社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

# 大域ウェブアクセスログとウェブコミュニティを用いたトピックに関連する検索語群の発見法

大塚　真吾†　　喜連川　優†

† 東京大学 生産技術研究所
E-mail: †{otsuka,kitsure}@tkl.iis.u-tokyo.ac.jp

**あらまし**　サイバー空間上では多くの人々が自分の欲しい情報を探すために検索サイトを利用する．検索技術の進歩により検索精度は向上し，自分が調べたい事柄を検索語として入力するだけで様々な情報を得ることが可能となった．多くの人が入力した検索語を解析することで，世の中のニーズや動向，流行などを把握することが可能である．そこで，本稿ではテレビ視聴率調査と同様，統計的に偏りなく抽出された日本人（パネル）を対象に URL 履歴の収集を行う大域ウェブアクセスログ（パネルログ）とウェブコミュニティの技術を用いてトピックに関連する検索語群の発見法の提案を行い「あるトピックにおけるユーザの興味」を把握するツールの実装を行う．
**キーワード**　関連語発見，検索語クラスタリング，ウェブアクセスログマイニング，ウェブコミュニティ

# A Method for Finding Search Keywords Associated with Topics using Global Web Access Logs and Web Communities

Shingo OTSUKA† and Masaru KITSUREGAWA†

† Institute of Industrial Science, The University of Tokyo
E-mail: †{otsuka,kitsure}@tkl.iis.u-tokyo.ac.jp

**Abstract**　In the cyberspace, users search for the desired information through search engine. With the improvement of searching accuracy due to the advancements in technologies, it has become possible for users to obtain kinds of information by inputting just search word(s) representing the topics which they are interested in. At the same time, it has become possible to grasp the needs, trends and fashions of the world by analyzing the search keywords used by the users. In this paper, by using web community technique and the web access logs (called panel logs), which are the collected URL histories of Japanese users (called panels) selected without statistic deviation similar to the survey on TV audience rating, we propose a method for finding search words associated with topics and introduce a tool to graph "a user's interest in a topic".
**Key words**　Related keywords finding, Search keywords clustering, Web access logs mining, Web community

## 1. はじめに

サイバー空間上では多くの人々が自分の欲しい情報を探すために検索サイトを利用する．検索技術の進歩により検索精度は向上し，自分が調べたい事柄を検索語として入力するだけで様々な情報を得ることが可能となった．多くの人が入力した検索語を解析することで，世の中のニーズ，興味，動向，流行などを把握することが可能である．

最近では，検索エンジンやポータルサイトの検索語の解析から今後の流行を発見することが可能となり，実際に検索サイトなどでは検索数が急上昇した検索語を「注目キーワード」として公開している．また，テレビ番組や雑誌の記事などではジャンル別の検索語ランキングや急上昇した検索語の集計を行い，その結果が公表されている．しかし，これらの情報は検索数が一定期間に急上昇した検索語を対象としているため，銀行，大学，温泉など普段から利用されている検索語やオリンピック，ワールドカップ，万博など長い期間開催されているものについてはユーザの興味を調べることは困難である．

そこで，本稿ではテレビ視聴率調査と同様，統計的に偏りなく抽出された日本人（パネル）を対象に URL 履歴の収集を行う大域ウェブアクセスログ（パネルログ）とウェブコミュニティの技術(注1)を用いてトピックに関連する検索語群の発見法

(注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

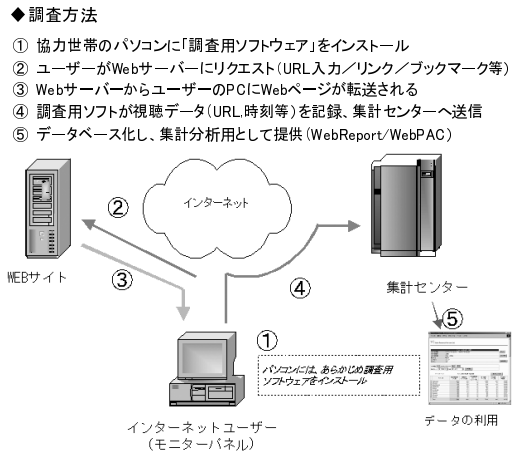



図 1　パネルログ収集の概要

の提案を行い「あるトピックにおけるユーザの興味」を把握するツールの実装を行う.

