がある．

森田ら[5] は，閲覧した履歴に加えてブックマークやプリントアウトなど，ページに対して行った行動も記録し，集中的に閲覧を行った期間（集中期間）を推定している．この集中期間をユーザに提示することで，閲覧行為で獲得した情報に関する記憶を想起させる支援を行なっている．集中期間では時系列を保持したまま，閲覧順通りにページが提示される．しかし時間が経過すると，「A ページの後に B ページを閲覧した」というような細かな閲覧順序は記憶に残っていない可能性が高い．それよりも，ページの記述内容などが類似したページ群を提示するほうが記憶想起に効果的である可能性が高い．

これらの事例の考察から，本研究では検索条件を満たすページに加えて，それらと似ているページも同時に提示することで，検索漏れを防ぎ，記憶想起を促す手法を構築した．

3. アノテーションと履歴分類

履歴とユーザが保存した情報を横断的に検索するために必要な支援の方針を検討した．次にその結果を述べる．

3.1 目的ごとに分類された履歴の提示

履歴の解析手法として，ページ同士の関連度を解析する手法を構築した．**図 1** 内の表は実際の閲覧履歴の一部を示している．この履歴では閲覧目的がこまめに入れ替わり，最初の「学則」を目的とした閲覧は後に再開されている．このような履歴では，単純に見たページを時系列順で羅列しても効果的に記憶を想起できないと考えられる．

そこで，ページ内のテキストや閲覧時刻を解析し，関連度に基づいて分類された履歴を提示する試みを行った．関連度の高いページ同士は閲覧目的が同じである可能性が高いため，記憶想起に効果的であると期待できる．

3.2 マーカーによるアノテーション保存

保存手法としてマーカーを採用し，ページ内のテキストに直接マーカーを引きブックマークの代わりとした．マーカーには，ページのコンテンツに直接マーカーを付加することで，ページ内のどの部分に着目したかを明示できるという利点がある．また，マウスのドラッグ等，簡単な操作で実現できる点も利点である．

3.3 分類結果の提示に用いる情報

分類結果を提示する際は，ページのスクリーンショットを用いる．ユーザが閲覧していたページのスクリーンショット群を図 1 のように目的ごとに並べ，一覧で提示する．これは，ページのレイアウトや掲載画像などの視覚的な情報を提示することで，効果的に記憶を想起できるためである．また Web ページにマーカーが引かれている場合，スクリーンショットと併せてマーカーを引いたテキストも提示する．こうすることで，閲覧履歴とユーザが保存したページを横断的に探すこ

とが可能となる．

スクロールしなければ全てを見ることができないページは，全体のスクリーンショットを縮小して見た場合に視認性が下がる．そこで，最も大きな画像が提示されている位置で切り出しを行う．過去の調査[6]にて，大きな画像を含んだスクリーンショットが記憶想起に最も効果的であると結論づけたためである．

4. 支援手順

3 節で述べた方針を基に，実際に行う支援の手順を踏まえながら述べる．支援は「初回閲覧」,「履歴分類」,「リファインディング」という三つのフェーズから成る．

4.1 初回閲覧フェーズ

初回閲覧フェーズは，リファインディングを行わない閲覧のフェーズである．閲覧している Web ページに関する情報の自動的な保存と，ユーザの手によるマーキングを行う．

表示したページのテキストとスクリーンショットを保存し，検索のためのリソースとして活用できるようにした．ユーザが Web ページを訪問する度に，自動的にページのテキストと画面のスクリーンショットを保存する．保存したテキストは，履歴の検索や関連度算出のためのリソースとして利用し，スクリーンショットはユーザに提示することで記憶想起を促す．

また，ユーザが他の Web ページと差別化するためにアノテーション（マーカーによる注釈情報の付加）を導入した．3.2 で述べた利点から，アノテーションにはマーカーが適切である．ユーザがページ内に重要な記述を発見した際は，ブックマークの代わりにマーキングを行う．

