大規模アクセスログを用いた検索語想起支援システムの提案とその評価

大塚 真吾 †　　喜連川 優 †

† 東京大学 生産技術研究所

要 旨

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）を用いて，与えられた検索語に関連する関連語 (検索語) 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．また，その評価法についての検討を行う．

The Proposal and Evaluation of The Search Support System 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. And the evaluation methods are considered.

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．

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，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 を知ることが可能となった．また，このログにはユーザが入力した検索語情報が含まれている．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．

本稿ではパネルログを用いてユーザが入力した検

索語に関連する関連語 (検索語) を提示し，ユーザに検索語の想起を促すシステムの提案を行う．また，システムが提示した関連語の評価法について検討を行う．

2　関連研究

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．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 [14] は大学の電算室にあるマシンのウェブ閲覧履歴を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ほとんど行われていない．

文献 [8] 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 [1, 13]．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[1]」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．Google サジェストは「検索語入力の簡略化 (検索語入力の手間を省く)」と「絞り込み検索語の提示」に重点を置いており，後者については本研究と類似する．この点については文献 [11] で述べているためここでは議論しない．

[1]http://www.google.co.jp/webhp?complete=1&hl=ja

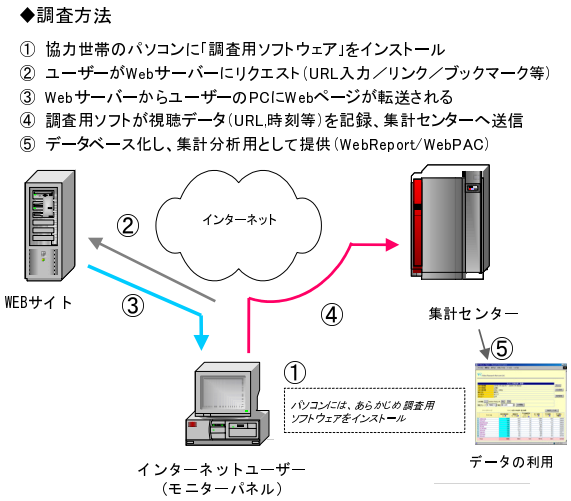

図 1: パネルログ収集の概要

3　関連語の発見に必要な技術

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

本稿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，本稿では「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

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）

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 [12]．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 [3, 5, 7].

ウェブコミュニティに関する研究の 1 つにハブとオーソリティの概念に基づいているものがある．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[4] 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．

本稿では 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．本稿ではコミュニティ間の関連

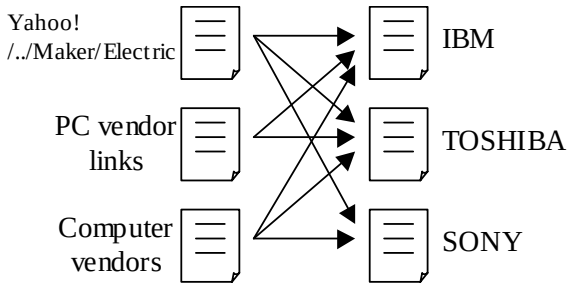

図 2: ハブとオーソリティからなる典型的なグラフ

度を必要としないため，コミュニティ部分のみ利用する．

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，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[2]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 [9] で述べている．

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

[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

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 (サマリー) 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，これを用いて関連語の抽出を行う．

また，本稿では「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．[3]

4.1 特徴空間の定義

我々は関連語を発見するために閲覧ページ集合から**コミュニティ空間，名詞空間，サイト空間**の 3 つの特徴空間を抽出した[4]．

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[5]を行い，その中から名詞だけ[6]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．

[3]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．

[4]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 [10] を参照)．

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

[6]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

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

(ただし，特徴空間がコミュニティの場合は t_x は Community ID[7]，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 $p_{xi\cap yi}$ 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[8]．

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 検索語想起支援システム

4.2 節で定義した関連度をもとに検索語想起支援システムの構築を行った．以下では，検索語を想起するためにシステムで入力した語を**検索語**と呼び，システムが提示した語を**関連語**と呼ぶ．システム利用画面を図 3 に示す．図中 (1) に検索語を入力するとその語に関連する関連語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．

