ウェブのアクセスパターンとリンク構造の同時可視化の一手法

川本真規子　（指導教員：伊藤貴之）

1. 概要

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いる．ウェブ可視化の対象は，リンク構造や文書内容などウェブサイト本体に関する情報と，アクセス統計をはじめとする閲覧者情報に大別される．この 2 種類の情報を一画面に同時に可視化することで，ウェブサイトの構築や管理に関する有用な知見が得られることが期待される．

本研究では，「FRUITS Net」[1]という可視化手法を用いて，アクセスパターンとリンク構造の同時可視化を試みる．ここで本研究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同一ウェブページ群と定義する．これをリンク構造と同時可視化することにより，アクセスパターンとリンク構造が適切に対応しているか，という知見を視覚的に得られると考えられる．そしてこの知見を，適切なリンク構造の再構築，ウェブサイトのページ構成やページ内容の検討，などに活用できると考えられる．

2. 関連研究

閲覧者の興味遷移の抽出手法として，文献[2]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．

ウェブサイトのリンク構造を可視化する手法として，ウェブサイトを階層型グラフデータとして表現し，力学モデルを用いたグラフデータの画面配置手法により可視化する手法が挙げられる[3]．また，ウェブサイトのアクセス分布の可視化手法として，文献[4]ではアクセス統計とリンク構造の同時可視化を試みているが，この手法でのリンク構造は 1 ページを根とした木構造に限定されている．

本研究で用いる「FRUITS Net」[1]は，ノード配置とノード着色の工夫により，リンク構造とカテゴリ情報（本報告ではアクセスパターン）の同時可視化を目指す手法である．「FRUITS Net」では，力学モデルに基づく画面配置アルゴリズムにより，

[条件 1] 共通のカテゴリを有するノードが画面上で近くに配置される

[条件 2] リンク長の総計とリンク間交差を減らす

の 2 条件を満たすような配置を実現する．さらに，テンプレートを用いた空間充填モデルに基づく画面配置アルゴリズムによって配置結果を修正することにより，

[条件 3] ノードが画面上で重ならない

[条件 4] 配置結果の画面占有面積を減らす

という 2 条件も同時に満たすような配置を実現する．

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「FRUITS Net」で同時可視化する．

3.1 入力情報の定義

本報告では，アクセスログファイルの入手可能な 1 ドメインを対象として，そのウェブサイトのトップページからクローラによってリンクを辿ることで，リンク構造を構築する．

また本報告では，標準的なアクセスログとして，閲覧者 IP アドレス，アクセス日時，アクセスされたファイル名，リンク元ページの URL，使用している OS 名やブラウザ名，などが記録されているアクセスログファイルの使用を前提とする．

ただし，このアクセスログファイルから，閲覧者の全てのページ遷移を抽出できるとは限らない．例えばウェブブラウザの「戻る」ボタンを押した場合などには，キャッシュされたウェブページを再表示するためにサーバへのアクセスが発生せず，結果として閲覧履歴がアクセスログファイルに記録されないことがある．そのため本研究では，以下の 2 種類の立場を想定するものとする．

[立場 a] 閲覧者がどのページからどのページへ辿ったというページ遷移を一切参照しない

[立場 b] 徹底的に閲覧者のページ遷移を記録する

[立場 a]では，アクセスパターンとリンク構造を同時に可視化することで，画面上で閲覧者のページ遷移を想像できるが，その正当性は保証されない．[立場 b]では，正当性のある形で閲覧者のページ遷移を可視化できる．しかし，そのためにはウェブサイトの各ページにトラッキングコードを埋め込む，あるいは閲覧者側のパソコンに特定のプログラムをインストールしてデータを採るなど，特殊な方法で閲覧者の全ページ遷移を記録する必要がある．だが，これらの方法を使用すると，他のウェブサイトでの応用が困難になる，あるいは限られた閲覧者のページ遷移しか採れない，といった制限が生じる．そのため現時点では，我々は[立場 a]を前提として研究を進めている．しかし原理的には，提案手法は[立場 b]を前提とすることも可能である．

