ウェブのアクセスパターンの対話的可視化の一手法

理学専攻　情報科学コース　川本真規子

1　はじめに

ウェブの可視化に関する研究は既に広く発表されており，その中でも著者らはこれまでアクセスパターンとリンク構造の同時可視化に着目してきた [1]．本報告では，ユーザが選択したアクセスパターンを可視化する対話的な可視化手法を提案する．本報告ではアクセスパターンを，複数の閲覧者からアクセスされるウェブページ集合と定義する．仮に，m 枚のウェブページ $P = \{p_1, p_2, ..., p_m\}$ があり，これらを閲覧する n 人の閲覧者 $B = \{b_1, b_2, ..., b_n\}$ がいるとする．このとき，P に属する一定枚数以上のウェブページ $P' = \{p_i, p_j, p_k, ...\}$ の全てに対して，B に属する一定人数の閲覧者 $B' = \{b_a, b_b, b_c, ...\}$ の全てからアクセスがあったとき，P' を構成するウェブページ集合を本論文ではアクセスパターンと称する．アクセスパターンを対話的に可視化することにより，ユーザ個々の視点でのアクセスパターンの比較が可能となり，各ユーザの関心に合わせた可視化結果の考察ができると考える．またパターン選択の際，長期間のアクセスログから抽出されたパターンの中からユーザの関心があるものを選べるようにする．これにより，同月の中でのパターンの比較，同パターンの年間を通しての比較など，可視化結果を考察するうえでより詳細な分析が可能となり，ウェブサイトの改善に役立つ有用な知見が得られると期待できる．

2　関連研究

ウェブに関する情報可視化の研究は，1990 年代中盤から非常に多く発表されており，いくつかのサーベイが報告されている [2]．先行研究では，リンク構造とアクセスパターンを同時に可視化するということに着目し研究を行ってきた [1]．先行研究においても典型的なアクセスパターンを発見するなど成果があったが，可視化できるパターン数が限られている，可視化の対象期間が 1ヶ月に固定されている，などの点でユーザ操作の自由度が小さかった．本報告の提案内容ではその点を拡張し，ユーザ自身が関心のあるパターンを選択し対話的に可視化できるようにしている．

閲覧者の興味や類似度を抽出する手法として，文献 [3] では，ウェブアクセスログデータを解析し，閲覧者の興味やアクセスしている情報が時間と共にどのように変化しているのかを抽出して可視化している．また，Nasraoui らは，閲覧者の行動に対して類似度を計算し，ファジークラスタリングを適用することでアクセスパターンを求めている [4]．

3　提案手法

本手法では前処理として，

- アクセスログからのアクセスパターン構築
- クローラを用いたリンク構造構築

により入力データを生成する．そして，これらのデータを可視化手法「FRUITS Net」を用いて可視化する．

3.1　アクセスパターン抽出

本研究におけるアクセスパターン抽出は，ウェブページへのアクセスに対する閲覧者の共起性に着目した手法である．本処理ではまず，アクセスログファイルを読み込み，閲覧者と URL の一覧を作成する．ただし我々の実装では，画像や音楽などのコンテンツファイルの URL を削除し，それ以外の URL だけを対象とする．続いて本処理では，閲覧者の IP アドレスの数を n，アクセスされた URL の数を m として，$n \times m$ の表を作成する．表の各欄には，各閲覧者から各 URL へのアクセス回数の集計結果を記録する．続いて本処理では，閲覧者のデンドログラムを構築する．このとき 1 閲覧者のアクセス回数を m 次元ベクトルとして，すべての閲覧者ペアについてベクトル間余弦を算出する．閲覧者 x および y からの各ウェブページへのアクセス回数を，m 次元ベクトル $x = (x_1, x_2, ..., x_m)$ および $y = (y_1, y_2, ..., y_m)$ と表したとき，ベクトル間余弦は以下の式で表される．

$$S_{cos}(x, y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \tag{1}$$

上記の式で算出される余弦値が最大となる 2 閲覧者を連結してクラスタを構成し，さらに重心法を用いてクラスタ間の余弦値が最大となる 2 クラスタを連結する，という処理を再帰的に反復することで，デンドログラムを構築する．

続いて，このデンドログラムを用いて閲覧者をクラスタリングする．この処理では閾値 α を決め，余弦値が閾値以上の閲覧者群を 1 つのクラスタとする．そして，全クラスタの中から構成員数が β 人以上のクラスタを抽出する．続いて各クラスタに対して，クラスタ構成員の γ %以上がアクセスしたページを抽出し，アクセスパターンのデータを構築する．α，β，γ はそれぞれユーザが定義した値を使用する．4 章で使用しているデータは，$\alpha = 0.6$,

$\beta = 3$, $\gamma = 60$ とした．また，現時点での実装では，3 ページ以上のページが抽出されたアクセスパターンのみを可視化の対象とした．

