2003－DBS－131　(84)
2003／7／18

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

大塚 真吾†　豊田 正史†　喜連川 優†

† 東京大学 生産技術研究所

要 旨

ウェブページを閲覧する人々の行動モデルを抽出すること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全ての行動を解析することは容易ではない．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．パネルを使って集められたログ（パネルログ）を解析することにより，ユーザが訪れた全てのウェブページ（URL）を収集することができる．パネルログは極めて広いページ空間を対象とするため，個々のページの参照履歴を見るだけでは大域的な行動を捉えることは難しい．そこで，本論文では類似したウェブページをまとめる手法であるウェブコミュニティ技術を用いることにより，大域的なユーザの行動パターンを把握すべく実験システムの構築を行った．

The System of Panel Logs Analysis with Web Communities

Shingo Otsuka†　Masashi Toyoda†　Masaru Kitsuregawa†

†Institute of Industrial Science, The University of Tokyo

Abstract

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. Here, by utilizing Web communities technique which clusters similar web pages into groups, a prototype system which extract users' behavior access patterns globally is proposed.

1　はじめに

ウェブ上でのユーザの行動を解析すること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 を知ることができる．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ユーザの行動を理解することは難しい．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]の技術を用いる．また，パネルログの統計的な解析から検索語の重要性を示す．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の行動を把握することを支援するシステムの提案を行う．

2　関連研究

アクセスログを用いた研究は今まで数多く行われており，その目的も様々である [4]．以下に主な研究分野を示す．

- ユーザの行動に関する研究 [9, 1, 13]
- ウェブページ間の関連に関する研究 [10, 11, 15]
- 検索サイトに関連する研究 [2, 14, 7]
- アクセスログの視覚化に関する研究 [6, 8]

従来の殆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 [16] はプロキシーサーバのアクセスログ用いておりやや類似するがログを用いることにより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．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詳細な研究は行われていない．

[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

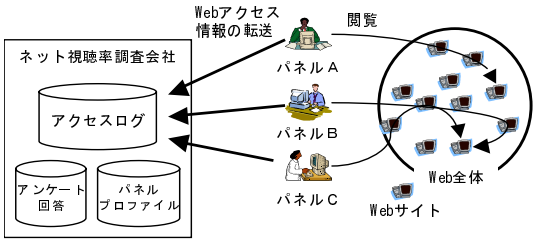


図 1: パネルログ収集の概要

3　パネルログとウェブコミュニティ

この節では本論文で提案するシステムを構築する上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．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てアクセスログを収集する．

図中のアンケート回答とパネルプロファイルには個人情報が含まれるため本論文では利用しない．このように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，

- ユーザ 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

などから構成されている．ユーザ 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 である．また，表 1 の (a) は検索語の情報を含む URL である．

セッションの抽出に関しては，パネルログはユーザの特定が容易なためユーザごとの行動解析が可能である．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のことである．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

表 1: パネルログの一部

UserID	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

(a)

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

ウェブコミュニティに関する研究の多くはハブとオーソリティの概念に基づいている．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ってい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．ウェブコミュニティを作成するにはウェブページのリンク解析によってハブとオーソリティを抽出する必要があり HITS[5] はこれらを効率良く抽出するアルゴリズムである．図 2 に HITS によって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って密に結合されている．

本論文では HITS を利用して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手法であるウェブコミュニティチャート [12] を利用する．この手法はコミュニティ間の関連性を考慮しているため，その構造は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である．また，この手法では 1 つの URL は 1 つのコミュニティのみに属する．本論文ではコミュニティ間の関連度を必要としないため，コミュニティ部分のみ利用する．

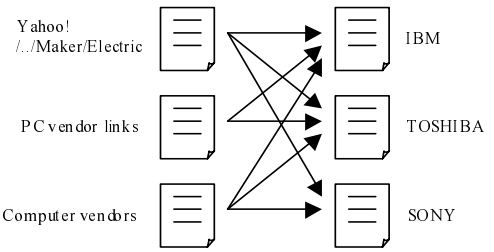


図 2: ハブとオーソリティからなる典型的なグラフ

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

パネルログは非常に多くの URL を含むため，URL ベースの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．そこで，我々はウェブコミュニティのような URL の集まりを利用することで，URL ベースの解析よりも抽象度が高い解析結果が得られると考えている．また，ウェブコミュニティの利用により URL ベースの解析では捉え難い現象を発見できると考えている．例えば，コミュニティに含まれる個々の URL の頻度は低いがその総数はある程度高くなる場合，コミュニティを用いた解析でなければユーザの行動パターンを発見することは難しい．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を解析することで，コミュニティの内容を表す単語群（コミュニティラベル）を自動的に抽出することが可能である．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見ることなくコミュニティの概要を把握することが可能となる．