3.2 アクセスパターン構築

アクセスパターン構築における我々の実装は，以下のとおりである．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し，これが最大であるペアを併合する．この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．そして，このデンドログラムを用いて閲覧者を階層的にクラスタリングする．続いて各クラスタに対して，所定人数を超える閲覧者がアクセスしたページを抽出することで，アクセスパターンのデータを構築する．現時点での我々の実装では，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ている．

3.3 リンク構造データ構築

リンク構造のデータ構築にはクローラを使用する．本処理では，アクセスログファイルを入手したサイトのトップページを指定し，そこからリンクで繋がっているページを抽出してリストを作り，得られた URL をもとにリンクのグラフを構築する．我々の実装では，オープンソースとして提供されている「JSpider」[5]というクローラを採用している．

3.4 可視化

本手法では 3.2 節および 3.3 節で構築されたデータを統合し，「FRUITS Net」を用いて可視化する．本処理ではまず，3.2

節および 3.3 節で抽出された URL を統合し，ディレクトリ構造に基づいて階層的に格納することで，各 URL を葉ノードとする木構造を生成する．そして各ノードにリンク構造を付加することで，階層型ネットワークを形成する．さらに，各 URL にアクセスパターン情報を付加することで，可視化のための入力データを構築する．

本手法にて「FRUITS Net」を適用する利点は，以下のとおりである．まず[条件 1][条件 2]により，同一アクセスパターンに属する URL や，リンクされた URL が，画面上の近い位置に配置されるので，アクセスパターンとリンク構造の関係を視覚的に理解しやすくなる．また[条件 3][条件 4]により，できるだけ多くの URL を一画面上で一覧できるようになる．

4. 適用事例

本章では，我々の所属研究室のウェブサイトに本手法を適用した事例を報告する．

本手法による可視化結果において，各ノードはウェブページ，カテゴリ情報を表す色はアクセスパターン，リンクはウェブページ間のハイパーリンクを表す．ノードの大きさはリンク数に比例する．図 1 は可視化結果画面の一例である．右端は GUI の操作部分であり，色一覧がアクセスパターンに使われている色を表している．

図 1 は，7 月のアクセスログファイルから抽出されたアクセスパターンを用いて可視化を行った結果である．この結果では，教員の担当講義の資料ページへのアクセスパターンに該当する部分を拡大表示している．この可視化結果から，教員のトップページから担当科目のページへアクセスし，そこから専門科目の資料のページへアクセスした，という閲覧者のページ遷移を想像できる．この結果から，目的のページにアクセスするために，トップページから順にリンクを辿ったアクセスが多数あったことがわかる．

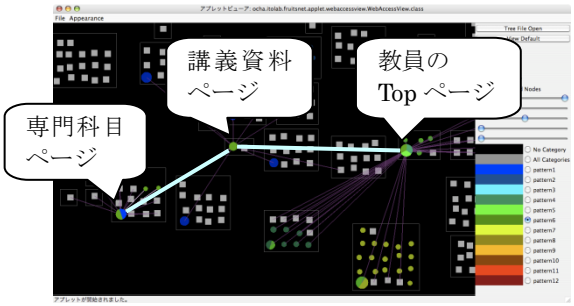


図 1：可視化結果 1（教員の講義科目の資料ページに関するアクセスパターン）

別の事例として，研究室のメンバーのページに関連するアクセスパターンに関する可視化結果を示す．図 2 は，メンバー一覧のページから，同学年全員のホームページにアクセスする閲覧者が複数いることを示している．図 3 は，同クラスタ内の大半のページにアクセスされ，特定の個人のホームページだけをアクセスする閲覧者が複数いることを示している．