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

[7]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．

[8]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．

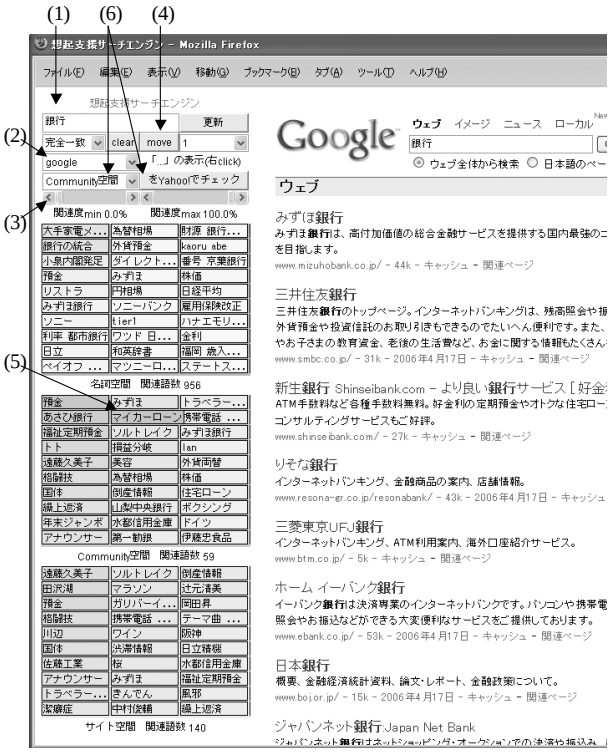

図 3: 検索語想起支援システム画面（「銀行」の例）

(a)サッカーの例　　(b)釣りの例

図 4: システムの実行例

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 30 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．語数が多い語は「...」のように省略された表示となるが，右クリックを押すと図中 (5) のように語全体が表示される．また，図が白黒のためわからないが，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ように表示される．

5.1 想起支援例

図 3 は検索語を「銀行」とした例である．特徴空間に名詞を用いた結果は「銀行」と関連がある語を数多く関連語として表示している．その他の例を図 4 に示す．図中 (a) 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を関連語として提示している．また，サイト空間を用いた結果では関連性のある語をあまり得ることができなかった．

次に「釣り」と入力した例を図中 (b) に示す．この例では名詞空間では関連性のある語を関連語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 システムが提示した関連語の評価

検索語想起支援システムにより提示された関連語が正しいかどうかの判断は，システムを利用したユーザにより異なるため，その評価を行うことは難しい．一般的に類似検索などのシステムの評価は，複数のユーザに利用してもらったアンケートをもとに行うが，今回我々が利用したデータ（パネルログ）の性質上，システムを一般のユーザに公開することが困難である．そのため，我々はアンケートと異なる評価法として Yahoo! API を用いる方法を検討した．

5.2.1 Yahoo! API を用いた評価ツール

通常，Yahoo!で検索を行うとその結果は「ページタイトル」と「簡単な説明文（サマリー）」から構成されている．これらはウェブページの特徴を説明しているため，その中に登場する名詞同士の関連性

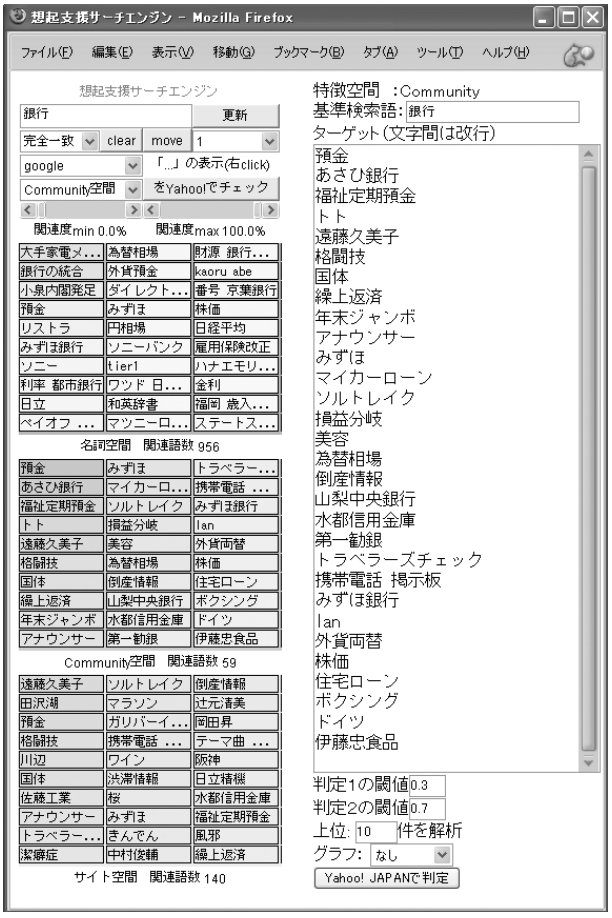

図 5: 評価ツール

は高いと考えられる．そこで，我々は検索語と関連語について Yahoo!で同時検索を行い，その検索結果にあるページタイトルまたはサマリーが両者の語を含む場合は関連性があると判断した．Yahoo API を用いて自動的に判定と行うツールを作成した．