3.2 ネットワーク可視化

本手法では，force-directed 法と空間充填モデルの 2 種類の画面配置手法を組み合わせたネットワーク可視化手法を適用した [5]．本手法では，各カテゴリ情報に独立した色を割り当てており，カテゴリ情報を有するノードは色付けられた円として描かれる．複数のカテゴリ情報を有するノードの場合は，円の内部を分割して 1 つのノードに複数の色を付けられるようにしている．本手法では，ウェブページをノード，ハイパーリンクをエッジ，ウェブサイトのディレクトリ構造をクラスタとし，アクセスパターンの情報を色で表す．

3.3 対話的可視化

3.1 節で抽出したアクセスパターンを図 1 のように GUI 操作画面上のアクセスパターン選択パネルにボタンとして設置する．このパネルは格子状にボタンが配列されており，横軸がアクセスパターンを形成するウェブページ群，縦軸が時期を示す．また，パターン選択の際のユーザの指標となるように，各パターンに含まれるページの平均アクセス数を算出し，アクセス数の量によって選択パネル上のボタンの色を 5 段階で色分けする．このパネル上でユーザが関心のあるパターンを選択することで，選択されたパターンに含まれるページのみがネットワーク描画の際に色付けされる．

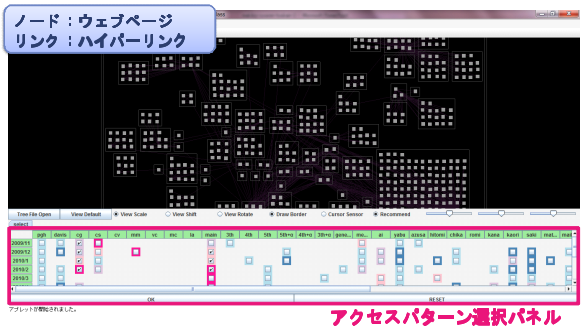

図 1: GUI 操作画面

4 適用事例

本報告では，所属研究室のウェブサイト (http://itolab.is.ocha.ac.jp/) に本手法を適用した事例を報告する．利用したアクセスログは 1 年分（2009 年 11 月〜2010 年 10 月）で，1 月当たりの平均アクセス総数は約 70000 件であり，当時のウェブサイトの総ページ数は 621 ページであった．我々は提案手法を Java JDK1.5.0 を用いて実装し，Windows7 (CPU 1.2GHz, RAM 4GB) を用いて実行した．

アクセスパターン選択パネルを見てみると，講義の資料 cg のパターンは年間を通して 2 月のみアクセス数が多いということがわかった．そこで，cg に関する全てのパターンを選択し可視化してみると図 2 のような可視化結果が得られた．各ページの色を詳しく見てみると，研究室紹介のページは緑色しかついていないことがわかる．緑色はパターン選択の際に 2 月のパターンに割り当てられた色であった．そのため，研究室紹介のページは 2 月のパターンにのみ含まれているということがわかる．2 月は研究室配属が実施される月であるが，閲覧者が講義資料と合わせて研究室紹介のページにもアクセスしている様子が可視化結果より読み取れる．このように，通常のアクセスパターンには含まれていないページを発見することは，イベント時などに閲覧者がどのページを求めてアクセスしてきているのかという把握につながり，ウェブサイトの更新時期やページ内容の検討などに役立つと考える．

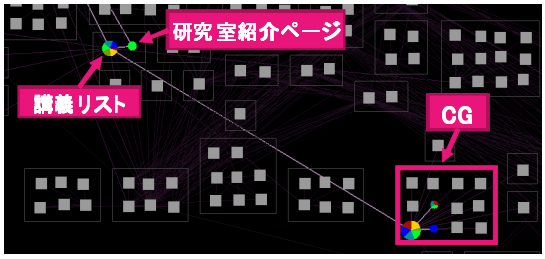

図 2: 可視化結果（cg に関するパターンを選択）

5 まとめ

本報告では，「FRUITS Net」を用いたアクセスパターンの対話的可視化の一手法を提案した．本手法を用いて，ユーザがアクセスパターンを対話的に可視化することによって，イベント時の閲覧者のアクセス傾向の把握などウェブサイト改善に役立つ知見が可視化結果より発見できることがわかった．

今後は，より大規模なウェブサイトでの本手法の適用や，現在実装しているアクセスパターン抽出手法についての再検討を行いたい．また，サイト構成上の問題点を発見しやすくなるように可視化手法を改良したいと考えている．

参考文献

[1] M. Kawamoto, T. Itoh, A Visualization Technique for Access Patterns and Link Structures of Web Sites, *International Conference on Information Visualization*, pp. 11-16, 2010.

[2] An Atlas of Cyberspaces, http://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/atlas/atlas.html

[3] 山田和明，中小路久美代，上田完次，Web ユーザの行動履歴解析のためのデータマイニング，電子情報通信学会 WI2 研究会資料，pp. 59-64, 2005.

[4] O. Nasraoui, H. Frigui, A. Joshi, R. Krishnapuram, Mining Web Access Logs Using Relational Competitive Fuzzy Clustering, *Eight International Fuzzy Systems Association World Congress*, 1999.

[5] T. Itoh, C. Muelder, K.-L. Ma, J. Sese, A Hybrid Space-Filling and Force-Directed Layout Method for Visualizing Multiple-Category Graphs, *IEEE Pacific Visualization Symposium*, pp. 121-128, 2009.