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

大塚真吾[†] 豊田正史[†] 喜連川優[†]

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，与えられた検索語と関連する検索語（関連語）を発見する方法について検討を行う．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，我々は検索語を入力した後に訪れたウェブコミュニティ（類似したウェブページの集まり）と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集合を特徴空間に利用する手法を提案する．実験結果から提案手法は特徴空間に URL を用いる手法よりも多くの関連語を抽出し，また，特徴空間に名詞を用いる手法とコミュニティを用いる手法では，抽出する関連語の性質が異なる傾向があることを示す．

A Study for Related Words Finding Method Using Global Web Access Logs

Shingo Otsuka,[†] Masashi Toyoda[†] and Masaru Kitsuregawa[†]

Web search engines are playing more and more important role for information retrieval in the cyberspace. 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 keyword(s) representing the topic which users are interested in. But it is not always true that users can hit upon keyword(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 study the methods of finding the related keywords associated with the keywords inputted by users. Different from the existing systems where the related keywords are extracted based on the set of URLs visited by the users after inputting their original keyword(s), we propose two methods to extract the related keywords. One is based on the Web communities (set of similar web pages); the other is based on the set of nouns obtained by morphological analysis of Web pages. According to evaluation results, the proposed methods can extract more related keywords than that based on URL. The results also show that the method based on the Web communities and the method based on nouns have different characters while extracting the related keywords.

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用いる．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できる．しかし，そのためにはユーザが自分の目的に適した検索語を入力する必要があるが，いつでも目的に適した検索語を思い付くとは限らない．

Google などの既存の検索エンジンではミススペルや表記の違いなどに関しては適応しているが，たとえば「銀行名を忘れてしまったが，その銀行についての検索をしたい.」や「銀行に関連するキーワードなのだが思い出せない.」など，ある単語と関連のあるものを検索することは難しい．このような場合,「銀行」を検索語として検索を行ったとしてもユーザが望む情報にたどりつける可能性は低い．我々は検索エンジンを使い慣れていない人をサポートするために，検索語のヒントとなる関連語の提示が重要だと考えている．

本論文では大域的なアクセスログを用いて，ユーザが提示する検索語と関連する検索語（関連語☆）の発見方法についての検討を行う．ユーザが入力した検索語

† 東京大学生産技術研究所
Institute of Industrial Science, The University of Tokyo

☆ 本論文では関連語を「ある語に対して表記的，意味的に類似する語」という意味で用いる．

とその直後に閲覧された URL の情報は検索エンジンなどのサイトで記録されている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すべての URL と検索エンジンなどで入力された検索語情報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．

アクセスログを用いて関連語の発見を行う研究は，検索語クラスタリングに関する研究分野で行われており，検索語とその直後に閲覧された URL の組合せを基にしている[2),21)]．本論文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☆を利用した手法と，ウェブページ内の文章に対する形態素解析から得られる名詞の集合を用いる手法を提案する．また，我々は関連した検索語の集合を発見するために必要な 4 つの関連度の提案も行い，実験結果から特徴空間に URL を用いる手法よりも多くの関連語を抽出し，特徴空間に名詞を用いる手法とコミュニティを用いる手法では，抽出する関連語の性質が異なる傾向があることを示す．

以下，2 章で関連研究について述べる．3 章では関連語の発見のために必要な技術である，パネルログ，ウェブコミュニティ，ウェブページアーカイブについて述べ，4 章では我々が提案するパネルログを用いた関連語の発見法について述べる．5 章では提案手法の評価と考察を行い，最後に 6 章でまとめを行う．

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)]．主な研究として，

- ユーザの行動に関する研究[1),13),16),20)]
- ウェブページ間の関連に関する研究[17),18),22)]
- 検索サイトに関連する研究[2),11),12),21)]
- アクセスログの視覚化に関する研究[7),14)]

などがあげられる．従来のほとん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 23) はプロキシサーバのアクセスログを用いておりパネルログとやや類似するが，本論文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[13)]．

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お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),21)]．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．

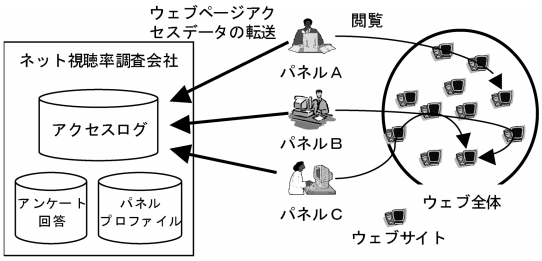

図 1 パネルログ収集の概要
Fig. 1 A method of collecting panel logs.

