2005－DBS－137（Ｉ）（26）
2005／7／13

大域ウェブアクセスログを用いた検索語クラスタリング

大塚 真吾 †　　喜連川 優 †

† 東京大学 生産技術研究所

要 旨

検索語はサイバー空間におけるユーザの目的や意思を表す重要な要素であり，ウェブページを閲覧する人々の行動を把握するために有用である．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人（パネル）を対象に URL 履歴の収集を行ったログ（パネルログ）の解析から検索語のクラスタリングを行う．先行研究では，検索語を入力した後に閲覧した URL を基にしているが，我々はコミュニティ技術とウェブページの形態素解析から得られる名詞空間を用いる手法を提案する．実験結果より提案手法は URL だけを用いた手法よりも良好な結果が得られた．

The Method of Search Keywords Clustering Using Global Web Access Logs

Shingo Otsuka†　　Masaru Kitsuregawa†

†Institute of Industrial Science, The University of Tokyo

Abstract

The search word is one of the important factor in representing users' purpose and utilized to analyzed behaviors of users who view web pages. Here, by analyzing logs (called panel log), which are collected URL histories of users (called panel) who are selected without static deviation similarly to survey on TV audience rating, and we study a method of clustering search keywords. Previous researches are implemented based on only visited URLs after inputting search words, here we propose a method based on search words in noun terms space gotten by Web communities techniques and morphological analysis of Web pages. According to evaluation result, our proposed method can get better results than URL only method.

1　はじめに

本稿ではサイバー空間におけるユーザの目的や意思を表す検索語に着目する．ある検索語から関連が深い単語群を獲得できれば，商品のイメージや競合商品の情報など，マーケティング分野での活用が期待できる．また，新語やシソーラスの発見などにも有効であると考えられる．

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．

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稿ではパネルログと呼ぶ．

ログを用いた検索語クラスタリングの研究では検索語とその後に閲覧したURLの組合せを基に行っているが，本稿では内容が類似しているURLをまとめたウェブコミュニティ[1]の技術を用いる手法と，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空間を用いる手法を提案する．

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, 19, 13]
- 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, 20, 12, 11]
- アクセスログの視覚化に関する研究 [7, 14]

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．また，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．

[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

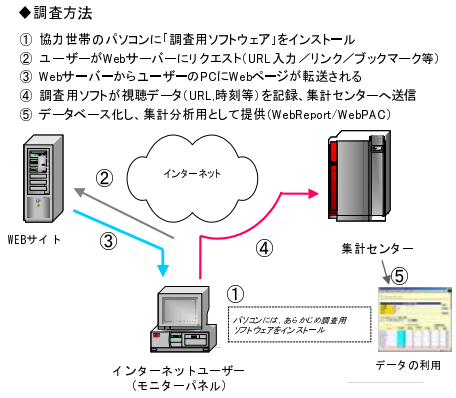

図 1: パネルログ収集の概要

3　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

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）

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

本稿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

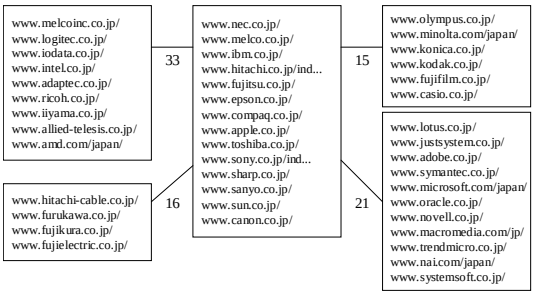

図 2: ウェブコミュニティチャートの一部

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 [18, 5] 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 [18, 5] 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

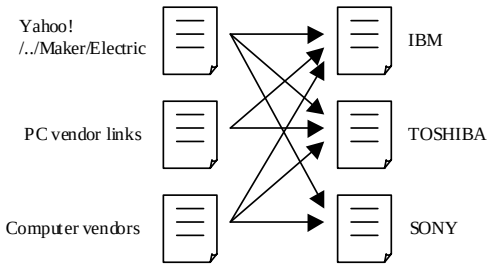

図 3: ハブとオーソリティからなる典型的なグラフ

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[2]．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[3]．

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[4]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 [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．

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%

[2]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．

[3]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．

[4]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

4 パネルログを用いた検索語のクラスタリング手法

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，本稿では閲覧ページ集合から特徴空間を生成し，これを用いて関連度の計算を行う．

また，本稿では「箱根 温泉」のような複数の検索語を同時に入力した場合については，これを 1 つの単語とみなした．[5]

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 4 つの特徴空間を抽出する．

- コミュニティ空間
- 名詞空間
- URL 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[6]を行い，その中から名詞だけ[7]を抽出して作成した特徴空間である．URL 空間は 2 節で述べたように先行研究で行われており，今回は比較対象としての特徴空間である．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りこれも比較対象とする..

