

## 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と適用事例

川本真規子 伊藤貴之

お茶の水女子大学大学院 人間文化創成科学研究科

## A Visualization Technique for Access Patterns and Link Structures of Web Sites and Its Case Study

Makiko Kawamoto Takayuki Itoh

Graduate School of Humanities and Sciences, Ochanomizu University

{makiko, itot} @ itolab.is.ocha.ac.jp

### 概要

ウェブの可視化手法は、ウェブサイト本体に関する情報（リンク構造や文書内容など）の可視化と、閲覧者のアクセス動向の可視化に大別される。この2種類の可視化を統合することで、ウェブサイト運営に関する有益な知見が得られると考えられる。

本論文では、カテゴリ情報を有するネットワークを効果的に画面配置するネットワーク可視化手法「FRUITS Net」を用いて、ウェブサイトのアクセスパターンとリンク構造を可視化する手法を提案する。本手法では、クローラによりリンク構造を、アクセスログファイルによりアクセスパターンを構築し、これらの情報を可視化する。我々は本手法が、ウェブサイトのページ構成やページ内容の再検討に役立つと考えている。本論文では、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、アクセスパターンとリンク構造との関係性、ウェブサイトに訪れる閲覧者の特徴、などを発見した事例を報告する。

キーワード 可視化, アクセスパターン, リンク構造, アクセスログ

### Abstract

There have been two types of Web visualization techniques: visualization of Web sites themselves based on link structures or lexical contents, and visualization of browsers' behaviors.

We think that integration of such two visualization techniques is very useful for Web site management, and therefore we are currently studying on visualization of access patterns and link structures on a single screen. This paper presents a Web visualization technique using 'FRUITS Net', a visualization technique for multiple-category-embedded network data. The presented technique constructs link structures using crawler software, and access patterns from access log files. We expect that users can utilize the knowledge obtained from the visualization results for design and management of Web sites. This paper presents several case studies of relationships between access patterns and link structures, and features of those who access to the Web site and so on.

**Keyword** Visualization, Access Pattern, Link Structure, Access Log

## 1 はじめに

ウェブに関する情報可視化の研究は、1990年代中盤から活発に報告されている。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。これら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。

我々はウェブサイトの閲覧者情報の中でも、アクセスパターンの可視化に着目している。本論文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  $P = \{p_1, p_2, \dots, p_m\}$  があり、これらを閲覧する  $n$ 人の閲覧者  $B = \{b_1, b_2, \dots, b_n\}$  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  $P' = \{p_i, p_j, p_k, \dots\}$  の全てに対して、 $B$ に属する一定人数の閲覧者  $B' = \{b_a, b_b, b_c, \dots\}$  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。多くの場合において、アクセスパターンとリンク構造には一定の関係があると考えられる。そして、アクセスパターン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との関係性、ウェブサイトに訪れる閲覧者の特徴、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、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。そこで本研究では、各アクセスパターンが、どのようなアクセスによるものなのかを理解するための可視化手法の構築を目標とする。

ウェブサイトの可視化技術の多くは、ウェブページをノード、ハイパーリンクをエッジに置き換えて、ネットワーク可視化技術を適用することでウェブサイトを表現している [1]。これに加えてアクセスパターンも可視化する場合に、リンク構造とアクセスパターンの双方の可読性を高めるための要件は、かなり複雑になる。そのため、汎用的なネットワーク可視化技術をそのまま適用したのでは、リンク構造とアクセスパターンの双方の可読性を高めることは難しいと考えられる。

本論文では、力学モデル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「FRUITS Net」[2]というネットワーク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た事例を報告する。本手法では、まずウェブページを葉ノード、ウェブサイトのディレクトリ構造をクラスタとして、ウェブページの木構造を構築する。続いてウェブページ間のハイパーリンクに基づいて、葉ノード間にリンクを生成する。以上の手順によって構築されたデータ構造を、本論文では「リンク付き木構造」と称する。そして、以下の4つの条件を満たすようにノードを画面配置する。

条件 1: 同じアクセスパターンに属するウェブページを画面上の近い位置に配置する

条件 2: ハイパーリンクで接続されたウェブページを画面上の近い位置に配置する

条件 3: ウェブページやディレクトリの重なりを回避する

条件 4: ウェブページやディレクトリの画面占有面積を低減する

本論文の提案手法には、リンク構造とアクセスパターンを同時に可視化した点、また、これら双方の可読性を高めるために上記の4条件を満たすネットワーク可視化手法「FRUITS Net」を採用している点に新規性があると考えられる。

本論文では、我々の所属研究室のウェブサイトを例として本手法を適用した結果を示し、我々が可視化した典型的なアクセスパターンについて議論する。

## 2 関連研究

### 2.1 ウェブのアクセスパターンの抽出

ウェブサイトにおける閲覧者の行動は、ウェブデザイナーや管理者にとって、とても興味深い情報である。そのため、そのような閲覧者の行動分析や抽出手法についての研究が活発に行われている。その一環として抽出される情報がアクセスパターンである。

