ウェブ空間におけるユーザ行動の理解支援に関する一考察

大塚真吾† 喜連川優†
†東京大学 生産技術研究所

1 はじめに

本稿ではウェブアクセスログとウェブコミュニティの技術を用いて，ウェブ空間上を動き回るユーザの行動を理解するための支援方法について検討を行う．通常，ウェブサーバのアクセスログなどを利用することで，自サイトを訪れたユーザの直前・直後の URL を知ることが可能だが，URL からではそのページがどのような内容であるか推測することは難しい．そこで本稿では，ウェブコミュニティの技術を用いて問題の解決を試みる．

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 [1]．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 [2, 3] が提案されているが，本稿では我々が以前提案したウェブコミュニティチャート [1] を用いる[1]．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り，良い authority ページおよび良い hub ページを元にコミュニティの作成を行っている．良い authority とは多くの良い hub からハイパーリンクを張られている著名なページを表し，良い hub とはリンク集およびブックマークなど多くの良い authority へハイパーリンクを張っているページを表す．この循環した定義により密に結合した hub と authority が抽出され，それらがよく関連したページを表すことが [1, 4] で示されている．

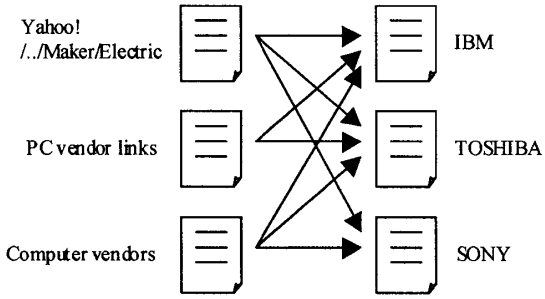

図 1: ハブとオーソリティからなる典型的なグラフ

表 1: コミュニティラベルの例

コミュニティID	コミュニティラベル
18	高知 県立 学校 高等 商業 知江 仁淀 伊野 ...
54	教育 大学 研究センター 高等 センター ...
110	検査 病院 臨床 大学 医 学部 附属 付属 ...
40876	銀行 バンク 住友 パソコンバンキング ...
145535	博物館 県立 東北 歴史 仙台 秋田 福島 ...

典型的な authority と hub のグラフ構造を図 1 に示す．このグラフの右側には大手のコンピュータ関連会社が authority としてあり，それらに密にリンクを張っているリンク集が左側に hub としてある．このようなグラフ構造はウェブ上に多々見られるものである．関連ページアルゴリズムは，図 1 のように密に結合された authority と hub を抽出するものであり，IBM, TOSHIBA, SONY のどれかひとつをシードとして与えることで，これらの会社のリストが結果として出力される．

各々のコミュニティに含まれるページに対して張られたリンクのアンカータグの解析から，充分に正確ではないものの表 1 に示すように，コミュニティの内容を表す単語群（コミュニティラベル）を自動的に抽出できており，解析者はコミュニティに含まれる個々の

A study for the understanding support of the user behavior in the webspace
Shingo OTSUKA† and Masaru KITSUREGAWA†
†Institute of Industrial Science, the University of Tokyo

[1]本稿ではウェブコミュニティチャートのエッジの部分は利用せず，コミュニティ部分のみ利用する．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．