## 2. 関 連 研 究

検索語に関連する研究はその成果がビジネスに直結するため外部に公開される機会が少なく，またデータの入手が困難であるなどの理由から研究成果はあまり公開されていない．文献 [7] では NTT DIRECTORY で入力された検索ログを用いて,「桜と花見」など時期に依存した類似性の抽出を行っている．この研究ではある一定の期間に於ける検索語の頻度や入力間隔を基に同義語の抽出を行うため我々の手法とは異なる．また，文献 [11] では検索結果のページに含まれる単語とオーバーチュアのキーワードアドバイスツールを用いて得られた検索語に関連する語を用いて検索語のクラスタリングを行っている．英語圏におけるアクセスログを対象とした検索語の研究に関しては Lycos と Microsoft がそれぞれ発表を行っている [1], [10]．これらの研究ではユーザが検索語を入力した後に閲覧されたディレクトリや URL を用いて検索語の分類を行っている．我々はユーザが閲覧したページの内容解析やウェブコミュニティ技術を利用するため研究手法が異なる.

また，最近では Google, goo, Yahoo!がユーザに対して想定される検索語や絞り込み検索語を提案する「サジェスト」(注2) と呼ばれるサービスを行っている．サジェストは入力中の検索語に対して想定される検索語や絞り込み検索語を提案する機能であり，検索語入力を開始した瞬間から候補語がドロップダウン表示される．候補語の選定方法については詳細な情報は公開されていないが，検索サイト上で頻繁に検索された言葉や，検索結果のリストの中で頻繁にクリックされる URL など，様々な要因を基に選ばれている．また，特定のユーザー，コンピュータ，Web ブラウザからの検索情報は使っていないとされている.

例えば，Google サジェストにおいて「ワイン」と入力する場合，まず「w」を入力すると「winny」「winmx」などが,「a」を入力して「わ」を表示すると「早稲田大学」「早稲田」などが提案され，さらに「わいん」の場合は「ワインセラー」「ワイングラス」が提案される.「ワイン」と変換した後にスペースを入力すると「ワイン　通販」「ワイン　ラベル」など絞り込み検索語が提示される.

前者の部分は検索語の入力の手間を省く事に重点を置くため本研究と目的が異なる．また，後者の「絞り込み検索語の提示」についても本研究はあるトピックに関連する検索語を提示するため目的が異なる.

(注2)：Google サジェストは http://www.google.co.jp/webhp?complete=1&hl=ja, goo サジェスト $\beta$ は http://suggest.search.goo.ne.jp/suggest/index.php，Yahoo! Japan は「入力補助版」という名のサービスを提供している.

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)

表 2　パネルログの詳細
Table 2　A detail of the panel logs.

| | |
|---|---|
| 総データ量 | 9,992 (Mbyte) |
| 今回利用したデータ量 | 2,377 (Mbyte) |
| データの収集期間 | 45 (週間) |
| アクセス数 | 55,415,473 (アクセス) |
| セッション数 | 1,148,093 (セッション) |
| URL の種類 | 7,776,985 (種類) |

## 3. トピックと関連する検索語群の発見に必要な技術の概要

### 3.1 パネルログ

本論文で利用するパネルログの概要を図 1 に示し，その調査方法を以下に示す.

- インターネット視聴率調査会社が所有する全国のインターネットユーザーの調査協力サンプル（パネル）により視聴されたウェブページの情報を収集・集計.
- パネルがインターネット利用に使用するパソコンに調査用ソフトウェアをインストールし，視聴状況をリアルタイムで収集.

