大規模アクセスログを用いた検索支援システム

大塚　真吾† 　　喜連川　優†

† 東京大学 生産技術研究所
〒 153-8505 東京都目黒区駒場 4-6-1
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エンジンを利用している．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．しかし，ユーザがいつでも検索目的に適した検索語を思い付くとは限らない．本論文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）を用いて，与えられた検索語に関連する検索語 (関連語) 群を表示し，ユーザに検索語を想起させるシステムの提案を行う．

キーワード　検索支援システム，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

The Search Support System Using Global Web Access Logs

Shingo OTSUKA† and Masaru KITSUREGAWA†

† Institute of Industrial Science, The University of Tokyo
4-6-1 Komaba Meguro-ku, Tokyo, 153-8505, Japan
E-mail: †{otsuka,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract　In cyberspace, users search their interested information by using search engine. Due to the improvement of searching accuracy with development of technologies, it the becomes possible that users can get kinds of information by just inputting search word(s) representing the topic which users are interested in. But it is not always true that users can hit upon search word(s) properly. In this paper, by using Web access logs (called panel logs), which are collected URL histories of Japanese users (called panels) selected without static deviation similar to the survey on TV audience rating, we propose search support system in order to show the related search words associated with the search words inputted by users.

Key words　Search Support System, Search words 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エンジンを利用している．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．しかし，ユーザがいつでも検索目的に適した検索語を思い付くとは限らない．そこで，本論文ではユーザが入力した検索語に関連する検索語 (関連語) 群を表示し，ユーザに検索語を想起させるシステムの提案を行う．

ユーザが入力した検索語とその後に閲覧した URL の情報は検索サイトのログから抽出できるが，この情報は一般に公開されておらず，データの収集が困難であったが，近年，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，個々のパネルが閲覧した全ての URL を知ることができる．また，パネルログはユーザが入力した検索語情報を保持している．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．

先行研究ではユーザが検索語を入力した後に閲覧された URL の集合を特徴空間として関連語の抽出を行っているが，本論文では以下の 3 つの手法を提案する．

- URL からファイル名とディレクトリ名を取り除いたサイト名を用いる手法
- 内容が類似している URL をまとめたウェブコミュニティ(注1) を用いる手法

(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

- ウェブページの文章に対して形態素解析を行いそこから得られる名詞を用いる手法

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]．主な研究として，

- ユーザの行動に関する研究 [1], [13], [19]
- ウェブページ間の関連に関する研究 [16], [17]
- 検索サイトに関連する研究 [2], [11], [12], [20]
- アクセスログの視覚化に関する研究 [7], [15]

などが挙げられる．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，文献 [21] 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，本研究で用いるパネルログを用いた研究は我々が知る限り，他では詳細な研究は行われていない．

検索語のクラスタリングに関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 [11] 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．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 [2], [20]．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 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「Google サジェスト(注2)」と呼ばれるサービスを行っている．Google サジェストは入力中の検索語に対し，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，Google 上で頻繁に検索された言葉や，その言葉が検索された場合に頻繁にクリックされる検索結果など，様々な要因を基に選ばれている．また，特定のユーザーやコンピュータ，Web ブラウザからの検索情報は使われていない．

例えば「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いている本研究と目的が異なるが，後者の「絞り込み検索語の提示」については本研究と類似する．

(注2)：http://www.google.co.jp/webhp?complete=1&hl=ja

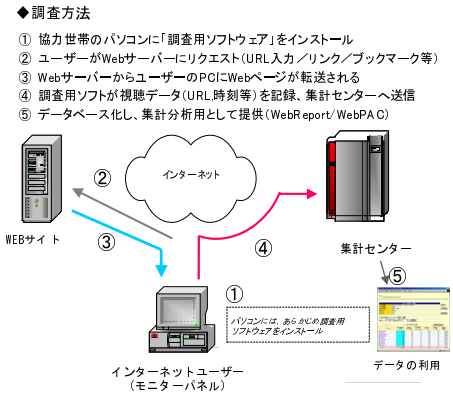


図 1　パネルログ収集の概要

表 1　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）
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. 関連語の発見に必要な技術の概要

