社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－DBS－140 (I) (22)
2006／7／12

大域ウェブアクセスログを用いた検索語想起支援システムの評価に関する検討

大塚 真吾† 喜連川 優†

†東京大学 生産技術研究所

要 旨

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．我々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語(関連語)群を提示し，ユーザに検索語を想起させるシステムの提案をの提案を行った．本稿では構築したシステムの評価方法についての検討を行う．

A Study for The Evaluation of Search Keywords Remembrance Support using Global Web Access Logs

Shingo Otsuka† Masaru Kitsuregawa†

†Institute of Industrial Science, The University of Tokyo

Abstract

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. 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 proposed search keywords remembrance support system in order to show the related search words associated with the search words inputted by users. In this paper, we perform examination about evaluation methods of our system.

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．

我々は以前パネルログを用いてユーザが指定した語に関連する語(関連語)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った．本稿では

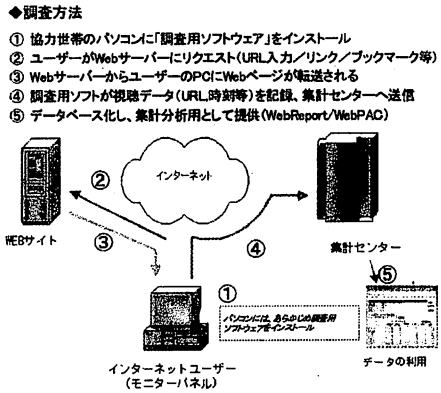

図 1: パネルログ収集の概要

システムが提示した関連語についての評価方法について検討を行う．

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 [6] 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, 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」[1]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．

[1]Google サジェストは http://www.google.co.jp/webhp?complete=1&hl=ja, goo サジェスト β は http://suggest.search.goo.ne.jp/suggest/index.php, Yahoo! Japan は「入力補助版」というサービスを提供している．

表 1: パネルログの詳細

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 検索語の想起支援に必要な技術の概要

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**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 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る．

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 [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

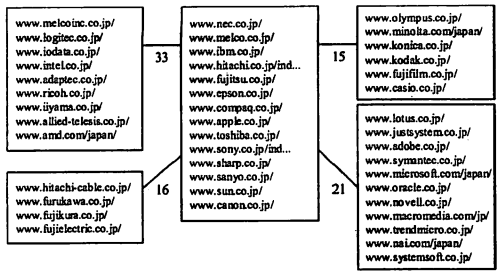

図 2: ウェブコミュニティチャートの一部

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 [2, 3, 5].

本稿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．本研究ではウェブコミュニティチャートのエッジ部分は利用せず，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 [7] 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し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我々のウェブアーカイブ内に格納されていることを確認した．

[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わない

4 検索語想起支援システム

4.1 関連語の抽出手法

検索サイトなどで検索語を入力した場合，通常は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．したがって，本システムにより提示される関連語は他のユーザ (または自分自身) によって入力された検索語となる．

なお，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]以下，本稿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．

4.2 特徴空間の定義

我々は関連語の発見を行うため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 つの特徴空間の抽出を行った[4]．コミュニティ空間

[3]「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 [8] を参照).

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．

4.3 関連度の定義

本稿では特徴空間の共通部分に着目して関連度の定義を行う．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．

4.4 検索語想起支援システム

4.3 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．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) で選択された検索サイトで検索を行い，その結果が右側に表示される．

図 3: 検索語の想起支援システムの実行例

図中 (3) の 2 つのスライドバーを使って最小関連度と最大関連度の調節ができる．2 つのスライドバーで指定された関連度を持つ関連語が表示される．各特徴空間において最大 30 語の表示が可能であるが，それ以上の関連語がある場合でも図中 (4) のボタンを押すことで各語が動き出し関連度が高い関連語から順番に見ることが可能である．語数が多い関連語は「...」のように省略された表示となるが，右クリックを押すと語全体が表示される．また，図が白黒のため分かり難いが関連度が高い関連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．この例では検索語を「銀行」としており，特徴空間に名詞空間やコミュニティ空間を用いて提示された関連語の多くは銀行と関連がある．その一方，サイト空間を用いて提示された関連語は関連のない語が多い．

5 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評価を行うことは難しい．

[5]実験では日本語形態素解析システム ChaSen「茶筌」[4] 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
[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．

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）の特性上，システムを一般のユーザに公開することが困難であるため，我々は主観的な評価と機械的な評価として Yahoo! API を用いた評価を行った．

5.1 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! Japan 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，Yahoo API を用いてシステムが提示した関連語の自動評価を行うツールを作成した．

