ウェブコミュニティを用いた大域ウェブアクセスログ解析法の一提案

大塚真吾[†] 豊田正史[†] 喜連川優[†]

ウェブページを閲覧する人々の行動モデルの抽出は重要であり多くの研究が行われている．既存の研究のほとんどはウェブサーバのログを用いたものであり，当該サイト上での挙動は把握できるものの，サイト外を含めたユーザの行動を解析することは容易でない．最近，テレビ視聴率調査と同様，統計的に偏りなく抽出された人（パネル）を対象に URL 履歴の収集を行う事業が登場している．パネルから集められたログ（パネルログ）の解析により，パネルが訪れたすべてのウェブページ（URL）を収集できる．ウェブサーバに対する従来のログ解析では解析対象となるページ空間が狭いのに対し，パネルログではきわめて広大なページ空間を対象とするため，個々のページの参照履歴から大域的な行動の把握は容易でない．本論文では類似したウェブページを抽出するウェブコミュニティ手法を用いたパネルログ解析システムを提案し，URL を基にした解析ではとらえ難い大域的なユーザの行動パターン抽出例を紹介する．

A Study for Analysis of Web Access Logs with Web Communities

Shingo Otsuka,[†] Masashi Toyoda[†] and Masaru Kitsuregawa[†]

To extract model of Web users' behavior is of decisive importance and there are a lot of work has been done in this area. As far as we know, most of the work utilize logs on server-side, even it can gain an understanding of behavior inside the server, but it is hard to analyze complete users' behavior (inside and outside the server). Recently, similar to survey on TV audience rating, a new kind of business appeared, which collects URL histories of users (called panel) who are selected without statistic deviation. By analyzing panel logs which are merged from panels, it becomes possible to collect all the web pages (URLs) accessed by the users. In contrast to Web server logs which have a limited page-space, panel logs have an extremely broad page-space. For this reason, it's difficult to get hold of behavior on global page-space by just checking reference histories. In this papaer, we propose a prototype system to extract user access patterns from panel logs and show users' global behavior patterns which are hard to be grasped for URL-based analysis using our proposed system.

1. はじめに

ウェブ上でのユーザの行動解析は重要な研究課題であり，様々な研究が行われている．これらの研究の多くはウェブサーバのアクセスログ（サーバログ）を利用している．

一方，テレビの視聴率調査と同様，統計的に偏りなく抽出された人（パネル）を対象に URL 履歴の収集を行う事業が登場している．パネルから集められたアクセスログの解析により，個々のパネルが訪れたすべての URL を把握できる．このようにして集められたログを本論文ではパネルログと呼ぶ．

パネルログは解析対象となるページの種類が多いため，URL に基づく解析からユーザの行動を把握することは難しい．そこで，我々は大域的なユーザの行動をとらえるためにウェブコミュニティ☆を用いる．我々はユーザの行動パターンをパネルログから抽出するには現時点でその自動化は容易でなく人間の解析が不可欠との判断から，本論文では解析者がパネルログから大域的なユーザ行動の把握を支援するシステムの提案を行い，その有効性を確認する．

以下，2 章で関連研究について述べる．3 章では本論文で利用したパネルログとウェブコミュニティについて述べ，4 章でパネルログの予備的な解析について述べる．5 章では我々が提案するウェブコミュニティを用いた大域的なユーザ行動を把握するためのシステムの提案を行う．6 章では本システムの利用例とその

† 東京大学生産技術研究所
Institute of Industrial Science, The University of Tokyo

☆ 以降,「コミュニティ」は「ウェブコミュニティ」の意味で使用.

有効性について述べる．

2. 関 連 研 究

アクセスログを用いた研究は今までに数多く行われており，その目的も様々である[4]．これらの研究で用いられるアクセスログはサーバログがほとんどであり，本論文で利用しているパネルログを用いた研究は我々が知る限りでは詳細な研究は行われていない．

- **ユーザの行動に関する研究**

  この研究は e コマースなどのビジネスに直結するためさかんに行われている．文献 1)，15) では，ユーザの行動パターンの抽出に関する研究を行っている．また，同じ行動パターンのユーザをグループ化するユーザクラスタリングに関する研究も行われている[6]．ショッピングサイトでのユーザのグループ化については文献 19) で述べられている．

  これらの研究はウェブサーバが置かれたサイト内のウェブページに限定されるため，局所的なユーザ行動の抽出に焦点を当てている．

- **ウェブページ間の関連に関する研究**

  この研究はリンク解析を用いた手法が主流であるが，最近ではアクセスログを用いた手法が提案されている．文献 16) では，アクセスログを用いたウェブページのクラスタリングについて述べられている．また，ユーザのアクセスパターンからウェブページ間の相関を抽出する研究も行われている[17]．文献 21) は OLAP を利用して解析の利便性を高めている．これらの研究はウェブページの関連性を調べるものである．

- **検索サイトに関連する研究**

  検索エンジンサイトを提供するベンダでは検索語に関する研究が行われている．lycos のサーバログを用いた検索語のクラスタリングについては文献 2) で述べられている．また，文献 20) では microsoft が提供するオンライン百科事典の Encarta のサーバログを用いてクラスタリングを行っている．文献 12) では検索向上の支援のためにアクセスログを利用している．これらの研究は検索語を入力したサイト内でのユーザの行動を解析するものである．

- **アクセスログの視覚化に関する研究**

  アクセスログの解析結果を分かりやすくするために，その視覚化に関する研究が行われている[8]．文献 14) では，電話番号案内サイトでのユーザ行動の視覚化について述べている．

- **その他の研究**

  本論文に密接な関連研究として，ユーザとウェブページの両方のクラスタリングに関する研究がある[22]．この研究ではアクセスログからユーザとウェブページのクラスタリングを行い，クラスタリングしたユーザ群とウェブページ群の相関抽出を目的としている．また，この研究ではウェブページを閲覧する側のサーバログであるプロキシーログを用いる．このログはユーザが訪れた URL をすべて保持し，また，プライベート IP の情報などからユーザの特定が容易であり，パネルログの性質と類似している．最近では研究用のデータとして大学のプロキシーログなどが無償で公開されているが，このログを使った研究はまだ少ない．

  そのほかに，大量なログの格納方法に関する研究[11] や，モバイル環境でのアクセスログ解析に関する研究[13] などがある．

以上のように従来のほとんどの研究はサイト内でのユーザ挙動の解析を対象としている．文献 22) はプロキシーログを用いており，やや類似するが参照したウェブページのクラスタリングを目的とするため研究の方向性が異なる．

