# Webにおけるリンク選択行動からユーザの時系列の興味空間を作成するシステム

**村上 晴美[†] 平田 高志[††]**

**2003年9月**

**JCSS-TR-47**

[†] 大阪市立大学 大学院 創造都市研究科
都市情報学専攻 情報メディア環境研究分野
兼 学術情報総合センター
〒 558-8585 大阪市住吉区杉本 3-3-138
Phone/Fax: 06-6605-3375
harumi@media.osaka-cu.ac.jp

[††] 防衛庁 陸上自衛隊
takash-h@h5.dion.ne.jp



日本認知科学会
事務局
名古屋大学大学院 人間情報学研究科 認知情報論講座内
Phone: 052-789-4891
FAX: 052-789-4752
jcss@jcss.gr.jp

# Webにおけるリンク選択行動からユーザの時系列の興味空間を作成するシステム

村上 晴美　　平田 高志

**Abstract** ユーザの興味の理解を支援するために, ユーザの Web におけるリンク選択時に, リンクを含む行のテキストからキーワードを抽出して, リンク先 URL と日時と共に新しい履歴に蓄積し, その履歴から, キーワードと, Web ページを表す Web ページアイコンを, 2 次元空間上に配置して時系列に提示する手法を提案する. キーワードと Web ページアイコンを 2 次元空間上に配置した空間を興味空間と呼び, 時系列の興味空間を通して過去に見た Web ページにアクセスするシステムを興味空間ブラウザと呼ぶ. 新しい履歴を作成する Web ブラウザと, 興味空間ブラウザを試作した. 日常生活における利用のシミュレーションと, 新聞記事の Web サイトを対象とした実験の結果, (1) ユーザの選択したリンクを含む行のテキストからユーザの興味を表す日本語のキーワードを抽出できること, (2) 興味空間ブラウザがユーザの Web 閲覧時の興味空間を表していること, (3) 興味空間ブラウザが自己の興味の理解, 過去の想起, Web ページの整理に役立つ可能性があること, がわかった.

## 1　はじめに

「自分の好きなこと, 興味のあることをやりなさい.」こう言われたことがある人は多いだろう. 一般に, 成功したり, 幸福になるためには, 自己を知り, 好きなことをやることが重要であると言われている. 本研究の長期的な目的は自己の理解である. これまでに心理学や精神医学等の分野において自己を理解するための数多くの方法論が提案, 実証されてきた. しかし, コンピュータを用いた方法は, まだあまり確立されているとは言えない. 我々は, コンピュータを用いて, 人間が自己を理解するための方法論の確立を目指している.

本研究は, 上記研究の一つとして, 自己のさまざまな行動履歴をコンピュータ上に蓄積することにより, 自己の興味の理解を支援するシステムの開発を目的としている. 行動履歴の情報源として, 本研究では, 我々の日常生活に浸透してきている World Wide Web(以下 Web) におけるブラウジング行動をとりあげる.

Web ブラウジング行動をコンピュータ上に蓄積するものとして, Microsoft Internet Explorer(以下 IE)[1] や Netscape Navigator[2] 等の既存の Web ブラウザでは, ユーザが過去に見た Web ページの URL, タイトル, アクセスした日時等を記録する「履歴」がある. 既存の履歴を利用して, ユーザの興味の理解を支援するシステムの開発が可能かもしれない. しかし, 既存の履歴をそのまま利用するには以下の二つの問題点がある. (1) 当日以前に見た Web ページの履歴に関しては, Web サイトの URL リストが提示されるだけであり, どのような Web ページを見たかを理解することが難しい. (2) タイトルは正確に内容を反映していない場合がある. たとえば, Yahoo!ニュースの「円急騰１２１円台、日銀が介入」という記事の Web ページの場合, タイトルは「Yahoo!ニュース - 経済総合 - 読売新聞社」であるが, このタイトルからどのような内容の記事を見たかはほとんどわからない. この場合, ユーザが選択したリンクのテキストである「円急騰１２１円台、日銀が介入」の方が参考になるであろう.

本研究では, Web ブラウジング行動の中でもリンク選択行動に焦点をあて, (1) ユーザの選択したリンク周辺に含まれるテキストからユーザの興味を表すキーワードを抽出できる, (2) 抽出したキーワードを時系列に提示することにより, ユーザの興味の理解を支援するシステムを開発できる, という仮説をたてる.

上記の仮説に基づき, ユーザの興味の理解を支援するために, ユーザのリンク選択時に, リンクを含む行のテキストからキーワードを抽出して, リンク先 URL と日時と共に新しい履歴に蓄積し, その履歴から, キーワードと, Web ページを表す Web ページアイコンを, 2 次元空間上に配置して時系列に提示する手法を提案する.

キーワードと Web ページアイコンを 2 次元空間

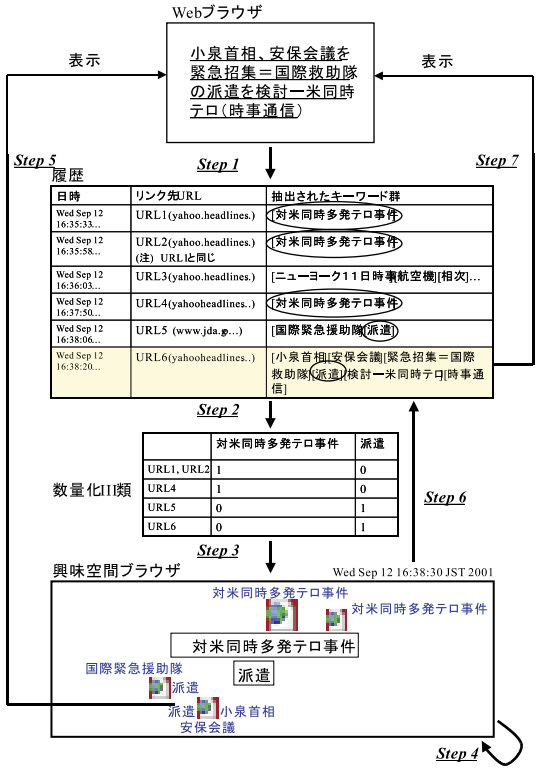


図 1: 提案手法の概要

上に配置した空間を興味空間と呼び, 時系列の興味空間を通して過去に見た Web ページにアクセスするシステムを興味空間ブラウザと呼ぶ.

我々は, 新しい履歴を作成する Web ブラウザと, 興味空間ブラウザを試作し, 実験的な評価を行った. 実験は主として新聞記事のサイトを対象とした部分的なものであるが, ある程度の有用性が示されたので報告する.