動作例を図 4(a) 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「釣り」）である．ターゲットとは基準検索語との関連性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同時に評価することが可能である．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 中の (5) を押すと選択された特徴空間（この例は名詞）に表示されている関連語が自動的に入力される．また，コミュニティ空間やサイト空間の場合についても評価を行うことが可能である．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．

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．我々は判断 2 の条件を厳しくする方法としてサマリー中の個々の文章中に基準検索語とターゲットが同時に出現するかの判定も行う．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．また，例えば「銀行」と「ソルトレイク」の検索結果の上位 10 件のページを調査したところ，多くのページで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「ソルトレイク」との関連性はほとんどないことがわかった．他にクイズのページも存在し「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もわかった．

図 4: 評価ツール

このような場合，判定 2 だけの評価では「関連性がある」と判定されてしまう．また，判定 1 だけでは条件が厳しすぎるため，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方法を考えた．

5.2 評価ツールの利用例

検索語を「釣り」としたときの関連語 (名詞空間を利用) の評価例を図 5 に示す．この例では判定 1 の閾値を 0.3 とし，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「○の数」を解析数[9]で割った値である．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．この例では，想起支援システムが「釣り」の関連語として提示した上位 30 語のうち 6 割の関連語が「関連性がある」と判定されたということを示し

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

基準検索語: 釣り
上位:10件を解析
単語数:20
解析対象: 名詞

	判定1 (0.3以上)	判定2 (0.7以上)	最終判定
○の数	18	25	16
割合	0.60	0.83	0.60

判定1: 緩い条件(タイトルに出現、または、「...」で分割した文章中に出現)
判定2: 厳い条件(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
タナゴ釣り 千葉県	10	4	0.40	○	8	0.80	○	○
栗山 アオリイカ	10	0	0.00	×	6	0.60	×	×
1月 シーバス ミノー	10	0	0.00	×	6	0.60	×	×
逗子 ポイント ハゼ	10	0	0.00	×	10	1.00	○	×
羽田 つり政	0	0	0	×	0	0	×	×
ミズイカ	10	6	0.60	○	8	0.80	○	○
しかけ いわし つり	10	0	0.00	×	7	0.70	○	×
PEPリール 通販	0	0	0	×	0	0	×	×
兵庫 名釣会	10	2	0.20	×	8	0.80	○	×
釣り 行橋	10	9	0.90	○	10	1.00	○	○
ルアー 北九州 シーバス	10	0	0.00	×	6	0.60	×	×
釣り 皇田湖 ポイント	10	2	0.20	×	7	0.70	○	×
マルキュウ ワーム	10	2	0.20	×	10	1.00	○	×
仕掛け ワラサ	10	8	0.80	○	9	0.90	○	○
仕掛け 釣り サヨリ	10	6	0.60	○	8	0.80	○	○
避釣り 英浜	10	5	0.50	○	10	1.00	○	○
チヌ 魔工センター 広島	10	5	0.50	○	8	0.80	○	○
魚図鑑	10	9	0.90	○	9	0.90	○	○
福岡海釣り公園	10	8	0.80	○	8	0.80	○	○
大阪湾 バチ抜け ポイント	10	0	0.00	×	8	0.80	○	×
岡山 桜	10	8	0.80	○	9	0.90	○	○
釣り方 海老	10	8	0.80	○	9	0.90	○	○
矢田川 スズキ	10	3	0.30	○	10	1.00	○	○
釣り パーツ	10	8	0.80	○	10	1.00	○	○
青野川 ハゼ	10	8	0.80	○	10	1.00	○	○
道具 ハゼ	10	8	0.80	○	10	1.00	○	○
青戸大橋 釣り	10	5	0.50	○	10	1.00	○	○
えさ 真鯛	10	4	0.40	○	10	1.00	○	○
大阪湾 サヨリ	10	0	0.00	×	7	0.70	○	×
ボート釣り 戸田	10	8	0.80	○	8	0.80	○	○

図 5: 検索語「釣り」とその関連語の評価例

ている.