4.2 履歴分類フェーズ

履歴分類フェーズは，3.1 で述べた提示手法を実現するためにページの関連度を評価し，分類を行うフェーズである．次の四つの項目で関連度を評価する．

(1) 頻出名詞の一致率
(2) URL ドメインの一致・不一致
(3) 閲覧時刻の近さ
(4) 遷移元と遷移先の関係

（1）では，一つのページ内に頻出する単語は，そのページを特徴づける単語であると仮定し，保存したテキスト毎に形態素解析を行い，頻出する名詞を重要語として抜き出す．この重要語が共通しているページ同士は，関連度が高いとみなす.（2）では，ページ URL のドメインが一致していれば関連度が高いとみなす．これは，同じブログ内の記事同士などは，同一目的で閲覧していた可能性が高いためである.（3）では，閲覧時刻が近いほど関連性が高いとみなす.（4）では，ページ内のリンクをクリックして次のページへ遷移した場合，それらは関連性が高いとみなす．

以上の項目を評価し，1 日の履歴を最短距離法によるクラスタリングを用いて関連性の高いページごとに分類した結果を記録する．つまり，日付ごとに複数のカテゴリと，そのカテゴリに所属するページが記録されることになる．

4.3 リファインディングフェーズ

リファインディングフェーズは，文字通りリファインディングを行うフェーズである．分類された履歴を，**図 2** のような 2 種類の手順によって提示し，リファインディングの支援を行う．各手順の説明を，利用シーンを交えて述べる．

自動アシスト 自動アシストは，Web ブラウザに表示しているページが，過去の閲覧実績とマッチした場合に分類を提示する手順である．これは，過去に何度もキーワードを変えながら検索した過程で閲覧したページをリファインディングする場合に効果的である．

例えば，複数の類似したキーワードで検索を行なっていた場合，求めるページが，どのキーワードを用いてたどり着いたのかわからなくなることが多くある．このような場合のリファインディングでは，目的のページが見つからなくとも，目的ページと同時期に見ていたページにたどり着く確率が高い．そこで，ユーザが過去に閲覧したことのあるページを表示した場合，過去に閲覧した日付を知らせる．ユーザが，目的ページと閲覧時期が近いと感じた場合，日付を選択することで，その日の閲覧履歴を分類した状態で提示する．ユーザはそこから目的ページを探し出す．

履歴検索 履歴検索は，ユーザが入力したキーワードで履歴検索を行うことで，分類を提示する手順である．これは，目的ページの検索キーワードが明確な場合や，自動アシストで閲覧実績のあるページにたどり着かない場合に有効である．

このような場合のリファインディングでは，保存した Web ページのテキストを対象に全文検索を行う．検索条件に該当するページがあった場合，そのページが所属している全ての分類先を提示する．

検索対象を履歴内のページに限定することで，汎用的な単語であっても，目的のページを発見する確率の向上が期待できる．また，未閲覧の Web ページが対象外になるため，検索エンジンを用いるよりも

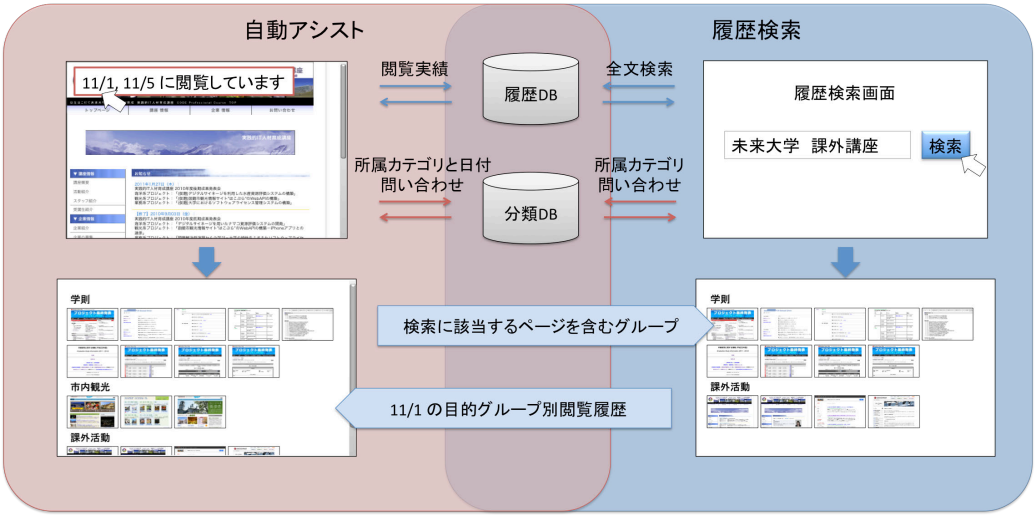

図 2　自動アシストと履歴検索の手順
Fig. 2　Processes of automatic assist and search from browsing history