動作例を図 5 に示す．基準検索語とは関連性を調べる語であり，ここではユーザが入力した検索語（この例では「銀行」）である．ターゲットとは基準検索語との関連度を判定するための評価語であり．リターンで区切ることで複数の語を一度に評価することができる．今回は，システムが提示した関連語の評価を行うため，図 3(6) を押すと選択された空間（この例は Community 空間）に表示されている関連語が自動的に入力されるようにした．この例は Community 空間の場合であるが，名詞空間やサイト空間についても同様に行うことが可能である．

基準検索語とターゲットの関連性の判定には以下の 2 つの閾値を設けた．

判定 1 基準検索語とターゲットの両方がタイトルに出現する．または，サマリー中の「...」で区切られた文章のどれかに基準検索語とターゲットの両方が出現する．

判定 2 基準検索語とターゲットの両方がタイトルに出現する．または，基準検索語とターゲットの両方がサマリーに出現する．

Yahoo!の検索結果にあるサマリーはウェブページの中のいくつかの文章をもとに作成され，各文章は「...」で区切られている．日記のようなページのサマリーは各文章の内容が異なる場合があり，基準検索語とターゲットの両方がサマリー中に出現しても，関連性が高いかどうかの判断は難しい．そこで，判断 2 の条件を厳しくした方法として，サマリー中の個々の文章中に同時に出現するかどうかも判定できるようにした．

上位件数とは，タイトルとサマリーの解析を行う検索結果数であり，今回の実験では上位 10 件を解析対象とした．

5.2.2 評価ツールの結果例

検索語を「銀行」としたときの関連語 (Community を利用) の評価例を図 6 に示す．この例では判定 1 の閾値を 0.3 に，判定 2 の閾値を 0.7 とした．図中の上の表にある「○の数」とは閾値以上のターゲット (関連語) の数であり割合とは解析数[9]で割った数である．最終判定とは判定 1 と判定 2 の両方を満たしてる語の数と割合を表示している．

下の表は各語についての判定結果を示している．例えば，表の中程にある「ソルトレイク」は判定 1 を満たしている結果数が 2 件あり，判定 2 を満たすものは 10 件ある．解析数が 10 のため，割合はそれぞれ 0.2, 1.0 となる．判定 1 は閾値 (0.3) 以下のため×となり，判定 2 は閾値 (0.7) 以上のため○となる．最終判断については判断 1 が閾値を満たしていないため×となる．

この評価ツールでは判定結果について詳細に見ることが可能である．そこで，判定 1 と判定 2 の値が極端に異なっていた「銀行」と「ソルトレイク」についての詳細な判定結果を図 7 に示す．検索結果の多くはソルトレイクオリンピックに関連する商品のオークションページであり,「銀行」という単語は代金の振り込み先の情報として用いられており,「ソルトレイク」との関連性はほとんどない．また，表の下から 2 番目はタイトルとサマリーからクイズのページであると推測でき,「ソルトレイク」と「銀行」は別の問いであることから，両者の関連性は低いことがわかる．

[9] この例では Yahoo!の上位 10 件を解析しているので 10 となる

Yahoo! JAPAN の結果を解析 - Mozilla Firefox

ファイル(F) 編集(E) 表示(V) 移動(G) ブックマーク(B) タブ(A) ツール(T) ヘルプ(H)

基準検索語:銀行
上位:10件を解析
単語数:30
解析対象:コミュニティ

	判定1 (0.3以上)	判定2 (0.7以上)	最終判定
○の数	25	28	24
割合	0.83	0.93	0.80

判定1:厳しい条件(タイトルに出現、または、「...」で分割した文章中に出現)
判定2:緩い条件(タイトルに出現、または、サマリーに出現)