[5]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.

[6]実験では日本語形態素解析システム ChaSen 「茶筌」[9] を用いた.

[7]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

4.2 関連度の定義

本稿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

$$T_x = \{t_{x1}, t_{x2}, \ldots, t_{xm}\}$$

(ただし，特徴空間が URL の場合は t_x は URL，コミュニティの場合は Community ID[8]，名詞の場合は名詞，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 の関連度 K_{xy} は

$$K_{xy} = \frac{|T_x \cap T_y|}{|T_x \cup T_y|}$$

と定義する[9].

また，パネルログを用いることで検索した後に閲覧したページの頻度 (閲覧頻度 (TF) と定義する) や,「価格.COM」や「楽天」など，どのような閲覧ページ集合にも含まれている URL，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 (DF が高いもの) を求めることが可能なため，TF*IDF を利用した関連度の定義を行う．[10]特徴空間 T_x に対応する TF*IDF 値の集合を H_x と定義すると t_{x1} に対応する h_{x1} は,

$$h_x1 = (t_{x1}\text{の } TF \text{ 値}) * (log\frac{n}{t_{x1}\text{の } DF \text{ 値}})$$

となる[11].

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_z とする．T_z の頻度 (TF) を考慮した TF*IDF 空間 H_z を

$$H_z = \{(h_{z1}, h_{z2}, \ldots, h_{zj}\}$$

(ただし，h_{z1} は T_x と T_y の TF*IDF 値を足したものである．また，j は T_x と T_y の共通部分における特徴量の総数である．)

[8]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.

[9]一般には Dice 係数と呼ばれている [15].

[10]テキスト解析の分野で利用されている TF(Term Frequency) と DF(Document Frequency) の概念に近いため本稿では TF, DF(IDF) と呼ぶ.

[11]今回の実験では log の底は 2 とした.

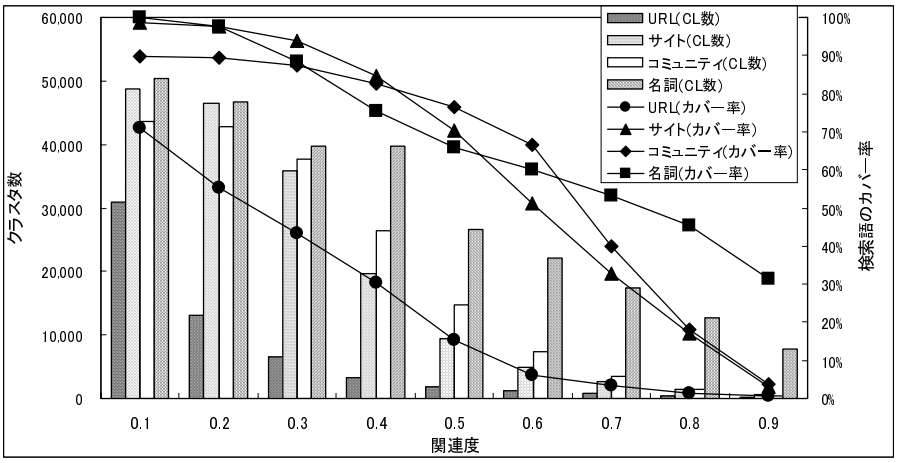

図 4: 関連度 K' とクラスタ数，カバー率の関係

のように定義すると TF*IDF を考慮した関連度 K'_{xy} は

$$K'_{xy} = \frac{H_z}{H_x + H_y}$$

となる．

関連度 K' を基に検索語のクラスタリングを以下の手順で行う．

1. 検索語の全体集合 A を TF が高い順に並べる．
2. 閾値 α を決める．
3. a_i(i は 1 から n まで増加させる) に対して a_{i+1} から a_n のすべての検索語に対し関連度 K' を求め α 以上のものを a_i に属するクラスタとする．

