大域ウェブアクセスログを用いた関連語の発見に関する一考察

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

大塚 真吾♥ 豊田 正史♥ 喜連川 優♥

Shingo OTSUKA Masashi TOYODA
Masaru KITSUREGAWA

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を基に解析を行っているが，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群を用いた手法を提案する．実験結果より，提案手法はURLだけを用いた手法よりも良い結果が得られた．

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 logs), which are collected URL histories of users (called panels) 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

検索エンジンなどで入力された検索語はユーザの目的や意思を表すため，ウェブ上でのユーザの行動分析に役立つ．例えば，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．

アクセスログを用いた検索語分類の研究では検索語とその直後に閲覧したURLの組合せを基に解析を行っているが，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[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群を用いる手法を提案する．また，実験結果より提案手法が関連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．主な研究として，

- ユーザの行動に関する研究[1][2]
- ウェブページ間の関連に関する研究[3][4]
- 検索サイトに関連する研究[5][6]
- アクセスログの視覚化に関する研究[7][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[9]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，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．

本論文と類似するものとして[5][6]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の内容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．

3. 関連語の発見に必要な技術の概要

3.1 パネルログ

本論文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図1で示す調査方法により集計を行ったデータである．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[10]はこれらを効率良く抽出するアルゴリズムである．本論文では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[11]を利用する．この

♥ 正会員 東京大学生産技術研究所
{otsuka, toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

[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

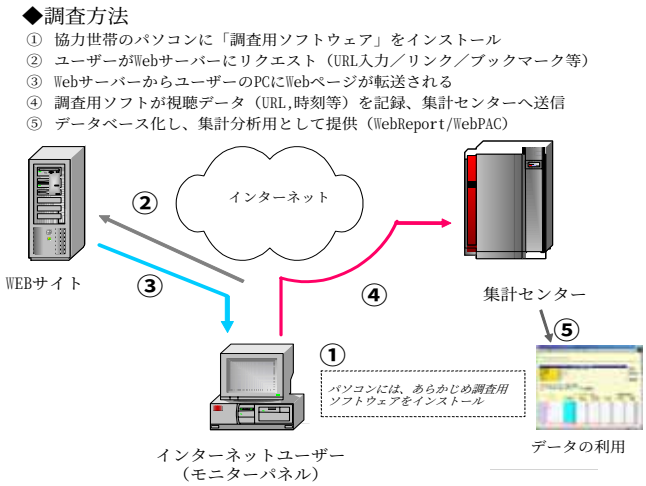

図１　パネルログ収集の概要

Fig.1　Collection method of panel logs

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．本論文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．

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タイトルと簡単な説明文のリストが表示される．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，これを用いて関連語の発見を行う．

表１　パネルログの概要

Table 1　The details of the panel logs

総データ量	9,992(Mbyte)
利用データ量	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)

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

Table 2　The matching factor between the URLs belonging to web-communities and the URLs included in 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（ドメイン）部分を削除して合致	7.7%
合致せず	35.7%

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

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

$$K_{1x}=\frac{|T_1\cap T_x|}{|T_1\cup T_x|}$$

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と頻度を考慮した関連度K'_{1x}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は特徴量の総数である．）

$$K'_{1x}=\frac{\sum_1^m H_1+\sum_1^{m'} H_x(\text{ただし，}T_1\cap T_x\text{空間内})}{\sum_1^m H_1+\sum_1^{m'} H_x}$$

（ただし，m ' はT_xにおける特徴量の総和である．）

[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

[3] 厳密には，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
Table 3　The feature of the evaluated search words

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

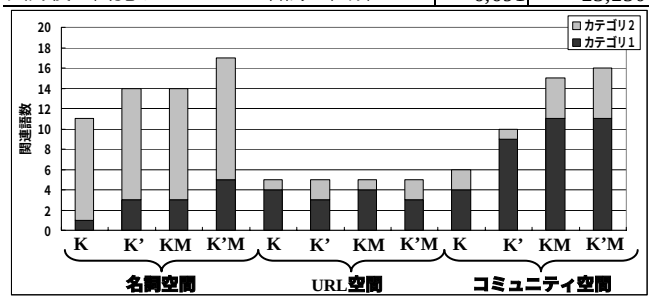

図 2　「銀行」と関連する検索語の評価
Fig.2　A evaluation of search words related to “bank”

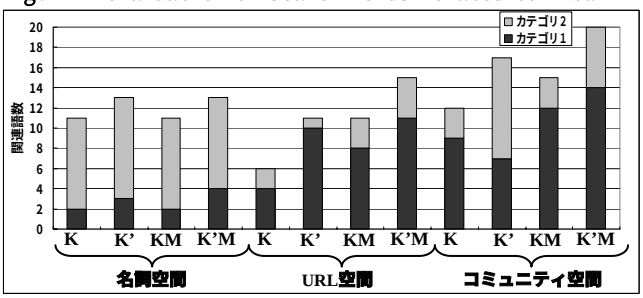

図 3　「携帯電話」と関連する検索語の評価
Fig.3　A evaluation of search words related to “Cellular phone”

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」で行った．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，関連語の判定には以下の2つ基準を設け，どちらのカテゴリにも属さない場合は不正解とした．