3. 関連語の発見のために必要な技術

この章では関連語の発見のために必要な技術である，パネルログ，ウェブコミュニティ，ウェブアーカイブについて述べる．

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

☆ 以降,「コミュニティ」は「ウェブコミュニティ」の意味で使用.

表 1 パネルログの一部
Table 1 A part of the panel logs.

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/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

検索語を含むURL

表 2 パネルログの概要
Table 2 The detail of our used 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）
検索語の種類	334,232（種類）

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 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持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[19]．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あげられる．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[19] 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き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，

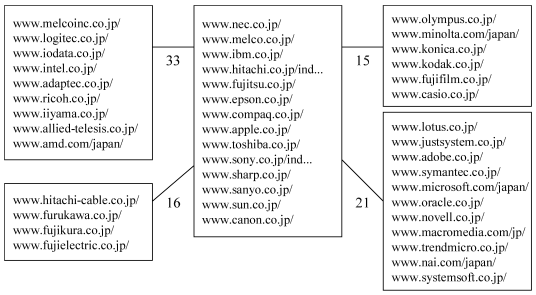

図 2 ウェブコミュニティチャートの一部
Fig. 2 A part of our web community chart.

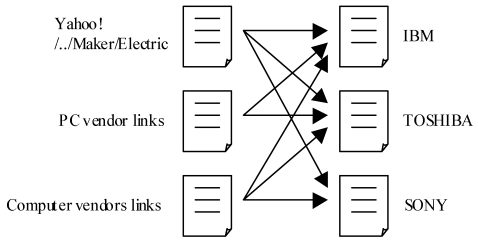

図 3 ハブとオーソリティからなる典型的なグラフ
Fig. 3 Typical graph of authorities and hubs.

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),19)] 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),19)] 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 1 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け取り，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どうしは，しばしば同じレベルのトピックを共有することを[19] 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どうしをコミュニティとして抽出する☆．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 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，十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 3** 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☆☆☆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．

表 3 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
Table 3 The adaptation ratio of the URLs belonged to web-communities and the URLs included panel logs.

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%

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，そのうちおよそ 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. パネルログを用いた関連語の発見法

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ともに表示される．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，4 セッション以上で使用された検索語約 3 万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，これを用いて関連度の計算を行う．

また，パネルログを解析した結果，ほとんどの検索語は 1 単語であったため，本論文では複数の検索語を同時に入力した場合については解析の対象外とした．複数検索語の入力を考慮した解析法については，ここ

☆ 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．
☆☆☆ 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

では論じない．

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 つの特徴空間を抽出する．

- 名詞空間
- コミュニティ空間
- URL 空間

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☆1を行い，その中から名詞だけ☆2を抽出して作成した特徴空間である．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．URL 空間は 2 章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．

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

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

（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☆3，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☆4．

また，パネルログから検索した後に閲覧されたページの頻度を求めることが可能なため，頻度を考慮した関連度の定義も行う．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 の頻度を考慮した頻度空間 H_z を

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

（ただし，h_{z1} は T_x と T_y での頻度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）

のように定義すると頻度を考慮した関連度 $K\mathrm{f}_{\mathrm{xy}}$☆5は

$$K\mathrm{f}_{\mathrm{xy}} = \frac{\mathrm{H_z}\text{の合計}}{\text{総頻度数}}$$

となる．

さらに，我々は「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を，「高出現要素」と定義する．約 3 万語の検索語についての高出現要素を調べたところ，全閲覧ページ集合の 1%以上にあたる 300 の閲覧ページ集合で存在した高出現要素の数を **表 4** に示す．我々は特徴空間 T_x と T_y の中から高出現要素を計算対象から除外して計算を行ったものを Kd☆6とし，頻度空間 H_z の中から高出現要素を計算対象から除外して計算を行ったものを Kfd と定義する．最後に，我々が定義した 4 つの関連度の比較を **表 5** に示す．

表 4 高出現要素の詳細
Table 4 The detail of high frequency elements.

