1T-2

Web アクセスログにおけるユーザ行動の分析

鈴木　康之†　　木村　昌臣†
芝浦工業大学工学部情報工学科†

1 はじめに

今日，インターネットの普及に伴い個人や企業の Web サイトが数多く開設されている．そのため，Web サイトに訪れるユーザの獲得が困難となっているが，Web サイトを改善することによって Web サイト訪問者(以下ユーザ)の獲得が期待できる．そのサイト改善の参考となるものの１つにユーザが Web サイト内でとった行動が Web サーバ内に記録されるログ（Web アクセスログ）がある．このアクセスログの分析結果を Web サイト管理者及び作成者が把握し，サイトを再構築することでユーザのニーズに合うサイトを提供することができると期待できる．図 1 にアクセスログの構造を示す．

```
172.21.37.32 - - [05/Sep/2006:13:11:32 +0900]
  IP アドレス          アクセス日時
"GET /access.html HTTP/1.1" 304 2378
   アクセスページ URL
http://shibaura-it.ac.jp/study.html
リンク元のページ URL（リファラー）
"Mozilla/4.0 (compatible; MSIE 6.0;
Windows NT 5.1; SV1; .NET CLR 1.1.4322)"
ユーザの使用環境
```

図 1.アクセスログの構造

本研究ではアクセスされた URL の推移，すなわちユーザの行動分析を行っていく．

2 提案手法

ユーザの行動を分析するにあたり，まずセッションと呼ばれるユーザ 1 人の一連の行動を確定し抽出する必要がある．本研究では抽出の際にページの遷移確率情報を付加したセッション同定について研究を行った．ページの遷移確率とはユーザの経路が一意に決まるのではなく確率的に推移するという考え方である．本来，IP アドレスはネットワーク上で 1 台の PC を識別するために割り当てられたものなのでその場合はユーザの経路が一意に決まるセッション同定ができる．しかし，ユーザがプロキシサーバを経由して Web サーバにアクセスしてきた場合，ユーザの IP アドレスがすべて同一となるので複数のユーザが同時にアクセスしてきた際にはセッションが一意に決まる同定はできない．図 2 は IP アドレスが同一のユーザ A , B のアクセス経路が

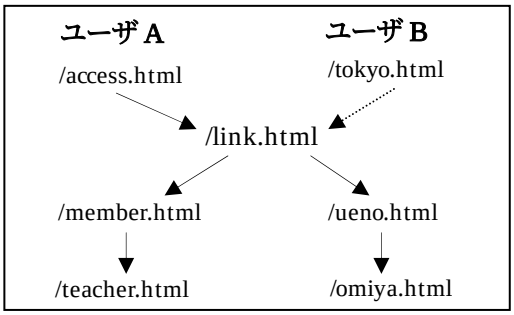

図 2.ユーザ A のアクセス経路

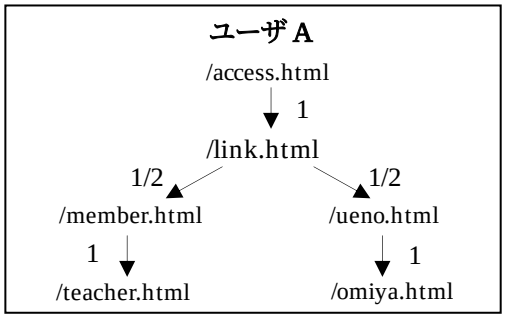

図 3.遷移確率を付加したユーザ A の経路図

link.html というページで交差している場合を表した例である．図中の矢印はあるページから次のページへ遷移したことを表す記号である．この遷移情報はアクセスページ URL とリファラーの対応関係から求める．実際にユーザ A が辿った経路はアクセスログから判断しただけでは link.html から member.html もしくは ueno.html のどちらへ遷移したのか判別できない．そこで本研究では図 3 のようにどちらのページへ遷移したのか分からないユーザの経路に対し，遷移確率を付加することにした．図 2, 3 の例では access.html から link.html へは分岐がないので遷移確率 1 を付加し, link.html から次のページへは 2 つのどちらかに分岐する可能性があるので遷移確率 1/2 を付加する．その後も分岐はないので 1 を付加する．

このように閲覧者があるページから次のページに遷移する確率を考えるとき 3 通りの場合が考えられる（図 4）．図中の n は「入」の数（閲覧者が現在のノードに入ってきた回数）, m は「出」の数（次に遷移するページの個数）を表している．n > m は n である「入」が m である「出」よりも多いことを意味している．n = 2 , m = 1 の場合を例にとって見てみると，n = 2 は閲覧者 2 人が「現在のノード」のページを見たことを意味する．一方 m = 1 は一人の閲覧者が次のページに遷移したことを表している．アクセスログ上からではこの 2 人のうちどちらの閲覧者が次のノ

Analysis of the user path in a Web access log
† Yasuyuki Suzuki, Masaomi Kimra
Shinaura Institute of Tecnology

