2004－DBS－134　(I)　(2)
2004／7／13

大域ウェブアクセスログを用いたユーザ行動の分析

大塚 真吾†　　豊田 正史†　　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）を解析し，ユーザが入力した検索語と関連する語の発見方法の検討を行う．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精度が良好な結果が得られた．

The Analysis of Users Bihavior using Global Web Access Logs

Shingo Otsuka†　　Masashi Toyoda†　　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 finding words related to specific search words. 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 more precise results than URL only method.

1　はじめに

ウェブ上でのユーザの行動分析は重要な研究課題であり，さまざまな研究が行われている．本研究ではサイバー空間での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分類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．実験結果より提案手法が検索語と関連する語の発見に有用な点についても述べる．

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 [3]．主な研究として，

- ユーザの行動に関する研究 [9, 1, 13, 7]
- ウェブページ間の関連に関する研究 [10, 11, 15]
- 検索サイトに関連する研究 [2, 14, 6]
- アクセスログの視覚化に関する研究 [5, 8]

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 [16] 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．

また，検索語に関する研究はデータの入手が困難などの理由からあまり行われていない．本研究と類似するものとして [2, 14] の研究があり，検索語を入力したユーザが閲覧したディレクトリや 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内容の解析やウェブコミュニティの利用を行うため研究の方向性が異なる．

[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

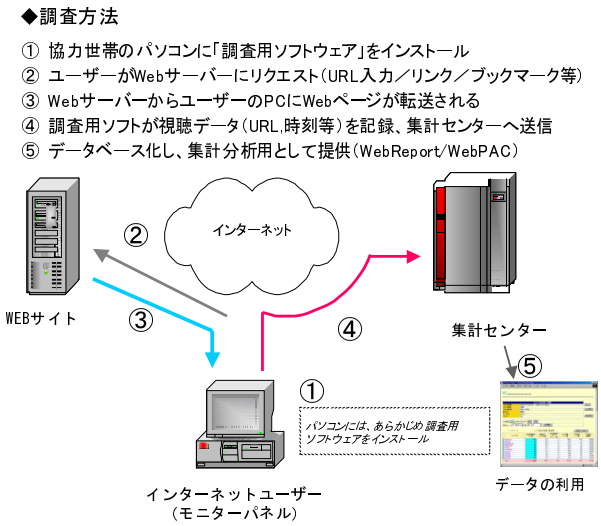

図 1: パネルログ収集の概要

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

本稿で利用するパネルログは (株) ビデオリサーチインタラクティブ社が以下の調査方法により集計を行ったデータである．

- 同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・集計したもの．
- 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．

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）
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）

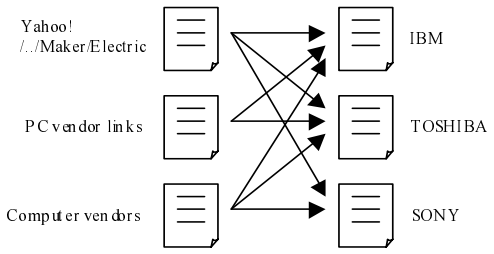

図 2: ハブとオーソリティからなる典型的なグラフ

また，その概念図を図 1 に示す．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．ユーザ ID とはパネル全員に対してユニークに割り当てられた ID である．最後にパネルログの基本情報を表 1 に示す．

3.2 ウェブコミュニティ

ウェブコミュニティに関する研究の多くはハブとオーソリティの概念に基づいている．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．

表 2: 全アクセスの中でウェブコミュニティに含まれる URL の割合

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%

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．パネルログを調べた結果，検索語を入力した後に閲覧した URL は約 100 万種類であり，その内およそ 68 万ページがウェブアーカイブ内に存在した．

また，我々はウェブアーカイブの一部（2002 年 2 月の国内 4,500 万のウェブページ）から 100 万個の有用なページを 17 万個のコミュニティに自動分類した．ウェブページの収集時期はパネルログ収集期間中のため，パネルがアクセスしたウェブページの変更や削除が行われている可能性がある．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 の適合度を測定し，その結果を表 2 に示す．無修正時における適合率はおよそ 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全体の約 40%をカバーした．また，サイト名を削除する処理[2]により適合率が 8%程度向上した．このように URL の修正により全アクセスの約 65%をカバーした．