このように収集されたパネルログは表 1 に示すように**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秒数，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た ID である．また，表中 (a) のように URL の中には検索サイトなどで入力された検索語についての情報も記録されている．次に我々が利用したパネルログの基本情報を表 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単位であり，本論文では「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」とし，閲覧の終了を「ウェブページを閲覧し終えてから，次の

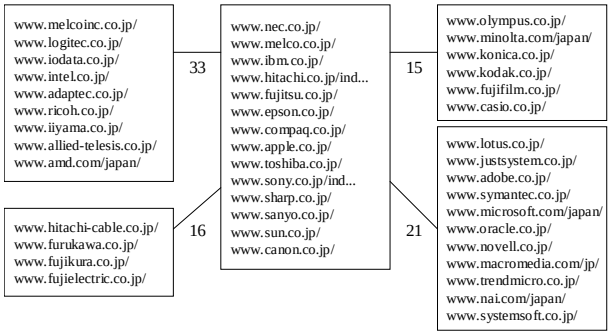



図 2　ウェブコミュニティチャートの一部

ウェブページをアクセスするまでに 30 分以上あるとき」と定義する [2].

### 3.2　ウェブコミュニティ

本論文ではウェブコミュニティを「同じトピックに関心をもつ人々や組織によって作成されたウェブページの集合」という意味で用いる [9]．ウェブコミュニティの例として，同じ業種に属する会社のホームページの集合やあるサッカーチームを応援するホームページの集合などが挙げられる．これまでに WWW をウェブページとその間に張られたハイパーリンクによるグラフと見なし，グラフ構造を解析することでウェブコミュニティを抽出する様々な手法が提案されている [4]〜[6].

本論文ではウェブコミュニティの抽出手法として，我々が提案したウェブコミュニティチャート [9] を用いる．ウェブコミュニティチャートはウェブコミュニティをノードとし，関連するコミュニティの間に重み付のエッジを張ったグラフである．図 2 に我々が作成したウェブコミュニティチャートの一部を示す．エッジの重みはコミュニティ間の関連度を表す．中央に大手コンピュータメーカのコミュニティがあり，その周りに関連するコミュニティとして，ソフトウェア，周辺機器，デジタルカメラなど関連業種の会社のコミュニティが抽出されている．

我々はウェブコミュニティチャートの作成のために以下に示す関連ページアルゴリズム [3], [9] を利用する．

（1） 1 つのシードページを入力として与える．

（2） シードページと近傍するウェブグラフから，良い authority ページおよび良い hub ページを抽出する．

（3） 上位の authority ページを関連ページとして出力する．

ここで良い authority とは多くの良い hub からハイパーリンクを張られている著名なページを表す．良い hub とはリンク集およびブックマークなど多くの良い authority へハイパーリンクを張っているページを表す．この循環した定義により密に結合した hub と authority が抽出され，それらがよく関連したページを表すことが [3], [9] で示されている．

典型的な authority と hub のグラフ構造を図 3 に示す．このグラフの右側には大手のコンピュータ関連会社が authority としてあり，それらに密にリンクを張っているリンク集が左側に hub としてある．このようなグラフ構造はウェブ上に多々見られるものである．関連ページアルゴリズムは，図 3 のように密に結合された authority と hub を抽出するものであり，IBM, TOSHIBA, SONY のどれかひとつをシードとして与えることで，これらの会社のリストが結果として出力される．

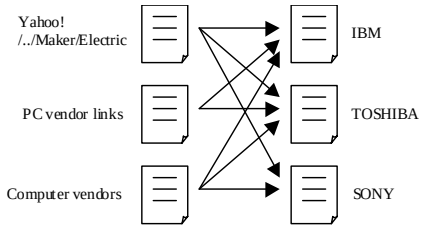



図 3　ハブとオーソリティからなる典型的なグラフ

表 3　コミュニティラベルの例

| コミュニティ ID | ラベル |
|---|---|
| 18 | 高知 県立 学校 高等 商業 知江 仁淀 伊野 ... |
| 54 | 教育 大学 研究センター 高等 センター 開発 ... |
| 110 | 検査 病院 臨床 大学 医 学部 附属 付属 ... |
| 40876 | 銀行 バンク 住友 パソコンバンキング ... |
| 145535 | 博物館 県立 東北 歴史 仙台 秋田 福島 山形 ... |