我々はパネルログの解析にウェブコミュニティの概念を取り入れることで解析結果を直感的に理解することができ，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 パネルログの基本的性質

パネルログを用いた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掴むために統計的な解析を行った．

表 2: パネルログの概要

総データ量	9,992（Mbyte）
今回利用したデータ量	2,377（Mbyte）
アクセス数	55,415,473（アクセス）
セッション数	1,148,093（セッション）
URL の種類	7,776,985（種類）
検索語の種類	334,232（種類）

表 3: 全アクセスの中でウェブコミュニティに含まれる URL の割合

無修正	18.8%
ディレクトリ（ファイル）部分を削除して合致	37.8%
サイト部分を削除して合致	7.7%
合致せず	35.7%

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

我々は 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を 17 万個のコミュニティに自動分類した．一方，今回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．パネルログの中には解析に利用しない情報があり，それらを除いたデータ量は 2 ギガ程度となる．ウェブページの収集時期はパネルログ収集期間中のため，パネルがアクセスしたウェブページの変更や削除が行われている可能性がある．そこで，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 の適合度を測定し，その結果を表 3 に示す．無修正時における適合率はおよそ 20%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全体の約 40%をカバーした．また，サイト名を削除する処理[2]により適合率が 8%程度向上した．このように URL の修正により全アクセスの約 65%をカバーした．

[2]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っていない

表 4: 検索語を抽出した検索エンジン（ポータル）サイト

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		

表 5: 全アクセスに対する検索エンジン群などの割合

1. 検索語を入力した検索エンジン群	4.1%
2.(1) の後に訪れたページ	11.8%
3. 検索語がない検索エンジン群	19.7%
4.(3) の後に訪れたページ	35.1%
5. 検索エンジン群の前に訪れたページ	6.4%
6. 検索エンジン群に行かないセッションのアクセス数	10.4%
7. オークションサイト	11.1%
8.Yahoo!のショップ or 楽天	1.5%

4.2 パネルログに含まれる検索エンジン・ディレクトリサイトの割合

インターネット視聴率会社[3]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，ディレクトリサイト，ポータルサイトが常に上位であるため，我々はユーザの行動にこれらのサイトが深く関与していると考えた．検索エンジンなどのサイトは多数存在するが，ここでは表 4 に示すサイトを対象として[4]パネルログ中に含まれるアクセス数を測定した．また，3.1 節で述べたようにパネルログ中の URL は検索語に関する情報を含むため検索語入力の有無についても調べ，その結果を表 5 に示す．パネルログ中の全アクセスに対する検索語入力の割合は 4.1%と低いが，検索語がないものは約 20%ある．また，これらのサイトから訪れたと思われるアクセスなどを含めた値（表 5 の 1〜4）は全体の 70%と

[3]http://www.vrnetcom.co.jp/など

[4]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から除外している．

非常に多い．さらに，全セッションを調べた結果，約 25%のセッションで検索語の入力があった．

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．検索語はユーザの目的となるためユーザ行動を捉える手助けとなると考えられる．そこで，5 節で我々が提案するシステムに検索語の解析を取り入れる．

5　ウェブログ解析システムの提案

この節では大域的なユーザ行動を捉えるためのパネルログ解析システムの提案を行う．

5.1　本システムの解析機能

本論文ではウェブコミュニティと検索語の切り口からパネルログの解析を行うことで大域的なユーザの行動を把握するシステムの構築を行う．また，解析結果の中で解析者が指定したウェブコミュニティについて URL レベルで結果が表示できる．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）で管理するため，指定した URL が属するコミュニティID を検索する機能を有する．主な解析機能を以下に示す．

- **検索語入力後に流入したコミュニティの表示**
  任意の検索語を入力すると，その検索語で検索を行ったユーザが訪れたコミュニティの一覧を表示する．
- **コミュニティに流入するために使用した検索語の表示**
  任意のウェブコミュニティを指定すると，そのコミュニティを訪れるために用いた検索語の一覧を表示する．
- **流入・流出コミュニティの表示**
  任意のウェブコミュニティを指定すると，そのコミュニティを訪れる前後のコミュニティの一覧を表示する．