以下では, 2 節で提案手法, 3 節で実験結果を述べ, 4 節で関連研究と比較して議論し, 5 節で今後の課題を述べる.

# 2 提案手法

## 2.1 概要

提案する手法は, 以下の Step 1 から Step 7 までの 7 段階で構成される. 図 1 に提案手法の概要を示す.

- Step 1: ユーザが Web ブラウザでリンクを選択すると,「日時, リンク先 URL, リンクを含む行から抽出されたキーワード群」を一組のデータとして新しい履歴に保存する.
- Step 2: ユーザが興味空間ブラウザで「興味空間の生成」を指示すると,「履歴において 2 度以上出現するキーワード」(興味語と呼ぶ) をカテゴリデータとし, 興味語と共起する URL をサンプルデータとして, 数量化 III 類の計算を行う.
- Step 3: ユーザが興味空間ブラウザで「興味空間の表示」を指示すると, Step 2 で計算された内容に基づき, 興味空間の表示を行う.
- Step 4: ユーザが興味空間ブラウザで, 表示対象の日時を変更する「日時変更スライドバー」と, 表示対象の期間を変更する「期間変更スライドバー」を操作すると, 日時情報に基づき興味空間の表示を変化させる.
- Step 5: ユーザが興味空間ブラウザで Web ページを表す「Web ページアイコン」をダブルクリックすると, インターネットに接続して Web ブラウザに該当の URL の Web ページを表示する.
- Step 6: ユーザが興味空間ブラウザで「履歴のオープン」を指示すると, 現在の日時を中心として履歴を表示する.
- Step 7: ユーザが履歴上で履歴データをダブルクリックすると, インターネットに接続して Web ブラウザに該当の URL の Web ページを表示する.

以下では, 提案手法の詳細を述べる.

## 2.2 リンク選択時の履歴蓄積

ユーザが Web ブラウザでリンクを選択する際に, リンクを含む行のテキスト[1] からキーワードを抽出し, 日時と, リンク先の URL と一緒に一組のデータとして履歴に蓄積する.

図 2 に既存の Web ブラウザの履歴と本研究の履歴の違いの例を示す. 既存の Web ブラウザでは, 日時と, Web ページの URL と, タイトル要素から抽出されるテキストが組として蓄積されるが, 本研究の履歴では, 日時と, リンク先 URL と, リンクを含む行のテキストから抽出されたキーワード群が

[1] 正確には, アンカー要素を含む, Java2 SDK の HTML-document クラスの中の getRawStart クラスで得られる位置を開始点, getRawEnd クラスで得られる位置を終了点とする行のテキスト.

図 2: 既存のブラウザと本研究の履歴の違い

組として蓄積される. たとえば, 1 節の例のように, ユーザが「Yahoo!ニュース- トップ[2]」から, リンク「円急騰１２１円台、日銀が介入 (読売新聞) [3]」を選択したとする. この場合, リンク先の Web ページに関して, 既存の Web ブラウザの履歴では, タイトルである「Yahoo! ニュース - 経済総合 - 読売新聞社」が保存されるが, 本研究の履歴では, リンク元の Web ページのリンクを含む行から抽出されたキーワードである「円急騰１２１円台」「日銀」「介入」「読売新聞」が保存される.

日本語の文字列からキーワードを抽出する手法として, 形態素解析を用いて名詞を中心に抽出する手法と, ひらがな以外を中心に抽出する字種による手法があり, 本研究では試験的に後者を採用している. 実験して確かめたわけではないが, 形態素解析を用いるよりも, 実装が容易, 日本語でも英語でも動作可能[4], 処理速度が速い, 品詞不明の未定義語の抽出が容易, キーワードに余分な語や記号が付着するが, それがかえってユーザにブラウジング時の文脈を想起させ, 興味の理解を助ける, と考えたからである.

本研究では, (1) ひらがなと特殊文字以外の文字が 2 つ以上続いた文字列をキーワード候補として抽出し, (2) 不要語判定ヒューリスティックと不要語リスト法を用いてキーワード候補から不要語を削除し, キーワードを確定する, という単純な手法を用いている.

不要語判定ヒューリスティックは, 文字列から不要語を判定するものである. 本研究では, (a) 数字不要語判定, (b) 日付不要語判定, (c) その他不要語判定, の 3 種類を実装している. たとえば, (a) では「10」「２００」のような文字列を数字の不要語, (b) では数字と「年」「月」「日」等で構成される文字列を日付の不要語, (c) では「具体的」というように最後に「的」のついた文字列をその他の不要語と判定している. 不要語ヒューリスティックの詳細は [3] を参照されたい.

不要語リスト法は, 不要語リストに登録された語をキーワードとしないというものである. 2.4 節で後述するとおり, ユーザは興味空間ブラウザに表示されるキーワードの中から不要と思うものを選択して不要語リストに登録できる.

図 1 の Step 1 の例では, ユーザが Web ブラウザ上で「小泉首相、安保会議を緊急招集＝国際援助隊の派遣を検討ー米同時テロ (時事通信)」というリンクを選択すると, 履歴に, 抽出されたキーワードである「小泉首相」「安保会議」「緊急招集＝国際援助隊」「派遣」「検討ー米同時テロ」「時事通信」が保存される. もし仮にユーザが「時事通信」を不要語リストに登録していた場合は, 履歴に「時事通信」は登録されない.

## 2.3 興味空間の生成と表示

興味空間の表示には, 「興味語のみ」「興味語と Web ページアイコン」の 2 種類の選択が可能である. これは, 最初は「興味語のみ」を表示して興味空間を探訪し, ある程度日時を定めたら, 「興味語と Web ページアイコン」に切り替えて表示できるようにするためである.

ユーザが興味空間ブラウザで「興味語のみ」または「興味語と Web ページアイコン」を選択してから, 「興味空間の生成」を指示すると, 履歴において 2 度以上出現するキーワード (興味語) をカテゴリデータとし, 興味語と共起する URL をサンプルデータとして, 数量化 III 類の計算を行う. 数量化 III 類は, 質的データに対する主成分分析の一種で, 要素が 2 値データであるときや頻度である場合にそれを数量化し, サンプルやカテゴリを分類したり, 特性を調べるために利用する手法である. 数量化 III 類を用いて散布図を描くことにより, サンプルデータ同士, カテゴリデータ同士, サンプルデータとカテゴリデータを, 類似するものが近くにくるように配置できる.