[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っていない

4 パネルログを用いた関連語の発見

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ニーズに合ったページをクリックしてウェブページを閲覧するため，そのページは検索語との関連が強いと考えることができる．我々は検索語を入力した後に閲覧したページの集合を「閲覧ページ集合」と定義する．

また，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発見を行う．

4.1 特徴空間の定義

我々は閲覧ページ集合から以下の 3 つの特徴空間を用いる．

- URL 空間
- コミュニティ空間
- 名詞空間

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．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を用いた特徴空間である．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を行い，名詞だけ[3]を取り出して作成した特徴空間である．ウェブアーカイブからユーザが閲覧した時期のページを取り出し名詞の抽出を行った．

4.2 検索語の関連度

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

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

(ただし，a_x は検索語，また，n は検索語の総数である．)
と定義し，a_1 の特徴空間 T_1 を

$$T_1 = \{t_1, t_2, \ldots, t_m\}$$

(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，名詞の時は名詞，また，m は特徴量の総数である．)
と定義する．検索語 a_1 と a_x の関連度 K_1x は

$$K_{1x} = \frac{(T_1 \cap T_x) \text{ の数}}{(T_1 \cup T_x) \text{ の数}}$$

と定義する．また，パネルログからユーザが検索語を入力した後に閲覧し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．a_1 の特徴空間 T_1 に対応する頻度空間 H_1 を以下のように定義する．

$$H_1 = \{h_1, h_2, \ldots, h_m\}$$

(ただし，h_1 は t_1 に h_x は t_x に対応する頻度である．また，m は同様に特徴量の総数である．)
検索語 a_1 と a_x の頻度を考慮した関連度 K'_{1x} は

$$K'_1x = \frac{(H_1 \cap H_x) \text{ の頻度の合計}}{\text{総頻度数}}$$

と定義する．

[3]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

表 3: 評価した検索語の特徴

検索語	銀行	携帯電話
入力後に閲覧した URL の種類	20	58
入力後に閲覧したコミュニティの種類	10	48
入力後に閲覧したページの名詞の種類	6,691	23,250

5 評価

我々は 4.2 節で述べた関連度を基にユーザが指定した検索語と関連する検索語群を表示するツールを作成した．実験はパネルログ中にある検索語の中で頻度が多いおよそ 4,000 語を対象に行った．評価は検索語を入力した後に閲覧した検索語の種類が少ない「銀行」と種類が多い「携帯電話」で行い，検索語の特徴を表 3 に示す．

評価は関連度の値が高い上位 20 件を対象に我々が判定を行った．また，関連語かどうかの判定には以下の 2 つ基準を設けた．

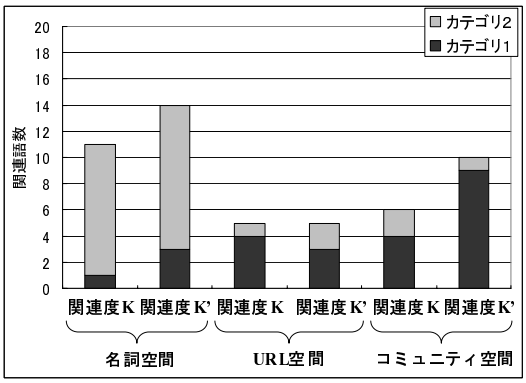

図 3: 「銀行」と関連する検索語の結果 (その 1)

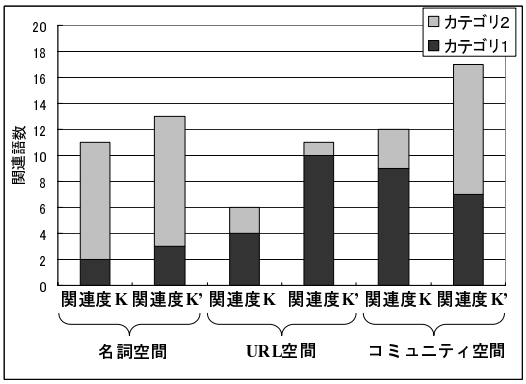

図 4: 「携帯電話」と関連する検索語の結果 (その 1)