また，下の表は各関連語についての判定結果を示している．例えば，表の上から 9 番目にある「兵庫名釣会」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 8 件ある．そのため，割合はそれぞれ 0.2, 0.8 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り，判定 1 と判定 2 の値が極端に異なっている関連語「逗子 ポイント ハゼ (下の表の上から 4 番目)」について調べた結果，全てのサマリーにおいて，ターゲット語は「釣り」と関連がある使い方がされているが，ターゲット語の数が多い (関連語が複数) ため判定 1 のように「... 」内にある文章中に全てのターゲット語が存在することは難しいため，その結果が 0 となった．そこで，図 4(b) のように，ターゲット (関連語) が複数の場合は語を分割して 1 つの関連語として評価する方法についても評価を行う．

5.3 主観的な評価

検索語想起支援システムに検索語を入力し，提示された関連語のうちで関連性の高い 30 語について主観的な評価を行った．検索語と関連性が強いと判断した関連語をカテゴリ 1，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．

5.4 評価結果

我々は一般に用いられる言葉と良く知られているゲームとアニメのタイトルの計 10 語の検索語について評価ツールを用いて実験を行った．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，評価ツールの判定 1 の閾値は 0.3，判定 2 の閾値は 0.7 とした．また，Yahoo! Japan の検索結果の上位 10 件を解析対象とした．

表中の他ページ閲覧頻度とは検索語を入力した後に訪れたウェブページの数である．セッション数とは 3.1 節で述べたように，ユーザがウェブページの閲覧を開始してから終了するまでの一連の行動の中で入力された検索語の頻度である．Y1 とは評価ツールを利用した結果であり図 4 の (a) ように複数語をそのまま評価したものである．Y2 は評価ツールを利用し図 4 の (b) ように複数語を 1 語に分割して評価したものである．cat1,2 はそれぞれカテゴリ 1,2 を示す．また，cat1+2 はカテゴリ 1 と 2 を足した値である．

実験結果から，特徴空間に名詞空間を用いて提示された関連語の平均値は Y2 や cat1+2 の評価では一番良く，Y1 の評価でも 2 番であった．また，Y1 と Y2 を比較すると約 2 割増加していることから，関連語が複数である場合が多く，また，分割された各語との関連性も高いと推測できる．

コミュニティ空間を用いて提示された関連語の平均値は Y1 の評価が一番良く，Y2 や cat1+2 の評価は 2 番であった．また，Y1 と Y2 を比較するとほとんど増加していないことから，関連語が 1 単語である場合が多いと推測できる．

サイト空間を用いて提示された関連語の平均はどの評価も 1 番悪い結果となった．

5.5 考察

名詞空間はユーザが検索した後に閲覧し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特徴空間の共通部分が大きくなる傾向がある．例えば，検索語「銀行」の名詞空間を

表 2: システムが提示した関連語の評価

検索語	名詞空間					コミュニティ空間					サイト空間				
	Y1	Y2	cut1+2	cut1	cut2	Y1	Y2	cut1!2	cut1	cut2	Y1	Y2	cut1+2	cut1	cut2
銀行	0.63	0.84	0.76	0.43	0.33	0.80	0.84	0.67	0.27	0.40	0.57	0.59	0.26	0.13	0.13
大学	0.80	0.95	1.00	0.97	0.03	0.89	0.93	0.61	0.57	0.04	0.67	0.71	0.37	0.30	0.07
サッカー	0.60	0.74	0.96	0.93	0.03	0.62	0.69	0.97	0.97	0.00	0.37	0.53	0.46	0.43	0.03
釣り	0.60	0.95	1.00	0.97	0.03	0.83	0.83	0.03	0.00	0.03	0.60	0.71	0.00	0.00	0.00
温泉	0.67	0.90	1.00	0.93	0.07	0.90	0.92	0.60	0.50	0.10	0.77	0.77	0.53	0.20	0.33
ガンダム	0.57	0.68	0.97	0.97	0.00	0.57	0.64	0.27	0.20	0.07	0.60	0.69	0.30	0.30	0.00
ドラクエ	0.64	0.73	0.90	0.83	0.07	0.55	0.55	0.70	0.70	0.00	0.28	0.44	0.30	0.27	0.03
競馬	0.57	0.71	0.83	0.53	0.10	0.77	0.83	0.43	0.30	0.13	0.57	0.64	0.24	0.17	0.07
映画	0.48	0.69	0.83	0.73	0.10	0.73	0.78	0.90	0.77	0.13	0.70	0.83	1.00	0.83	0.17
カレンダー	0.48	0.71	0.60	0.23	0.37	0.63	0.73	0.20	0.13	0.07	0.60	0.79	0.10	0.10	0.00
平均	0.60	0.79	0.87	0.75	0.11	0.73	0.77	0.54	0.44	0.10	0.57	0.67	0.36	0.27	0.08

調査したところ，ほとんどの名詞が「金融」と関連していることがわかった．検索語「銀行」と入力しその後に閲覧されたウェブページを調べた結果，その多くは「銀行のホームページ」であることがわかり，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い．また，銀行の以外の検索語でも同様な結果が得られており，このことが評価の向上に影響していると考えている．