図 1 の Step 2 の例では,「対米同時多発テロ事件」「派遣」を興味語及びカテゴリデータとし, それらの興味語を持つ URL1, 4, 5, 6 をサンプルデータとする. この際, URL の文字列が同じ場合は同じ URL

[2] http://headlines.yahoo.co.jp/hl
[3] http://www.headlines.yahoo.co.jp/hl?a=2002062400000105-yom-bus_all
[4] ただし, 英語の場合, 語単位で切り出されるだけである.

として取扱う. たとえば, URL1 と URL2 は URL1 として扱い, URL1 の頻度は 2 となる. URL1 では「対米同時多発テロ事件」を持つので「1」,「派遣」を持たないので「0」を数量化 III 類のマトリックスに入れる.

ユーザが「興味空間の表示」を指示すると興味空間が表示される. 最初に表示される興味空間は履歴上の最新の日時であり, 期間は過去 30 日以内である. すなわち, 最新の日時から 30 日以内に履歴中に出現した興味語と URL が表示対象となる. 数量化 III 類の計算結果の 1 軸を X 軸, 2 軸を Y 軸とする. 数量化 III 類の 1 軸と 2 軸の計算方法については [4] 等を参照されたい.

興味語は四角で囲まれたキーワードとして表示される. Web ページアイコンは地球マークのアイコンとして表示される. Web ページアイコンは, 履歴中の URL の頻度に応じて, 大 (3 回以上), 中 (2 回), 小 (1 回) の 3 種類に変化する. また, 表示対象とする日付から過去に遡る日数に応じて, 青 (3 日以内), 濃いグレー (7 日以内), 薄いグレー (8 日以前) の 3 種類に変化する.

ユーザが興味空間ブラウザでカーソルを Web ページアイコンにあわせると, 抽出されたキーワードが「Web ページキーワード」として, Web ページアイコンの周囲に表示されるので, 実際にどのような Web ページを見たのか概要の把握が容易となる.

図 1 の Step 3 の例では, ユーザが興味空間ブラウザで「興味空間の表示」を指示すると,「対米同時多発テロ事件」と「派遣」が四角に囲まれたキーワードとして表示され, これらの興味語を持つ 4 つの異なる URL が青色の Web ページアイコンとして表示される. 類似する興味語と Web ページアイコンが近くに表示されるため, 興味語を参考にしながら Web ページを探すことができる. 選択頻度に応じてアイコンのサイズが変わるため (たとえば, URL1 の Web ページアイコンは他よりも大きく表示される), よく見た Web ページがどれかわかりやすくなる.

図 3 に実験 1(後述) における第一著者の 2001 年 9 月 12 日 16:38:30 時点での興味空間ブラウザの画面例を示す. 興味語として「対米同時多発テロ事件」「日本図書館情報学会」「久保純子アナ」「産休」等が表示されている. Web ページキーワードに記をつけた Web ページアイコンは図 1 の URL1, 2 である. この画面を見て第一著者は, この時期に起きた対米同時多発テロ事件に関連するニュースをよく見ていたことと, 10 月に開催される日本図書館

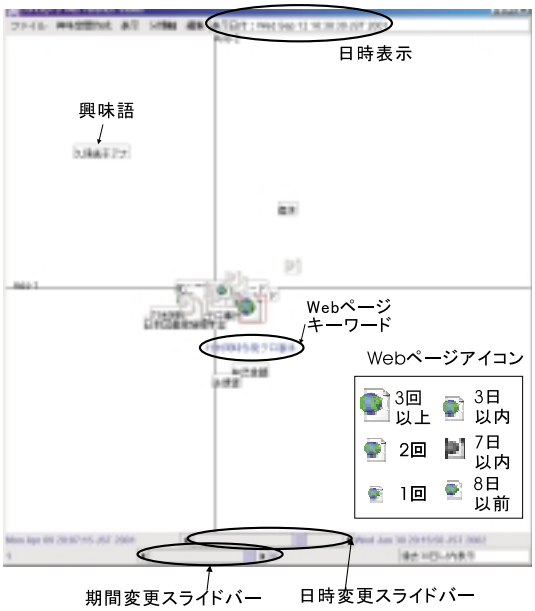


図 3: 興味空間ブラウザの画面例

情報学会研究大会の原稿を作成していたことを思い出すとともに, 女性タレントの芸能ニュースに興味を持って見ていたことにあらためて気づいた.

## 2.4 興味空間の変更

ユーザが「日時変更スライドバー」「期間変更スライドバー」を操作すると, 日時情報に基づき, 興味語と Web ページアイコンの表示を変更する.

日時変更スライドバーを操作することにより, システムの使用初めから現在までの間で表示対象の日時を変更できる. 期間変更スライドバーでは表示対象の日付から過去に遡る期間を 1 日から 30 日の間で変更できる. 両スライドバーを操作することにより興味空間が変化する. 日時に応じて Web ページアイコンの色が変わるため, 興味空間の理解が容易になる. IE の履歴では, 2 週間前, 3 週間前等の一定の単位で Web サイトのリストが提示されるが, 本研究では, より柔軟な単位で日時と期間の設定ができる.

図 1 の Step 4 で, 日時変更スライドバーを操作して表示対象の日時を少しずつ過去に戻すと, Step 3 の時点で薄いグレーであった Web ページアイコンが, 濃いグレー, 青色へと変化してゆく.

なお, ユーザは興味語を選択して「不要語リストに登録」を指示することにより, 2.2 節で述べた不要後登録が行える. また, Web ページアイコンを選択して「不要 URL に登録」を指示することにより, 指定した Web ページアイコンを今後非表示に

できる.

### 2.5 Web ページの表示

ユーザが興味空間ブラウザから Web ページを表示する方法は二種類ある.

一つは, 図 1 の Step 5 で, ユーザが興味空間ブラウザで「小泉首相」「安保会議」「派遣」などの Web ページキーワードを参考にして, 該当の Web ページアイコン (URL6) をダブルクリックすると, Web ブラウザに該当する Web ページが表示される.

もう一つは, 図 1 の Step 6, 7 で, ユーザが「履歴のオープン」を指示すると, 興味空間ブラウザの現在の日時を中心とする履歴が表示されるので, ユーザは履歴 (URL6) を選択して, Web ブラウザに該当する Web ページを表示できる.

## 3 実験

試作した Web ブラウザは, Java の実装上の問題により, CGI やフレーム等を利用したページを中心として, 全く表示できなかったり, 表示がずれたり, 文字化けをおこしたり, 非常に時間がかかる等, 表示に関わる問題があり, 既存の Web ブラウザのような自然なブラウジングが困難であった. そのため, 一般のユーザに自然な状態で利用してもらう実験の実施は難しいと判断し, 以下の 2 つの実験を計画した.