カテゴリ 1 指定した検索語と関連性が強い検索語

カテゴリ 2 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検索語

また，どちらのカテゴリにも属さない場合は不正解とする．

5.1 実験結果

実験結果を図 3,4 に示す．関連度 K と頻度を考慮した関連度 K' を比較すると，どちらの結果でも頻度を考慮した方が良い結果であった．検索語を入力した後に閲覧したページの種類が少ない「銀行」の場合，特徴空間に名詞空間を用いる手法が最も良い結果となった．一方，閲覧したページの種類が多い「携帯電話」の場合は特徴空間にコミュニティ空間を用いる手法が一番良い．名詞空間，コミュニティ空間共に，URL 空間を用いる手法よりも良い結果が得られ，提案手法が関連検索語の発見に有効である．また，名詞空間を用いるとカテゴリ 2 に該当する結果が多く得られることが分かる．

表 4: 高出現要素の詳細

特徴空間	高出現要素の名詞, URL, コミュニティの数	存在する閲覧ページ集合
名詞空間	約 1,000 (単語)	400
URL 空間	10 (URL)	40
コミュニティ空間	25 (コミュニティ)	40

5.2 精度向上のための改良

特徴空間に名詞とコミュニティを用いることで，上位 20 件のうちのおよそ半数が指定した検索語と関連する語であった．我々は上位 20 件に占める関連語の数を増やすために，不正解の検索語の関連度を詳しく調べた．その結果，価格.COM や楽天など，どんな検索語の閲覧ページ集合にも含まれている名詞，URL，コミュニティが原因であった．そこで，我々はこれらの割合を調べその結果を表 4 に示す．名詞空間では全体の 10%以上の検索語に含まれる名詞が 1,000 語程度あり，URL 空間とコミュニティ空間では全体の 10%以上の検索語に存在するものは無く 1%以上のものが 10URL,25 コミュニティあった．本稿ではこれらの名詞，URL，コミュニティを高出現要素と呼ぶことにする．

我々は特徴空間 T_x の中で高出現要素を計算対象から除外して計算を行う KM と KM' を提案する．KM と KM' の有効性を検証するために K, K' との比較を行った．その結果を図 5,6 に示す．高出現要素を考慮することで，特徴空間や閲覧したページの種類に関わらず関連語抽出の精度が上昇し，さらに頻度を考慮すると一番良い結果が得られる．

特徴空間に名詞を用いた場合は検索語を入力した後に閲覧したページの種類が少ない時（「銀行」の例）に精度が良く，カテゴリ 2 に該当する関連語を多く含む性質に変化は無かった．URL を用いた場合は閲覧したページの種類が少ない時に精度が悪く，種類が多い時（「携帯電話」の例）でもコミュニティより精度が劣っていた．コミュニティを用いた場合は頻度と高出現要素を考慮することで，閲覧

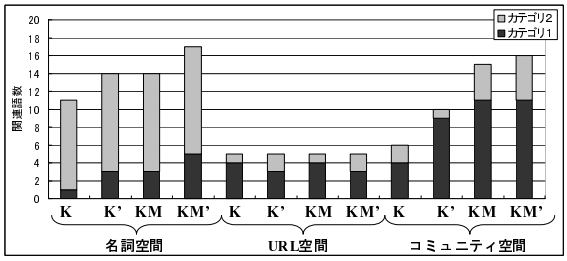

図 5: 「銀行」と関連する検索語の結果 (その 2)

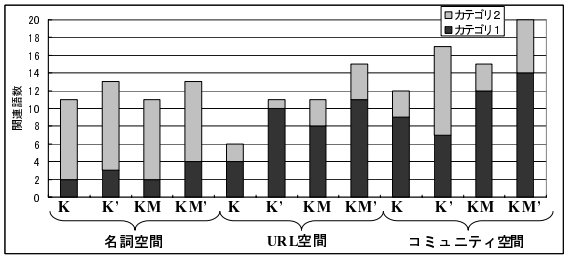

図 6: 「携帯電話」と関連する検索語の結果 (その 2)

