大規模アクセスログを用いた検索支援システムの提案

The Search Support System Using Global Web Access Logs

大塚 真吾♥ 喜連川 優♥
Shingo OTSUKA Masaru KITSUREGAWA

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，与えられた検索語に関連する検索語(関連語)群を表示し，ユーザに検索語を想起させるシステムの提案を行う．

In cyberspace, users search their interested information by using search engine. 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. 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 propose search support system in order to show the related search words associated with the search words inputted by users.

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論文ではパネルログと呼ぶ．

本論文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．

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

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[1]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あまり行われていない．

文献[2]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[3,4]．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．

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．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を抽出する必要があり HITS[5]はこれらを効率良く抽出す

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

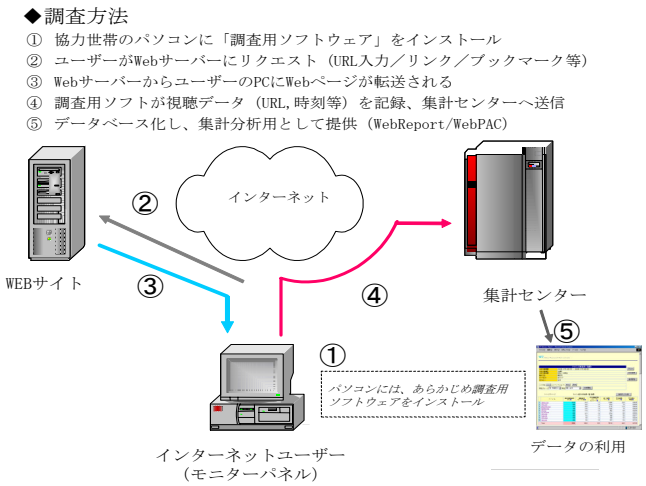

図 1　パネルログ収集の概要

Fig.1　Collection method of panel logs

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[6]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 の適合度を測定した．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．詳細に行いては文献[7]で述べている．

4. パネルログを用いた関連語の発見

4.1　特徴空間の定義

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

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)

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抽出を行う．また，本論文では「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]．

4.2　関連度の定義

我々は関連語を発見するために閲覧ページ集合から**名詞空間，コミュニティ空間，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を抽出して作成した特徴空間である[6]．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．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）

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

$$K_{xy} = T_{x\cap y}/2$$

と定義する．

---

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 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[8]を参照)．

5 実験では日本語形態素解析システム ChaSen (茶筌)[9]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 2　検索語想起支援システム画面（「温泉」の例）

Fig.2　The search support system.

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支援システムの構築を行った．その画面を図 2 に示す．図中(1)に検索語を入力するとその語に関連する語群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．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．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．語数が多い場合は「...」のように省略された表示となるが，右クリックをすると語全体が表示される．

5.1　想起支援例

図 2 は検索語に「温泉」を入力した例である．特徴空間に名詞を用いた結果は「温泉」と関連がある語群を数多く候補として表示している．コミュニティ空間では関連度が低いものはあまり良い結果とはならなかった．また，サイト空間に関しては地名が多く提示される．

その他の例を図 3 に示す．図中の(a)は「携帯電話」を入力した例であるが，サイト空間を用いた場合に関連性のない語が若干表示されるが，そのほかの空間では関連のある語群が検索語候補として表示されている．「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群を候補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．

(a)携帯電話の例　(b)釣りの例

図 3　システムの実行例

Fig.3　The example of search support.