5 実験

まず，パネルログの中から閲覧ページ集合が存在する検索語およそ 26 万語を対象に特徴空間と TF*IDF 空間を作成した．その中から閲覧頻度が 5 以上の検索語約 5 万語についてクラスタリングを行った[12]．

5.1 関連度とクラスタ数などの関係

関連度 K' を 0.1 から 0.9 まで増加させたときのクラスタ数とクラスタに含まれた検索語の割合（検索語のカバー率）[13]を図 4 に示す．クラスタ数に着目すると，URL 空間では関連度 K' が 0.2 で急激に減少し，0.5 以上ではクラスタがほとんど存在しない．サイト空間では URL 空間よりも緩やかに減少し 0.6 まではある程度のクラスタを生成している．コミュニティ空間はサイト空間と同じ傾向を示しているが，サイト空間より若干ではあるがクラスタを多く生成している．最後に名詞空間では，他の特徴空間よりも緩やかに減少し K' が 0.9 でも約 8,000 個のクラスタを生成する．

検索語のカバー率に関しては，URL 空間では関連度 K' が 0.6 まで単純に減少していき，それ以降はほぼ 0%となった．サイト空間とコミュニティ空間は同様な傾向を示し，K' が 0.6 あたりまでは 50%程度のカバー率があり，0.9 になるとほぼ 0%となる．一方，名詞空間は他の空間よりもカバー率が良く K' が 0.9 でも約 30%ある．

生成されたクラスタに含まれる検索語の評価の詳細については今後の課題とするが，特徴空間に URL を用いると各クラスタの大きさは小さいものの関連性が高い検索語の集合となる傾向がある．サイト空間に関してはテレビ局や新聞社など大きなサイトの場合が存在する場合は良いクラスタを生成するが，それ以外では良い結果を得られなかった．コミュニティ空間に関しては URL 空間と比べて若干ノイズがあるものの，関連性が高い検索語の多く生成し，また，各クラスタの大きさは URL 空間を用いるよりも大きくなる傾向がある．一方，名詞空間につい

[12] 1 文字のアルファベット，記号，ひらがな，カタカナと，閲覧頻度が 1,000 以上の 3 語（2ch, zard, internet explorer）を解析対象から外した．

[13] カバー率 = クラスタに含まれた検索語の種類÷全検索語 (実験では 5 万)

図 5: 検索語「旅行」の例

てはノイズが多いものの関連性が少しある検索語の集団を多く生成する傾向がある．

5.2 クラスタの生成例

本手法により抽出されたクラスタの例を図 5 に示す．この例では検索語を「旅行」，特徴空間をコミュニティ，関連度 K' の最低値を 0.4 とした．図の左の表は閲覧頻度が「旅行」よりも多い検索語により生成されたクラスタの中で「旅行」を含むものを表示している．図中の「検索語数」とはこのクラスタに含まれる検索語の数である．

中央の表は閲覧頻度が「旅行」よりも少ない検索語を用いて生成したクラスタである．また，右の表は左の表の中で「旅行」を含む行と中央の表を統合した表であり，関連度 K' が 0.4 以上の検索語の一覧となっている．

この結果では「旅行」は若干ノイズがあるものの，「JTB」や「近畿日本ツーリスト」など，旅行と関連する検索語が形成するクラスタに含まれている．また，「旅行」より閲覧頻度が低い検索語から作成したクラスタを見るとホテル名や場所のような検索語を抽出している．

6 おわりに

本稿ではサイバー空間上でユーザが入力した検索語のクラスタリング手法の提案を行った．先行研究

では検索語を入力した後に閲覧した URL を基にしているが，我々はサイトやコミュニティ技術とウェブページの形態素解析から得られる名詞空間を用いる手法の提案とユーザが入力した検索語の入力回数に着目した関連度の提案を行った．実験結果より，URL を用いた既存の手法では検索語同士の関連度を 0.5 以上にするとクラスタをほとんど生成しないのに対し，提案手法では多くのクラスタを生成することができた．今後は，生成されたクラスタの評価について検討を行う予定である．

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] 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 cybercommunities. *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] 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.

[15] G. Salton and M. J. McGill. Introduction to modern information retrieval. 1983.

[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.