したページの種類に関係なく，良好な結果が得られた，特に「携帯電話」の例では上位 20 件全ての検索語が関連語であった．

5.3 実際に抽出された関連語群

実験結果の中で一番精度が良い頻度と高出現要素を考慮して発見された検索語群を図 7,8 に示す．画面上の検索語 (ノード) はランダムに表示されるため位置に意味は無いが，今回は便宜上カテゴリ 1 に属する検索語を下に，カテゴリ 2 に属する検索語を右側に，関連が低い検索語を左上に配置した．

図の左側は名詞空間を用いた結果である．「銀行」と「携帯電話」どちらの場合も，カテゴリ 2 に属する検索語を多く抽出している．また，不正解の検索語はどちらの例でも多少は関連性がある．

一方，図の右側はコミュニティ空間を用いた結果であり，どちらの場合もカテゴリ 1 に属する検索語を多く抽出している．検索語を入力した後に閲覧したページの種類が多い「携帯電話」の場合は「電報」は若干問題があるものの，全てカテゴリにも属していると判断したが，閲覧したページの種類が少ない「銀行」場合は全く関連が無い検索語が抽出された．関連が無い 2 つの検索語を調べた結果，銀行関連のコミュニティを 1 回だけ閲覧していた．「銀行」との 2 つの検索語のどちらも閲覧したページの種類が少ないため，1 回だけの閲覧でも関連度が高くなるのが原因だと考えられる．このような場合の対処法として，

- 閲覧したページの種類が少ない検索語は関連度の計算を行わない．
- 閲覧頻度が低いものを関連度の計算対象から外す．

などの方法が考えられる．この 2 つを考慮して予備的な実験を行った結果，逆に精度が低下してしまった．これは，関連がある検索語の中にも同様な特徴を持ったものが存在するため，本来ならば関連がある検索語の関連度の低下や関連度自体が計算されなかったことが原因であった．この問題の解決については今後の研究課題とする．

6 おわりに

本稿ではウェブページを閲覧する人々の行動を把握するために，ユーザが指定した検索語と関連する語の発見法の検討を行った．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た．また，本稿では関連語の抽出性能を向上させるために高出現要素を考慮した手法の提案を行い，実験結果から関連語の抽出性能が向上した．