3. パネルログとウェブコミュニティ

この章では本論文で提案するシステムを構築するうえで基礎となるパネルログとウェブコミュニティについて述べる．また，ウェブコミュニティを用いた理由についても述べる．

3.1 パネルログ

本論文で利用するパネルログの収集法を**図 1** に示す．インターネット視聴率調査会社はエリア別のインターネット利用率を基に RDD（Random Digit Dialing）方式により無作為抽出した世帯を決定しパネルの依頼を行う．パネルとなる人のパソコンに，その人が訪れた URL 履歴などのウェブページアクセスデータを定期的に送信するプログラムをインストールし，アクセスログの収集を行う．図中のアンケート回答とパネルプロファイルは個人情報を含むため本論文では利用しない．収集されたパネルログの形式を**表 1** に示す．パネルログは，

- パネル ID
- ウェブページにアクセスした時刻
- ウェブページを閲覧した時間
- アクセスしたウェブページの URL

などから構成される．パネル ID とはパネル全員に対

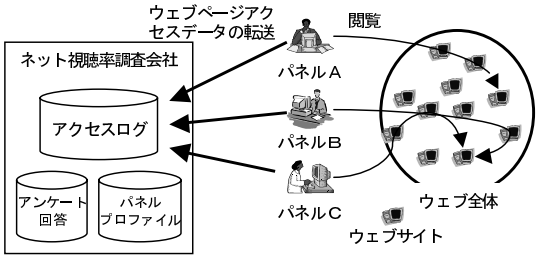


図 1 パネルログ収集の概要
Fig. 1 A method of collecting panel logs.

表 1 パネルログの一部
Table 1 A part of the panel logs.

Panel ID	AccessTime	RefSec	URL
1	2002/9/30 00:00:00	4	http://www.tkl.iis.u-tokyo.ac.jp/welcome_j.html
2	2002/9/30 00:00:00	6	http://www.jma.go.jp/JMA_HP/jma/index.html
3	2002/9/30 00:00:00	8	http://www.kantei.go.jp/
4	2002/9/30 00:00:00	15	http://www.google.co.jp/
1	2002/9/30 00:00:04	6	http://www.tkl.iis.u-tokyo.ac.jp/Kilab/Welcome.html
5	2002/9/30 00:00:04	3	http://www.yahoo.co.jp/
6	2002/9/30 00:00:05	54	http://weather.crc.co.jp/
2	2002/9/30 00:00:06	11	http://www.data.kishou.go.jp/maiji/
3	2002/9/30 00:00:08	34	http://www.kantei.go.jp/new/kousikiyotei.html
5	2002/9/30 00:00:07	10	http://search.yahoo.co.jp/bin/search?p=%C5%B7%B5%A4
1	2002/9/30 00:00:10	300	http://www.tkl.iis.u-tokyo.ac.jp/Kilab/Members/members-j.html

検索語を含むURL (1)

してユニークに割り当てられた ID であり，個々のパネルが特定できる．また，表中 (1) は検索語の情報を含む URL である．

通常，アクセスログの解析ではセッションの概念を導入している．セッションとはウェブサイトを訪れたユーザが行う一連の行動である．本論文ではセッションを「パネルがウェブページの閲覧を開始してから，閲覧を終了するまでに訪れた URL の集合」と定義する．また，閲覧の終了をウェブページを閲覧し終えてから，次のウェブページにアクセスするまでに 30 分以上あるときと定義する[3]．

3.2 ウェブコミュニティ

ウェブ全体をグラフ構造と見なしてウェブコミュニティを発見する手法は，

(1) 密な部分グラフを抽出する手法[5]
(2) 完全二部グラフを抽出する手法[9]

の 2 つに大別できる[10]．また，本論文では文献 10) に従いウェブコミュニティを「ハイパーリンクによって密に結合した関連ウェブページの集合」という意味で用いる．(1) の手法はネットワーク理論における最大流最小切断定理をウェブに適用し，ウェブコミュニティの内側と外側を分ける境界を発見する手法である．(2) の手法は興味を共有するページ集合のリンクは完全二部グラフを構成することに注目し，ウェブのスナップショットデータからサイズを固定した完全二部グラフを探索する手法である．

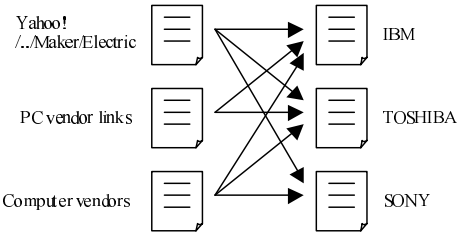


図 2 ハブとオーソリティからなる典型的なグラフ
Fig. 2 Typical graph of authorities and hubs.

また，特定のトピックのページに関するランキングアルゴリズムの代表的なものに HITS[7] がある．これはウェブページの有効性の評価基準としてハブとオーソリティという概念を用いる．ハブとはあるトピックに関連するリンク集やブックマークなどのページを指し，多くの良質なオーソリティにリンクを張るページと定義される．一方，オーソリティとはあるトピックについて良質な内容を持ったページであり，多くの良質なハブからリンクが張られていると定義される．**図 2** に HITS により抽出される例を示す．図の右側のオーソリティは大手のコンピュータメーカのページである．これらのページはコンピュータメーカリンク集などのハブにより密に結合されている．

我々の研究室では (2) の手法と HITS を基本とし，大量なウェブページから自動的にコミュニティの抽出を行うウェブコミュニティチャート[18] なる手法を提案している．当該手法はウェブのスナップショットデータからコミュニティを頂点とし，コミュニティ間の関連度を重み付きの辺で表したグラフを抽出する．2002 年 2 月に国内 4,500 万のウェブページの収集を行い，100 万個の有用なページから自動的な処理により 17 万個のコミュニティを生成した．本研究では我々が生成したコミュニティを利用してパネルログの解析を行う．

3.3 パネルログ解析のためのウェブコミュニティの利用

パネルログはきわめて多くの URL を含むため，URL に基づく解析結果からユーザの行動を把握することは容易でない．我々はウェブコミュニティの利用により抽象度の高い解析結果が得られ，個々の URL の解析だけではとらえ難い現象を発見できると考えている．たとえば，あるトピック X に関するページ A とトピック Y に関するページ B，C，D がありそれぞれのアクセス数は 5，2，2，2 とする．この場合，個々のページに基づく解析手法ではトピック Y に関するアクセスが多いという事実を抽出することは容易でない．そこで，ウェブコミュニティを導入することで，より抽象度の高いユーザ挙動を取り出せることが期待される．さら

表 2 パネルログの概要
Table 2 The detail of our used panel logs.