ノイズとなるページが減少すると期待できる．

5. アドオンの実装

4 節で述べた手法を Web ブラウザ Firefox のアドオンとして実装する．以降，実装したアドオンを本システムと呼ぶ．本システムは，5 つの機能によって構成される．

(1)　ログ保存機能
(2)　マーキング機能
(3)　クラスタリング機能
(4)　リファインディング機能

（1）および（2）は 4.1 で述べた手法，(3）は 4.2 で述べた手法をそれぞれ実装したものである．(4）は 4.3 で述べた手法をひとつの機能として実装したものである．

（1）は Web ページの読み込みが完了した時とブラウザのタブが切り替えられた時に，表示しているページの全テキストとスクリーンショットを保存する．(2）は Wired-Marker[7] という既存のアドオンを導入することで実現している．テキストをマウスでドラッグし，右クリックすることによりメニューが現れ，マーカーの色を選択することができる．

（3）は履歴に対してクラスタリングを行う．4.2 で述べた項目を評価し，最短距離法でクラスタリングを行う．クラスタリングは 1 日分の履歴毎に行い，分類結果を記録する．

（4）では，4.3 に述べた手順で分類結果の提示を行う．ページを表示する度に，履歴を参照し，閲覧実績と所属カテゴリの有無を判定する．また，ユーザが履歴検索を行った場合，キーワードによる全文検索を行う．

6. 履歴分類結果の検証

4.2 で述べた履歴分類の方法を実際の履歴に適用し，分類の精度を次のような手順で評価した．まず，被験者 A，B，C の 3 名から履歴を 1 週間程度収集し，一定の条件を満たした履歴 1 日分を分類対象として抽出した．抽出する履歴は，1 日に二つ以上の目的で 50 件以上のページを閲覧していることを条件とした．

次に，抽出した履歴に対し最短距離法によるクラスタリングを行った．次に，実験者がそれぞれの履歴を手動で目的ごとに分類した．分類は，履歴内のページのコンテンツやページ遷移の過程をもとに判断した．最後に，クラスタリングによる分類結果と手動による分類結果を比較し，どの程度一致しているかで精度を測った．

6.1　クラスタリングによる分類

4.2 の四つの評価項目に基づき，各ページの関連度を距離として算出し，クラスタリングを行った．まず，一つのページともう一つ別のページを組とし，その 2 ページ間に対する 4 項目の評価点を求める．次に 2 ページ間の距離を，四つの評価点の相加平均で算出する．評価対象の組あわせををページ 1，ページ 2 とし，各評価項目をもとにした距離の算出方法を説明する．

頻出名詞の一致率　ページ内の頻出名詞をそれぞれ 20 個ずつ選出し，ページ 1 とページ 2 間の一致率で評価した．ページ間の評価点を x_1，一致した個数を n とすると