今後は，検索語を入力した後に閲覧したページの種類が少ない検索語の場合の精度の向上を目指す．また，検索語の数を増やして実験を行う予定である．

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

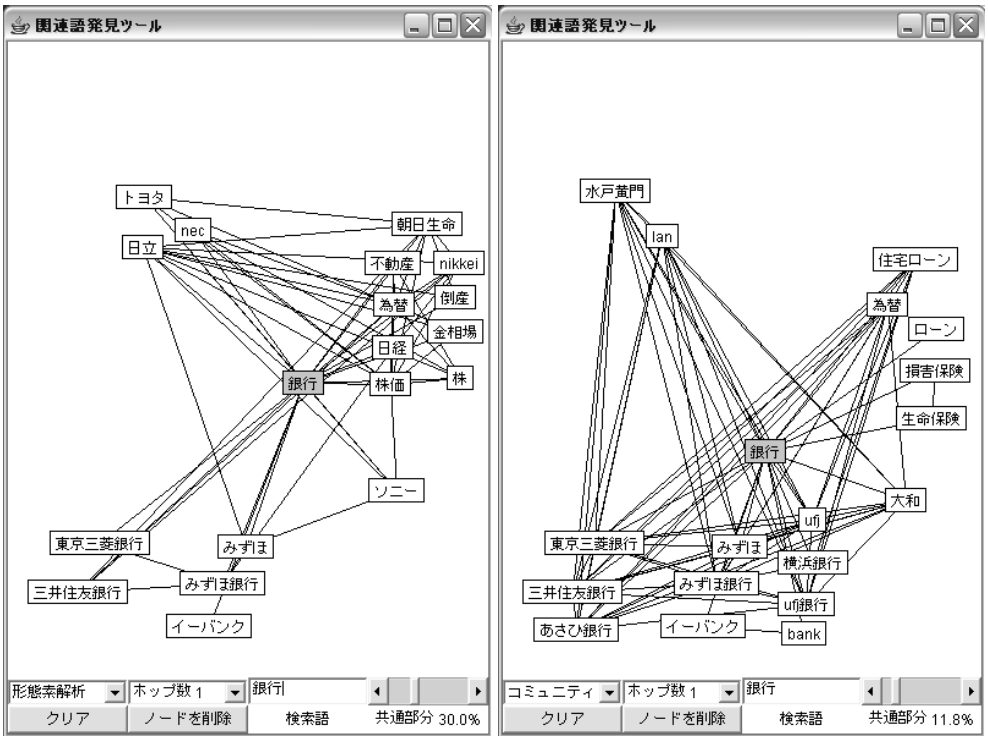

図 7: 「銀行」と関連する検索語の表示例

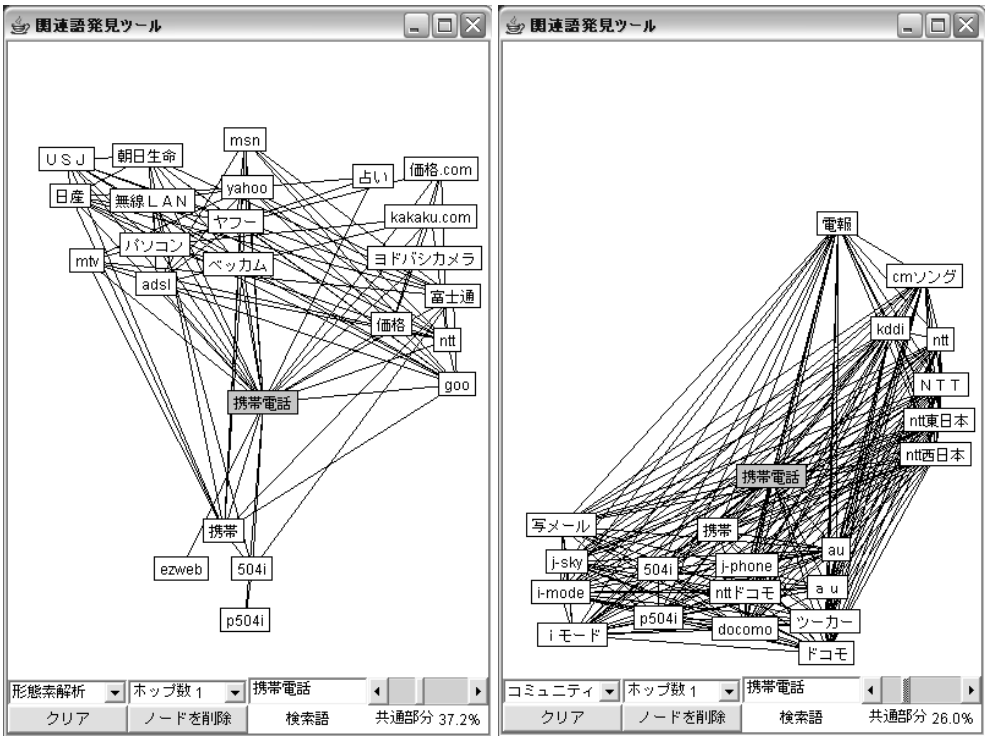

図 8: 「携帯電話」と関連する検索語の表示例

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] 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.

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

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

[6] 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.

[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, 12 2003.

[8] 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.

[9] C. Shahabi, A.M. Zarkesh, J. Adibi, and V. Shah. Knowledge discovery from users web-page navigation. *In Proceedings of the IEEE RIDE97 Workshop*, April 1997.

[10] 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.

[11] 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.

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

[14] 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.

[15] O.R. Zaiane, M. Xin, and J. Han. Discovering web access patterns and trends by applying olap and data mining technology on web logs. *in Proc. Advances in Digital Libraries (ADL'98)*, April 1998.

[16] 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.