5.2　システムの構成

本システムではウェブブラウザを利用して解析を行う．ウェブブラウザ上で入力された問合せはウェブサーバを介してパネルログ解析サーバに送られる．

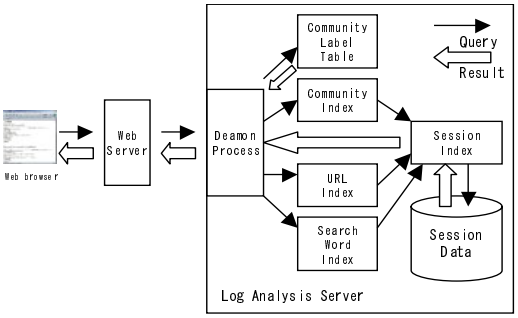


図 3: システム構成図

ログ解析システムの構成を図 3 に示す．パネルログの実データはセッション毎にまとめてからディスクに格納する（Session Data）．Session Data へのアクセスは Session Index を介して行う．Community Label Table は URL からコミュニティID の検索を行う場合や，コミュニティID に対応する Community ラベルを獲得する場合に利用される．

また，本システムでは URL，ウェブコミュニティ，検索語の視点から解析を行うために各々がインデックスを持つ．インデックスはコミュニティID とその ID を含むセッション ID から構成される．このインデックスから得られたセッション ID を用い Session Index にアクセスした後，Session Data からセッション情報を得る．

5.3　システムの実装

本システムでは繰り返し解析を行うことを想定しているため，解析処理を素早く行う必要がある．そこで，パネルログデータをメモリ上に載せることで処理の高速化を図った．我々はコミュニティID，検索語，URL からセッション ID の検索を行う必要があることを考慮して実装を行った．その概要を図 4 に示す．Sesseion Table の配列の大きさはセッション数と同じである．Session Table からリンクされたノードは 1 つのアクセスを示している．各ノードは，

- Search Word Table へのリンク (図 4-a)
- URL Table へのリンク (図 4-b)
- 実データへのポインタ (図 4-c)
- 閲覧時間

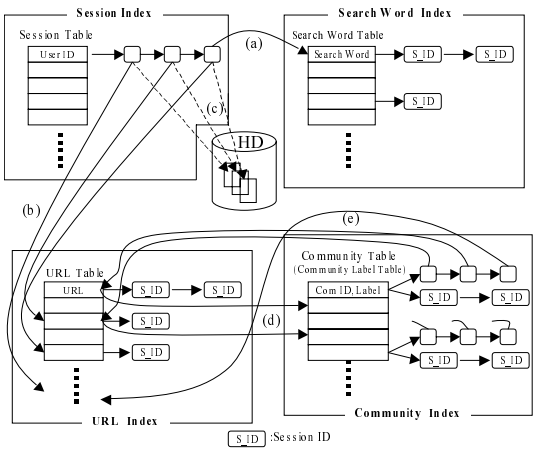


図 4: 実装の詳細

を保持する．また，このノードから URL Table を辿り Community ID を得ることができる．このように，Session Table からリンクされたノードからそのセッションの内容を知ることができる．

Session Table 以外のテーブルからリンクされるノードには，それを利用した Session の ID が格納されており，ノードを辿ることで利用された全セッションがわかる．例えば,「赤ちゃん」という検索語を入力したセッションについて解析を行う場合は，Search Word Table からこの検索語を探し，そこから張られたセッション ID ノードを辿ることで，この検索語が使用された全てのセッション ID を得る．このセッション ID から検索語が使われたセッションの解析を行うことができる．URL Table の中にコミュニティに属する URL がある場合は Community Table へリンクが張られる (図 4-d)．Community Table からはそのコミュニティが属する URL に対してリンクが張られており，コミュニティID からそのコミュニティに属する URL を検索することが可能である (図 4-e)．また，Community Table はコミュニティの内容を表すラベルを保持している．

実際に利用したパネルログのデータ量は約 2.3 ギガであるが，実装を行った結果プログラム中のメモリの使用量は約 3.3 ギガ[5]であった．

[5]UNIX の top コマンドを使用

6 システムの利用例

我々はプロトタイプシステムを実装を行った．本システムの開始画面の一部を図 5(a) に示す．検索語またはコミュニティID の入力と各パラメータの設定を行うだけで容易に解析ができる．