本事例を通して我々が発見したアクセスパターンは，主として以下の 3 種類の分布に分類できる．

[分布 1] 直線型のアクセスパターン（図 1 参照）

[分布 2] 1 つのノードを中心とした放射型のアクセスパターン（図 2 参照）

[分布 3] 同クラスタ内だけにアクセスが集中している同クラスタ型のアクセスパターン（図 3 参照）

直線型のアクセスパターンをとる閲覧者は，目的の 1 ページにアクセスするためにリンクを辿ってアクセスしているということが想像される．このアクセスパターンをとる場合には，目的のページまでに，どのくらいのページを経由していったかということが重要である．多くのページを経由している場合はリンクの再構築を検討する必要があると考えられる．

放射型のアクセスパターンをとる閲覧者は，あるページを中心として紹介されている複数のページに関心があるということが想像される．このアクセスパターンの場合は，一緒にアクセスされたページがどのようなページなのかということを知ることで，閲覧者の興味を把握することができ，ページ内容の充実に役立てられると考えられる．また，同クラスタ型のアクセスパターンをとる閲覧者は，同じディレクトリ内のページにだけ関心があるということが想像される．このアクセスパターンの場合は，同クラスタ内にあるアクセスされなかったページに着目することが重要である．アクセスされなかったページの内容は適切か，そのページへのリンクはきちんと張られているかということの再確認に役立つと考えられる．このように，可視化結果に現れたアクセスパターンの分布から，サイトを訪問した閲覧者がどのような意図を持ってアクセスしているのかということを視覚的に捉え，想像することによって，リンクの再構築やページ内容の再考に役立てることができる．

5. まとめ

本報告では，「FRUITS Net」を用いたアクセスパターンとリンク構造の同時可視化の一手法を提案した．本手法を用いて，アクセスパターンとリンク構造を対応させて同時可視化することで，アクセスパターンごとに異なる分布を画面上で発見できることがわかった．本報告で紹介した適用事例では，3 種類の分布を発見することができた．

今後は，各ページのアクセス数を高さで表示できるようにするなど「FRUITS Net」の機能の拡張や，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

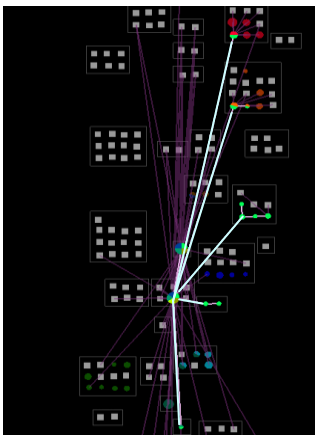
[1] T. Itoh, C. Muelder, K.-L. Ma, J. Sese, A Hybrid Space-Filling and Force-Directed Layout Method for Visualizing Multiple-Category Graphs, 2009 IEEE Pacific Visualization Symposium, pp. 121-128, 2009.

[2] 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.

[3] 土井淳，伊藤貴之，力学モデルを用いた階層型グラフデータ画面配置手法の改良手法とウェブサイト視覚化への応用，芸術科学会論文誌，Vol. 3, No. 4, pp. 250-263, 2004.

[4] 山縣修，中村泰明，アクセス確率による Web サイトのリンク構造可視化ツール，可視化情報学会論文集，Vol. 26, No. 6, pp. 43-50, 2006.

[5] 「JSpider」 http://j-spider.sourceforge.net/



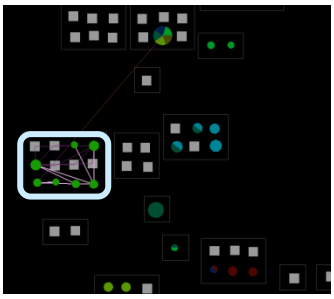


図 3：可視化結果 3（1 人のメンバーのウェブページ群に関するアクセスパターン）

図 2：可視化結果 2（同学年の各メンバーのページに関するアクセスパターン）