特徴空間	高出現要素の数
名詞空間	4,565（単語）
コミュニティ空間	9（コミュニティ）
URL 空間	4（URL）

表 5 関連度の比較
Table 5 Comparison between similarities.

	特徴空間の頻度情報	高出現要素の除外
K	なし	なし
Kf	あり	なし
Kd	なし	あり
Kfd	あり	あり

5. 評　　価

我々は，4 章で述べた 4 つの関連度と 3 つの特徴空間を用いて，ユーザが指定した検索語に対して関連度が高い検索語群を表示するツールを作成し，約 3 万語の検索語に対して特徴空間と関連度の評価を行った．

5.1 評価方法

一般的によく使われる検索語「銀行，大学，占い」についての実験を行った．検索語の詳細を **表 6** に示す☆7．それぞれの検索語から 4 つの関連度と 3 つの特徴空間を用いて関連語を抽出し，その関連性について評価を行った．抽出された関連語と検索語（銀行，大学，占い）に関連があるかどうかの判定には以下の基

☆1 実験では日本語形態素解析システム ChaSen「茶筌」[9] を用いた．
☆2 厳密にいうと，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
☆3 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．
☆4 一般には Dice 係数と呼ばれている[15]．
☆5 f は Frequency の意味である．
☆6 テキスト解析の分野で利用されている DF（Document Frequency）の概念を取入れたため Kd とした．
☆7 表 6 中の「入力回数の順位」とは，入力された全検索語を入力頻度の多い順に並べたときの順位である．

表 6 評価を行った検索語の特徴
Table 6 Characters of evaluated keywords.

検索語	入力頻度	入力回数の順位	検索語を入力した後に閲覧された		
			ページの名詞	コミュニティ	URL
銀行	330（回） 94（セッション）	679 位	6,591（回） 1,725（種類）	24（回） 10（種類）	24（回） 20（種類）
大学	799（回） 195（セッション）	168 位	5,255（回） 483（種類）	31（回） 5（種類）	31（回） 8（種類）
占い	1,741（回） 417（セッション）	51 位	90,749（回） 4,321（種類）	395（回） 62（種類）	395（回） 125（種類）

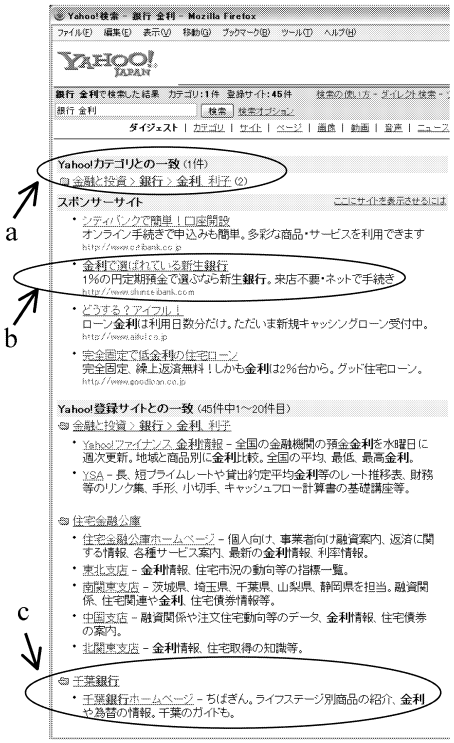

図 4 Yahoo!サイトで「銀行 金利」と入力した例
Fig. 4 An example of input 'bank' and 'interest rate' in the Yahoo! site.