この節では検索語に関連する語の発見のために必要な技術の概要について述べる．

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**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* などから構成されている．パネル ID とはパネル全員に対してユニークに割り当てた ID である．また，URL に加え検索エンジンサイトなどで入力された検索語についての情報を保持している．最後に我々が利用したパネルログの基本情報を表 1 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 [3]．

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 [18]．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

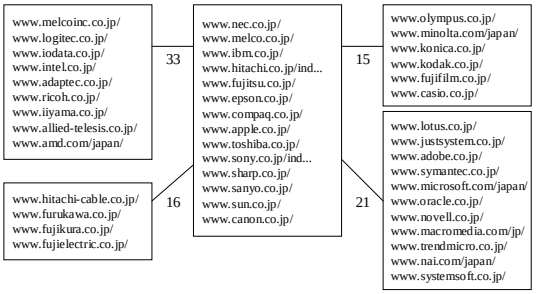


図 2　ウェブコミュニティチャートの一部

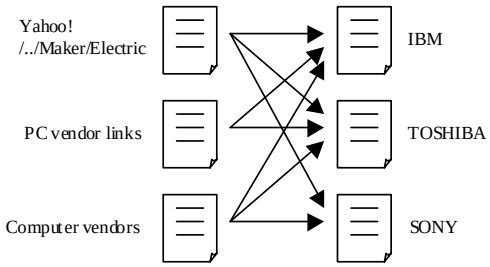


図 3　ハブとオーソリティからなる典型的なグラフ

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 [6], [8], [10].

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 [18] 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．

ウェブコミュニティチャートの作成のために，我々は以下に示す関連ページアルゴリズム [5], [18] 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 [5], [18] 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と，これらの会社のリストが結果として出力される．

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 [18] 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)．

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．また，各々のコミュニティは「コミュニティラベル」と呼ばれる，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を保持している．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

と定義して測定を行い，その結果を表 2 に示す．無修正時は約 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5)に

(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)：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 2　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

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%

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 [13] に示す．

また，我々の提案手法ではユーザが検索語を入力した後に閲覧されたページのテキストを解析するため，パネルログ収集当時のウェブページが必要となる．パネルログを調べた結果，検索した後に閲覧されたウェブページは約 100 万種類であり，その内およそ 68 万ページがパネルログ収集当時のままの状態でウェブアーカイブ内に格納されていることを確認した．

4. 関連語の抽出手法

検索エンジンなどで検索語を入力した場合，通常，その語との関連性が高いウェブページの一覧がタイトルと簡単な説明文と共に表示される．ユーザは検索結果の一覧の中から自分の目的に合ったページをクリックしウェブページを閲覧するため，このページは検索語と関連性が強いと考えられる．検索語は様々なユーザにより何回も入力されるため，パネルログの解析により検索語とその後に閲覧したページの集合を数多く抽出することができる．我々はこのようなページの集合を「閲覧ページ集合」と定義し，閲覧ページが 3 つ以上ある検索語約 125,000 語について閲覧ページ集合の抽出を行った．検索語の関連度を求める手法には意味空間ベクトルなどいくつかの手法が考えられるが，本論文では閲覧ページ集合から特徴空間を生成し，これを用いて関連語の抽出を行う．

また，本論文では「箱根 温泉」のように同時に複数の検索語を入力した場合については，これを 1 つの単語とみなした．(注6)