データ量	約 10 Giga byte
データ収集期間	45 週間
アクセス数	55,415,473 アクセス
セッション数	1,148,104 セッション
パネル数	約 1 万人
パネルの抽出法	RDD（Random Digit Dialing）方式
検索語の種類	約 30 万種類

に，各々のコミュニティに含まれるページに対して張られたリンクのアンカータグの解析から，十分に正確ではないもののコミュニティの内容を表す単語群（コミュニティラベル）を自動的に抽出できており，これにより，解析者はコミュニティに含まれる個々のウェブページを閲覧することなくコミュニティの概要を把握できる．

我々はウェブコミュニティを用いたパネルログの解析より，解析者が結果を直感的に理解でき，さらに大域的なユーザ行動を把握する手がかりになると考えている．

4. パネルログの予備的解析

ウェブコミュニティを用いたパネルログ解析システムの構築の前に，本論文で用いるパネルログの基本的な性質をつかむために予備的な解析を行った．

4.1 ウェブコミュニティとパネルログに含まれる URL のマッチング

今回利用するパネルログのパネルはすべて日本人であり，その詳細を**表 2** に示す．ウェブコミュニティ生成のためにウェブのスナップショットデータを収集した時期はパネルログ収集期間中であった．パネルがアクセスしたウェブページに変更や削除の可能性があるため，パネルログに含まれる URL とウェブコミュニティに登録されている URL の適合率を測定した．ここでは，

$$\text{適合率} = \frac{\text{コミュニティ } URL \text{ と合致するパネル } URL \text{ の数}}{\text{パネル } URL \text{ の数}}$$

ただし，コミュニティ URL = コミュニティに属する URL

パネル URL = パネルログに含まれる URL

と定義し，結果を**表 3** に示す．無修正時における適合率は 18.8%と低いが，ファイル名やディレクトリ名を削除する処理により 36.3%に増加した．また，サイト名を削除する処理☆1によりさらに 7.7%向上した．このように URL の修正から最終的にアクセスログ全体の 63%の URL をコミュニティに適応させることができた．

☆1 http://xxx.yyy.com/で合致しない場合は xxx を削除し，http://yyy.com/で再びチェックを行う．また，.com や co.jp などの組織名についての照合は行わない．

表 3 ウェブコミュニティに登録されている URL とパネルログに含まれる URL の適合率
Table 3 The adaptation ratio of the URLs belonged to web-communities and the URLs included panel logs.

無修正	18.8%
ディレクトリ（ファイル）部分を削除して合致	36.3%
サイト部分を削除して合致	7.7%
合致せず	37.2%

表 4 検索語を抽出した検索エンジン（ポータル）サイト
Table 4 The search (portal) sites which extracted search words.

yahoo.co.jp	nifty.com	biglobe.ne.jp
infoseek.co.jp	msn.co.jp	ocn.ne.jp
so-net.ne.jp	dion.ne.jp	lycos.co.jp
goo.ne.jp	hi-ho.ne.jp	odn.ne.jp
excite.co.jp	google.co.jp	fresheye.co.jp
altavista.com		

4.2 アクセス数が多いサイトの解析

インターネット視聴率会社☆2が公表するインターネットアクセスランキングでは

- 検索エンジン（google など）
- ディレクトリサイト（Yahoo!など）
- ポータルサイト（nifty，biglobe など）
- ショッピングサイト（楽天など）

がつねに上位である．また，Yahoo!と同一サイト内にある Yahoo! auctions などのオークションサイトもアクセスが多いと思われる．我々はパネルログの性質をつかむためにパネルログに含まれるこれらのサイトの割合を測定した．検索エンジンやディレクトリサイトは多数存在するが，ここでは**表 4** に示すサイトを含む URL を対象とし☆3，検索エンジン群と呼ぶ．また，3.1 節で述べたようにパネルログに含まれる URL は検索語に関する情報を含むため検索エンジン群を対象に検索語入力の有無についても調べた．ショッピングサイトに関しては Yahoo! shopping と楽天について，オークションサイトは Yahoo! auctions と楽天のオークションについて調べた☆4．

☆2 http://www.vrnetcom.co.jp/など．

☆3 yahoo に関しては，http://shopping.yahoo.co.jp/と http://auctions.yahoo.co.jp/は除いた．また，nifty などのポータルサイトの場合，個人や企業のホームページは検索エンジン群から除外する．

☆4 http://shopping.yahoo.co.jp/
http://www.rakuten.co.jp/
http://auctions.yahoo.co.jp/
http://www.rakuten.co.jp/auction/

表 5 パネルログに含まれる URL に対する検索エンジン群，ショッピングサイト，オークションサイトの割合
Table 5 The ratio of the group of the search sites, shopping sites and auction sites in the URLs included in panel logs.

検索語が入力された検索エンジン群	4.1%
検索語がない検索エンジン群	19.4%
ショッピングサイト	1.5%
オークションサイト	10.9%
その他のサイト	64.1%

表 6 パネルログに含まれる URL に対する検索エンジン群の割合など
Table 6 The ratio of the group of the search sites in the URLs included in panel logs.

(1) 検索語が入力された検索エンジン群*	4.1%
(2) 検索語がない検索エンジン群**	19.4%
(3) *の後に訪れた URL	12.3%
(4) **の後に訪れた URL	43.4%
(5) 検索エンジン群の前に訪れた URL	7.7%
(6) 検索エンジン群を含まないセッションの URL	13.1%

表 7 検索エンジン群，ショッピングサイト，オークションサイトが含まれるセッションの割合
Table 7 The ratio of the sessions included in the group of the search sites, shopping sites and auction sites.

検索語が入力された検索エンジン群を含むセッション	23.3%
検索語がない検索エンジン群を含むセッション	69.6%
ショッピングサイトを含むセッション	5.7%
オークションサイトを含むセッション	12.4%

パネルログに含まれる全 URL に対するそれぞれのサイトの割合を**表 5** に示す．検索語の入力がある検索エンジン群とショッピングサイトの結果は 4.1%，1.5%と低い．検索語の入力がない検索エンジン群の割合は約 20%であり，オークションサイトに関しては約 10%である．

次に，検索エンジン群についてさらに詳細な解析を行いその結果を**表 6** に示す．表中の (3) と (4) は検索エンジン群から訪れたと予測される URL の割合である．表中 (1) から (4) の合計から，パネルログに含まれる URL の約 80%が検索エンジン群またはそこから訪れた URL であることが分かる．また，検索語が入力された検索エンジン群とその後に訪れた URL の合計は 16.4%（表中の (1) と (3) の合計）である．

最後に全セッションに対するそれぞれのサイトの割合を測定し，その結果を**表 7** に示す．検索語の入力がある検索エンジン群を含むセッションの割合は約 23%である．また，検索語の入力がない検索エンジン群は約 70%と高い．ショッピングサイトを含むセッションの割合は表 5 の結果に比べ増加している．一方，オークションサイトを含むセッションの割合は表 5 が示す結果とほとんど変わらない．これはオークションサイトを訪れるユーザは品物の検索や入札のためにサイト内の滞在時間が長くなり，1 セッションあたりのアクセス数が多くなるためだと思われる．

このようにユーザの行動には検索エンジンなどのサイトや検索語の入力が深く関与している．そこで，5 章ではコミュニティだけでなく検索語にも着目し，ユーザ行動をコミュニティならびに検索語のいずれからも柔軟に解析可能なシステムの構築を目指すこととした．また，解析結果から Yahoo! shopping，Yahoo! auctions，楽天の 3 サイトと検索エンジン群に関してはアクセス数やセッション数が多く，かつ，これらのサイトの内容はコミュニティを用いなくても理解できるため，5 章以降で述べる解析システムでは Yahoo! shopping と楽天は「ショップ」，Yahoo! auctions と楽天のオークションは「オークション」，検索エンジン群は「検索エンジン・ポータルサイト」とする．