- 実験 1: システムを熟知している著者らが被験者となり, 日常生活における利用のシミュレーションを行い, 提案手法の問題点と有効性を検討する.
- 実験 2: 一般のユーザを被験者とし, 彼らがシステムを利用できる Web サイトに限定して, システムの有用性を検討する.

### 3.1 実験 1

#### 3.1.1 方法

被験者は第一著者と第二著者である. 以下それぞれ被験者 A, B と呼ぶ. 被験者 A は 37 歳女性であり, 情報関連の大阪の大学教員である. 被験者 B は 31 歳男性であり, 防衛庁職員兼情報科学研究科の大学院生である[5].

以下の例外を除いて, 原則として, 開発した Web ブラウザを用いて日常生活で行うブラウジングのシミュレーションを行う.

- 急いでいる場合には IE を利用してよい.
- あるページが表示できなかった場合, 該当ページを IE で表示し, その後のリンク選択には IE を利用してよい.
- 表示に問題があることが経験上わかったサイトやページの表示には IE を利用してよい.
- 履歴を他人 (A の場合は B, B の場合は A) に見られると恥ずかしいと思うページの表示には IE を利用してよい.

興味空間の表示は利用日に毎回行うこととする.

実験期間は, 被験者 A が 2001 年 4 月 9 日から 2001 年 12 月 27 日, 被験者 B が 2000 年 5 月 26 日から 2000 年 11 月 13 日までである. 実験は, 原則として, 平日の仕事時間中に研究室で行った.

#### 3.1.2 結果と考察

**概要** 被験者 A に関しては, 選択したリンク数が 1,485, 内訳は, Yahoo! Japan[6] 611 (41%), 大阪市立大学[7] 94 (6%), 大阪大学[8] 27 (2%), その他 753 (51%) であった. Yahoo! Japan と近隣の大学の Web サイトをよく見ていることがわかる. 抽出されたキーワード数は 2,488 であった. 一つのリンクに対して平均 1.68 のキーワードが抽出された. 不要語リストに登録された語は 117 であった.

被験者 B に関しては, リンク数が 808, 内訳は, 朝日新聞[9] 549(68%), Yahoo! Japan[10] 72 (9%), goo[11] 28(3%), その他 159(20%) であった. 新聞記事の Web サイトと Yahoo! Japan をよく見ていることがわかる. キーワード数が 2,112, 平均 2.61 のキーワードが抽出され, 不要語リストに登録された語は 748 であった. ただし, 被験者 B の不要語リストは, 今回の研究とは異なる目的で登録された[12]た

[5] 実験当時の年齢と職業である. 以下同様.
[6] 第 3 レベルで判定, 以下同じ, yahoo.co.jp
[7] osaka-cu.ac.jp
[8] osaka-u.ac.jp
[9] www.asahi.com
[10] yahoo.co.jp
[11] goo.ne.jp
[12] ブラウザ上で Web ページをなぞることによりキーワードを抽出する機能があり, その際の不要語リストを兼ねている. 748 個ともこの目的で登録された.

め, 748 個の中でどれが興味空間に影響を与えたかは不明である.

**キーワード抽出**　不要語リストを利用せずに, 抽出した頻度 2 以上のキーワードが語として妥当かどうか判定したところ, 被験者 A で 91%, B で 93%が妥当であった. キーワード抽出手法自体には大きな問題はないと考える.

キーワード抽出に関する問題点は以下のとおり分類できる. まず日本語に関しては, 再現性の問題として, (a) ひらがなを含むキーワードを抽出できない (例:「つくば市」「神の国」「のぞみ」), (b) 一文字のキーワードを抽出できない (例:「産」「森」) があり, 適合性の問題として, (c) 不要な記号が付着する (例:「結果→」), (d) 記号列や記号列の一部が抽出される (例:URL や ISBN), (e) 記号等をはさんで単語が長くなりやすい, (f) ひらがなで終わる品詞のひらがな部分が切れる (例:「見直」), (g) 理由なく途中で切れる, があげられる. 英語の場合は, (h) 全ての単語が抽出される, (i) 不要な記号が付着する (例:「market,」), が問題である.

上記の中で, (a), (b), (h), (i) は字種による手法のみでは解決できず, 形態素解析を併用する等の検討が必要である. (c), (d), (e), (g) は不要な文字列を除去するようアルゴリズムを改善することにより解決できると考える. (f) は語としては妥当でないが興味語としては利用可能であるためこのままでも大きな問題ではない.

上記より, 日本語のキーワード抽出に関しては, 現在の手法がある程度有効であると考える.

**興味語抽出**　キーワードの高頻度語 10 語を観察し, 提案手法によりどの程度興味語抽出が可能か検討する.

不要語リストを利用しない場合, 被験者 A の高頻度語 10 語は, 「メッセージ」(258)[13], 「最初」(130), 「最新」(128), 「一覧」(89), 「...」(49), 「ショッピング」(42), 「http」(39), 「メッセージリスト」(39), 「エンターテインメント」(32), 「ホーム」(32), 「旅行 (32)」であった. B では, 「asahi」(229), 「ニュース」(147), 「Yahoo」(76), 「一覧」(61), 「テスト」(59), 「goo」(56), 「ページ」(33), 「mainichi」(32), 「毎日テストデータ」(24), 「スポーツ」(19) どちらもあまり興味を表していない.

不要語リストを利用すると, A では, 「日本」(22), 「経済」(12), 「地域情報」(12), 「地方」(12), 「コンピュータ」(11), 「大学」(11), 「旅行」(10), 「Java」(10), 「教育」(10), 「インターネット」(9), 「企業」(9), 「研究」(9), 「スポーツ」(9), 「ビジネス」(9) となり, これらは, 情報に関する教育, 研究を本業とする大学教員としての興味を示していると考える. 同様に B では, 「中国」(13), 「北朝鮮」(11), 「森首相」(11), 「ロシア」(9), 「アラファト議長」(6), 「イスラエル」(6), 「クリントン大統領」(6), 「プーチン大統領」(6), 「首相」(5), 「フィジー」(5), 「防衛庁長官」(5) となり, これらは, 国際関係を中心とする防衛庁職員としての興味をよく表していると考える.

このように, 約 6-8 か月程度の利用では, 高頻度語に不要語が多くなるが, 不要語リストを用いることにより除去できることがわかった. 不要語がどの程度の割合になるかはっきりとはわからないが, 被験者 A, B の「不要語リストに登録された語数/抽出された全キーワード数」より, 概ね 5%から 30%の間になることが予想される.