4.1 特徴空間の定義

我々は関連語集合の発見を行うため，閲覧ページ集合から以下の 3 つの特徴空間の抽出を行った(注7)．

- コミュニティ空間
- 名詞空間
- サイト空間

コミュニティ空間は 3.2 節で述べたように，類似する URL をまとめたコミュニティ技術を用いて作成した特徴空間である．名詞空間は閲覧ページ集合内の文章に対して形態素解析(注8)を行い，その中から名詞だけ(注9)を抽出して作成した特徴空間である．サイト空間は URL からファイル名とディレクトリ名を取り除いた特徴空間である．

4.2 関連度の定義

本論文では特徴空間の共通部分に着目し，関連度の計算を行った．検索語の全体集合 A を

$$A = \{a_1, a_2, \ldots, a_x, \ldots, a_n\}$$

(ただし，a_x は任意の検索語，また，n は検索語の総数である．)
と定義し，a_x の特徴空間 T_x を

$$T_x = \{(t_{x1}, p_{x1}), (t_{x2}, p_{x2}), \ldots, (t_{xp}, p_{xi}), \ldots, (t_{xm}, p_{xm})\}$$

(ただし，特徴空間がコミュニティの場合は t_x は Community ID(注10)，サイトの場合はサイト名，名詞の場合は名詞であり，p_x は検索した後に閲覧したページの頻度 (閲覧頻度) を T_x における全閲覧頻度で割った数である．また，m は特徴量の総数である．)
と定義する．

任意の検索語 a_x と a_y の特徴空間をそれぞれ T_x と T_y とし，その共通部分を $T_{x\cap y}$ とする．このとき $T_{x\cap y}$ の $p_{x\cap yi}$ は p_{xi} と p_{yi} の合計となる．ここで,「yahoo!」「価格.COM」「楽天」など，どのような閲覧ページ集合にも含まれているサイト，コミュニティや,「私」や「今日」など，どのようなウェブページにも含まれている名詞については $T_{x\cap y}$ から除外した(注11)．

任意の検索語 a_x と a_y の関連度 K_{xy} は

$$K_{xy} = \frac{T_{x\cap y}}{2}$$

と定義する．K_{xy} は 0 から 1 の間の値を取る．

5. 検索支援システム

前節で定義した関連度をもとに検索支援システムの構築を行った．その画面を図 4 に示す．図中 (1) に検索語を入力するとその語に関連する語群が特徴空間ごとに表示される．候補として表示された語を左クリックすると図中 (2) で選択した検索エンジンで検索を行い，その結果が右側に表示される．語数が多い場合は「...」のように省略された表示となるが，右クリックをすると語全体が表示される．

図中 (3) の 2 つのスライドバーで関連度の調節ができ，左側のスライドバーで最小関連度を指定し，右側で最大関連度を指定する．スライドバーで指定した関連度の範囲にある関連語が関連度が高い順に表示される．各特徴空間で最大 39 語を表示できるが，図中 (4) のボタンを押すと各語が動き出し関連度が高いものが押し出されて消える代わりに関連度が低い語が新たに表示される．また，図中 (5) のように関連度が高い語ほど赤く表示され，関連度が低くなるにつれて色が薄くなる．

(注6)：なお,「箱根 温泉」と「温泉 箱根」のように順番が異なる場合は同じ検索語として扱う．
(注7)：先行研究などで行われている URL を用いた手法は精度が良くないため対象外とした (詳細については文献 [14] を参照)．
(注8)：実験では日本語形態素解析システム ChaSen 「茶筌」[9] を用いた．
(注9)：厳密に言うと，名詞・一般，名詞・固有名詞，名詞・副詞可能，名詞・形容動詞語幹，名詞・サ変接続である
(注10)：各コミュニティにユニークな ID が割り当てられているものとする．
(注11)：実験では検索語全体のうちで 0.5%以上に含まれているものを除外した．