ウェブコミュニティチャートの作成アルゴリズムは分類したいシードページの集合を入力として受取り，チャートを結果として出力する．シードページとしてはウェブ上で著名なページを抽出して使用する．判断基準は外部のサーバから IN 本以上リンクが来ていることとした．IN はチャートのサイズを決めるパラメータとなる．

シードセットを受け取ると各シードページについて別々に上記の関連ページアルゴリズムを適用し，各シードが他のシードをどのように関連ページとして導出するかを調べる．この際，関連ページアルゴリズムの結果のうち上位 N 個を使用する．N はコミュニティの粒度を決めるパラメータとなる．我々はシード a がシード b を関連ページとして導出し，かつその逆も成り立つという対称関係に注目し，この関係で密に結合されたシード同士は，しばしば同じレベルのトピックを共有することを [9] で示した．これに従って，対称関係で密に結合されたシード同士をコミュニティとして抽出する(注3)．さらに 2 つのコミュニティのメンバ間に導出関係がある場合には，その間にエッジを張ることでコミュニティのグラフ (チャート) となる(注4)．

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表 3 に示すように，コミュニティの内容を表す単語群（コミュニティラベル(注5)）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，ラベル内の単語は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形態素解析して名詞や未知語を取り出したものであり，左から頻度が多い順に並んでいる(注6)．したがって，

(注3)：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．
(注4)：本論文ではウェブコミュニティチャートのエッジの部分は利用せず，コミュニティ部分のみ利用する．
(注5)：以降,「ラベル」は「コミュニティラベル」の意味で使用．
(注6)：多くのラベルに含まれている単語はストップワードとし，ラベル内の単語から削除している．

ラベルの上位にある単語はそのコミュニティの内容を良く表している単語といえる．

### 3.3 ウェブページアーカイブ

我々は定期的に国内のウェブページの収集を行っている．パネルログ収集期間中にも国内 4,500 万のウェブページの収集を行い，ウェブコミュニティチャートの手法を用いて 100 万個の有用なページから自動処理により 17 万個のコミュニティを生成した．パネルログの収集期間はウェブページの収集期間に比べ長いため，パネルが閲覧したウェブページに変更や削除の可能性がある．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 の適合率を

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

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 $URL$

と定義して適合率の測定を行った．無修正時の適合率は約 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約 40%となった．また，サイト名を削除する処理(注7)により適合率がさらに 8%程度向上し，最終的にパネルログに含まれる URL の約 65%をウェブコミュニティに登録されている URL に適合させることができた．詳細については文献 [8] で述べている．

## 4. トピックに関連する検索語群の発見

ユーザはウェブ上で自分の欲しい情報を探す場合，検索サイトなどで検索語の入力を行い検索結果のリストから自分が閲覧したいウェブページをクリックする．一方，閲覧したいページが存在しない場合は次の結果リストを見るか，または，検索語の追加や変更を行う．検索結果のリストはそのページのタイトルと簡単な説明文から構成されているため，検索結果のリストでクリックされたウェブページは自分の目的とする情報を保持していると考えられる．目的のページを訪れるために入力した検索語を集めることで，そのページに対するユーザの興味やニーズを知ることができるが，アクセスログ中には URL の情報しかないため，例えば「あるトピック関するページを訪れるために用いた検索語群の抽出」のような処理を行うことは容易ではない．

そこで，我々は前節で述べたウェブコミュニティの技術を用いて検索語群を抽出する手法を考えた．たとえば，ユーザが検索直後に閲覧したウェブページがウェブコミュニティに属している場合，コミュニティのラベルを用いることでページの内容（または，トピック）をいくつかの単語群で表現することができる．この単語群をトピックの名称としてラベルを検索することで,「あるトピックに関するウェブコミュニティの集合」を得ることができ，このコミュニティ集合に含まれる URL を閲覧するために用いた検索語を抽出することで，ユーザの興味やニーズを知ることができると考えた．我々は以下の手順でトピックに関連する検索語群の抽出を行うツールを作成した．