興味を表していない語は, 主として, Yahoo! Japan の掲示板にある表現 (例:「メッセージ」, 「最初」, 「最後」) やトップページにある表現 (例:「ショッピング」, 「旅行」, 「音楽」) 等の, ユーザがよく見る Web ページのリンク周辺の表現である. また, ここの例には現れていないが, 現状のアルゴリズムではすべての英単語を抽出してしまうため, 英文によく含まれる英単語 (例:「The」, 「to」, 「of」) の頻度が高くなることが観察された.

上記より, 高頻度語の観察を通して, ユーザによる不要語の除去を前提として, ユーザが選択したリンクを含む行から, ユーザの興味を表す日本語のキーワードを抽出できることが示唆された.

**興味空間**　実験の結果, 提案手法を用いて興味空間が生成, 表示できることがわかったが, 以下の問題点が観察された. (a) 履歴の量が少ないと頻度 2 以上のキーワードがないため, 興味空間が表示されない, (b) 履歴の量が多いと興味空間の生成に時間がかかる, (c) 直観的に, 興味語と興味語を含む Web ページアイコンは比較的近くに表示されることが多いが, 類似する興味語同士, Web ページアイコン同士があまり近くに表示されない, (d) X 軸と Y 軸の意味を判別できない, (e) 頻度が 1 の興味語や, 興味語を持たない Web ページアイコンは表示されない, (f) 興味語と Web ページアイコンが中心に固まり, 見づらく操作しにくい.

これらは, Web ブラウジング行動から蓄積した新

[13] () 内は頻度. 以下同様

しい履歴と数量化 III 類の組合せによる提案手法に問題を含むことを示している. (a) から (e) に関しては, 履歴の作成方法を改善する, 数量化 III 類以外の空間配置手法を試す, 等の検討が必要である. (b) に関しては, 部分的に計算する等の方法による改善も可能であろう. (f) に関しては, オブジェクトを適当にちらばらせる等の方法による改善が考えられる.

実験 1 では, 被験者 B のリンクあたりのキーワード数が A よりも多いことや, B の高頻度語に固有名詞を含むキーワードが A よりも多いことより, ユーザが一般的な Web サイトよりも新聞記事の Web サイトをよく見る場合に, より良い興味語を抽出できる可能性が示唆された. さらに, 特定の新聞記事の Web サイトでは表示上の問題が少ないことが確認されたため, 一般のユーザによる実験が可能であると判断した. そこで, 実験 2 では新聞記事の Web サイトを対象として検討する.

## 3.2 実験 2

### 3.2.1 方法

**被験者** 被験者は 32, 37, 37, 24, 20, 33, 57, 22 歳の女性 8 名であり, 順番に A, B, ... , H と呼ぶ. A と C と F は主婦, B はシステムエンジニア, D は秘書, E と H は学生, G は事務員であり, 全員 Web 利用経験は 2, 3 年程度である.

**ブラウジング** 毎日インタラクティブ[14]から 2000 年 6 月から 8 月の 3 か月の HTML ファイル (1 ファイル 1 記事) をダウンロードしておく. Web ブラウザを用いて, 見たい記事を自分のペースで 1 日分につき 1 つ以上を見るよう指示する.

毎日インタラクティブの新聞記事では, リンクを含む行はアンカー要素から抽出されるテキスト (<a href=”URL”> と </a> で囲まれた文字列) と一致するため, ここではリンクテキストと呼ぶ. 新聞記事であるため, リンクテキストは記事の見出しと同じである. たとえば, 「宮内庁：参与に経団連名誉会長の平岩外四氏」というものである. リンク先の Web ページは記事内容の Web ページであり, たとえば, タイトルは「毎日インタラクティブ・記事全文」である.

1 節の Yahoo!ニュースの例と同じように, 記事の Web ページのタイトルには記事内容に関する情報がない. この場合では, キーワードを抽出する情報源としてリンクテキストとタイトルのどちらが良いかは自明である. 以下では, リンクテキスト (見出しに相当) と Web ページ (見出しと記事本文を含む) のどちらが情報源として良いか検討する.

**興味語抽出手法の評価** 興味語抽出手法に関しては, 履歴を回収した後に, 以下の判定を行う.

(1) リンクテキストと Web ページとどちらから抽出したキーワード群が直観的に興味語としてふさわしいかを調べるために, 被験者にリンクテキストと Web ページからのすべての抽出キーワード群[15]を見せ, どちらが直観的に興味語としてふさわしいか選択させる.

(2) 本研究の提案である「リンクを含む行から抽出される頻度 2 以上のキーワード」が興味語としてふさわしいかどうかを調べるために, リンクテキストから抽出された頻度 2 以上のキーワードの一つ一つに対して, 被験者に「Web 閲覧時の興味を表していると思うか」を 5 段階 (5: 非常に思う, 4: やや思う, 3: どちらともいえない, 2: あまり思わない, 1: 全く思わない) で評価させ, 平均をリンクテキストの興味度 X とする. 比較のために, Web ページから抽出されるキーワードの中, 最高頻度語から, リンクテキストから抽出される頻度 2 以上のキーワード数と同じ順位の頻度までのキーワードを対象として評価させ, 平均を Web ページの興味度 X とする[16].

(3) リンクテキストと Web ページから抽出した興味語の高頻度語の性質を調べるために, 高頻度語 10 語に関して, (2) と同様に被験者の評価を行い, 興味度 Y とする.

(4) 興味語の中には非常に良い興味語とやや良い興味語があると考えられる. 非常に良い興味語を抽出できるのは, リンクテキストと Web ページのどちらかを調べるために, 被験者に, リンクテキストと Web ページの興味度 X のキーワード群の中から最も自分が興味を持った 10 語を選択するよう求める.

**興味空間ブラウザの評価** 興味空間ブラウザの有用性に関しては, ユーザに「興味空間の生成」と「興味空間の表示」をさせてから, 図 4 のように教示する.

[14] http://www.mainichi.co.jp

[15] 頻度順, 以下同様

[16] Web ページから抽出される全てのキーワードを評価させることも考えられるが, この場合のキーワード数は膨大であり, 被験者の負荷を揃えることも考慮し, この方法は採用しなかった.

表示される画面は, Web 閲覧履歴から自動的に生成された空間を示しています. □内の単語はあなたの興味として抽出されたキーワードです.
□と地球マークで示されたアイコンは, 実際に見た Web ページを示しています. 表示が薄くなっているものは過去のものを, 大きくなっているものは何度も見たものを表しています.
マウスカーソルを移動させた時に, アイコン上に表示される青字は, その Web ページから抽出されたキーワードを表しています.
一番下に 2 つあるスライドバー (上を A, 下を B と呼びます) のうち, 上にある方 (A) を, 左から一番右までスライドさせて下さい. スライドバーを左から右へ動かすと過去から現在の状態を表示します.
スライドバー A を左右に適当に動かしながら, 表示を見て下さい.
注意: 単語やアイコンが重なって見えにくかったら, マウスで移動させて下さい.