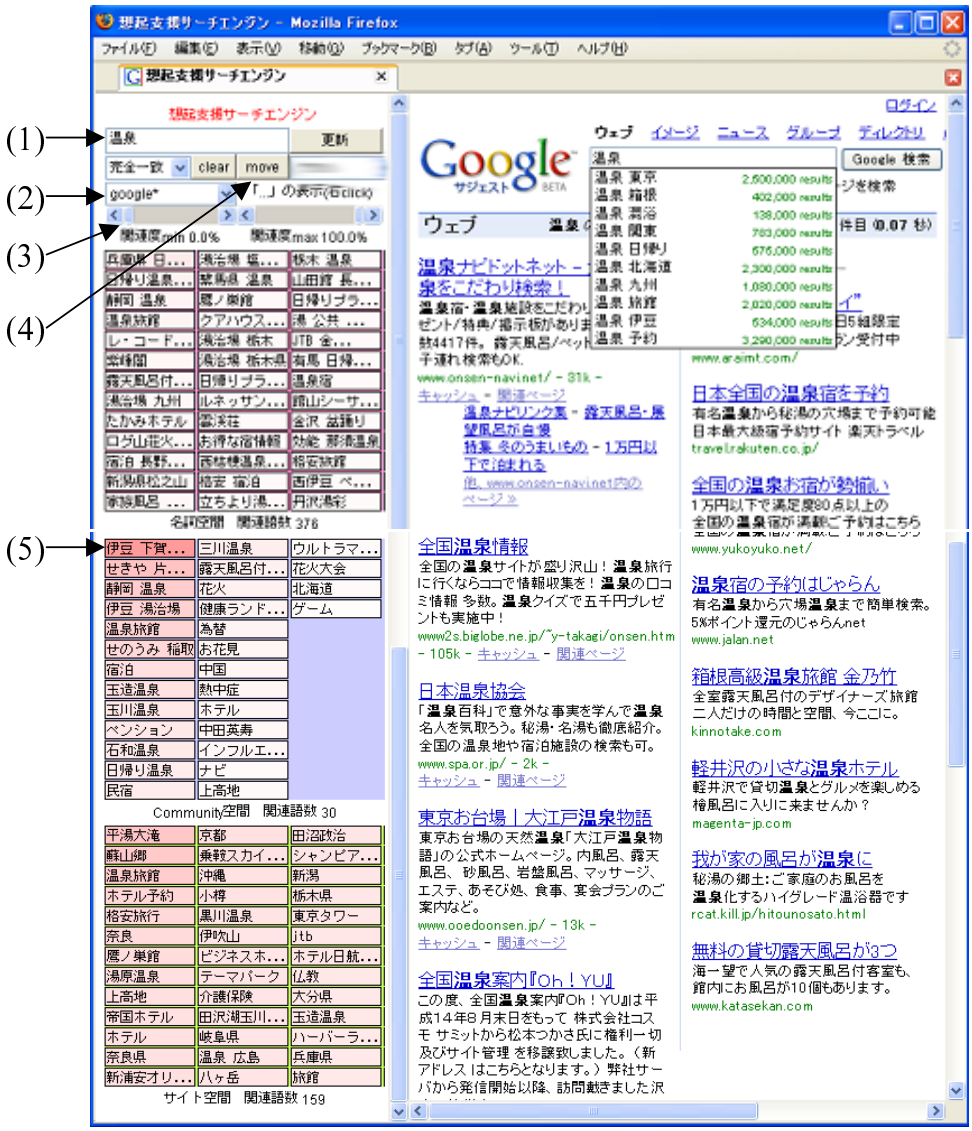


図 4　検索支援システム画面 (「温泉」の例)

5.1　検索支援例

図 4 は検索語に「温泉」を入力した例である．特徴空間に名詞を用いた結果は「温泉」と関連がある語群を数多く候補として表示している．コミュニティ空間では候補となる語数は少なくまた，関連度が低いものはあまり良い結果とはならなかった．サイト空間に関しては関連度が低くなると関連のない語が多くなる．