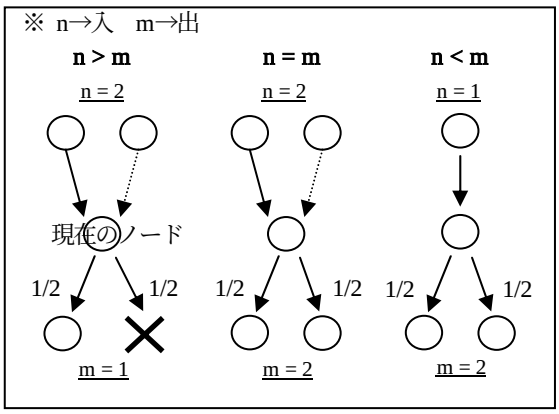

図 4.ノードの遷移図（3 パターン）

ードに遷移したのか分からないので 2 人の次の行動に対して次のページへ遷移するものと閲覧を停止するものの2つの分岐にそれぞれ1/2の確率を付加する．また n = m, n < m の場合も同様の考え方で解釈できる．この操作をデータベースに格納してあるアクセスログの各レコードに対して適用することによりセッション同定処理を行う．

3 閲覧行動の可視化

セッション同定抽出結果により得られる「あるノード」から「次のノード」への全ての組み合わせと,「先頭ページからの遷移確率の合計」を抽出した結果を用いる．遷移確率を合計することでトップページからあるノードへ遷移したユーザ数を伴った遷移確率を把握できるので，全ユーザのアクセス経路の傾向を掴むことができる．また，全経路を描画してしまうとノードの数が多くなり出力された結果が煩雑になってしまうので「先頭ページからの遷移確率の合計」に対する閾値を用いてノードの数を限定した表示をする．

4 実験

本研究で実験するアクセスログは芝浦工業大学の Web サーバから取得した Web アクセスログを用いる．解析対象データの詳細について表 1 に記載する．

表 1　解析対象データ

対象 Web サイト	芝浦工業大学 Web サイト
URL	http://www.shibaura-it.ac.jp/
期間	2006 年 6 月 13 日 ～2006 年 9 月 25 日
平均件数/日	93165 件
ファイルサイズ	2,206,279,614 バイト

本実験ではこのデータを用いて「条件無」に対しては全レコード，「時期別」に対しては夏休み前（6/13～7/31），夏休み期間中（8/1～8/31），夏休み後（9/1～9/25）の 3 シーズンで，「組織別」に対しては企業（co.jp），学校（ac.jp），政府機関（go.jp）の 3 組織で分析を行った．図 5 は政府機関からのアクセスを対象とした結果を示している．

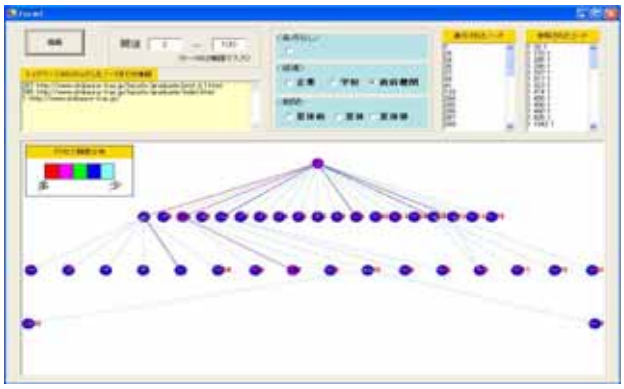

図 5.ユーザ経路の可視化システム

5 結果・考察

高い確率で遷移するノードのみに限定した経路の描画を行った結果，各条件で 共通して描画された経路があった．その経路はトップページからアクセスマップページ(/access/index.html)への遷移だった．また,アクセスマップページから芝浦工大豊洲校舎アクセスマップページ（/access/toyosu_map.html）へと遷移する確率も高い結果となった．また，組織別のみで比較を行ったところ学校関係者は「公募ページ」に高い確率で遷移していることがわかり，企業と学校関係者は工学部ページなどの研究や授業内容が記載されたページに高い確率で遷移していることもわかった．政府機関関係者は広報ページへ高い確率で遷移していた．このように特に「組織別」によって閲覧するパターンが異なる傾向を持つことがわかった．一方,「時期別」は 3 期間中で高い確率で遷移しているページに大きな差異はなかったが詳細に調べると夏休み中，夏休み後の 2 シーズンでは夏休み前の結果ではなかったオープンキャンパス情報を見ていることがわかった．

6 おわりに

本研究では Web アクセスログから遷移確率を用いたユーザのアクセス経路の抽出と，全体もしくは各条件に絞ったアクセスログのユーザ経路動向を視覚的に表現するシステムの開発を行った．これによって, Web サイトの管理者及び作成者がサイト内での閲覧者行動傾向を簡単に掴むことが可能になると考えられる．

参考文献

[1] 宇根田純治・横田治夫, Web ログの共通シーケンス解析，信学技報, 102(64), 7-12 2002
[2] 加藤久慶・平石広典・溝口文雄，データマイニング技術を利用した Web アクセスログへの適用，人工知能学会, 1D1-01, 2001
[3] 戸田誠二・横田治夫, LCS を用いたアクセスログ解析の並列処理による性能向上，DEWS2004, 7B05