「アクセスパターン」という単語の定義は必ずしも厳密に一意ではないが、多くの論文において、アクセスパターンの抽出とは「頻出するアクセス遷移パターン」「アクセスの共起性が高いウェブページの集合」「アクセス遷移に類似性が見られる閲覧者群」などの抽出を意味する。

頻出するアクセス遷移パターンを抽出する手法として、以下のようなものが発表されている。文献 [3] では、閲覧者の1セッション内においてアクセスされた一連のURLの推移について、Longest Common Subsequence(LCS) アルゴリズムを用いて頻出のアクセスパターンを抽出するという方法を提案している。また文献 [4] では、各ページに閲覧時間の長さに応じて重みを付け、グラフマイニングによりアクセスパターンを抽出する方法を提案している。文献 [4] では、アクセスパターンとして分岐遷移を含むユーザのページ遷移をグラフ構造で抽出している。Pitkowらは、マルコフモデルに基づいたユーザ閲覧パターンの予測モデルを提案している [5]。Davisonは、ユーザが最近訪れたページの内容からユーザの関心を見つけることによって、ユーザの次の行動を予測する手法を提案している [6]。

閲覧者の興味や類似度を抽出する手法として、以下のようなものが発表されている。文献 [7] 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。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 [8]。

このように、ウェブのアクセスパターンの抽出には多くの考え方に基づいた多くの手法が発表されているが、それらは優劣関係を有するというより、互いに相補的な関係にあるといえる。本研究で適用したアクセスパターン抽出手法(4.1節参照)は、閲覧者間のアクセス類似性に基づいて、共起性の高いウェブページ集合を抽出しているが、上述の関連手法に置き換えて適用することも可能である。

## 2.2 ネットワーク構造の可視化

本論文の提案手法が採用しているウェブサイトの可視化手法は、ネットワーク構造の可視化手法の一種である。ウェブサイト限定しない汎用的なネットワーク構造の可視化手法は、グラフ描画 (Graph Drawing) [9] という理論を応用した技術として、長い間に渡って活発に研究されている。

ネットワーク構造の可視化手法の多くは、ネットワークを構成するノードを点群として描画し、ノード間のリンクを直接または曲線で表現する、いわゆる「ノード=リンク型」を採用している。ノード=リンク型のネットワーク構造可視化手法の最も大きな課題は、ネットワークを構成するノードの効果的な画面配置によって、どのようにネットワーク構造の可読性を高めるか、という点にある。これを解決する最も汎用的な手法の一つとして、ノード間を連結するリンクにバネ等の力学モデル<sup>1</sup>を仮想し、運動方程式の反復処理によって最適なノード配置を求める、という手法がある。この手法は force-directed 法 [10] [11] と呼ばれ、グラフ描画のための力学モデルの最も有名なものとして広く実用されている。

それに対して本研究が対象とするネットワーク構造は、ノード間を連結するリンク構造だけでなく、各ノードがカテゴリ情報を有するものである。ここでカテゴリ情報とは、あらかじめ指定された数種類のカテゴリのうち、各ノードがどのカテゴリに属するかを表すものであり、本論文では各ウェブページのアクセスパターンがカテゴリに相当する。このようなネットワーク構造において、リンク構造とカテゴリ情報の両方の可読性を高めるための要件は複雑になる。本論文で採用する可視化手法「FRUITS Net」<sup>2</sup>は、force-directed 法と空間充填法の2種類の画面配置手法を組み合わせたノード配置、およびノード着色の工夫により、リンク構造とカテゴリ情報の両方に対して可読性を高める、という点においてネットワーク構造可視化の新しい手法であるといえる。「FRUITS Net」については3章にて詳しく説明する。

## 2.3 ウェブサイトの可視化

1章でも述べたとおり、ウェブに関する情報可視化の研究は、1990年代中盤から活発に報告されており、いくつかのサーベイが報告されている [12]。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。ウェブ可視化の初期の研究は、リンク構造や文書内容の可視化が多かったのに対して、ウェブ可視化を本格的にウェブサイト管理やウェブビジネスに活用したいという発想から、アクセス統計をはじめとする閲覧者情報の可視化の事例が増

<sup>1</sup>手法によっては、リンクで連結されていないノード間にも分子間力などの力学モデルを仮想する。

<sup>2</sup>この可視化手法の名称のうちFRUITSは、FRamework and User Interface for Tangled Segmentsの略である。ネットワーク可視化手法に限らず多くの可視化手法において、画面上で線分群が絡まって可読性が低下する、という問題を解決するために提案されている手法群の総称である。FRUITS Netはその一環として、ネットワーク可視化手法として提案されたものである。

えてきた。以下、ウェブサイトの可視化手法の中でも、本研究に関連する例として、以下があげられる。

ウェブサイトのリンク構造の可視化において重要な点に、以下があげられる。

- 可読性の高い形でリンク構造を表現するために、良好なネットワーク構造可視化手法を適用することが望ましい。
- 大量のウェブページで構成されるウェブサイトを、よく整理された形で可視化するために、ウェブページを内容的に分類して可視化することが望ましい。

これらの要件を満たす手法として土井らは、ウェブサイトをリンク付き木構造として表現し、force-directed型の力学モデルを改良して可視化した [1]。

アクセスログファイルから得られる情報を可視化した例として、文献 [13] では、ウェブサイトを提供されるサービスの関連性分析のための可視化方法を提案している。また、ウェブサイトのアクセス統計の可視化手法 [14][15] もいくつか提案されている。このうち文献 [15] 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。