5.2　Google サジェストとの比較

Google サジェストは候補が 10 件のみのため，我々の結果と単純に比較することはできないが，図 2 の「温泉」の例では主に地名が多い．それに対して，我々の結果では「石和温泉」など温泉地の名称や「立ち寄り湯」「お得な宿情報」など温泉と関連性の高い検索語を提示している．図 3(a)の「携帯電話」の例では，Google サジェストと同様な結果の他に「着メロ」「写メール」などを候補として提示している．図 3(b)の「釣り」の例では Google サジェストでは関連がないものが多いのに対して，名詞空間の結果では「釣り」と関連がある語を提示していることがわかる．

5.3　考察

今回の例では閲覧ページ数（検索語を入力した後に閲覧したページの数）が一番少ない語は「釣り」であり，「温泉」や「携帯電話」の 1/5 であった．サイト空間においては閲覧ページ数が多い「携帯電話」では他の検索語と比べて関連性がある語を提示する一方で，閲覧ページ数が少ない「釣り」では関連性のある語をほとんど提示しなかった．このことからサイト空間では閲覧ページ数が多いと提示された語の関連性の高いことがわかる．

コミュニティ空間に関してはコミュニティの精度の影響が強いと考えられ，「温泉」のように閲覧ページ数がほぼ同じであっても提示された語の質が異なっている．

名詞空間に関しては閲覧ページ数に関係なく，どの検索語でも関連がある語を提示していることがわかった．Google

表 2　システムが提示した関連語の評価
Table 2　Evaluation of the related keywords.

cat はカテゴリーの略

検索語	名詞		コミュニティ		サイト	
	cat1	cat2	cat1	cat2	cat1	cat2
銀行	0.43	0.33	0.27	0.40	0.13	0.13
大学	0.97	0.03	0.57	0.04	0.30	0.07
サッカー	0.93	0.03	0.97	0.00	0.43	0.03
釣り	0.97	0.03	0.00	0.03	0.00	0.00
温泉	0.93	0.07	0.50	0.10	0.20	0.33
ガンダム	0.97	0.00	0.20	0.07	0.30	0.00
ドラクエ	0.83	0.07	0.70	0.00	0.27	0.03
競馬	0.53	0.10	0.30	0.13	0.17	0.07
映画	0.73	0.10	0.77	0.13	0.83	0.17
カレンダー	0.23	0.37	0.13	0.07	0.10	0.00

サジェストで「釣り」の例があまり良くない理由として，ネット上で「釣り」と入力するユーザはゲームやアスキーアートなどに興味があり，一般的に連想される「魚を釣る」とは異なっているためではないかと考えられる．

5.4　評価

最後に，我々は検索語想起支援システムにいくつかの検索語を入力し，提示された関連語のうちで関連性の高い 30 語について特徴空間の比較を行った．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．

結果を表 2 に示す．表中の値はカテゴリに該当した関連語を評価した語の数(この実験では 30 語)で割った値である．実験結果から特徴空間に名詞空間を用いるとほとんどの検索語で一番良い結果となった．また，どの検索語についてもカテゴリに含まれる関連語の割合は 0.5 以上であった．コミュニティ空間は良い場合と悪い場合の値が極端であるが，サイト空間より結果が良かった．

6. おわりに

本論文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検索語(関連語)群を表示し，ユーザに検索語を想起させるシステムの提案を行った．関連する検索語群の発見のため，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．利用例から我々のシステムが関連性のある検索語群を提示していることを示し，さらに，既存のサービスとの比較を行った結果 Google サジェストと同等またはそれ以上の関連語を提示していることを示した．最後に主観的ではあるが，システムが提示した関連語の評価を行い，特徴空間に名詞を用いる方法が一番良いという結果が得られた．今後はシステムの有効性を示すために，より客観的な評価を行う．

[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，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] 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.

[2] 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.

[3] 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.

[4] 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.

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

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

[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.

大塚 真吾　Shingo OTSUKA

1996 千葉工業大学工学部情報工学科卒．2002 同大大学院工学研究科情報工学専攻博士後期課程修了．博士(工学)．同年，東京大学生産技術研究所 学術研究支援員．2006 同大同研究所 産学官連携研究員 特任助手．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．

喜連川 優　Masaru KITSUREGAWA

1978 東京大学工学部卒．1983 年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，日本データベース学会理事，情報処理学会，電子情報通信学会 各フェロー．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, IEEE データ工学国際会議(ICDE2005) General Chair.