準を設けた．

カテゴリ 1，2 検索語と関連性が強いと判断した関連語をカテゴリ 1 とし，カテゴリ 1 ほど関連性は強くないが，何らかの関連があると判断した関連語をカテゴリ 2 とした．

Yahoo!登録 Yahoo!のサイトで検索を行うと「カテゴリとの一致」「スポンサーサイトとの一致」「登録サイトとの一致」「ページとの一致」の 4 項目の結果が表示される．**図 4** に示すように「スポンサーサイト」と「登録サイト」はタイトルと簡単な説明文から構成されている．各々のサイト（スポンサー）はページタイトルと短い説明文で，自サイトの特徴を説明するため，その中に登場する名詞どうしの関連性は高いと考えられる．そこで，検索語と関連語の両者を Yahoo!で検索した結果にこの 2 語を含むスポンサーサイトまたは登録サイトが存在する場合は，検索語と関連性が強いと判断する．また，該当例は少ないがカテゴリと一致した場合も同様に関連性が強いと判断する．図 4 の例は Yahoo!に「銀行」と「金利」を入力した結果であり，a の部分はカテゴリの中で両検索語が存在している例である．同様に，b はスポンサーサイトに，c は登録サイトで両検索語が表示されている例である．

カテゴリ 1，2 の判断は我々が行うため主観的な評価であるが，Yahoo!登録はカテゴリ 1，2 よりも客観的な判断といえる．

5.2 実験結果

我々が作成した関連語表示ツールの結果例を **図 5**，**図 6** に示す．図中の中央下の関連度を変化させるスライドバーがあり，表示させる関連語の数を調節することができる．また，ノード（抽出された関連語）間で関連度が高い場合にはエッジが表示され，ノード間の関連度が高いほどエッジは短くなるように設定した．各ノードの位置に意味はなく，関連語を表示させた直後はランダムに表示されるが，時間が経つと関連度が高いノードどうしが近くになる．図 6 の例では抽出された関連語が，都市銀行系，地方銀行系，ネット銀行系，金融関連ごとに集まっている．

関連度と特徴空間の比較

検索語「銀行」において関連語の数を 10 語，100 語にしたときの結果を **図 7** に示す．図中の関連度 K は 4.2 節で述べたように，頻度と高出現要素を考慮しない関連度である．Kf は頻度を顧慮，Kd は高出現要素を考慮，Kfd は両方を考慮した関連度である．また，適合率はカテゴリの場合は「カテゴリ 1，2 に該当した関連語の数を抽出したすべての関連語の数で割ったもの」と定義し，Yahoo!登録の場合は「Yahoo!登録サイトまたはスポンサーサイトで登録されていた検索語の数を抽出したすべての関連語の数で割ったもの」

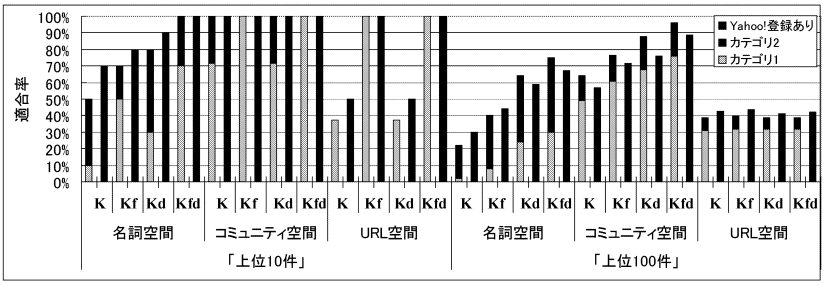

図 7 「銀行」の適合率
Fig. 7 Precision of 'Bank'.

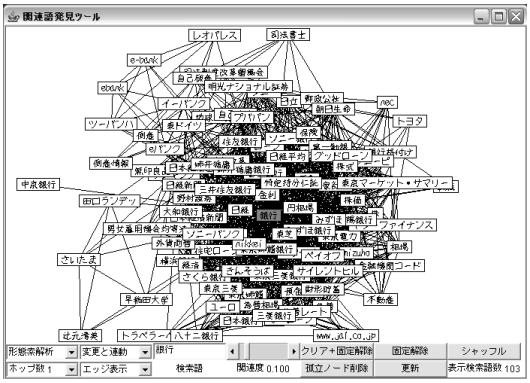

図 5 名詞空間を用いた時の表示例（関連度 Kfd，検索語は「銀行」）
Fig. 5 An example of expression using nouns space.