## 3 可視化手法「FRUITS Net」

「FRUITS Net」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可視化手法である。

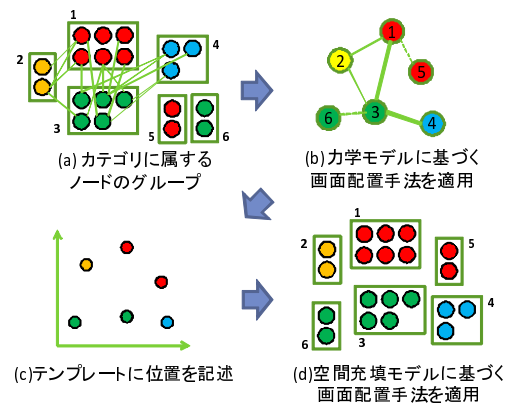


図 1: 画面配置処理手順

「FRUITS Net」によるネットワーク可視化の処理手順を、図 1 に示す。図 1(a) は色付けされたノードのグループの集合の例である。図 1(b) はサブネットワークの例である。ここで、サブネットワークのエッジは、少なくとも1つ以上のノードがリンクで結ばれているグループ間、もしくは、同じ色のノードで構成されるグループ間を結ぶ。続いて、力学モデルに基づく画面配置手法を適用することで、各グループの位置を計算する。具体的には、グループ間を連結する各エッジにバネを仮想し、反復的解法によってバネ間の力に関する運動方程式を解くことで、グループを連結

するエッジの長さを適正化する．続いて図 1(c) に示すように，力学モデルによる配置結果をテンプレートと呼ばれるデータとして記述する．最後に，テンプレートに記述されている位置を理想位置として，空間充填モデルに基づく画面配置手法を適用し，力学モデルによる配置結果を修正する．具体的には，グループを表現する各長方形が，

空間充填条件 1: 互いに重ならない

空間充填条件 2: 画面占有面積を最小化する

空間充填条件 3: テンプレートに記述された理想位置にできるだけ近い位置に配置する

という条件をできるだけ満たすような位置に，各長方形を配置する．図 1(d) は空間充填モデルによる配置結果を描いたものである．なお「FRUITS Net」における画面占有面積とは，画面に表示される全ての長方形を内挿する長方形領域の面積のことである．

以上の配置手法により，1 章で述べた条件 1~4 をできるだけ満たす画面配置を実現する．具体的には，力学モデルに基づく画面配置手法により [条件 1] と [条件 2] を満たし，空間充填モデルに基づく画面配置手法により [条件 3] と [条件 4] を満たしている．そのため，同じカテゴリ情報を有するノードや，リンクで接続されたノードを近くに配置し，また，重なりを回避しながら，できるだけ多くのノードを一画面上に表示することが可能となっている．

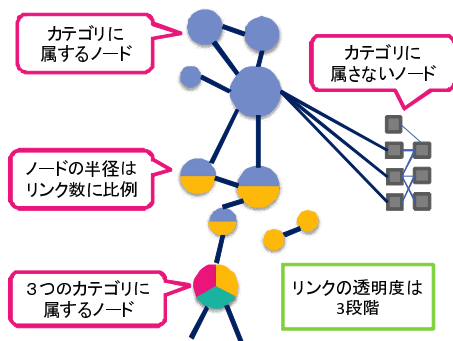


図 2: ノードとリンクの描画

図 2 は「FRUITS Net」でどのようにノードやリンクが描かれているかを示している。「FRUITS Net」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．ノードの半径は接続されたリンク数に比例しており，たくさんリンクを持つノードほど大きく描かれる．また，カテゴリ情報を持たないノードは，小さな灰色のドットで描かれる．

さらに，リンクの描画に際して，両端のノードに着目してリンクを

1. 色づけされたノード同士を連結するリンク

2. 色づけされたノードと色づけされていないノードを連結するリンク

3. 色づけされていないノード同士を連結するリンク
- の 3 種類に分類し，1. を最も明るい色で描画し，3. を最も透明度の高い状態で目立たないように描画している．

## 4 アクセスパターンと リンク構造の同時可視化手法

本論文で提案する手法の処理手順は以下の通りである．本手法ではまず，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，入力データであるリンク付き木構造を生成する．そして，このリンク付き木構造を「FRUITS Net」で可視化する．

本論文では，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により，リンク構造を構築する．また本論文では，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，リファラ，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 3 種類の立場を想定するものとする．なお，この「立場」というのは，我々可視化手法の開発者のとる立場である．

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: アクセスログファイルから抽出できるページ遷移のみを参照する