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) と一致すれば良いため，サイト空間を用いるよりも特徴空間の共通部分に属する URL が多くなる．例えば「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，サイト空間を用いた場合はサイト名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．また，コミュニティの中にはあまり良くないものも存在するため常に良い結果になるとはいえず，表 2 の「釣り」の例のように主観的な結果があまり良くない場合もある．

5.6 評価ツールにおける判定 1,2 の閾値に関する考察

今回の実験では評価ツールの判定 1,2 の閾値を予め定めている．我々は判定 1,2 の閾値による評価の傾向を調べるためのツールを作成し，その動作例を図 6 に示す．この例では検索語を「釣り」として名詞空間を用いて提示された関連語の上位 30 語を対象に評価した結果 (Y1) である．

各グラフは判定 2 の閾値を固定した時の判定 1 の傾向を示しており，横軸の数値は判定 1 の閾値を表している．また，縦軸は関連があると判定された関連語を全関連語で割った値（以下,「関連語の割合」とする.）を示しており，全ての関連語が関連があると判定された場合は 1 となる．例えば，1 番目のグラフは判定 2 の閾値を 0 にしたものであるため，実際には判定 1 だけの傾向を示している．グラフから判定 1 の閾値を大きくすると関連語の割合は減少し，閾値が大きくなるにつれて急激に減少する．また，グラフ同士を比較すると，判定 2 の閾値が大きくなるにつれて判定 1 の結果が反映されなくなる (グラフの傾きが緩やかになる) ことがわかる．

その他の検索語についても同様に調査をした結果，関連語の割合の高低は若干あるがグラフの傾向はほぼ同じであった．

今回の実験では，1 番目のグラフから関連語の割合がほぼ中間である 0.3 を判定 1 の閾値とし，グラフ同士の比較から判定 1 の結果が反映されていて閾値の値が比較的大きい 0.7 を判定 2 の閾値とした．

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評価方法について検討を行った．実際にプロトタイプシステムを作成を行い，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

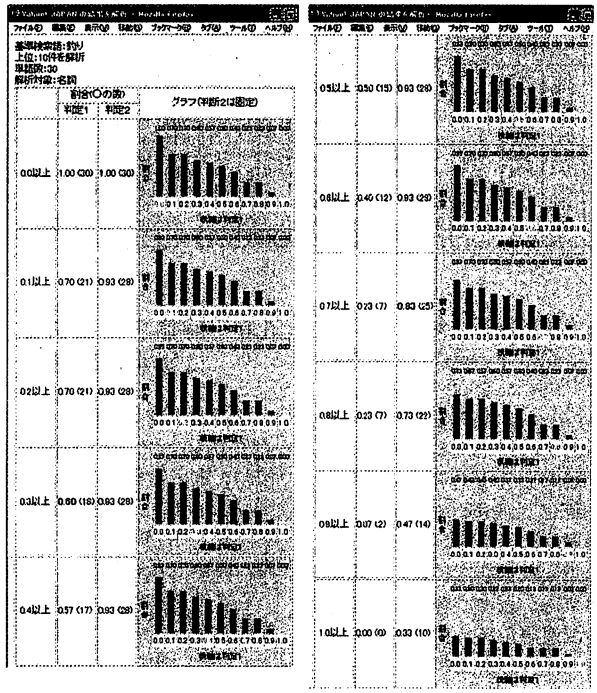

図 6: 判定 2 を固定した時の判定 1 の傾向

ムが提示した関連語の評価を行った結果，特徴空間に名詞空間やコミュニティ空間を用いる手法が良いことが分かった．今後は判定 1,2 の閾値をどのように定めるかについて検討を行う予定である．

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．

参考文献

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

[2] Flake, G., Lawrence, S., Giles, C. L. and Coetzee, F.: Self-organization and identification of Web communities, *IEEE Computer*, Vol. 35, No. 3, pp. 66–71 (2002).

[3] Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A.: Trawling the Web for Emerging Cyber-Communities, in *Proc. of the 8th WWW conference*, pp. 403–416 (1999).

[4] 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」.

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

[6] 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).

[7] 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).

[8] 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 (2005).

[9] Toyoda, M. and Kitsuregawa, M.: Creating a Web Community Chart for Navigating Related Communities, in *Conference Proceedings of Hypertext 2001*, pp. 103–112 (2001).

[10] Wen, J., Nie, J. and Zhang, H.: Query Clustering Using User Logs, *ACM Transactions on Information Systems (ACM TOIS)*, Vol. 20, No. 1, pp. 59–81 (2002).