その他の例を図 5 に示す．図中の (a) は「携帯電話」を入力した例であるが，サイト空間を用いた場合に関連性のない語が若干表示されるが，そのほかの空間では関連のある語群が検索語候補として表示されている．

図中 (b) は「サッカー」を入力した例である．名詞空間，コミュニティ空間ともに関連性のある語群を提示している．また，サイト空間を用いた結果では関連性のある語群をあまり得ることができなかった．

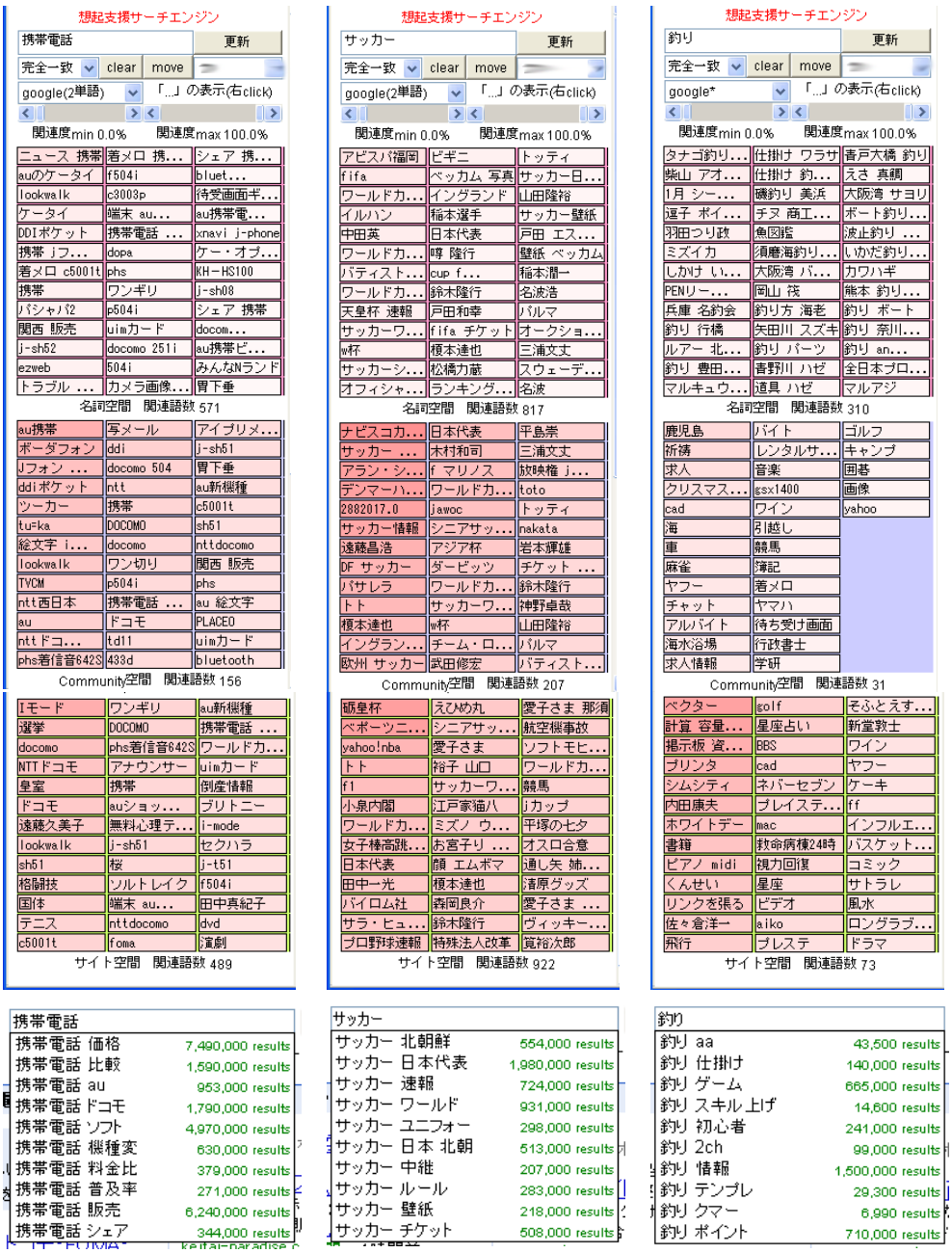
最後に「釣り」と入力した例を図中 (c) に示す．この例では名詞空間では関連性のある語群を候補として表示しているが，その他の特徴空間では良い候補を提示することができなかった．