6.1 検索語入力後に流入したコミュニティの表示

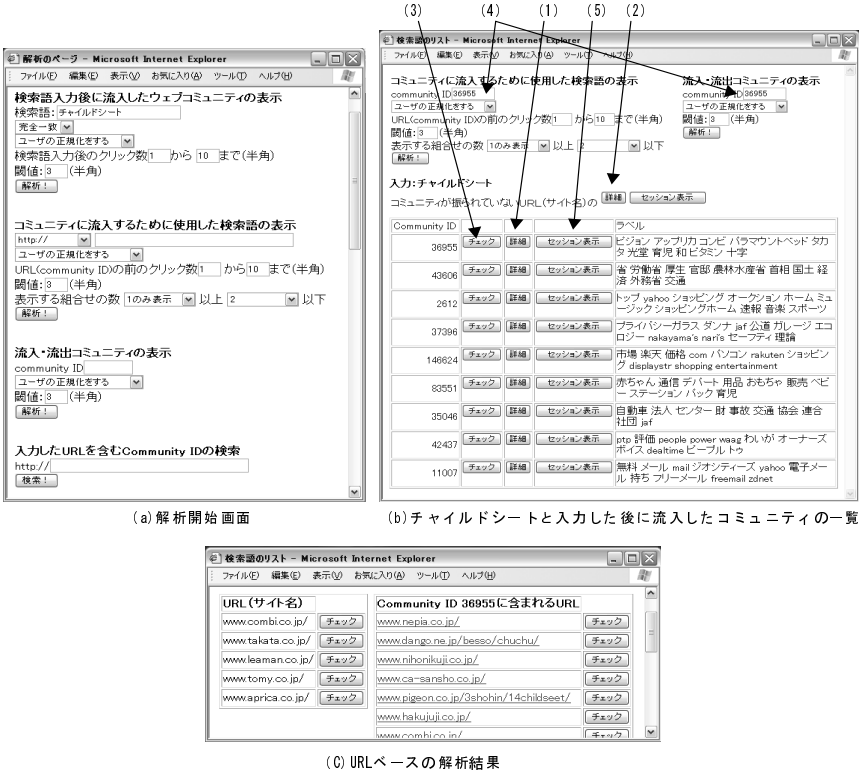
「チャイルドシート」と入力した後に流入したコミュニティの表示例を図 5(b) に示す．解析結果は頻度順に表示される．詳細ボタン (1) を押すとコミュニティに属するページについて URL ベースの結果が閲覧できる（図 5(c)）．また，(2) を押すとコミュニティに属さないページの結果が閲覧できる．解析者が結果の中で興味があるコミュニティのチェックボタン (3) を押すと上部の入力フォームにその値が代入され新たな解析を行うことができる (4)．さらに，セッション表示ボタン (5) を押すと実際のユーザの行動を閲覧できる．その他の機能については，紙面の都合上省略する．

6.2 パラメータの設定

本システムでは，解析時にパラメータの設定が可能である．主なパラメータについて以下に示す．

- ユーザの正規化
- 指定した検索語と完全一致，部分一致
- 検索結果表示の最低値（閾値）
- 検索語とウェブコミュニティの距離

一人のユーザが何回も同じ行動を行った場合は解析結果にそのユーザの行動が大きく反映するため，ユーザの正規化の選択を可能とした．また，ユーザが検索語を入力した後に訪れるコミュニティの解析を行う場合，検索語を入力した何クリック後までを解析対象とするかは解析時の状況によって異なる．そこで，検索語とウェブコミュニティの距離を設定するパラメータを作成した．



(a) 解析開始画面　(b) 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一覧

(C) URLベースの解析結果

図 5: 「チャイルドシート」と入力した後に閲覧したウェブコミュニティの解析例

6.3 本システム利用によるユーザ行動の把握

図 5 では「チャイルドシート」と入力したユーザは,

- チャイルドシートのベンダー
- 行政
- ショッピング
- オークション

のコミュニティを訪れるユーザが多く，この単語との関連が深い．さらに，ユーザの行動を詳しく解析した結果，ユーザは図 6 に示すような行動パターンに大別できる．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ーと e ショップで性能と販売価格の調査を行う傾向がある．一方，行政系のコミュニティを訪れるユーザの目的は「取り付け方，安全基準」などで，ベンダー系，行政，非営利団体などのコミュニティへ流出している．このように，本システムでは解析結果から大域的なユーザの行動を把握することができる．