$$x_1 = n/20$$

となる．

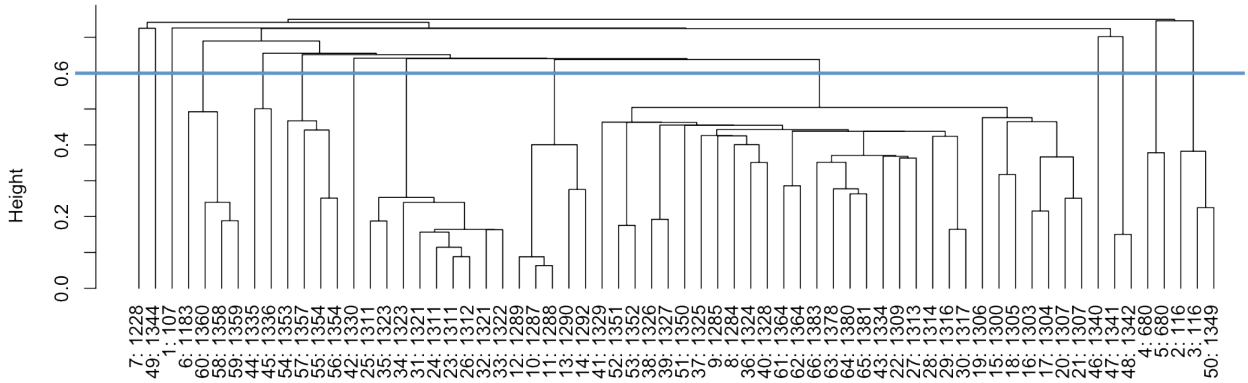

図 3 クラスタリング例: 被験者 A
Fig. 3 Example of clustering: user A

URL ドメインの一致・不一致 ページ 1 とページ 2 のドメインを抜き出し，等しいか評価した．ページ間の評価点を x_2，ページのドメインをそれぞれ dom_1，dom_2 とすると，

$$x_2 = \begin{cases} 0 & (dom_1 = dom_2) \\ 1 & (otherwise) \end{cases}$$

となる．

閲覧時刻の近さ ページ 1 とページ 2 の閲覧時刻の差を取り，最長時刻とどれほど離れているか評価した．ページ間の評価点を x_3 とし，ページ 1 とページ 2 の閲覧時刻をそれぞれ a，b，時刻の差の最大値を max とすると，

$$x_3 = |(a - b)|/max$$

となる．

遷移元と遷移先の関係 ページ 1 のリンクをクリックしページ 2 へ遷移した，またはその逆というように，ページ間で遷移関係があるかを評価した．ページ間の評価点を x_4 とすると，

$$x_4 = \begin{cases} 0 & (\text{遷移関係あり}) \\ 1 & (otherwise) \end{cases}$$

となる．

四つの項目の相加平均 x_1〜x_4 の相加平均で，四つの評価項目による距離を求めた．距離は 0 から 1 の間の値をとり，0 に近いほど関連性が高いとみなした．ページ間の距離を X とすると，

$$X = \left(\sum_{k=1}^{4} x_k \right) /4$$

となる．

次に各組のページ間距離を用い，最短距離法でクラスタリングを行った．例として，被験者 A の履歴に対しクラスタリングを行った結果を**図 3** に示す．図の末端の各ノードが，被験者 A が閲覧したページとなる．

最後に距離 0.6 を閾値とし，閾値以下で結合されているノード群をそれぞれグループとして分けた．

6.2 手動による分類

それぞれの履歴をページのコンテンツや閲覧過程をもとに，関連性が高いとみなしたページ毎に分類を行った．関連性は，同じサイト内のページや似通ったキーワードで検索しながら閲覧したページ群は高いと判断した．分類結果は被験者に確認し，誤りがあれば修正を行った．

6.3 分類結果の評価方法

6.2 の分類結果を正解とし，6.1 の分類結果の正解率を用いて精度を評価した．それぞれの分類方法で形成されるグループの数が異なるため，次のような方法で正解率を算出した．

6.1 にて作成したグループを自動グループ，6.2 にて作成したグループを手動グループとする．一つの手動グループに属するページ群と，それらが最も多く属している自動グループへの所属件数を正解数として求める．**図 4** の場合，G_1 のうち二つのページが g_2 に属しているため，正解数は 2 となる．全ての手動グループに対して同様に調べ，自動グループの正解数の合計と閲覧件数（母数）の比率で正解率を算出した．

また，一つの自動グループに，手動グループの所属を問わず多くのページが分類されてしまう結果となったため，それを考慮した正解率も算出した．結果を**表 1** に示す．

6.4 結果と考察

1 日の閲覧件数が 70 件に満たない被験者 A と B の正解率は比較的高く，6 割を超えた．手動で作成した

グループの数と，自動で作成したグループ数にも大きな差はなかった．

一方，閲覧件数の多い被験者 B の正解率は 5 割以下と低いものになった．手動と自動で作成したグループ数にも大きな違いがあった．この原因は二つあると考察した．一つ目は，被験者 B の閲覧目的が JavaScript や jQuery などの実装方法を調べるというものが大半であり，目的の差別化が困難であったためである．二つ目は，同じ目的で閲覧を行なっている間に多数のサイトのページを訪れたため，頻出名詞と URL ドメインによる距離の評価が低く出たためである．