立場 c: 徹底的に閲覧者のページ遷移を記録したログファイルを参照する

[立場 a] では，ページ遷移に関する情報を一切表示せず，アクセスパターンとリンク構造だけを表示する．それでもこれらを同時に可視化することで，ページ遷移に関する情報がなくても，可視化結果から画面上で閲覧者のページ遷移を想像できる．しかし，その想像の正当性は保証されない．[立場 b] では，視覚的に認識できるページ遷移を可視化することはできるが，キャッシュされたページを表示した場合など，ログファイルに記録されていないため，閲覧者のすべてのページ遷移を正確に可視化することはできない．[立場 c] 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等を埋め込む，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，などの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，上述の措置がとられていないウェブサイトへの適用が困難になる，あるいは限られた閲覧者のページ遷移しか記録できない，といった制限が生じる。そのため現時点では，我々は [立場 c] だけでなく [立場 a] や [立場 b] も前提にして研究を進めている。5章に示す適用事例では，[立場 a][立場 b] を前提にした可視化結果を示している。

#### 4.1 アクセスパターン抽出

本研究における我々の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。本章は，アクセスパターン抽出に関する我々の実装について述べる。図 3 はアクセスパターンを抽出する処理手順である。

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  $x = (x_1, x_2, \dots, x_m)$  および  $y = (y_1, y_2, \dots, y_m)$  と表したとき，ベクトル間余弦は以下の式で表される。

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

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 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 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 %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。はそれぞれユーザが定義した値を使用する。我々は 5 章で使用しているデータに対して， $\theta = 0.6$ ， $n = 3$ ， $\alpha = 60$  とした。また，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た。

#### 4.2 リンク構造構築

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

いるページを再帰的に探索してリストを作り，得られた URL をもとにリンクのネットワークを構築する。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」[16] というクローラを採用している。

#### 4.3 データ統合

本手法では 4.1 節および 4.2 節で構築されたデータを統合する。本処理ではまず，4.1 節および 4.2 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。続いて，この木構造を構成する葉ノード間にリンク構造を付加することで，リンク付き木構造を生成する。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。

[立場 b] を前提とする場合には，本手法ではもう一度アクセスログファイルを解析し，アクセスされた各々のページに対して，URL とアクセス元 URL (このページからアクセスされたか) のペアを抽出する。続いて，ウェブサイトを構成する各リンクに対して，その両端にあるページの URL ペアが何回抽出されたかを集計し，閲覧者の通過頻度としてその集計値の情報を記録しておく。

#### 4.4 「FRUITS Net」の適用

本研究では，可視化手法「FRUITS Net」を，リンク構造とアクセスパターンを統合したデータに適用する。我々の実装ではこのデータを「FRUITS Net」のノードでウェブページを，グループでディレクトリを，色でアクセスパターンを可視化する。さらに，我々はリンクの描画に際して，3 章で示した色算出方法とは別に，[立場 b][立場 c] に基づいてデータを生成した場合，閲覧者の通過頻度によってリンクの色を計算する，という色算出方法も実装した。この色算出方法は，通過頻度の高いものから，赤，緑，青，そして通過頻度がゼロのリンクは薄紫色，というように色相を段階的に変化させながらリンクに色を割り当てる。

### 5 適用事例

本章では，我々の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 ヶ月間 (2009 年 7 月) のアクセスを記録した約 70000 行のものであり，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，Apple MacBook (CPU 2GHz, RAM 1GB) および MacOS 10.4.11 を用いて実行した。

図 4 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，右下部分にある色とボタンの一覧がアクセスパターンに対応している。色の横のボタンを押すと，そのアクセスパターンを有するウェブページに対応するノードがハイライトされるとともに，対応するノード同士を接続するリンクもハイライトされる。図 4 の例では，pattern1 ~ pattern12 の 12 種類

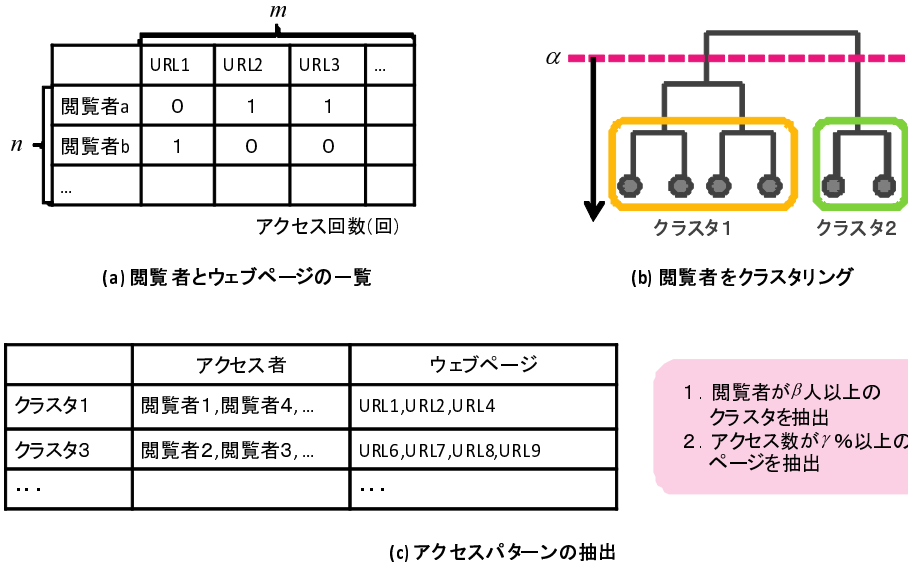


図 3: アクセスパターン抽出の処理手順

のアクセスパターンをボタンで選択できるようになっている。

我々の所属研究室のウェブサイトの内容は、図 5 のようになっている。研究テーマに関する分野の紹介であるプロジェクトページ、メンバーのウェブサイトへのリンクが張ってあるメンバーページ、発表論文が置いてあるレポートページ、大学までの地図が出ているアクセスページ、研究室に関連するページへのリンク集であるリンクページから成り立っている。メンバーページからは教員、学生 A、学生 B、学生 C... というように個人のウェブサイトへのリンクが張っており、教員のウェブサイトには、講義の資料、経歴、論文、海外滞在記などのページがある。