検索語	解析数	判定1 出現件数	判定1 割合	判定1 判定(0.3以上)	判定2 出現件数	判定2 割合	判定2 判定(0.7以上)	最終判定
預金	10	6	0.60	○	10	1.00	○	○
あさひ銀行	10	7	0.70	○	7	0.70	○	○
福祉定期預金	10	4	0.40	○	6	0.60	×	×
トト	10	4	0.40	○	8	0.80	○	○
遠藤久美子	10	2	0.20	×	7	0.70	○	×
格闘技	10	4	0.40	○	8	0.80	○	○
国体	10	8	0.80	○	10	1.00	○	○
繰上返済	10	7	0.70	○	8	0.80	○	○
年末ジャンボ	10	5	0.50	○	8	0.80	○	○
アナウンサー	10	5	0.50	○	7	0.70	○	○
みずほ	10	9	0.90	○	9	0.90	○	○
マイカーローン	10	9	0.90	○	10	1.00	○	○
ソルトレイク	10	2	0.20	×	10	1.00	○	×
損益分岐	10	5	0.50	○	9	0.90	○	○
美容	10	6	0.60	○	7	0.70	○	○
為替相場	10	8	0.80	○	8	0.80	○	○
倒産情報	10	1	0.10	×	3	0.30	×	×
山梨中央銀行	10	10	1.00	○	10	1.00	○	○
水都信用金庫	10	8	0.80	○	9	0.90	○	○
第一勧銀	10	6	0.60	○	7	0.70	○	○
トラベラーズチェック	10	8	0.80	○	8	0.80	○	○
携帯電話 掲示板	10	1	0.10	×	7	0.70	○	×
みずほ銀行	10	9	0.90	○	9	0.90	○	○
lan	10	9	0.90	○	9	0.90	○	○
外貨両替	10	10	1.00	○	10	1.00	○	○
株価	10	10	1.00	○	10	1.00	○	○
住宅ローン	10	3	0.30	○	7	0.70	○	○
ボクシング	10	1	0.10	×	7	0.70	○	×
ドイツ	10	10	1.00	○	10	1.00	○	○
伊藤忠食品	10	5	0.50	○	9	0.90	○	○

図 6: 検索語「銀行」とその関連語の評価例

このように，判定 2 だけでは上記のような場合も「関連性がある」と判定されてしまうため，我々は判定 1 と判定 2 の両方に閾値を設けて，関連性の判断を行うのが良いと考えた．

5.3 評価ツールを用いた実験結果

最後に，我々は評価ツールを用いていくつかの検索語について実験を行った．その結果を表 2 に示す．実験は各特徴空間で提示された上位 30 語を対象とし，判定 1 の閾値は 0.3，判定 2 の閾値は 0.7 とした．また，Yahoo!の検索結果の上位 10 件を解析対象とした．

実験結果から，特徴空間にコミュニティを用いる手法が一番良く，名詞とサイトを用いる手法は概ね同じ結果となった．

表 2: 評価ツールを用いた実験結果

検索語	名詞	コミュニティ	サイト
銀行	0.60	0.83	0.61
大学	0.77	0.89	0.73
サッカー	0.70	0.63	0.40
釣り	0.63	0.90	0.53
温泉	0.63	0.90	0.90
ガンダム	0.53	0.60	0.50
ドラクエ	0.79	0.66	0.30
競馬	0.50	0.70	0.53
映画	0.40	0.73	0.77
カレンダー	0.53	0.70	0.63

6 おわりに

本稿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関連語を提示し，ユーザに検索語を想起させるシステムの提案と構築を行った．関連語の発見のため，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．また，ユーザが入力した検索語とシステムが提示した関連語の関連性を調べるために Yahoo! API を用いた評価ツールを作成し，システムが提示した関連語の評価を行った結果，ウェブコミュニティを用いる手法が良いことが分かった．

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．

参考文献

[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.

Yahoo! JAPAN の解析詳細 – Mozilla Firefox

ファイル(F) 編集(E) 表示(V) 移動(G) ブックマーク(B) タブ(A) ツール(T) ヘルプ(H)

基準検索語:銀行
ターゲット語:ソルトレイク
上位:10件を解析

	判定1	判定2
○の数	2	10
割合	0.20	1.00

判定1:厳しい条件(タイトルに出現、または、「...」で分割した文章中に出現)
判定2:緩い条件(タイトルに出現、または、サマリーに出現)