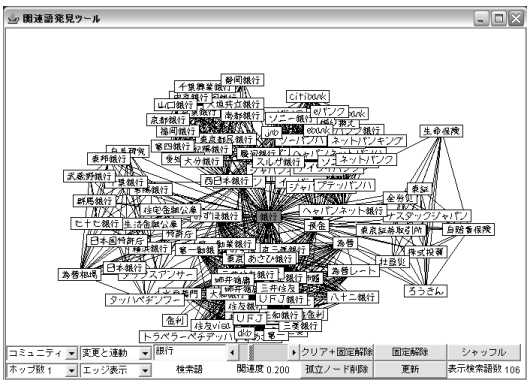

図 6 コミュニティ空間を用いた時の表示例（関連度 Kfd，検索語は「銀行」）
Fig. 6 An example of expression using communities space.

と定義した．

カテゴリ 1，2 と Yahoo!登録の適合率を比較すると，若干の誤差があるもののほぼ同じ傾向であった．名詞空間の特徴として他の特徴空間と比べてカテゴリ 2 に該当する結果を多く抽出する傾向があることが分かる．また，特徴空間や抽出した関連語の数に関係なく関連度に Kfd を用いると良い適合率が得られる．

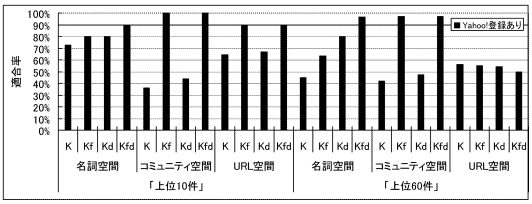

図 8 「大学」の適合率
Fig. 8 Precision of 'University'.

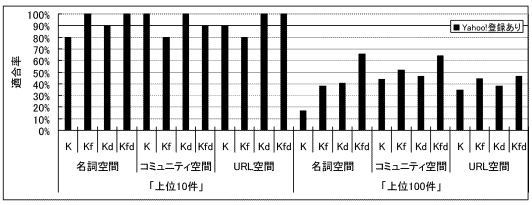

図 9 「占い」の適合率
Fig. 9 Precision of 'Fortune-telling'.

次に，「大学」と「占い」の評価を行い，その結果を **図 8**，**図 9** に示す☆．「銀行」の例と同様，関連度に Kfd を用いると良い適合率が得られる．

関連度 Kfd に着目すると，既存の手法である URL 空間の場合は関連語の数を増加させると適合率が急激に減少するのに対し，名詞空間では緩やかな減少であった．一方，コミュニティ空間では「銀行」と「大学」の場合，名詞空間よりさらに緩やかな減少となり，関連語を 100 語抽出しても適合率が約 90%であり良い結果となった．また，入力頻度が多い「占い」の場合でも URL 空間より良い結果となった．

5.3 考　　察

実験結果の考察を行うために，抽出した関連語の数（10 語，100 語）とそのときの関連度について調べ，結果を **表 7** に示す．また，検索語「銀行」について

☆ 「カテゴリ 1，2」の結果については傾向が同じであることから省略した．また，「大学」は特徴空間が URL のときに関連度を 0.001 にしても関連語の数が 60 であったため，名詞空間とコミュニティ空間でも関連語の数を 60 語として実験を行った．

＊は高出現要素

$$\text{割合}(\%) = \frac{\text{URL（名詞，コミュニティ）の頻度}}{\text{全てのURL（名詞，コミュニティ）の頻度を足した合計}}$$

図 10 検索語「銀行」の特徴空間
Fig. 10 Feature spaces of keyword 'Bank'.

図 11 検索語「銀行」の特徴空間（高出現要素を削除）
Fig. 11 Feature spaces of keyword 'Bank' (excluding high frequency elements).