図 4: 被験者への教示

表 1: 抽出された興味語に対する興味度

| 対象 | リンクテキスト | Web ページ |
|---|---|---|
| キーワード数 | 1,524 (SD=842) | 11,209 (SD=6,870) |
| 選択したリンク数 | 355 (SD=256) | 同左 |
| キーワード数/リンク数 | 3.63 (SD=0.19) | 33.49 (SD=3.77) |
| 興味度 X | 3.30 (SD=0.52) | 3.22 (SD=0.47) |
| 興味度 Y | 3.89 (SD=0.34) | 3.19 (SD=0.47) |

興味度 X: 5 段階評価の平均. リンクテキストについては頻度 2 以上のキーワード. Web ページについてはリンクテキストとほぼ同数のキーワード.; 興味度 Y: 高頻度語 10 語における 5 段階評価の平均; SD: 標準偏差

次に, 以下の質問に 5 段階評価 (5:非常に思う, 4:やや思う, 3:どちらともいえない, 2:あまり思わない, 1:全く思わない) で答えさせ, 理由を記述させる.

- 質問 1: このブラウザは Web 閲覧時のあなたの興味空間を表していると思いますか.
- 質問 2: このブラウザは自分の興味を理解するために役にたつと思いますか.
- 質問 3: このブラウザは自分自身について知るために役にたつと思いますか.
- 質問 4: このブラウザは閲覧した Web ページの整理に役立つと思いますか.
- 質問 5: このブラウザは過去を思い出すことに役立つと思いますか.

実験は, 被験者 A-D は 2000 年 12 月に, E-H は 2001 年 7-12 月に実施した.

表 2: 最高頻度語

| 被験者 | リンクテキスト | Web ページ |
|---|---|---|
| A | サミット (3), [シドニー五輪](3), NT 株 (3), パソコン (3) | 見込 (11) |
| B | [シドニー五輪](7) | 確認 (28) |
| C | [露原潜事故](22) | 死亡 (112) |
| D | 逮捕 (14) | 死亡 (47) |
| E | 偽造 (7), [爆発](7) | 製造 (30), 調査 (30) |
| F | [異物混入](17), 自主回収 (17) | 製造 (74) |
| G | [訃報](6) | 自宅 (20) |
| H | [サッカー](25) | 説明 (58) |

() 内は頻度, [] は被験者が最も興味を持ったと答えた語.

### 3.2.2 結果と考察

**概要** 表 1 に示すとおり, 選択されたリンク数は平均 355 であり, 抽出されたキーワードは, リンクテキストを対象とする場合で平均 1,524, リンク先の Web ページを対象とする場合で 11,209 であった. 一つのリンクに対して, リンクテキストで平均 3.63, Web ページで 33.49 のキーワードが抽出された.

**興味語抽出手法の評価** 被験者にリンクテキストと Web ページからのすべての抽出キーワード群を見せ, どちらが直観的に興味語としてふさわしいか聞いたところ, 全員がリンクテキストから抽出したキーワード群と回答した. 表 2 にリンクテキストと Web ページの最高頻度語を示す. 一見してわかるように, リンクテキストから抽出された最頻度語は Web ページから抽出されたものよりも, 固有名詞や時事を表すキーワードが多く, 何に興味を持ったか理解しやすい.

リンクテキスト及び Web ページから抽出されたキーワード群の一つ一つに対して興味度を評価したところ, 興味度 X は, リンクテキストで平均 3.30, Web ページで 3.22 であった. 1 要因の分散分析を行なった結果, 有意差が見られなかった [F(1,15)=0.10]. 興味度 Y は, リンクテキストで平均 3.89, Web ページで 3.19 であり, リンクテキストに有意差が見られた [F(1,15)=11.67, p=<.01].

リンクテキストに関して興味度 X と興味度 Y の平均に関して 1 要因の分散分析を行なったところ, 興味度 Y に有意差が見られた [F(1,15)=7.20, p=<.05].

被験者に興味度 X の語群の中から最も自分が興味を持った 10 語を選択するよう求めたところ, 最も興味を持った語については, 全員がリンクテキストの中から選択し, すべてが頻度が 1 ないし 2 位の語であった.

上記の結果より, 新聞記事の場合は, Web ページ

表 3: 興味空間ブラウザの主観的評価

| 質問 | 平均 | 標準偏差 |
|---|---|---|
| 1 このブラウザは Web 閲覧時のあなたの興味空間を表していると思いますか. | 4.13 | 0.83 |
| 2 このブラウザは自分の興味について知るために役にたつと思いますか. | 4.25 | 0.50 |
| 3 このブラウザは自分自身について知るために役にたつと思いますか. | 3.38 | 1.30 |
| 4 このブラウザは閲覧した Web ページの整理に役立つと思いますか. | 3.75 | 0.71 |
| 5 このブラウザは過去を思い出すことに役立つと思いますか. | 4.00 | 0.93 |

よりもリンクテキストを対象とした方が興味語として良いこと, リンクテキストにおいて高頻度語が低頻度語よりも興味語として良いことが確認された. これは, 本研究におけるリンクに着目する手法のメリットと 提案手法の有効性が示されたと考える. また, 3 か月分の新聞記事のみを情報源とした場合, 不要語リストを利用しなくても高頻度語から興味語を抽出できることがわかった.

**興味空間ブラウザの評価**　被験者の質問に対する評価の結果を表 3 に示す. 質問 1, 2, 4, 5 に関しては肯定的, 質問 3 に関しては中立的な結果であると考える.

質問 1 の興味空間の表示に関しては, 概ね肯定的であり, その理由として「キーワードに興味がよく表れている (B)[17]」「何度も見たものがよくわかる (D)」等, 主として興味語に関する評価があげられたが, 否定的な理由として「空間の見方がよくわからなかった (B)」こと等があげられた.

質問 2 の自己の興味の理解に関しては, 肯定的な理由として「改めて自分の興味のあるものを再認識でき, 以前興味があったものを思い出すことができる (H)」「偏った思考に気づかされる (G)」, 否定的な理由として「分類がわからないので役にたつかどうかもよくわからない (E)」等があった.