7 おわりに

本論文ではウェブコミュニティを用いたパネルログの解析システムの提案と実装を行った．システムの実装と利用例から，ユーザの大域的な行動や特徴のある行動を把握でき，システムの有効性を確認できた．

今後はより多くのデータを用いた検証を行いたいと考えている．また，新たな機能の追加によってシステムを充実させる予定である．

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．

本研究を進めるにあたり御協力頂いた株式会社東芝 e-ソリューション社 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御協力頂いた株式会社ビデオリサーチネットコム社に深

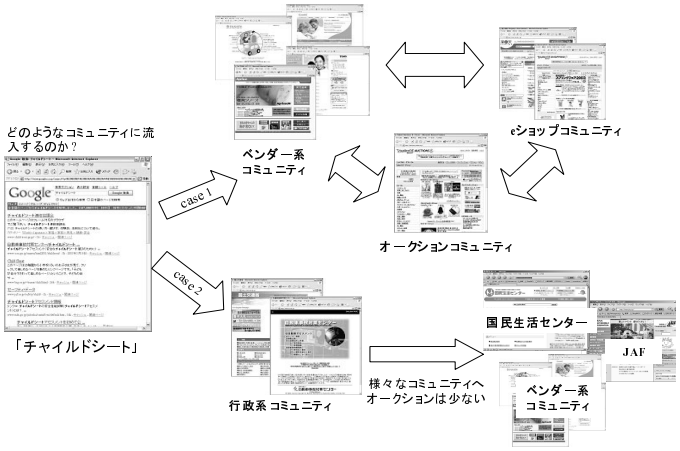


図 6: 「チャイルドシート」と入力したユーザの行動

謝致します.

参考文献

[1] P. Batista and M.J. Silva. 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] D. Beeferman and A. Berger. Agglomerative clustering of s earch engine query log. *The 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2000)*, August 2000.

[3] L. Catledge and J.E. Pitkow. Characterizing browsing behaviors on the world-wide web. *Computer Networks and ISDN Systems*, No. 27(6), 1995.

[4] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web. *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, November 1997.

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

[6] N. Koutsoupias. Exploring web access logs with correspondence analysis. *Methods and Applications of Artificial Intelligence, Second Hellenic*, April 2002.

[7] Y. Ohura, K. Takahashi, I. Pramudiono, and M. Kitsuregawa. Experiments on query expansion for internet yellow page services using web log mining. *The 28th International Conference on Very Large Data Bases (VLDB2002)*, August 2002.

[8] B. Prasetyo, I. Pramudiono, K. Takahashi, and M. Kitsuregawa. Naviz: Website navigational behavior visualizer. *Advances in Knowledge Discovery and Data Mining 6th Pacific-Asia Conference (PAKDD2002)*, May 2002.

[9] C. Shahabi, A.M. Zarkesh, J. Adibi, and V. Shah. Knowledge discovery from users web-page navigation. *In Proceedings of the IEEE RIDE97 Workshop*, April 1997.

[10] Z. Su, Q. Yang, H. Zhang, X. Xu, and Y. Hu. Correlation-based document clustering using web logs. *34th Hawaii International Conference on System Sciences (HICSS-34)*, January 2001.

[11] P. Tan and V. Kumar. 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*, January 2002.

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

[13] L.H. Ungar and D.P. Foster. Clustering methods for collaborative filtering. *AAAI Workshop on Recommendation Systems*, July 1998.

[14] J. Wen, J. Nie, and H. Zhang. Query clustering using user logs. *ACM Transactions on Information Systems (ACM TOIS)*, Vol. 20, No. 1, pp. 59–81, January 2002.

[15] O.R. Zaiane, M. Xin, and J. Han. Discovering web access patterns and trends by applying olap and data mining technology on web logs. *in Proc. Advances in Digital Libraries (ADL'98)*, April 1998.

[16] H. Zeng, Z. Chen, and W. Ma. A unified framework for clustering heterogeneous web objects. *The Third International Conference on Web Information Systems Engineering (WISE2002)*, December 2002.