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

5.1　実験結果

実験結果を図2,3のK,K'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

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，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．本論文ではこれらの名詞，URL，コミュニティを高出現要素と呼ぶ．

表 4　高出現要素の詳細
Table 4　The details of the high appearance factors

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

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った結果を図2,3のKM,K'M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K'Mを用いて発見された検索語群を図4,5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．

6. おわりに

本論文では関連語の発見法について検討を行った．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，本論文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

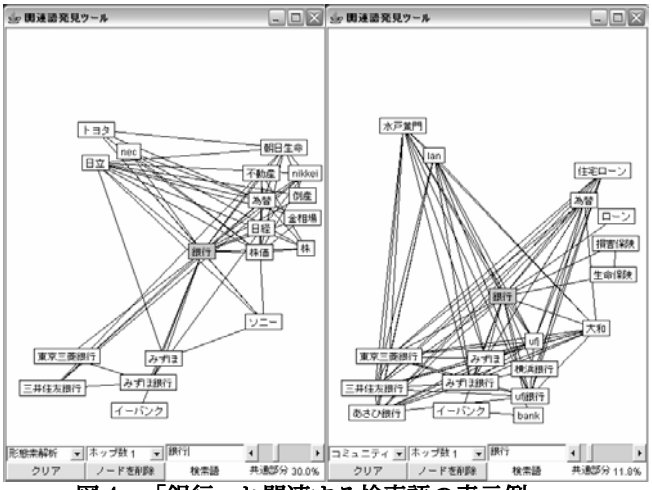

図 4 「銀行」と関連する検索語の表示例
Fig.4 Examples of search words related to “bank”

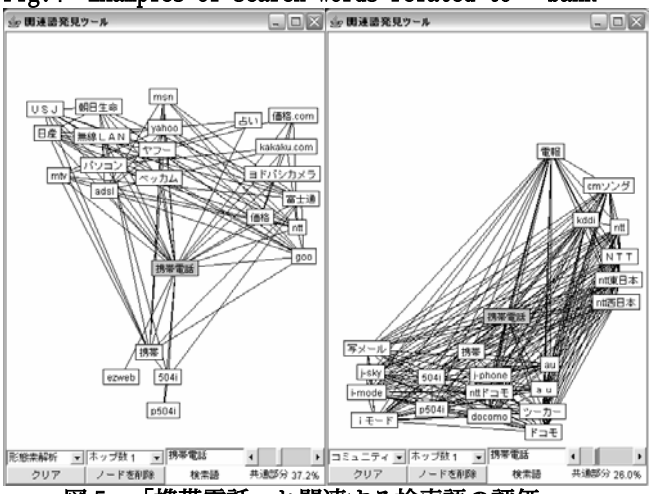

図 5 「携帯電話」と関連する検索語の評価
Fig.5 Examples of search words related to “Cellular phone”

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．また，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，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.

[文献]

[1] 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, pp.32-44 (2003).

[2] Ungar, L.H. and Foster, D.P.: “Clustering Methods for Collaborative Filtering”, AAAI Workshop on Recommendation Systems, 1998.

[3] 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), 2001.

[4] 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, 2002.

[5] 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), 2000.

[6] Wen, J., Nie, J. and Zhang, H.: “Query Clustering Using User Logs”, ACM Transactions on Information Systems (ACM TOIS), Vol. 20(1), pp. 59-81, 2002.

[7] Koutsoupias, N.: “Exploring Web Access Logs with Correspondence Analysis”, Methods and Applications of Artificial Intelligence, Second Hellenic, 2002.

[8] 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), 2002.

[9] Zeng, H., Chen, Z. and Ma, W.: “A Unified Framework for Clustering Heterogeneous Web Objects”, The Third International Conference on Web Information Systems Engineering (WISE2002), 2002.

[10] Kleinberg, J.M.: “Authoritative Sources in a Hyperlinked Environment”, In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.

[11] Toyoda, M. and Kitsuregawa, M.: “Creating a Web Community Chart for Navigating Related Communities”, Conference Proceedings of Hypertext 2001, pp. 103-112, 2001.

大塚 真吾 Shingo OTSUKA

1996 千葉工業大学工学部情報工学科卒．2002 同大学大学院工学研究科情報工学専攻博士後期課程修了．工学博士．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．情報処理学会，日本データベース学会正会員.

豊田 正史 Masashi TOYODA

1994 東京工業大学理学部情報科学科卒．1999 同大学大学院情報理工学研究科博士後期課程修了．理学博士．同年，科学技術振興事業団計算科学技術研究員．2001 東京大学生産技術研究所学術研究支援員．2003 同大 産学官連携研究員．2004 同大 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，日本データベース学会正会員.

喜連川 優 Masaru KITSUREGAWA

1978 東京大学工学部卒．1983 同大大学院工学系研究科情報工学博士課程了．工学博士．同年同大生産技術研究所講師．現在，同教授．2003 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．平成 11-14 年 ACM SIGMODJapan Chapter Chair．平成 9,10 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(97-02), IEEE ICDE, PAKDD,WAIM などステアリング委員会委員.