質問 3 の自己の理解に関しては,「自分の興味についてはわかるが, 自分自身について分析する上で参考になるかもしれないが, それでわかるとは言い難いような気がする (F)」「今の自分自身を知る手掛かりになると思う. ただ, 自分の興味はすぐに変わったり, ここにはないものもたくさんあるので, これだけでは判断できない (G)」等と, 部分的には役立つがこのままでは役に立たないという意見が多かった.

質問 4 の Web ページの整理に関しては,「もしかなり興味がある内容とかで後日また思いだしたいと思ったときに役にたつのでは (B)」「何度も見たいというデータがわかるため重要度で分類することができるため. 関心がどこにあるか明確になるので整理しやすいと思う」等, 興味に関わる整理の可能性が指摘されたが, 興味空間と同様に「意味が理解できない (A)」「表示が見にくい (H)」等の問題が指摘された.

質問 5 の過去の想起に関しては,「過去の興味やニュース等役にたつことと思う (C)」「自分が過去に何に興味があったのか思い出すことができると思う. ただこれだけで過去が思い出せるかどうかはわからない (H)」等, 過去の興味と新聞記事を中心とした過去の想起に役立つことが指摘されたが, 興味に関わらない部分については肯定的なコメントは特になかった.

実験 2 では, 興味空間ブラウザが, Web 閲覧時の興味空間を表しており, 自己の興味の理解, Web ページの整理, 過去の想起の役立つ可能性が示唆されたが, 主として表示と操作に関する問題が指摘された. これらは, 新聞記事の Web サイトを対象として, ユーザの興味の理解を支援するシステムの開発が可能であるが, 手法の改善が必要であることを示していると考える.

# 4　関連研究と議論

今回試作したシステムは, 個人の記憶や行動履歴等をコンピュータ上に蓄積するシステム Memory-Organizer[5] のサブシステムとして構成されている. Memory-Organizer は, Web ブラウジング履歴だけではなく, ブックマークやメモの統合, Web ページへの上書き, 検索結果の利用等も行える, 統合環境を目指している. 本稿では Memory-Organizer における履歴の作成と興味空間ブラウザについて述べた.

Web ブラウザの履歴の可視化に関する研究の多くは, 既存の Web ブラウザの履歴を対象としており, 本研究のように独自の履歴を作成するものはあまりない.

代表的な研究に, Web ページ (またはサイト) をノードとして, ノード間の関連を木やグラフ構造を用いてネットワーク表示することによりナビゲーションを支援する, MosaicG[6] や PadPrints[7]) 等

[17] () 内は被験者. 以下同様.

がある．これらは既存の Web ブラウザと同様につい最近見たページに「戻る」場合に効果があると考えられる．しかし, 基本的な考え方は, Web ページやサイトをリンクでつなぐのみであり, ユーザの興味を表現したり, 過去に見た Web ページにアクセスするためにどの程度有効かは不明である．

WebWatcher[8] や Letizia[9]) は, ユーザのナビゲーションや情報収集の支援, すなわち, 未来においてユーザが興味を持つであろう Web ページへのアクセス支援を目指し, ユーザの Web 利用行動を学習する手法に焦点をあてているが, ユーザが実際に何を見たかをわかりやすく提示するものではない．

履歴の可視化ではないが, サーチエンジン等への検索結果を分類, 視覚化する研究として, 類似する Web ページ (またはサイト) を分類してキーワードとともに領域化表示するもの (たとえば自己組織化を用いた WEBSOM[10]) がある．検索結果の領域化表示はユーザの興味を表現すると考えられるため本研究と類似するが, 本研究とは情報源と表示手法が異なっている．領域化表示と 2 次元空間表示のどちらが有効かどうかは今後の検討課題とする．

ユーザの興味空間を 2 次元空間上に表現することにより情報検索, コミュニケーション, 発想などを支援する研究がある (たとえば [11], [12])．これらとは, ユーザの興味を 2 次元空間上に表現している点で類似しているが, 目的, 興味空間を表現するための情報源, 2 次元空間上への表示手法が異なっている．

本研究の主目的は, ユーザが自己の興味を理解するためのシステムの開発である．従来, ユーザの興味の理解に関する研究は, 情報工学の分野では, 情報収集・検索, 発想等のユーザの知的情報活動を支援するために, 経営学の分野では, マーケティングのための消費者の興味の理解というように, ある目的のための手段であることが多かった．本研究では, 自己の興味の理解自体が, 人間が自己の人生を行き抜くために, 重要な研究課題であることを主張しておきたい．

さて, 有名な「ジョハリの窓」によると, 自己には 4 つの領域, すなわち, A: 自他にオープンな領域 (自己も他者も知っている領域), B: 自己盲点の領域 (自己は知らないが他者は知っている領域), C: 人にかくしだてしている秘密の領域 (自己は知っているが他者は知らない領域), D: 自分にも人にもわからない無意識界 (自己も他者も知らない領域) があり, A の領域が広く D の領域が狭い方が好ましいパーソナリティであると言う [13]．本システムではどの領域の興味の理解を支援できるのであろうか．実験 2 における「興味がよく表れていると思う (B)」「改めて自分の興味あるものを再認識できる (H)」等の多くの被験者のコメントから判断すると, 既に自己が知っている領域を中心として再確認する効果があったと考える．ただし, 履歴が実験者に見られることを被験者が知っていたため, C の領域に関しては明らかではない．「偏った思考に気づかされる (G)」というコメントがあるとおり, 自己が知らない領域の支援もある程度行っていることがわかる．

残念ながら, 本システムが直接自己の理解に役立つかどうかに関してはあまり肯定的な結果は得られなかった．これは, 「今の自分自身を知る手がかりになると思う．ただ, 自分の興味はすぐに変わったり, ここにはないものもたくさんあるので, これだけでは判断できないと思う．(H)」という被験者のコメントにあるとおり, 「自己」というものが非常に広範囲の概念であり, 新聞記事サイトにおける Web ブラウジング行動から生成された限定的な興味空間のみからはうかがいしれないことを意味していると言えよう．

本研究では, (1) ユーザの選択したリンク周辺に含まれるテキストからユーザの興味を表すキーワードを抽出できる, (2) 抽出したキーワードを時系列に提示することにより, ユーザの興味の理解を支援するシステムを開発できる, という仮説をたてた．この仮説は, 部分的にではあるが, 実験によりある程度実証されたと考える．以下に今後の課題を述べる．

## 5 今後の課題

前述したとおり, Web ブラウザの実装上の問題により, あらゆる Web サイトを対象とした自然な状態での利用実験はできなかった．現実問題として, 実用レベルの Web ブラウザの実装は非常に時間がかかる仕事であり, 本研究の対象外とする．以下では, 本研究の枠組の中で実証可能な課題について述べる．

### 5.1 興味語を抽出する対象となるテキスト

新聞記事のWebサイトに関して, 頻度に基づく単純な方法を用いて, リンクテキスト (見出し) を対象とした方が, Webページ (見出しと本文を含む) を対象とするよりも, ユーザの興味を表すキーワードを抽出できることがわかった. ただし, テキストからキーワードを抽出する手法に関しては, TDxIDFをはじめ, 多様な手法がある. これらの手法を用いてWebページからキーワードを抽出した場合との比較は, 今後の課題である.