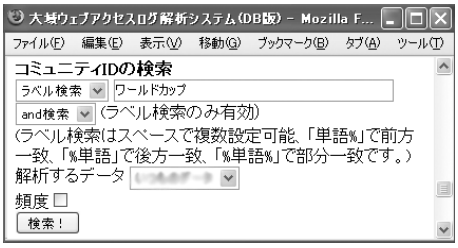



図 4　コミュニティラベル検索の入力画面

（1） ユーザの興味を知りたいトピックの名称をキーワードとして入力

（2） 入力されたキーワードをラベル中の単語の上位に含むコミュニティを検索

（3） 該当するコミュニティへ訪れる為に入力した検索語を表示

図 4 はトピックの名称を「ワールドカップ」としてラベルを検索した例である．図 5 は「ワールドカップ」をラベルに含むウェブコミュニティをラベル中の名詞の上位に含むものから順番に表示している．また，ラベル中の名詞の下位になるとそのコミュニティとの関連性が低くなるため，このツールでは上位 5 番までに含むコミュニティのみ対象としている．

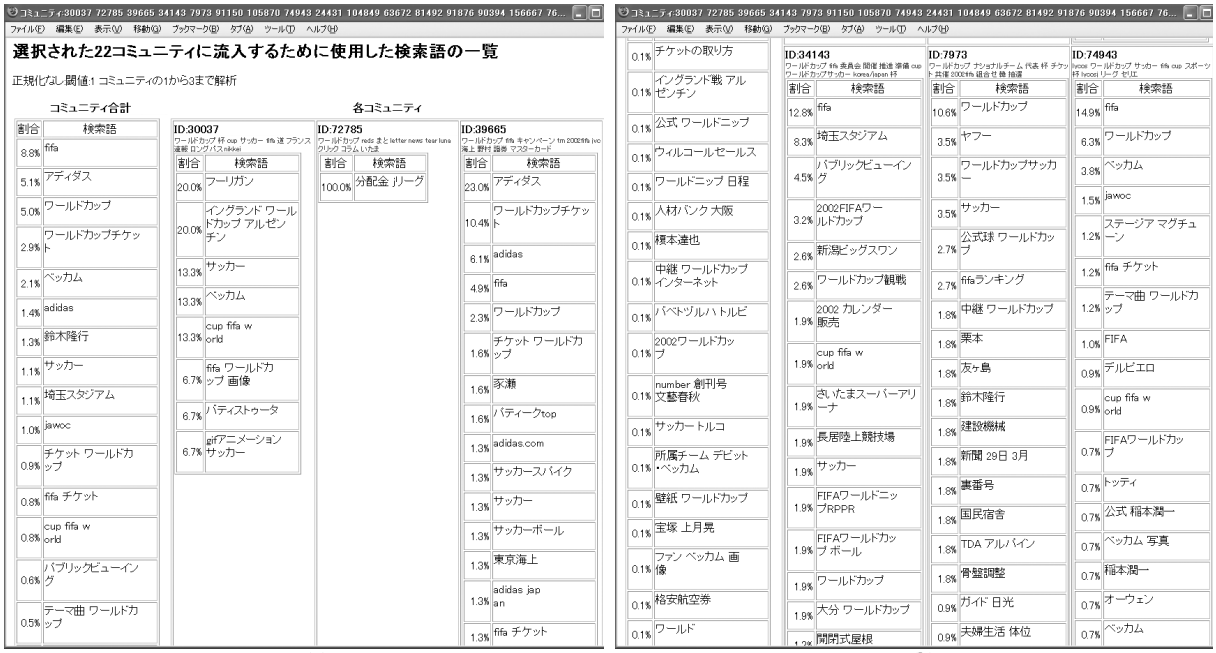
図 6 はラベル中の名詞の上位 5 番までに「ワールドカップ」を含むコミュニティに含まれる URL を閲覧するために入力した検索語群を示している．図 6(a) は入力頻度が多い検索語であり，(b) はある程度入力された検索語を表示いている．表中の「割合」とはその検索語の入力回数を全検索語の入力回数で割った値である．図の左側の表は「ワールドカップ」をラベルに含むコミュニティの合計であり，右側の表は各コミュニティ毎に結果を表示している．パネルログの収集時期に日本でサッカーワールドカップが開催されたため，検索語のほとんどはこれに関連している．また,「チケットの取り方，パブリックビューイング，中継 ワールドカップ インターネット」など，ワールドカップの観戦に関しての興味が意外と多いといった興味深い結果を得られた．