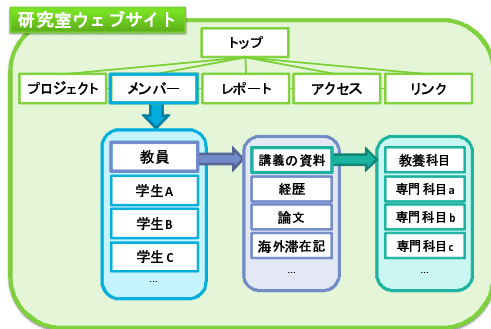


図 5: 適用事例に用いたウェブサイトの構成

### 5.1 可視化結果の全体像

研究室のウェブサイト全体を可視化すると、各ページは図 4 のように配置された。ノードの色に着目してみ

ると、研究室のトップページが pattern9, pattern10, pattern11 の色で色付けられていることに対して、教員のトップページは pattern4, pattern5, pattern6 の色で色付けられており、この 2 つのページは共通のアクセスパターンに属していないことがわかる。これにより、同じ研究室のページの中でも研究室のトップページと教員のトップページは一緒にアクセスされることはあまりないということが推察される。また、図 4 の真ん中の白い境界線より右側は教員のウェブサイトに含まれるページ、左側は学生のウェブサイトや研究紹介のページなどその他の研究室のページが配置されている。各アクセスパターンを見てみると、どれも同じ側のページ内で色付けられており、線をまたいで色付けられているアクセスパターンはほぼない。このことから教員のページは教員のページ内で、研究室のページは研究室のページ内で、それぞれ独立してアクセスされていることがわかる。

### 5.2 他の結果表示方法

図 6 は閲覧者の通過人数によってリンクに色をつけて表示した結果である。右下の拡大部分からは、閲覧者がディレクトリ内のページを順番にリンクを辿ってアクセスしているということがわかる。また、左上の拡大部分からは、このディレクトリ内のページにアクセスするために多くの閲覧者がリンクを辿ってアクセスしてきているということなどを考察できる。通過頻度の高いリンクを見つけることで、ウェブサイトを訪れる閲覧者にとっての重要なトピックを発見でき、また、リンクの色の有無によって、張られたリンクがきちんと機能しているかという確認に役立つと考えられる。