5. パネルログ解析システムの提案

この章では大域的なユーザ行動をとらえるためのパネルログ解析システムの提案を行う．

5.1 システムの解析機能

検索語はユーザの目的となるためコミュニティと検索語の関連から，コミュニティを訪れた理由を知ることができる．そこで，本システムではコミュニティと検索語の関連を解析する機能を保持する．また，検索語の入力がないセッションでもページ間の関連から，そのページを訪れた経緯が分かる．3.3 節で述べたように，URL に基づく解析では結果を直感的に理解することは困難なため，本システムではコミュニティ間の関連を解析する機能を保持する．

5.2 解析機能の詳細

本システムではウェブコミュニティを通し番号（コミュニティID）で管理しており，指定した URL が属するコミュニティID を検索することができる．加えて以下の機能を有する．

(1) **検索語入力後に流入したコミュニティの表示**
　指定した検索語を用いて検索を行ったユーザが訪れたコミュニティの一覧が表示される．

(2) **コミュニティに流入するために使用した検索語の表示**
　指定したコミュニティを訪れるために用いた検索語の一覧が表示される．

(3) **流入・流出コミュニティの表示**
　指定したコミュニティを訪れる前後のコミュ

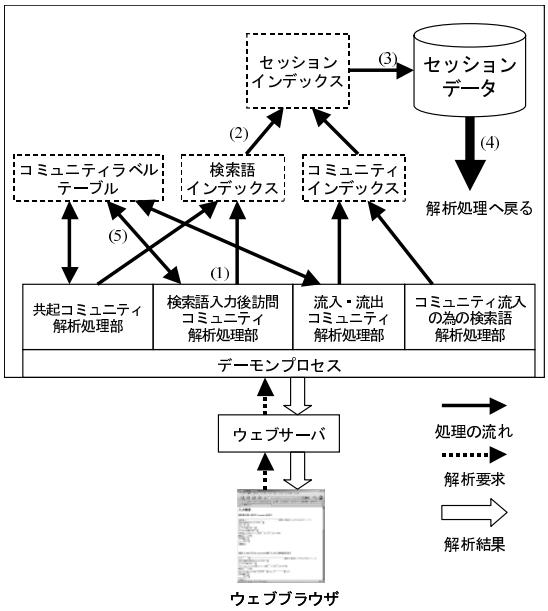


図 3 システム構成図
Fig. 3 The architecture of our proposed system.

ニティの一覧が表示される．

(4) **共起コミュニティの抽出**

ウェブコミュニティと検索語を指定しそれらを含むセッション中に共起するコミュニティの一覧が表示される．

(1) と (2) はコミュニティと検索語の関連を解析する機能であり，(3) と (4) はコミュニティ間の関連を解析する機能である．また，それぞれの解析ではいくつかのパラメータの設定が可能である．主なものにユーザの正規化があり，特定のユーザが解析結果に影響を及ぼす場合にその影響を除去する．そのほかに解析範囲の設定などがある．

5.3 システムの構成

本システムの構成図を**図 3** に示す．ウェブブラウザ上で入力された解析要求はウェブサーバを介してデーモンプロセスに送られる．デーモンプロセスはウェブサーバから解析要求を受け付ける部分である．解析処理を行う部分は図中

(a) 共起コミュニティ解析処理部
(b) 検索語入力後訪問コミュニティ解析処理部
(c) 流入・流出コミュニティ解析処理部
(d) コミュニティ流入のための検索語解析処理部

の 4 つに分かれており，受け付けた解析要求により選択される．ウェブサーバは解析結果を受け取り動的に HTML ファイルを生成する．

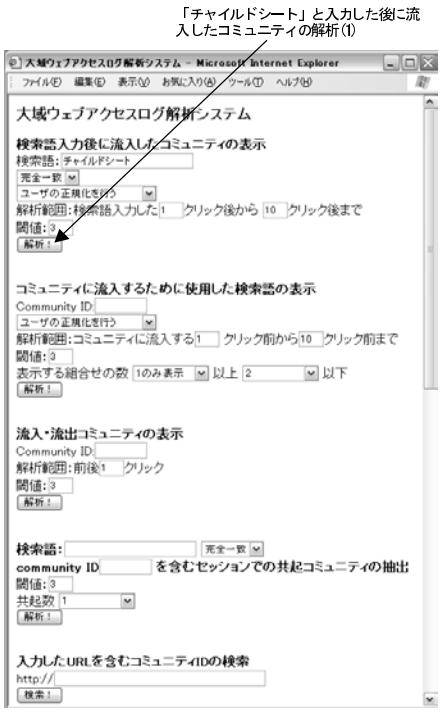


図 4 解析開始画面
Fig. 4 Starting page of our system.

本システムではパネルログの実データをセッション単位で 2 次記憶上に格納し（セッションデータ），各セッションはユニークな ID（セッション ID）を保持する．セッションインデックスはセッション ID からそれに対応するセッションの実データを得るために利用される．また，検索語やコミュニティ ID からセッション ID を検索するために 2 つのインデックスがある．コミュニティラベルテーブルは URL からコミュニティ ID の検索を行うときや，コミュニティ ID に対応するコミュニティラベルを獲得するときに利用される．

例を用いて処理の概要を述べる．たとえば**図 4** に示されるようにチャイルドシートを検索語として図 4 (1) の解析ボタンを押すと，

- デーモンプロセスは解析要求を図 3 の「検索語入力後訪問コミュニティ解析処理部」に渡す．
- 「チャイルドシート」と入力したセッション ID を得るため検索語インデックスにアクセスする（図中 (1)）．
- 得られたセッション ID を用いてセッションインデックスにアクセスする（図中 (2)）．
- セッションインデックスを用いてセッションデータにアクセスし「チャイルドシート」と入力したセッションのデータを得る（図中 (3)(4)）．
- 図中「検索語入力後訪問コミュニティの解析処理