判定1	判定2	Title	Summary
○	○	千葉**銀行**｜ニュースリリース	平成14年1月23日「**ソルトレイク**シティー冬季オリンピック公式記念コイン」および「... 千葉**銀行**(頭取 早川 恒雄)は、アメリカ合衆国造幣局の発行する「**ソルトレイク**シティー冬季オリンピック...
×	○	YAHOO！オークション – 希少 **ソルトレイク**オリンピック聖火ランナーユニフォーム トリノ	YAHOO！オークション – 商品詳細. 希少 **ソルトレイク**オリンピック聖火ランナーユニフォーム トリノ... – クレジットカード決済. – **銀行**ネット決済. ・**銀行**振込. ・郵便振替... **ソルトレイク**オリンピックの聖火リレーのユニフォームです。. **ソルトレイク**オリンピックのオフィシャルスポンサーであり...
×	○	YAHOO！オークション –【新品DVD】2002**ソルトレイク**オリンピック フィギュア2枚組	YAHOO！オークション – 商品詳細【新品DVD】2002**ソルトレイク**オリンピック... 郵便振替・**銀行**振込み(三井住友、新生)・イーバンク・かんたん決済 以上よりご指定下さい...
×	○	YAHOO！オークション – ハードロックカフェ ピンバッジ **ソルトレイク**オリンピック 桃色	YAHOO！オークション – 商品詳細 ハードロックカフェ ピンバッジ **ソルトレイク**オリンピック 桃色 このオークションは終了しています。... ハードロックカフェ ピンバッジ **ソルトレイク**オリンピック 桃色 [利用者からのアドバイス] [友だち... 郵便振替(ばるる送金) イーバンク**銀行**へ振込み ジャパンネット**銀行**へ振込み...
×	○	YAHOO！オークション – ハードロックカフェ ピンバッジ **ソルトレイク**オリンピック 青	YAHOO！オークション – 商品詳細 ハードロックカフェ ピンバッジ **ソルトレイク**オリンピック 青 このオークションは終了しています。... ハードロックカフェ ピンバッジ **ソルトレイク**オリンピック 青 [利用者からのアドバイス] [友だちに... 郵便振替(ばるる送金) イーバンク**銀行**へ振込み ジャパンネット**銀行**へ振込み...
×	○	YAHOO！オークション –【スキー】だから、いつも笑顔で 上村愛子 目指せ！**ソルトレイク**	YAHOO！オークション – 商品詳細【スキー】だから、いつも笑顔で 上村愛子 目指せ！**ソルトレイク** このオークションは終了しています。... YAHOO！かんたん決済 新生**銀行**・みずほ**銀行** イーバンク**銀行**...
○	○	U. S. FRONTLINE	US FRONT LINE... 小売り最大手ウォルマート・ストアーズが再び**銀行**業務への参入を試みている。ロサンゼルス・タイムズによると、同社はユタ州**ソルトレイク**シティで**銀行**免許を申請した...
×	○	YAHOO！オークション – ◆新品ビデオ2本**ソルトレイク**オリンピック フィギュアスケート	YAHOO！オークション – 商品詳細 新品ビデオ2本**ソルトレイク**オリンピック フィギュアスケート 出品者の情報 発送方法 ・落札者が送料を負担 ・支払い終了時に発送... で、ヤフーかんたん決済、郵便局ばるる、東京三菱**銀行**へ。落札価格に、円ドル換金手数料...
×	○	2002年ニュースの塔解答	2002年ニュースの塔 問題)次の問題の答えを選択肢より選びなさい 問1) 2002年冬季オリンピックが開催された都市は？... が開催された都市は？ **ソルトレイク**シティ トリノ レイクプラシッド 問2) マイクロソフトが発売した... 問題のあった**銀行**は？ UFJ**銀行** みずほ**銀行** 三井住友**銀行** 問8) 出直し選挙...
×	○	YAHOO！オークション – **ソルトレイク**オリンピック フィギュアスケートDVD全2枚 トリノ	YAHOO！オークション – 商品詳細 **ソルトレイク**オリンピック フィギュアスケートDVD全2枚 トリノ この... 決済 – クレジットカード決済 – **銀行**ネット決済 ・**銀行**振込 ・郵便振替... **ソルトレイク**オリンピックの競技のDVD(80分)とエキシビジョンのDVD(60分)の2枚セット(新品)に...

図 7: 「銀行」とソルトレイク」の判定例

[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] 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.

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

[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] 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」.

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

[8] 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.

[9] 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.

[10] 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. SIG8(TOD26), pp. 82–92, 6 2005.

[11] 大塚真吾, 喜連川優. 大規模アクセスログを用いた検索支援システム. 第 17 回データ工学ワークショップ, 第 4 回日本データベース学会年次大会 (DEWS2006), 3 2006.

[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] 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.

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