図 7 は各ページのアクセス数を高さで表示した結果である。この結果からは、研究室のトップページ・

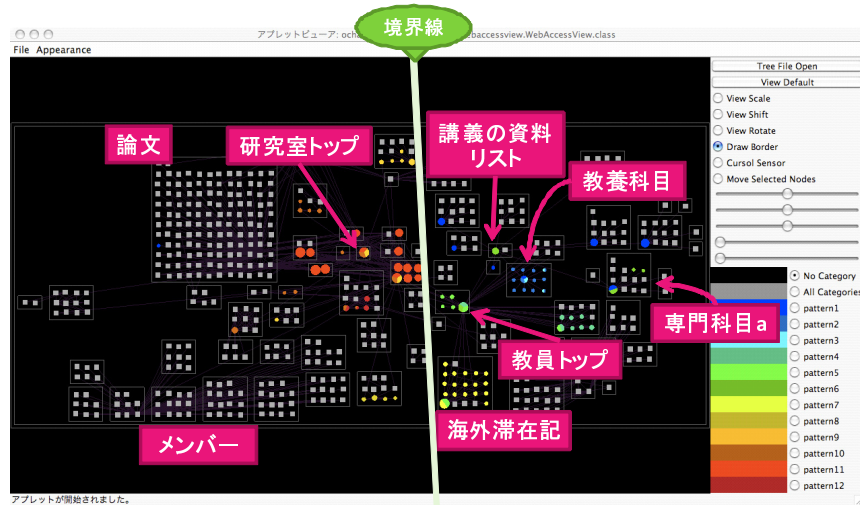


図 4: 可視化結果の全体像

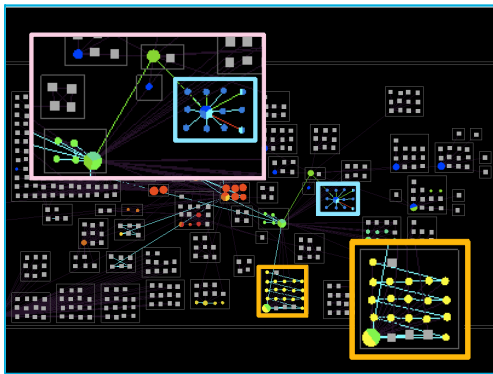


図 6: リンクに色をつけて表示

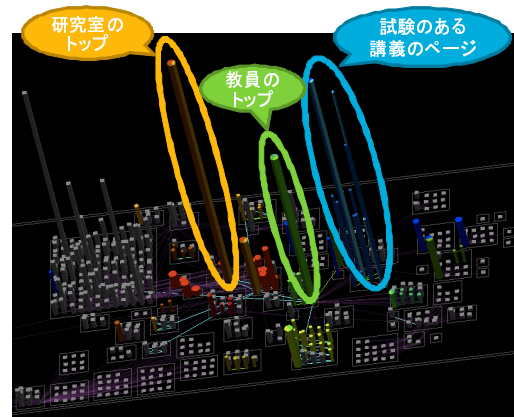


図 7: アクセス数を高さで表示

教員のトップページに加えて、試験のある講義の資料ページにも同じくらいアクセスがあることがわかる。トップページは他のページへアクセスするための軸となるページなのでアクセスが多いことは自明だが、講義のページにも同じくらいのアクセスがあるということから、このページに掲載されている情報が多くの閲覧者にとって必要であることが、この結果から示唆される。この例のような、多くの閲覧者を集めるページは、トップページから直接リンクを張るなど、少ないアクセスで目的のページまで辿りつけるような位置に置くことで、閲覧者のアクセスがより効率良いものになると期待できる。

### 5.3 本適用事例から発見される典型的なアクセスパターン

我々は可視化結果における、アクセスパターンに含まれるページのページ分布、及び、そのページ間に張られているリンク構造から、図 8 に示すような 3 つ

の典型的なアクセスパターンが効果的に表現されていることを見つけた。

1. 直線型のリンク構造を有するアクセスパターン
2. 1 つのノードを中心とした放射型のリンク構造を有するアクセスパターン
3. 同一ディレクトリ内のウェブページを多数含むアクセスパターン

図 8(左) 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、という閲覧者の軌跡を想像できる。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。また、図 8(中央)(右) は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 8(中央) は、メ

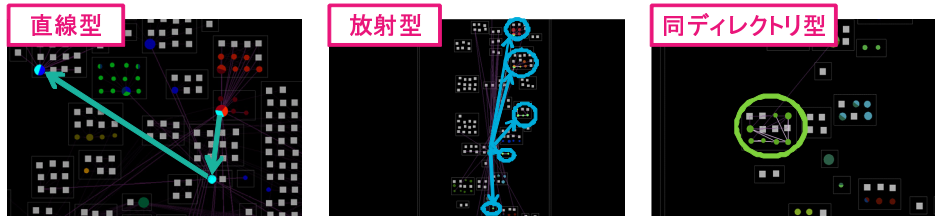


図 8: 典型的な 3 パターン

メンバー一覧のページから、同学年全員のウェブサイト  
にアクセスする閲覧者が複数いることを示している。  
図 8(右) は、同じディレクトリ内の大半のページに  
アクセスされ、特定の個人のウェブサイトだけをアクセ  
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 3 つの典型的なア  
クセスパターンのうち、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と考えられる。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。

また、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  
スパターンの場合には、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、同ディレクトリ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  
ンの場合には、同じディレクトリ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と考えられ  
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目視的  
にとらえ、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。

## 6 ユーザテスト

我々の構築した手法による可視化結果から得られる  
知見を、第 3 者が的確に理解することができるかを検  
証するために、本研究ではユーザテストを実施した。

本テストでは、まず被験者に「FRUITS Net」の操  
作方法を説明し、その後一人ずつ、実際に「FRUITS  
Net」を 5 分程度操作してもらい、その上で我々が用  
意した質問に回答してもらった。被験者数は 15 人、  
質問は以下の 5 問であった。

- 質問 (1) パターン 4 はどのようなアクセスか
- 質問 (2) パターン 3 とパターン 4 の違いは何か
- 質問 (3) パターン 7 とパターン 9 の違いは何か
- 質問 (4) パターン 8 はどのようなアクセスか
- 質問 (5) パターン 11 はどのようなアクセスか

質問は、5 章で挙げた典型的なパターンを判別でき  
るか、パターンの違いによって閲覧者のアクセスの仕  
方が違うことを理解できるか、ということ意識して  
出題した。各質問の詳しい出題意図を以下に述べる。

質問 (1) の「パターン 4」は、トップページ、メン  
バー一覧ページ、学部 4 年生全員のウェブサイトにあ  
クセスしているパターンである。放射型のリンク構造  
を有するパターンを判別できるか、という確認のため  
に出題した。

質問 (2) の「パターン 3」は、トップページ、メン  
バー一覧ページ、一人のメンバーのウェブサイトにあ  
クセスしているパターンである。個人を目的としたア  
クセスと、研究室の複数メンバーを目的としたアクセ  
スとの違いを読み取ることができるか、という確認の  
ために出題した。

質問 (3) の「パターン 7」は、教員のウェブサイト  
のトップページ、講義一覧のページ、専門科目 x にア  
クセスしているパターンであるのに対し、「パターン  
9 は」、専門科目 y のページのみへのアクセスである。  
目的のページにリンクを辿ってのアクセスと、目的の  
ページへの直接のアクセスの違いを読み取ることがで  
きるか、という確認のために出題した。