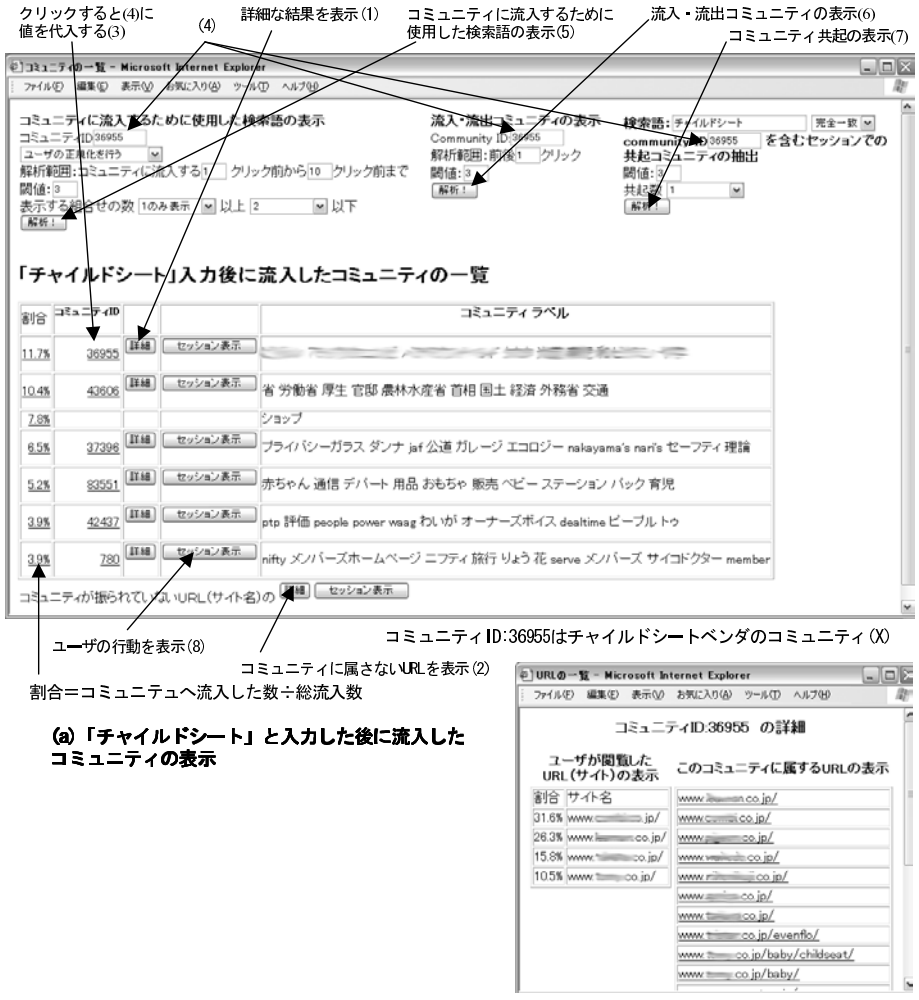


図 5 「チャイルドシート」と入力した後に流入したコミュニティの表示例
Fig. 5 Expression of Web communities with input 'child car seat'.

部」においてセッションデータの解析を行う．

- 解析結果のコミュニティID に対応するコミュニティラベルを得る（図中 (5)）．
- 結果をウェブサーバに送る．

6. システムの利用とその有効性

この章では，我々が提案したシステムの利用例とウェブコミュニティを用いたパネルログ解析の有効性について述べる．

6.1 本システムの利用例

解析者は検索語またはコミュニティID の入力と各パラメータを設定し解析する．また，解析者が解析結果から興味を持つ項目について，簡単な操作で新たな解析ができる．

6.1.1 検索語入力後に流入したコミュニティの表示

「チャイルドシート」と入力した後に流入したコミュニティの解析を行った結果を**図 5** (a) に示す（図 4 の (1) を押した結果）．解析結果は流入が多いコミュニティ順に表示される．流入が 2 番目に多いコミュニティ（ID: 43606）のラベルからこのコミュニティは行政関連であることが容易に想像できる．また，1 番多いコミュニティ（ID: 36955）のラベルは企業名が多いため伏せたが，ラベルからチャイルドシートベンダに関連するコミュニティであることを推測でき，これをチャイルドシートベンダのコミュニティ（X）とする．

図中 (1) を押すとコミュニティに属するページについての詳細な結果を閲覧できる（図 5 (b)）．また，4.1 節で述べたように，パネルログにはコミュニティに属さ

図 6 赤ちゃん関連のコミュニティに流入するために使用した検索語の表示例
Fig. 6 The list of search words used for view of the community related to baby.

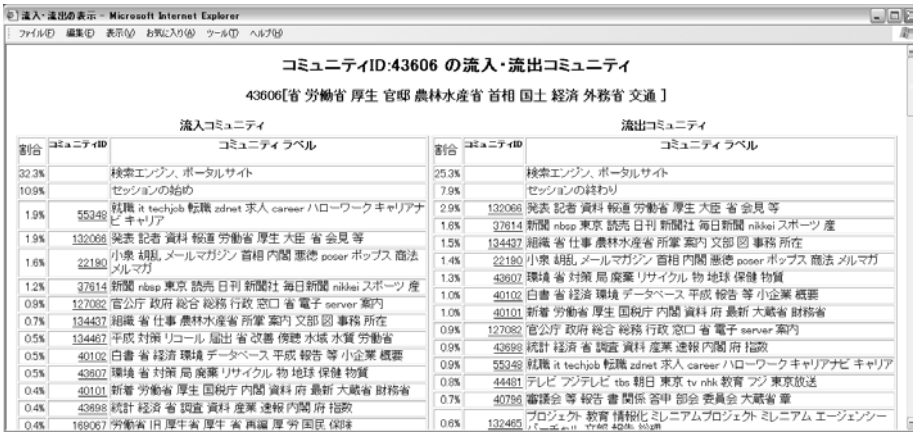


図 7 流入・流出コミュニティの表示例
Fig. 7 The list of inflow and outflow Web community.

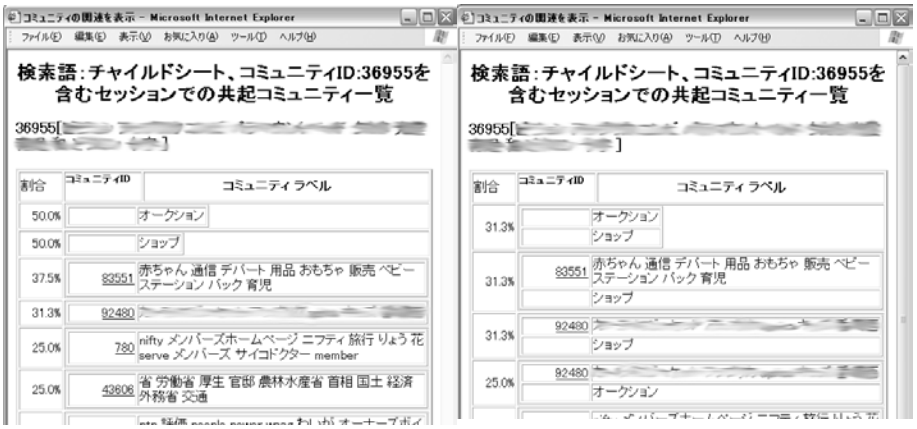
ない URL が約 37%あり，これらのページについての結果を閲覧することができる（図中 (2)）．

6.1.2 コミュニティに流入するために使用した検索語の表示

赤ちゃん関連のコミュニティに流入するために使用した検索語を**図 6** (1) に示す．「赤ちゃん」と関連した検索語が多い．また，URL が表示された表（図中 (2)）はコミュニティに含まれるページごとの結果であり，各ページへの流入に特徴ある検索語が使われている．図 5 (a) と同様，結果から新たな解析ができる．