5.2　Google サジェストとの比較

Google サジェストでは候補を 10 件のみ表示するため我々の結果と比較することは難しいが，図 4 の「温泉」の例では Google サジェストの結果は主に地名が多いのに対して，我々の結果では「石和温泉」など温泉地の名称や「立ち寄り湯」「お得な宿情報」など温泉と関連性の高い検索語を提示している．図 5(a) の「携帯電話」や (b) の「サッカー」の例では，Google サジェストと同様な結果の他に携帯電話の例では「着メロ」，サッカーの例では「ベッカム」「バティストゥータ」などの選手名を候補として提示している．最後に，図 5(c) の「釣り」の例では Google サジェストでは関連がないものが多いのに対して，名詞空間の結果では「釣り」と関連がある語を提示していることがわかる．

5.3　考　　察

今回の例では閲覧ページ数（検索語を入力した後に閲覧したページの数）が一番少ない語は「釣り」であり,「サッカー」が「釣り」の 4 倍,「携帯電話」は「釣り」の 5 倍であった．サイト空間では閲覧ページ数が多い「携帯電話」では他の検索語と



(a)携帯電話の例　(b)サッカーの例　(c)釣りの例

図 5　検索支援システムの実行例

比べて関連性がある語を提示しており，閲覧ページ数が少ない「釣り」では関連性のある語をほとんど提示されなかった．このことからサイト空間では閲覧ページ数が多いと提示された語の関連性の高いことがわかる．

コミュニティ空間に関してはコミュニティの精度の影響が強いと考えられ,「サッカー」と「温泉」のように閲覧ページ数がほぼ同じであっても提示された語の数が異なっている．

名詞空間に関しては閲覧ページ数に関係なく，どの検索語でも関連がある語を提示していることがわかった．

最後に，Google サジェストで「釣り」の例があまり良くない理由として，ネット上で「釣り」と入力するユーザはゲームやアスキーアートなどに興味があり，一般的に連想される「魚を釣る」とは異なっているためではないかと考えられる．

6. おわりに

本論文では大域ウェブアクセスログ（パネルログ）を用いて，与えられた検索語に関連する検索語 (関連語) 群を表示し，ユーザに検索語を想起させるシステムの提案を行った．関連する検索語群の発見のため，ユーザが検索語を入力した後に閲覧された URL のサイト名，ウェブコミュニティ，ウェブページに対する形態素解析処理により得られた名詞，の 3 つを用いた．利用例から我々のシステムが関連性のある検索語群を提示していることを示し，さらに，既存のサービスとの比較を行った結果 Google サジェストと同等またはそれ以上の関連語を提示して

いることを示した．今後はシステムの有効性を示すために，客観的な評価を行う．

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] 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. Dean and M. R. Henzinger. Finding related pages in the world wide web. 1999.

[6] 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.

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

[8] 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.

[9] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸. 日本語形態素解析システム chasen「茶筌」. *http://chasen.naist.jp/hiki/ChaSen/*.

[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] 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.

[12] 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.

[13] 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.

[14] 大塚真吾, 豊田正史, 喜連川優. 大域ウェブアクセスログを用いた関連語の発見法に関する一考察. 情報処理学会論文誌：データベース, Vol. 46, No. SIG18(TOD26), pp. 82–92, 6 2005.

[15] 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.

[16] 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.

[17] 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.

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

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

[20] 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.

[21] 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.