質問 (4) の「パターン 8」は、教員のウェブサイト  
のトップページ、講義一覧のページ、専門科目 z にア  
クセスしているパターンであり、また、これらのペ  
ージ間のリンクの色が赤くなっている。多くの人が目的  
のページにリンクを辿ってアクセスしたことが読み取  
れるか、という確認のために出題した。

質問 (5) の「パターン 11」は、教員の海外滞在出張  
記のディレクトリ内のページのみへのアクセスパター  
ンである。単一ディレクトリ内のウェブページのみで  
構成されるアクセスパターンを判別できるか、という  
確認のために出題した。

我々は被験者からの回答を確認し、的確な回答を  
、微細な説明不足や曖昧性を有する回答を、明ら  
かな誤りを含むものや無回答なものを x と評価した。  
評価の集計結果は表 1 のとおりである。全ての問いに  
対して 3 分の 2 以上の人が正解しており、概ね良好な  
結果だといえる。



表 1: ユーザテスト集計結果

| 質問 | (1) | (2) | (3) | (4) | (5) |
|----|-----|-----|-----|-----|-----|
|    | 13  | 11  | 13  | 10  | 10  |
|    | 1   | 3   | 0   | 2   | 4   |
| ×  | 1   | 1   | 2   | 3   | 1   |

各被験者の回答の内訳は表 2 のとおりである。行番号が被験者番号，列番号が質問番号を表している。正解以外の部分の回答内容を検証したところ，全体的には良好な結果であるものの，いくつかの問いで同じような間違いをしている人がいることがわかった。

表 2: ユーザテスト集計結果 (内訳)

| 質問     | (1) | (2) | (3) | (4) | (5) |
|--------|-----|-----|-----|-----|-----|
| 被験者 1  |     |     |     |     |     |
| 被験者 2  |     |     |     |     |     |
| 被験者 3  |     |     |     |     |     |
| 被験者 4  |     |     |     |     |     |
| 被験者 5  |     |     |     |     |     |
| 被験者 6  |     |     |     |     |     |
| 被験者 7  |     |     |     |     |     |
| 被験者 8  |     |     |     |     |     |
| 被験者 9  |     |     |     |     |     |
| 被験者 10 |     |     |     |     |     |
| 被験者 11 |     |     |     | ×   |     |
| 被験者 12 | ×   |     |     |     |     |
| 被験者 13 |     | ×   |     |     | ×   |
| 被験者 14 |     |     | ×   | ×   |     |
| 被験者 15 |     |     | ×   | ×   |     |

質問 (3) において被験者 14,15 は，カーソルの動きに沿ったリンクのハイライト表示が原因で間違えたのではないかと考えられる。我々の現時点での実装では，カーソルをノードに合わせると，そのページの URL が表示されると同時に，そのノードとつながっているリンクが全てハイライトされる。これにより，トップページなどリンクがたくさん張られているページにカーソルを合わせた場合，必要以上にリンクが目立ってしまい，誤った判断につながったと思われる。

また質問 (4) において被験者 11,14,15 は，アクセスパターンの色が見にくかったことが原因で不正解だったと考えられる。問題として出題したパターンに割り当てられていた色を，被験者によっては他の色と判別しにくい色であると感じたために，誤って他のパターンに含まれているノードまで同じパターンのノードだと認識してしまったのだと思われる。

質問 (5) において被験者 8,9,10,11 が正解出来なかった原因は，リンクの色に影響を受けたことだと思われ

る。リンクに色が付いていたことにより，アクセスパターンに含まれていないページまで，アクセスパターンに含まれるページと一緒にアクセスされたページであると判断し，誤りにつながった，という被験者が複数いた。

## 7 考察

本章では，前章のユーザテスト結果から得られた問題点，およびそれ以外の問題点を考察し，提案手法の課題について議論する。

現在，アクセスパターン 1 つに対して 1 色の色を割り当てている。本論文の適用事例では 12 色を用いてアクセスパターンを表現していたが，ユーザテストの質問 (4) からわかるように，既に数人の被験者に色の誤認が見られている。この方法では，非常に多数のアクセスパターンが抽出されたウェブサイトでも可視化したときに，色が多すぎて可読性が低下し，アクセスパターンの識別が大変困難になるという問題点がある。そのため，より多くのアクセスパターンを一画面上で表現できるように，色だけに頼らない他のデザインを考える必要がある。また別の問題点として，ユーザテストの質問 (5) にもあったように，リンクにも色をつけることで，可視化結果上の色が 2 種類の意味を有することになり，さらにユーザを混乱させる可能性がある。このことから，色だけに頼らないアクセスパターン表現のためのデザインが重要であると考えられる。