6.1.3 流入・流出コミュニティの表示

国土交通省などを含むコミュニティの流入・流出コミュニティを**図 7** に示す．図中のコミュニティ ID からそのコミュニティの詳細が閲覧できる．解析結果から流入と流出のトップは検索エンジンやポータルサイトである．流入にはセッションの初めが多く，他のコミュニティには流出せずにこのコミュニティでセッションを終了するユーザが多い．また，流入と流出のほと



コミュニティID:92480はチャイルドシートベンダのコミュニティ(Y)

$$\text{割合} = \frac{\text{コミュニティが出現したセッション数}}{\text{検索語「チャイルドシート」を入力し，コミュニティID:36955を含むセッション数}}$$

(a) 共起数1の結果　　(b) 共起数2の結果

図 8 「チャイルドシート」と入力しチャイルドシートベンダのコミュニティ（X）を含むセッションでの共起コミュニティの表示例

Fig. 8 The list of co-occurrence of Web community in the session with search words 'child car seat' and community 'child car seat vendors (X)'.

んどは行政機関やマスコミに関連するコミュニティが多いことが分かる．

6.1.4 共起コミュニティの抽出

検索語「チャイルドシート」の入力がありチャイルドシートベンダのコミュニティ（X）（コミュニティID: 36955）を訪れたすべてのセッションに対し頻出コミュニティを求め，その結果を表示する（**図 8** (a)）．チャイルドシートについて検索を行いベンダページをチェックするユーザは同時にオークションやショッピングサイトを訪れる可能性が高いことが分かる．また，表中の3番目コミュニティ（ID: 83551）はラベルからショッピング関連のコミュニティと分かり，ラベルを伏せている4番目のコミュニティ（ID: 92480）はチャイルドシートベンダのコミュニティである．以後このコミュニティをチャイルドシートベンダのコミュニティ（Y）とする．

この結果ではセッション中にオークションとショッピングサイトの両者を同時に訪れたかを判断できない．そこで，共起数を洞察することにより複数個のコミュニティをセッション内で同時に訪問しているかを調べることも可能となっている．共起数を 2 にした例を図 8 (b) に示す．すでに 1 つのコミュニティ（この例ではチャイルドシートベンダのコミュニティ（X））を指定しているため，これを含めると解析結果は 3 つのコミュニティを同一セッションで訪問する頻出パターンを示していることとなる．

この結果から検索語にチャイルドシートと入力しベンダのコミュニティ（X）を訪れるユーザは，

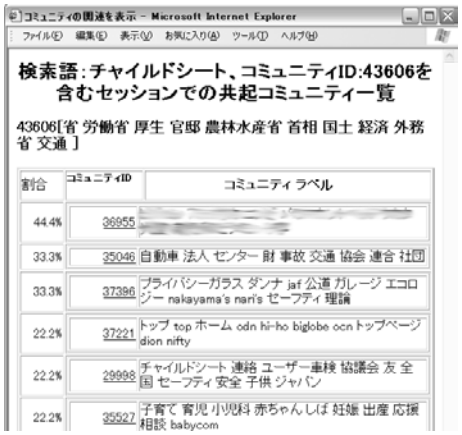
- オークションサイト＋ショッピングサイト
- 複数のショッピングコミュニティ（ショッピングのコミュニティ＋ショッピングサイト）
- ベンダのコミュニティ（Y）＋ショッピングサイト
- ベンダのコミュニティ（Y）＋オークションサイト

を同時に訪れる可能性が高いことが分かる．

次に，「チャイルドシート」の入力があり行政関連のコミュニティ（コミュニティID: 43606）を訪れるセッションについての解析結果を**図 9** に示す．図から行政関連に流入するユーザはチャイルドシートベンダのコミュニティ（X）や自動車の協会と思われるコミュニティを訪れる可能性が高いことが分かる．図中のコミュニティの中で実際に訪れたページを調べた結果，2 番目のコミュニティは「自動車事故対策センター」，3 番目は「JAF」である．図 8 (a) とは異なり，オークション，ショッピングサイトを訪れるセッションはほとんどないという特徴が明らかになった．

6.1.5 連続的な解析の支援

本システムでは解析者が解析結果の中で興味を持つ事柄について新たな解析を即座に行うことができる．たとえば，図 5 (a) の解析結果の中で興味があるコミュニティのチェックボタン（図中 (3)）を押すと上部の入力フォームにその値が代入される（図中 (4)）．ここで，図中 (5) からコミュニティに流入するために使用



コミュニティID:36955はチャイルドシートベンダのコミュニティ(X)

図 9 「チャイルドシート」と入力し行政関連のコミュニティを含むセッションでの共起コミュニティの表示例

Fig. 9 The list of co-occurrence of Web community in the session with search words 'child car seat' and community 'administrative organs'.

した検索語の解析を行うことができ，今回の例では特徴ある事例を示すためにコミュニティIDは異なるが図 6 のような結果が表示される．同様に，図中 (6), (7) を押すと新たな解析が行われ，図 7，図 8，図 9 はこのような操作を行った解析結果である．また，セッション表示ボタン（図中 (8)）を押すと実際のユーザの行動を閲覧できる．

このように，使用した検索語と訪れたコミュニティという観点からユーザの挙動をインタラクティブに解析する環境を実現している．

6.2 本システムの成果

この節では，本システムの利用から把握できた大域的なユーザ行動について述べる．次に，本システムで発見した事例の一部を示し，最後にウェブコミュニティ利用による利点について述べる．

6.2.1 本システム利用による大域的なユーザ行動の把握

図 5 (a) から「チャイルドシート」と入力したユーザは，

- チャイルドシートベンダのコミュニティ
- 行政関連のコミュニティ
- ショッピングサイト

を訪れるユーザが多く，この検索語との関連が深いことを解析者が発見できる．また，6.1.2 項で示したコミュニティに流入するために用いた検索語の結果からユーザがそのコミュニティを訪れる目的（検索語）や，競合するページ間での目的の違いについて理解できる．コミュニティ間の関連については 6.1.3 項で示した流入・流出コミュニティの解析から，大域的なユーザの行動を把握できる．

さらに，「チャイルドシート」と入力したユーザが流入したコミュニティの解析結果（図 5 (a)）とセッション中の共起コミュニティの解析結果（図 8，図 9）から，解析者は**図 10** に示すような大域的なユーザの行動パターンを抽出できる☆．チャイルドシートの使用期間は短いためオークションなどで中古品を探すユーザが多く，同時にチャイルドシートベンダとショッピングサイトで性能と販売価格の調査を行う傾向がある．一方，行政関連のコミュニティを訪れるユーザはベンダや JAF などを含むコミュニティを訪れることから，チャイルドシートの安全性などの調査が目的だと推測できる．