また自動グループの中で，ページの所属件数が最も多いグループ $g_{\max}$ の所属件数は，各被験者とも 20 件以上であった．$g_{\max}$ 内のページが所属している自動グループ数は 5 以上であり，多数の閲覧目的が混在したグループとなった．そこで $g_{\max}$ に所属しているページを不正解として，先ほどの正解率を計算し直したところ，どの被験者も正解率が 4 割程度に下がった．この原因は，Google の検索結果ページなどの汎用性の高いページのドメインを距離の評価に入れたためであると考察できる．実際，各被験者の $g_{\max}$ に所属するページは全て，Google 検索を起点として閲覧したページであった．

6.5 精度向上のための対策

以上から分類精度を向上させるために，閲覧時刻や遷移関係の評価を頻出名詞と URL ドメインより重くする，Google 検索結果などドメインが同じでも，目的が多様であるものはドメインの評価をしない，という 2 点の対策を行う．また今回は考慮しなかったが，マーカーを引いたテキストは閲覧目的を推定に有効であると考えられる．そこでマーカーを引いたテキストから抽出した名詞も，評価に加えていきたい．

また，1 日の閲覧件数が多い履歴やページの所属件数が多いグループは，提示するページの数が多くなりすぎる．その結果リファインディングの達成に時間がかる，または達成できない可能性がある．そのため，PC をシャットダウンした時刻や，ブラウザを長時間使用していない時間などで分類するページの母数を減らす必要がある．加えて所属件数が多いグループは閾値を下げ，より細かいグループに分割するなどの対策も必要である．

表 1 分類精度の比較
Table 1 Comparing accuracy of classification

	被験者 A	被験者 B	被験者 C
閲覧件数	66	152	67
自動グループ数	14	45	13
手動グループ数	12	16	13
正解数	47	74	44
正解率	0.71	0.49	0.66
所属件数が最も多い自動グループ $g_{\max}$	g_6	g_3	g_3
$g_{\max}$ の所属件数	30	36	23
$g_{\max}$ 内の手動グループ数	8	11	5
$g_{\max}$ の正解所属件数	19	6	17
差し引いた正解数	28	68	27
差し引いた正解率	0.42	0.45	0.40

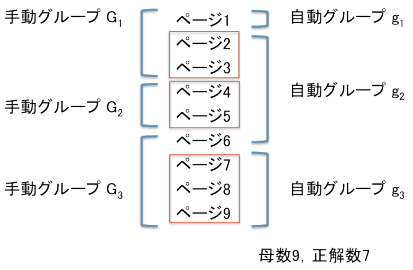

図 4 正解数の計算例
Fig. 4 How to count the number of correct imtems

7. ま と め

本論文では，履歴分類結果の提示とマーカーという二つの特徴を持つリファインディング支援手法を提案し，分類結果の精度を評価した．分類の精度は低くなってしまったが，原因が推察でき改善の見込みもあるため，対策を施した後再び評価を行う．今後は，本手法をリファインディングの場面に適用し，達成率や時間について評価を行いたい．

参 考 文 献

1) Robert, C. et al: Refinding Is Not Finding Again. Technical report, TR-05-10, Computer Science, Virginia Tech (2005).
2) W. Jones. et al: Keeping found things found on the web, Proceedings of the tenth international conference on Information and knowledge management, pp.119–126 (2001).
3) 五味渕大賀: ScrapBook, http://amb.vis.ne.jp/mozilla/scrapbook, (2011).
4) 馬芙榕ら: Web ブラウザの ScrapBook・履歴・ブックマークを横断的に検索可能なツールの開発と評価, ET2006-122 (2007).
5) 森田哲之ら: Memory-Retriever：体験獲得情報を想起させる行動検索手法, 情報処理学会論文誌 48(3), 1197-1208 (2007).
6) 飯野亜耶ら: アノテーションを用いたリファインディング支援手法の提案と検証, WI2-2011-36 (2011).
7) Wired-Marker, http://www.wired-marker.org/, (2011).