表 7 抽出した関連語数と関連度の関係
Table 7 The relation between a number of related words and similarity.

特徴空間	関連度	銀行		大学		占い	
		TOP10	TOP100	TOP10	TOP100	TOP10	TOP100
名詞空間	K	0.219	0.132	0.239	0.069(93)	0.251	0.155
	Kf	0.604	0.353	0.626	0.346	0.727	0.601
	Kd	0.087	0.034	0.152	0.015(74)	0.070	0.027
	Kfd	0.364	0.103	0.288	0.049(74)	0.475	0.121
コミュニティ空間	K	0.112	0.009	0.250	0.100(57)	0.074	0.016
	Kf	0.636	0.166	0.782	0.090(65)	0.227	0.022
	Kd	0.126	0.100	0.250	0.074(57)	0.076	0.017(93)
	Kfd	0.694	0.200	0.782	0.090(65)	0.238	0.047
URL空間	K	0.052	0.010(75)	0.125	0.055(57)	0.028	0.006
	Kf	0.272	0.008(78)	0.264	0.071(58)	0.182	0.009
	Kd	0.052	0.007(85)	0.125	0.055(59)	0.023	0.008
	Kfd	0.294	0.008(88)	0.264	0.055(62)	0.146	0.008(99)

（関連度を0.001にしても関連語の数が100語にならなかったものについては，括弧内にその時に抽出された関連語の数（最大関連語数）を示す．また，括弧がある時の関連度は最大関連語数を抽出したときの値である．）

の特徴空間の一部を **図 10** に，高出現要素を除いたものを **図 11** に示す．図中の URL，名詞，コミュニティは頻度が多い順に並んでいる．「ラベル」とは 3.3 節で述べたように，コミュニティの内容を表す単語群である．

5.3.1 URL 空間において関連語数の増加にともない適合率が急激に減少する理由について

表 7 から，URL 空間において 100 語の関連語を抽出したときの関連度（K，Kf，Kd，Kfd）は，他の特徴空間と比べて低いことが確認できる．ある 2 つの検索語における関連度は，4.2 節で述べたように，特徴空間と共通部分の大きさに影響される．URL 空間を用いて関連度を求める場合，URL が厳密に一致しなければ共通部分とならないため，他の特徴空間と比べて共通部分が小さくなり，関連度の平均値が低くなる傾向がある．したがって，URL 空間において多くの関連語を抽出するためには関連度の値をより小さくする必要があるが，この場合，URL 空間内にあるごく少数の URL が一致しただけでも関連語として抽出されるため，結果として関連性の低い関連語が多く抽出

される．以上のことから，URL 空間で関連語の数を増やすと適合率が急激に減少したと推察される．

また，URL 空間で関連度 Kd を用いた場合に適合率が低い原因として，図 10 に示すように「銀行」の URL 空間には高出現要素が存在しないため，関連度 K とほぼ同じ結果になったと推察される．他の 2 つの検索語を調べた結果，これらの検索語の URL 空間には高出現要素が存在しており，これが適合率が向上の要因になったと推察される．

5.3.2 コミュニティ空間を用いて抽出された関連語の適合率が他の特徴空間よりも高い理由について

表 7 から，コミュニティ空間で 100 語の関連語を抽出したときの関連度は URL 空間と比べて高いことが確認できる．コミュニティ空間を利用する場合，コミュニティに属する URL〔トピックが似ているページ（URL）の集合〕と一致すればよいため，URL 空間よりも共通部分に属する URL が多くなる．たとえば「都市銀行のコミュニティ」が存在し「A 銀行」や「B 銀行」などのホームページが含まれていたと仮定すると，URL 空間を用いた場合は URL が異なるため共通部分とならないが，コミュニティ空間の場合は「都市銀行のコミュニティ」として共通部分となる．このように，コミュニティ空間を用いることで，トピックが類似する異なる 2 つの URL を共通部分として扱うことができる．