本研究では「リンクテキストを含む行」を対象としたが, 実験2ではサイトの特徴上, 実際はリンクテキストが対象となった. Webサイト全体において, リンクテキストを含む行と, リンクテキストで, どちらが良いかは今後の検討課題である.

今回実験した新聞記事のWebサイトでは, タイトルにWebページの内容に関する情報が含まれていないため, リンクテキストからキーワードを抽出する方が有用であることは自明であったが, Web全体で実験を行ったわけではなく, タイトルよりもリンクテキストからキーワードを抽出する方が良いかどうかは不明である.

本研究は, 既存のWebブラウザの履歴をおきかえようとする試みではなく, 既存のWebブラウザの履歴に「リンク選択行動から抽出される情報」を追加しようとするものである. 今後はどちらがテキストとして良いかを深く調べるよりもむしろ, 両方のテキストをうまく扱える方法を検討していきたい.

### 5.2 キーワード抽出手法

本研究では, テキストからのキーワードの抽出に字種による方法を採用した. 今後は実験1で指摘した問題点を解決するために, 本手法の改善及び, 形態素解析を併用する方法を検討していきたい.

### 5.3 不要語リスト作成

実験1では, キーワードの増加に従い高頻度語に不要語が増加するが, 不要語リスト機能により不要語の排除が行えることが示された. 将来的には, ユーザのブラウジング行動等から自動的に不要語リストを作成する機能の追加が望ましい.

### 5.4 興味語の網羅性

興味空間ブラウザでは, ユーザにとって興味のあるキーワードを抽出することに焦点をあて, 数量化III類の利用を前提として, 頻度2以上のキーワードを興味語とした. しかし, 頻度が1のキーワードや興味語を持たないWebページを見ることはできない. これらに関しては興味空間ブラウザを参考にして履歴自体をオープンする機能によって補うことができるが, 興味空間ブラウザの手法の枠組の中でも解決できればなお良く, 今後の課題とする.

### 5.5 表示手法とユーザインターフェース

実験1で述べたように, 新しい履歴と数量化III類の組合せによる興味空間の表示手法に問題を含んでおり, 履歴の作成方法の改善, 数量化III類以外の空間配置手法の試用等の検討が必要である. オブジェクトが中心に固まる問題点については適当にちらばらせる等による改善方法や, 履歴の量が増えると表示に時間がかかる問題点に関しては部分的に計算する等による方法を検討する必要がある.

## 6 おわりに

ユーザの興味の理解を支援するために, ユーザのリンク選択時に, リンクを含む行のテキストからキーワードを抽出して, リンク先URLと日時と共に新しい履歴に蓄積し, その履歴から, キーワードとWebページアイコンを2次元空間上に配置して時系列に提示する手法を提案した.

本手法の特徴は, (a) ユーザのリンク選択行動に基づく新しい履歴を作成すること, (b) ユーザのリンク選択行動から抽出したキーワードとWebページアイコンを2次元空間上に配置して時系列に提示する表示手法である.

本研究では, 新しい履歴を作成するWebブラウザと, 興味空間ブラウザを試作した. 日常生活における利用のシミュレーションと, 新聞記事のWebサイトを対象とした実験の結果, (1) ユーザの選択したリンクを含む行のテキストからユーザの興味を表す日本語のキーワードを抽出できること, (2) 興味空間ブラウザがユーザのWeb閲覧時の興味空間を表していること, (3) 興味空間ブラウザが自己の興味の理解, 過去の想起, Webページの整理に役立つ可能性があること, がわかった.

興味語の抽出手法, 表示手法等の一つ一つのアルゴリズムには改善すべき点があり, 今後の課題とする. また, 異なる情報源も対象として検討していきたい.

## 参考文献


[1] http://www.microsoft.com/

[2] http://www.netscape.co.jp/

[3] 平田高志: 外化記憶の構築と共有支援に関する支援, 奈良先端科学技術大学院大学情報科学研究科博士論文 (2001).

[4] 木下栄蔵: わかりやすい数学モデルによる多変量解析入門, 啓学出版, (1987).

[5] 村上晴美, 平田高志: Memory-Organizer: 個人の外化記憶構築システム, 2001 年度人工知能学会全国大会 (第 15 回) 論文集, 3F1-03 (2001).

[6] Ayers, E. Z. and Stasko, J. T.: Using Graphic History in Browsing the World Wide Web, *Proceedings of WWW4* (1996).

[7] Hightower, R. R., Ring. L. T., Helfman, J. L., Bederson, B. B., and Hollan, J. D.: Graphical Multiscale Web Histories: A Studey of Padprints, in *Proceedings of ACM Hypertext'98*, pp.58–65 (1998).

[8] Armstrong, R., Freitag, D., Joachims, T. and Mitchell, T.: WebWatcher: A Learning Apprentice for the World Wide Web, *Proceedings of AAAI Symposium on Information Gathering from Distributed, Heterogeneous Environments*, pp.6–12 (1995).

[9] Lieberman, H: Letizia: An Agent That Assists Web Browsing, *Proceedings of IJCAI95*, pp.924–929 (1995).

[10] Kohonen, T., Kaski, S., Lagus, Sakojvi, J., Paatero, V., and Saarela, A.: Self Organization of a Massive Document Collection, *IEEE Transactions on Neural Networks, Special Issues on Neural Networks for Data Mining and Knowledge Discovery*, Vol.11, No. 3, pp.574–585 (2000).

[11] 門林理恵子, 西本一志, 角康之, 間瀬健二: 学芸員と見学者を仲介して博物館展示の意味構造を個人化する手法の提案, 情報処理学会論文誌, Vol. 40, No. 3, pp.980–989 (1999).

[12] 角康之, 間瀬健二: 実世界コンテキストに埋め込まれたコミュニティウェア, 情報処理学会論文誌, Vol. 41, No. 10, pp.2679–2688 (2000).

[13] 國分康孝: カウンセリング辞典, 誠信書房, pp.287-288 (1990).