6.2.2 本システムの利用により発見された事例の紹介

本システムを用いて発見した事例の一部を**図 11** に示す．図 11 (a) は「新幹線」と入力したユーザについての解析結果である．時刻表コミュニティへの流入など解析者が容易に想像できる結果のほかに「新幹線の建設推進，鉄道ファン，駅弁集」など興味深いコミュニティ間遷移を見い出すことができた．

図 11 (b) は検索語が「時刻表」の例である．ほとんどの場合は時刻表関連のコミュニティに流入するが，共起コミュニティの解析を行った結果，関東・関西など地域色が強い動きが発見された．最後に検索語が「阪神」の例を図 11 (c) に示す．「阪神」という単語は球団名，鉄道名，高速名とその意味が多岐にわたる．検索語に阪神と入力した後に流入するコミュニティの解析結果から，ユーザがどのような意図で検索語を入力したのかを把握できた．

6.2.3 ウェブコミュニティ利用による利点

図 5 (a) の解析結果にはコミュニティのラベルが表示され，2 番目のコミュニティのラベルには，労働省，外務省などがある．これらは省庁の名前であり行政関連のコミュニティであると推測できる．ラベルからコミュニティの内容が推測できない場合でも詳細ボタンから内容を理解できる．このように，ウェブコミュニティは解析結果の理解に役立つ．

また，チャイルドシートベンダのコミュニティの詳

☆ 図 10 自体はユーザの全体的な挙動をまとめて概観するべく人手で描いたものであるが，個々の流動たとえばチャイルドシートを検索語として入力した後 12%の割合でベンダコミュニティへ，また 10%の割合で行政関連のコミュニティへ流出するという解析結果は本システムにより直接得ることができる．

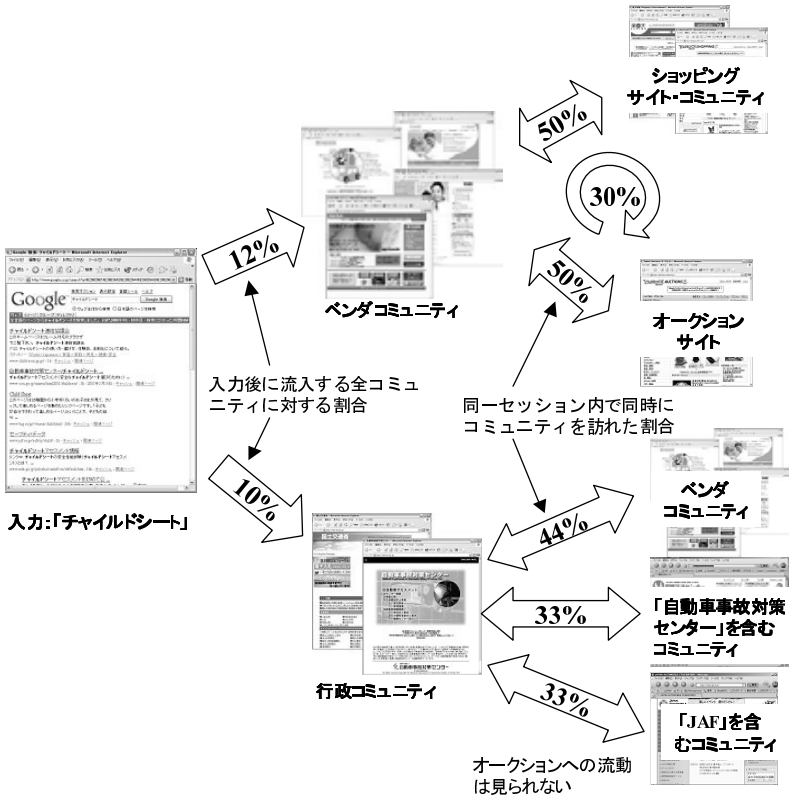


図 10 「チャイルドシート」と入力したユーザの行動
Fig. 10 The users' behaviors with input 'child car seat'.

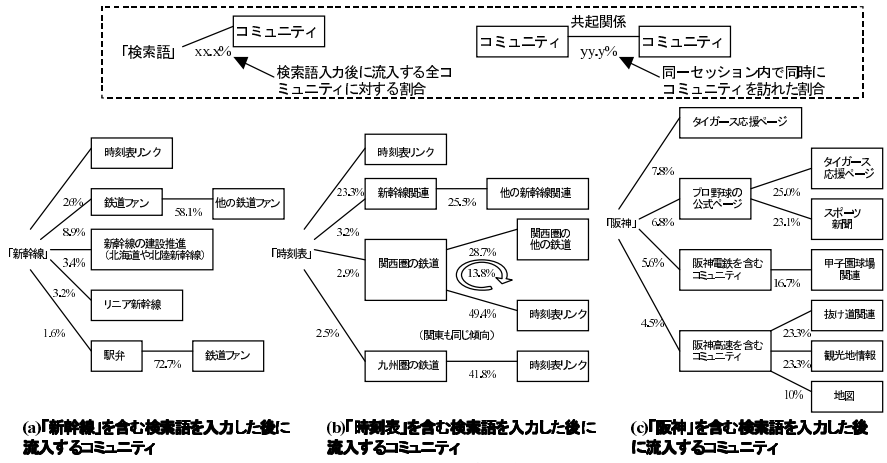


図 11 ユーザ行動の発見例
Fig. 11 The other examples of users' behaviors.

細（図 5 (b)）を見ると，実際にユーザが閲覧したページは分散しており各々のページの頻度はあまり高くない．URL を基に解析を行った場合は国土交通省のページが最も多く，ベンダに訪れた人が多いという情報は発見できない．このように，ウェブコミュニティの利用により URL を基にした解析ではとらえ難いユーザの行動パターンを把握できた．

6.3 本システムの有効性と今後

図 6 ではコミュニティに属するメンバそれぞれを検索語によって特徴付けた例を示したが，同様に流入・流出コミュニティにより特徴付けを行うことができる．たとえば競合する企業への流入経路を解析し，対象と

するサイトに対して流入を増加させる工夫を考えるためのツールとして本システムは有効なことが分かった．ブランドが確立している場合には検索語よりもむしろ流入経路の特徴付けがより有効であることが実験により判明した．このように，ウェブ空間全体の参照ログであるパネルログの解析により，従来のサーバログの解析とは異なる種々のマーケティングのための有益な情報が得られる．

サイト運営者にとって訪問者がどのような検索語を用いて自サイトに至ってくるか，どのような遷移をしながら自サイトに到達するか，その際どのような他サイトも同時に訪問しているかは有用な情報である．本システムにより膨大なパネルログからその行動特性をある程度把握することはできたが，実際の利用局面では解析結果を基に具体的な対応策が求められる．そのような取扱いについては本論文の範囲外であるが，今後さらに研究を進めていきたいと考えている．その過程における本システムのフィードバックも期待される．今後パネルログに対するより高度な解析手法の開発が望まれる．