また，我々が用いたコミュニティは有用な（質の高い）ページを対象に作成したため，あまり有用でないページ（URL）はコミュニティに属さないという性質を保持している．したがって，コミュニティ空間を用いることで，暗黙のうちに有用でないページ（URL）を特徴空間から削除し，有用なページ（URL）のみを対象にした関連度を得ることができる．

これらの要因から，コミュニティ空間を用いて抽出された関連語の適合率が他の特徴空間よりも高いと推察される．

5.3.3 名詞空間においてカテゴリ 2 を多く抽出する理由について

名詞空間はユーザが検索した後に閲覧されたページの名詞情報を利用するため，これらのページの内容が類似するほど共通部分が大きくなる傾向がある．図 10，図 11 の名詞空間を見ると，どちらの場合もほとんどの名詞が「金融」と関連しており，特に高出現要素を除いた 図 11 ではその傾向が顕著なことが分かる．図 10，図 11 の URL 空間から「銀行」と入力しその後に閲覧されたページの多くは「銀行のホームページ」であることが分かり，このページには金融関連の用語が多く使われている可能性が高いため，このような名詞空間になったのだと考えられる．同様に「証券」や「保険」などのホームページでも金融関連の用語が多く使われているため，名詞空間を用いた場合では銀行と密接な関連がある関連語以外にも，証券，経済，保険など「金融」と関連がある関連語が多く抽出される可能性が高くなり，カテゴリ 2 に属する関連語が多くなったのだと推察される．

銀行の以外の検索語でも同様な結果が得られており，我々が調べた限りでは，名詞空間を用いると他の特徴空間を用いるよりもカテゴリ 2 に属する関連語を多く抽出する傾向があることを確認した．

以上の結果から，我々はある検索語と何らかの関連がある語を多く抽出する名詞空間を使うことで,「銀行に関連するキーワードなのだが思い出せない.」など与えられた単語と何らかの関連があるものを検索したい場合に「ペイオフ，財形貯蓄」などの関連語をユーザに提示することが可能となる．また,「銀行名を忘れてしまったが，その銀行について調べたい.」など，銀行と直接関連がある単語について調べたい場合は，ある検索語と関連が深い関連語を多く抽出するコミュニティ空間を用いて，ユーザが忘れている銀行名を提示することが可能となる．

6. おわりに

本論文では大域的なアクセスログを用いて，ユーザが提示した語と関連する検索語の発見方法について議論した．先行研究では検索語を入力した後に閲覧された URL を特徴空間として関連語の抽出を行っているが，我々はコミュニティとウェブページの形態素解析から得られる名詞を用いる手法の提案を行った．また，抽出した関連語の精度を向上させるために，特徴空間の頻度情報の利用と高出現要素の除外を行った．実験結果より，提案手法が既存の URL を用いた手法より有用なことを示し，また，特徴空間に名詞を用いる手法とコミュニティを用いる手法では，抽出する関連語の性質が異なる傾向があることを示した．

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．

本研究を進めるにあたりご協力いただ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ご協力いただいた株式会社ビデオリサーチインタラクティブに深

謝いたします.