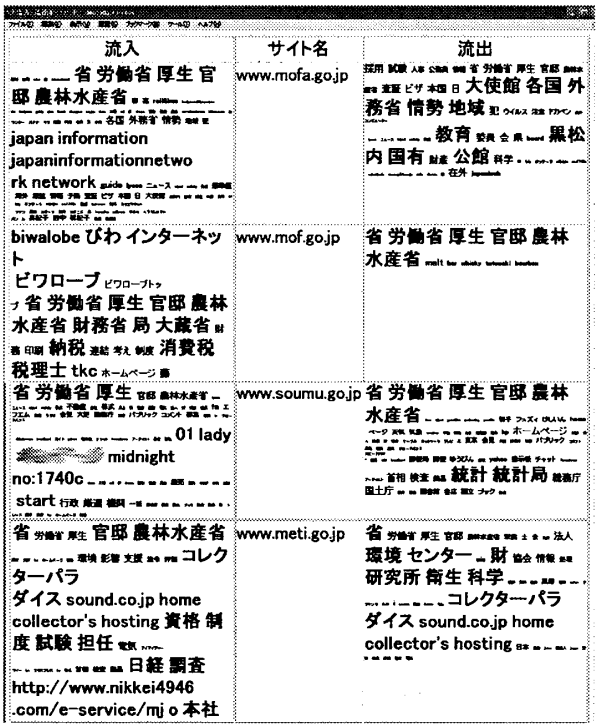

図 2: 動作例

ウェブページを閲覧することなくコミュニティの概要を把握できる．また，ラベル内の単語はコミュニティに含まれるページに対して張られたリンクのアンカータグを形態素解析して名詞や未知語を取り出したものであり，左から頻度が多い順に並んでいる[2]．したがって，ラベルの上位にある単語はそのコミュニティの内容を良く表している単語といえる．

3 ユーザ行動の理解支援

ユーザが自サイトを閲覧する直前の URL(以後，流入 URL と呼ぶ．) や自サイトから他のサイトへ移動した直後の URL(以後，流出 URL と呼ぶ．) は Web サーバのアクセスログや JavaScript などを用いて収集することが可能である．流入・流出 URL からユーザがどのような興味を持って自サイトを訪れたかを調べることは可能だが，URL のみからではページ内容がわからず実際に個々のページを閲覧しなければならないため，その調査には多くの時間と労力を要する．

そこで，我々は前節で述べたコミュニティラベルを用いて流入・流出 URL の内容を解析者に提示する試作システムの構築を行った．動作例を図 2 に示す．図中の「流入」と「流出」はタグクラウドのように重要な単語ほど大きな文字で表示され，文字サイズは以下の手順により決定される．

1. あるサイトに対し，その流入 URL(流出 URL) が所属するコミュニティの検索を行い，そのコミュニティラベルに含まれる単語とその頻度を記録する．
2. 単語の頻度とコミュニティラベルの出現位置をもとに表示させる単語の文字サイズを計算する．[3]

この例では省庁関連（一番上から外務省，財務省，総務省，経済産業省）のサイトに関するユーザの行動を示している．外務省の例から，ユーザは省庁に関連するページから外務省のサイトへ流入し，「大使館」「情勢」などに関連するページへ移動していることがわかる．また，経済産業省の例では，省庁の他に「日経」など経済に関連するページや資格に連関するページから流入し，研究所や資格に関連するページへ移動するユーザが多いことがわかる．このように，試作システムを用いることで各サイトに対してのユーザの行動を大まかではあるものの理解することが可能である．

4 おわりに

本稿では，ウェブサーバのアクセスログとウェブコミュニティの技術を用いてあるサイトに対するユーザの行動を理解するための支援方法を提案し，その試作システムの構築を行った．今後は生成されたタグクラウドの評価について考察を行う予定である．

参考文献

[1] M. Toyoda and M. Kitsuregawa: Creating a Web Community Chart for Navigating Related Communities, Hypertext 2001, 2001

[2] G.W. Flake, S. Lawrence, et al.: Self-organization and identification of Web communities, IEEE Computer, 2002

[3] R. Kumar, P. Raghavan, et al.: Trawling the Web for Emerging Cyber-Communities, Proc. of the 8th WWW conference, 1999

[4] J. Dean and M. R. Henzinger: Finding related pages in the World Wide Web, Proc. of the 8th WWW conference, 1999

[2]多くのラベルに含まれている単語はストップワードとし，ラベル内の単語から削除している．

[3]前節で述べたように，ラベルの上位にある単語はそのコミュニティの内容を良く表しているため文字サイズの計算にも利用した．