7. おわりに

パネルログは非常に多くの URL を含むため，URL を基にした解析結果からユーザ行動の把握は容易でない．そこで，本論文ではウェブコミュニティを利用した解析手法を提案した．また，我々はユーザが入力した検索語にも着目し，ウェブ上でのユーザの行動をコミュニティならびに検索語のいずれからも柔軟に解析可能なシステムの構築を行った．システムの利用例からユーザの大域的な行動や特徴のある行動を把握でき，提案システムはパネルログからユーザの行動パターンを抽出するのに有効であった．

謝辞 本研究の一部は，文部科学省科学研究費特定領域研究（C）課題番号：13224014 による．ここに記して謝意を表します．

本研究を進めるにあたりご協力いただいた東芝ソリューション株式会社 SI 技術開発センター平井潤様に，また，論文中で利用したデータの提供にご協力いただいた株式会社ビデオリサーチネットコム社に深謝いたします．

参考文献

1) Batista, P. and Silva, M.J.: Mining on-line newspaper web access logs, *12th International Meeting of the Euro Working Group on Decision Support Systems* (*EWG-DSS 2001*) (May 2001).
2) Beeferman, D. and Berger, A.: Agglomerative clustering of s earch engine query log, *The 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD2000*) (Aug. 2000).
3) Catledge, L. and Pitkow, J.E.: Characterizing browsing behaviors on the world-wide web, *Computer Networks and ISDN Systems*, Vol.27, No.6 (1995).
4) Cooley, R., Mobasher, B. and Srivastava, J.: Web mining: Information and pattern discovery on the world wide web, *Proc. 9th IEEE International Conference on Tools with Artificial Intelligence* (*ICTAI'97*) (Nov. 1997).
5) Flake, G.W., Lawrence, S., Lee Giles, C. and Coetzee, F.M.: Self-organization and identification of web communities, *IEEE Computer*, Vol.35, No.3, pp.66–71 (2002).
6) Fu, Y., Sandhu, K. and Shih, M.: Clustering of web users based on access patterns, *Proc. 1999 KDD Workshop on Web Mining* (*WEBKDD'99*) (Aug. 1999).
7) Kleinberg, J.M.: Authoritative sources in a hyperlinked environment, *Proc. ACM-SIAM Symposium on Discrete Algorithms* (1998).
8) Koutsoupias, N.: Exploring web access logs with correspondence analysis, *Methods and Applications of Artificial Intelligence, 2nd Hellenic* (Apr. 2002).
9) Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A.: Trawling the web for emerging cyber-communities. *Proc. 8th WWW Conference*, pp.403–416 (1999).
10) 村田剛志：Web コミュニティ，情報処理，Vol.44, No.7, pp.702–706 (2003).
11) Nanopoulos, A., Manolopoulos, Y., Zakrzewicz, M. and Morzy, T.: Indexing web access-logs for pattern queries, *4th ACM CIKM Nternational Workshop on Web Information and Data Management* (*WIDM2002*), pp.63–68 (Nov. 2002).
12) Ohura, Y., Takahashi, K., Pramudiono, I. and Kitsuregawa, M.: Experiments on query expansion for Internet yellow page services using web log mining, *The 28th International Conference on Very Large Data Bases* (*VLDB2002*) (Aug. 2002).
13) Pramudiono, I., Shintani, T., Takahashi, K. and Kitsuregawa, M.: User behavior analysis of location aware search engine, *Proc. International Conference On Mobile Data Management* (*MDM'02*), pp.139–145 (Jan. 2002).
14) Prasetyo, B., Pramudiono, I., Takahashi,

K. and Kitsuregawa, M.: Naviz: Website navigational behavior visualizer, *Advances in Knowledge Discovery and Data Mining 6th Pacific-Asia Conference* (*PAKDD2002*) (May 2002).
15) Shahabi, C., Zarkesh, A.M., Adibi, J. and Shah, V.: Knowledge discovery from users webpage navigation, *Proc. IEEE RIDE97 Workshop* (Apr. 1997).
16) Su, Z., Yang, Q., Zhang, H., Xu, X. and Hu, Y.: Correlation-based document clustering using web logs, *34th Hawaii International Conference on System Sciences* (*HICSS-34*) (Jan. 2001).
17) Tan, P. and Kumar, V.: Mining association patterns in web usage data. *International Conference on Advances in Infrastructure for e-Business, e-Education, e-Science, and e-Medicine on the Internet* (Jan. 2002).
18) Toyoda, M. and Kitsuregawa, M.: Creating a web community chart for navigating related communities, *Conference Proceedings of Hypertext 2001*, pp.103–112 (2001).
19) Ungar, L.H. and Foster, D.P.: Clustering methods for collaborative filtering, *AAAI Workshop on Recommendation Systems* (July 1998).
20) Wen, J., Nie, J. and Zhang, H.: Query clustering using user logs, *ACM Trans. Info. Syst.* (*ACM TOIS*), Vol.20, No.1, pp.59–81 (2002).
21) Zaiane, O.R., Xin, M. and Han, J.: Discovering web access patterns and trends by applying olap and data mining technology on web logs, *Proc. Advances in Digital Libraries* (*ADL'98*) (Apr. 1998).
22) Zeng, H., Chen, Z. and Ma, W.: A unified framework for clustering heterogeneous web objects, *3rd International Conference on Web Information Systems Engineering* (*WISE2002*) (Dec. 2002).

(平成 15 年 6 月 20 日受付)
(平成 15 年 10 月 6 日採録)

（担当編集委員　福島 俊一）

大塚　真吾（正会員）

1996 年千葉工業大学工学部情報工学科卒業．2002 年同大学大学院工学研究科情報工学専攻博士後期課程修了．工学博士．同年東京大学生産技術研究所学術研究支援員．ログマイニング，テキスト処理，ウェブマイニングに興味を持つ．

豊田　正史（正会員）

1994 年東京工業大学理学部情報科学科卒業．1999 年同大学大学院情報理工学研究科博士後期課程修了．理学博士．同年科学技術振興事業団計算科学技術研究員．2001 年東京大学生産技術研究所学術研究支援員．2003 年同大学産学官連携研究員．ウェブマイニング，ユーザインタフェース，ビジュアルプログラミングに興味を持つ．ACM，IEEE CS，日本ソフトウェア科学会各会員．

喜連川　優（正会員）

1978 年東京大学工学部卒業．1983 年同大学大学院工学系研究科情報工学博士課程修了．工学博士．同年同大学生産技術研究所講師．現在，同教授．2003 年より同所戦略情報融合国際研究センター長．データベース工学，並列処理，Web マイニングに関する研究に従事．現在，本学会理事，日本データベース学会理事，1999 年〜2002 年 ACM SIGMOD Japan Chapter Chair 1997 年，1998 年電子情報通信学会データ工学研究専門委員会委員長．VLDB Trustee（1997〜2002），IEEE ICDE，PAKDD，WAIM 等ステアリング委員．