参 考 文 献

1) Batista, P. and Silva, M.J.: 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) Beeferman, D. and Berger, A.: Agglomerative clustering of s earch engine query log, *The 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD2000*) (Aug. 2000).
3) Catledge, L. and Pitkow, J.E.: Characterizing browsing behaviors on the World-Wide Web, *Computer Networks and ISDN Systems*, Vol.27, No.6 (1995).
4) Cooley, R., Mobasher, B. and Srivastava, J.: Web mining: Information and pattern discovery on the World Wide Web, *Proc. 9th IEEE International Conference on Tools with Artificial Intelligence* (*ICTAI'97*) (Nov. 1997).
5) Dean, J. and Henzinger, M.R.: Finding related pages in the World Wide Web (1999).
6) Flake, G.W., Lawrence, S., Lee Giles, C. and Coetzee, F.M.: Self-organization and identification of web communities, *IEEE Computer*, Vol.35, No.3, pp.66–71 (2002).
7) Koutsoupias, N.: Exploring web access logs with correspondence analysis, *Methods and Applications of Artificial Intelligence, 2nd Hellenic* (Apr. 2002).
8) Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A.: Trawling the web for emerging cyber-communities, *Proc. 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) Murata, T.: Web community, *IPSJ Magazine*, 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 (1998).
12) Ohura, Y., Takahashi, K., Pramudiono, I. and Kitsuregawa, M.: Experiments on query expansion for internet yellow page services using web log mining, *28th International Conference on Very Large Data Bases* (*VLDB2002*) (Aug. 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 (2003).
14) Prasetyo, B., Pramudiono, I., Takahashi, K. and Kitsuregawa, M.: Naviz: Website navigational behavior visualizer, *Advances in Knowledge Discovery and Data Mining 6th Pacific-Asia Conference* (*PAKDD2002*) (May 2002).
15) Salton, G. and McGill, M.J.: *Introduction to modern information retrieval* (1983).
16) Shahabi, C., Zarkesh, A.M., Adibi, J. and Shah, V.: Knowledge discovery from users webpage navigation, *Proc. IEEE RIDE97 Workshop* (Apr. 1997).
17) Su, Z., Yang, Q., Zhang, H., Xu, X. and Hu, Y.: Correlation-based document clustering using web logs, *34th Hawaii International Conference on System Sciences* (*HICSS-34*) (Jan. 2001).
18) Tan, P. and Kumar, V.: 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* (Jan. 2002).
19) Toyoda, M. and Kitsuregawa, M.: Creating a web community chart for navigating related communities, *Conference Proceedings of Hypertext 2001*, pp.103–112 (2001).
20) Ungar, L.H. and Foster, D.P.: Clustering methods for collaborative filtering, *AAAI Workshop on Recommendation Systems* (July 1998).
21) Wen, J., Nie, J. and Zhang, H.: Query clustering using user logs, *ACM Trans. Inf. Syst.* (*ACM TOIS*), Vol.20, No.1, pp.59–81 (2002).
22) Zaiane, O.R., Xin, M. and Han, J.: Discovering web access patterns and trends by applying olap and data mining technology on web logs, *Proc. Advances in Digital Libraries* (*ADL'98*) (Apr. 1998).
23) Zeng, H., Chen, Z. and Ma, W.: A unified framework for clustering heterogeneous web objects, *3rd International Conference on Web Information Systems Engineering* (*WISE2002*) (Dec. 2002).

(平成 16 年 12 月 20 日受付)
(平成 17 年 4 月 6 日採録)

(担当編集委員 岩山 真)

大塚　真吾（正会員）

1996 年千葉工業大学工学部情報工学科卒業．2002 年同大学大学院工学研究科博士後期課程修了．博士（工学）．同年東京大学生産技術研究所学術研究支援員．ログマイニング，テキスト処理，ウェブマイニングに興味を持つ．

豊田　正史（正会員）

1994 年東京工業大学理学部情報科学科卒業．1999 年同大学大学院情報理工学研究科博士後期課程修了．博士（理学）．同年科学技術振興事業団計算科学技術研究員．2001 年東京大学生産技術研究所学術研究支援員．同大学産学官連携研究員．2004 年同大学特任助教授．ウェブマイニング，ユーザインタフェース，ビジュアルプログラミングに興味を持つ．ACM，IEEE CS，日本ソフトウェア科学会各会員．

喜連川　優（正会員）

1978 年東京大学工学部卒業．1983 年同大学大学院工学系研究科情報工学博士課程修了．工学博士．同年同大学生産技術研究所講師．現在，同教授．2003 年より同所戦略情報融合国際研究センター長．データベース工学，並列処理，Web マイニングに関する研究に従事．現在，本学会理事，日本データベース学会理事，1999〜2002 年 ACM SIGMODJapan Chapter Chair，1997 年，1998 年電子情報通信学会データ工学研究専門委員会委員長．VLDB Trustee（1997〜2002 年），IEEE ICDE，PAKDD，WAIM 等ステアリング委員，IEEE データ工学国際会議（ICDE2005）General Chair．