上述の問題点に伴って，現時点の我々の実装では，数百個単位のアクセスパターンが抽出されてしまった場合に，抽出されたアクセスパターンの中から手動で 10~20 個を選んで可視化している。一方で，可視化技術の改良によって数百個単位のアクセスパターンを表示できたとしても，ユーザにとって情報提示過多である場合があることも否定できない。よって可視化技術の改良の有無に関わらず，アクセスパターンの抽出個数を制御することは有用であると考えられる。そこで今後の課題として，抽出されたアクセスパターンの優先度を算出し，適切な個数のアクセスパターンを自動選択できるようにしたい。また，現在のアクセスパターン抽出処理は 4.1 節に示した非常に単純な方法をとっており，意味のあるアクセスパターンをもれなく抽出できているのかという点是不確定である。そのため，アクセスパターン抽出処理に関しても再検討し，改善を図りたい。

また，現在の実装では，閲覧者がどこのページからこのページへ移動したというアクセスの方向については可視化していない。今後の課題として，アクセスログファイルから解析できる範囲のみ閲覧者のアクセスの方向を可視化する，もしくは第 4 章で想定した 3 つの立場のうち [立場 c] を想定して研究を進める，などの形でアクセス方向の可視化を試みたい。

さらに別の課題として，サイト構成上の問題点を発見しやすくなるように，本可視化手法を改良したいと考えている。具体的には，期待に反してアクセス数の少ないページや，たくさんのリンクを辿らなければ目的のページまで到達できないページなど，問題のある場所が強調表示されるような可視化機能を設けること

で、ウェブサイト上の問題点をより発見しやすくしたいと考えている。

また、何万ページもあるような大きなウェブサイトでもスムーズな表示操作を実現できるように、計算速度の向上を図るとともに、インタラクション技術の向上にも努めたい。その上で、ユーザテストの質問(3)にあったような誤読を避けるために、さらに洗練されたインタラクション技術を開発したい。

## 8 まとめ

本論文では、「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、その適用事例を紹介した。本手法では、クローラを用いてリンク構造を、アクセスログファイルからアクセスパターンを得る。そして、それらを統合することによってリンク付き木構造を構築する。続いて、このリンク付き木構造に対して、力学モデルと空間充填モデルを組み合わせたネットワーク画面配置手法により可視化する。

本論文で紹介した適用事例では、3種類の典型的なアクセスパターンが表現されていることがわかった。また、これらのアクセスパターンがウェブサイトの改良にどのように役立つかということを議論した。

今後の課題として、前章で示した問題点の解決に努めるとともに、「FRUITS Net」以外の可視化手法でのウェブサイト可視化結果との比較、などを通して本研究結果の有用性を示したい。

## 参考文献

- [1] 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.
- [2] 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.
- [3] 宇根田純治, 横田治夫, Web ログの共通シーケンス解析, 電子情報通信学会信学技報, DE2002-2, 2002.
- [4] 三原宏一郎, 寺邊正大, 橋本和夫, ページ閲覧時間を考慮した Web ログマイニング手法の提案, 情報処理学会, pp. 39-44, 2007.
- [5] J. Pitkow, P. Pirolli, Mining Longest Repeating Subsequences to Predict World Wide Web Surfing, *2nd conference on USENIX Symposium on Internet Technologies and Systems*, pp. 139-150, 1999.
- [6] B. D. Davison, Predicting Web Actions from HTML Content, *13th ACM Conference on Hypertext and Hypermedia*, pp. 159-168, 2002.
- [7] 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.
- [8] 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.
- [9] G. D. Battista, P. Eades, R. Tamassia., I. G. Tollis, Graph Drawing - Algorithms for the Visualization of Graphs, *Prentice Hall*, ISBN-13-978-0133016154, 1999.
- [10] P. Eades A Heuristic for Graph Drawing, *Congressus Numerantium*, Vol. 42, pp. 149-160, 1984.
- [11] S. Hachul, M. Junger, An Experimental Comparison of Fast Algorithms for Drawing General Large Graphs, *Graph Drawing 2005*, pp. 235-240, 2005.
- [12] An Atlas of Cyberspaces, <http://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/atlas/atlas.html>
- [13] 末永高志, 岡田崇, 石打智美, Web アクセスログデータの系列情報を利用したサービスの関連性の分析, 電子情報通信学会信学技報, DE2005-17, 2005.
- [14] 山口裕美, 伊藤貴之, 池端裕子, 梶永泰正, 階層型データ視覚化手法「データ宝石箱」とウェブサイトの視覚化, 画像電子学会誌, Vol. 32, No. 4, pp. 407-417, 2003.
- [15] 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.
- [16] JSpider, <http://j-spider.sourceforge.net/>



川本 真規子  
2010 年お茶の水女子大学理学部情報科学科卒業。現在、お茶の水女子大学大学院人間文化創成科学研究科理学専攻博士前期課程在学中。



伊藤 貴之

1990年早稲田大学工学部電子通信学科卒業．1992年早稲田大学大学院理工学研究科電気工学専攻修士課程修了．同年日本アイ・ピー・エム(株)入社．1997年博士(工学)．2000年米国カーネギーメロン大学客員研究員．2003年から2005年まで京都大学大学院情報学研究科 COE 研究員(客員助教授相当)．2005年日本アイ・ピー・エム(株)退職，2005年よりお茶の水女子大学理学部情報科学科助教授(准教授)．2011年より同大学教授．ACM, IEEE Computer Society, 情報処理学会，芸術科学会，画像電子学会，可視化情報学会，他会員．