ウェブアクセスログ解析による大域的ユーザ行動パターンの抽出

○大塚 真吾，喜連川 優
（東京大学 生産技術研究所 戦略情報融合国際研究センター）

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，さまざま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全ての URL を把握できる．このようにして集められたログを本稿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捉えるためにウェブコミュニティ[1]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，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と構築を行う．

2　パネルログとウェブコミュニティ

パネルログ　本稿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．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してユニークに割り当てられた ID であり，

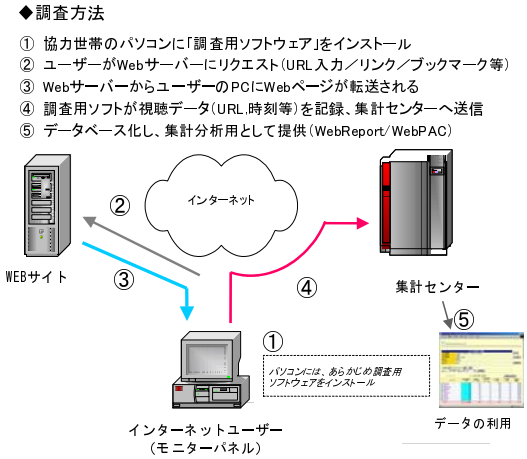

図 1: パネルログ収集の概要

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．

今回の実験で利用したパネルログのパネルは全て日本人であり，その詳細を表 2 に示す．表中の「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り，アクセスログの解析ではこの概念を導入している．本稿ではセッションを「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[2]．

ウェブコミュニティ　ウェブ全体をグラフ構造とみ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 [1]
2. 完全二部グラフを抽出する手法 [3]

の 2 つに大別される [4]．また，本稿では文献 [4] 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

[1]以降,「コミュニティ」は「ウェブコミュニティ」の意味で使用

[2]実験では，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した

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/kousikiyo tei.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)

表 1: パネルログの詳細

表 2: パネルログの概要

データ量	約 10Giga byte
データ収集期間	45 週間
アクセス数	約 5,500 万アクセス
セッション数	約 100 万セッション
パネル数	約 1 万人
パネルの抽出法	Random Digit Dialing 方式
検索語の種類	約 30 万種類

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．

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[2] 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．

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 [7] 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．本稿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 パネルログ解析のためのウェブコミュニティの利用

パネルログは極めて多くの 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 の解析だけでは捉え難い現象を発見できると考えている．例えば，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，充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．

また，パネルログの予備的な解析から，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ことがわかった．そこで，我々が提案するシステム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．さらに，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，且つ，これらのサイトの内容はコミュニティを用いなくとも理解できるため，提案システムでは 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」とした．

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掛かりになると考えている．

4 パネルログ解析システム

我々はウェブコミュニティの技術とユーザが検索エンジンサイト検索語を利用したパネルログ解析システムの構築を行った．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

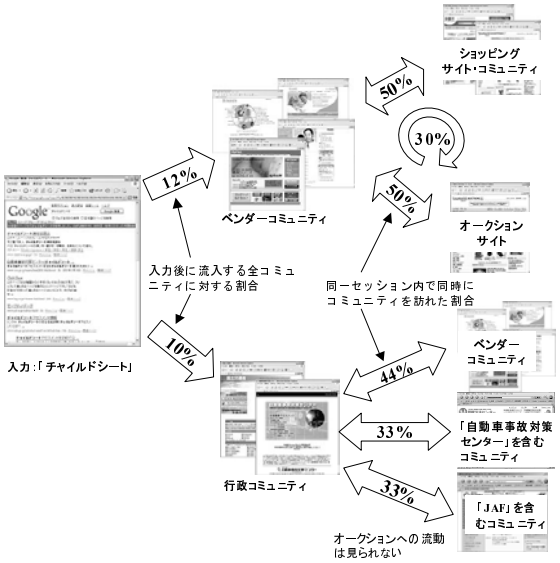

図 2: 「チャイルドシート」と入力したユーザの行動

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，コミュニティを訪れた理由を知ることができる．(3) と (4) はコミュニティ間の関連を解析する機能であり，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理解できる．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．その他に解析範囲の設定などがある．

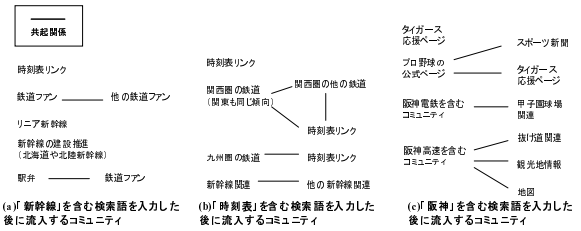

図 3: ユーザ行動の発見例

4.1 本システム利用による大域的なユーザ行動の把握

本システムの利用から「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ーのコミュニティ
- 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が発見できた．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とセッション中の共起コミュニティの解析結果から，我々は図 2 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た[3]．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ー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ーや 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．

4.2 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 3 に示す．図 3(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出すことができた．

図 3(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

[3]図 2 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，個々の流動例えばチャイルドシートを検索語として入力した後 12%の割合でベンダーコミュニティへ，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．

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 3(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渡る．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.

4.3 ウェブコミュニティ利用による利点

チャイルドシートベンダーのコミュニティの詳細を見ると，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，ベンダー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捉え難いユーザの行動パターンを把握できた.

その他にも，例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わかった．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到ってくるか，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もとに具体的な対応策が求められ，今後更に研究を進めていきたいと考えている.

5 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，本稿ではウェブコミュニティを利用した解析手法についての提案を行った．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)「ウェブマイニングの為のウェブウェアハウス構築に関する研究」（課題番号：13224014）による．ここに記して謝意を表します．また，本研究を進めるにあたり御協力頂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，実験で利用したデータの提供に御協力頂いた株式会社ビデオリサーチインタラクティブに深謝致します.

参考文献

[1] G.W. Flake, S. Lawrence, C. Lee Giles, and F.M. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, Vol. 35, No. 3, pp. 66–71, 2002.

[2] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. In *In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[3] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *Proc. of the 8th WWW conference*, pp. 403–416, 1999.

[4] 村田剛志. Web コミュニティ. 情報処理, Vol. 44, No. 7, pp. 702–706, 2003.

[5] 大塚真吾, 豊田正史, 喜連川優. ウェブコミュニティを用いた大域 web アクセスログ解析法の一提案. 情報処理学会論文誌：データベース, Vol. 44, No. SIG18(TOD20), pp. 32–44, 12 2003.

[6] S. Otsuka, M. Toyoda, J. Hirai, and M. Kitsuregawa. Extracting user behavior by web communities technology on global web logs. *Proc. of 15th International Conference on Database and Expert Systems Applications (DEXA'2004)*, pp. 957–968, 9 2004.

[7] M. Toyoda and M. Kitsuregawa. Creating a web community chart for navigating related communities. In *Conference Proceedings of Hypertext 2001*, pp. 103–112, 2001.

要旨

ウェブページを閲覧する人々の行動パターン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，サイト外を含めユーザの大域的な行動を解析することは困難であった．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，パネルから集められたログ（ウェブ視聴率データ）の解析により，パネルが訪れた全てのウェブページ（URL）を収集することが可能となった．本講演ではウェブ視聴率データを用いてた大域的なユーザの行動パターンを発見する技術について述べる．