また，その他の例について図 7 に示す．キーワードを「占い」にしたの例では「動物占い，夢占い」など様々な占いに関する興味が多く，また,「姓名判断，命名」など名前に関連したものもあるのが興味深い．キーワード「温泉」の例では「有馬温泉，湯布院」など温泉地への関心が多く，また,「日帰り温泉，スーパー銭湯」など温泉施設に関連する語も見受けられる.「大学」をキーワードとした例ではユーザは大学名や受験に興味があることがわかる.「銀行」の例では銀行名がほとんどであることから，ユーザは銀行のホームページを訪れるために銀行名を入力する以外は銀行にあまり興味がないと思われる．

最後に，Yahoo! JAPAN や MSN などの広告を行ってるオー

(注7)：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


図 5　コミュニティラベルの検索結果 (ワールドカップの例)




(a)　(b)

図 6　コミュニティラベルに「ワールドカップ」を含むコミュニティを閲覧するために入力した検索語

バーチュアが提供する「キーワードアドバイスツール(注8)」が提示する検索語との比較を行った．このツールは広告主に対して一般ユーザがどんな検索語を使っているかの情報を提供し，広告主がアドワーズの想起を促すために用いられる．例えば，図 8 のようにある言葉を入力すると，関連性が高い検索語が入力回数が多い順に表示される．詳しい解析手法については非公開であるが，同時に入力された検索語や閲覧したページ情報をもとに関連度の計算が行われていると思われる．我々のツールとキーワードアドバイスツールの目的は異なるが，入力した語 (トピック) に対するユーザの興味を表す検索語の提示という点では類似するため本稿では比較を行った．

図 8 より,「ワールドカップ」に関しては我々の結果と同様にサッカーのワールドに対する関心が高いことがわかる，これは原稿執筆時がサッカーのワールドカップの直前のためである．同様に「銀行」に関しても我々のツールの結果と同様に銀行の名称が多い.「占い」については様々な占いを提示しているが,「命名」など名前に関連したものはなかった．また,「温泉，大学」の例では温泉地や大学の名称のみを提示しているが，我々のツールでは「スーパー銭湯，露天風呂付き個室」や「河合塾,

(注8)：http://inventory.jp.overture.com/


図 8　キーワードアドバイスツールの結果


図 7　指定したキーワードをラベルに含むコミュニティへ流入するために用いた検索語

駿台」など，少し変わった視点からの検索語を提示している．このように，我々の提案ツールでは入力されたトピックに対する様々な興味を把握することが可能であることがわかった．

## 5. おわりに

本稿ではパネルログとウェブコミュニティの技術を用いてトピックに関連する検索語群の発見法の提案を行い，この手法をもとに「あるトピックにおけるユーザの興味」を把握するツールの実装を行った．結果例から色々なトピックに対するユーザの興味を抽出することができた．今後はこのツールにより提示された検索語の評価を行う予定である．

**謝辞**　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．

## 文　献

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

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

[3] J. Dean and M. R. Henzinger. Finding related pages in the world wide web. In *In Proceedings of the 8th World-Wide Web Conference*, 1999.

[4] 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.

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

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

[7] 大久保雅且, 杉崎正之, 井上孝史, 田中一男. WWW検索ログに基づく情報ニーズの抽出. 情報処理学会論文誌, Vol. 39, No. 7, pp. 2250–2258, 8 1998.

[8] 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.

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

[10] 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.

[11] 安川美智子, 内山智文, 横尾英俊. 検索語の関連語を用いたクラスタ型メタサーチ. 第 5 回 Web インテリジェンスとインタラクション研究会 (SIG-WI2), pp. 117–122, 3 2006.