知能と複雑系　132－10
(2003. 3. 14)

Web 視聴率データからの Web ユーザコミュニティ発見に向けて

村田 剛志† ‡

† 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2
‡ 科学技術振興事業団 〒151-0053 東京都渋谷区代々木 1-11-2 代々木コミュニティビル 3F
E-mail: † tmurata@nii.ac.jp

あらまし　著者は関連する内容の Web ページ集合(Web コミュニティ)の発見手法について検討を行なってきたが、その Web ページを閲覧するユーザ集合(ユーザコミュニティ)も存在すると考えられる。そのような興味を共有するユーザコミュニティを発見して分析することは、Web における視聴者の振る舞いを明らかにする上で重要であるとともに、その動的変化を検出することによって現実の人間社会における動向を見出すことが期待できる。本稿では、そのような人間のコミュニティの発見を最終目標とする。そのためのデータとして、Web 視聴者の振る舞いについてのログデータである Web 視聴率データに注目し、このデータから Web ユーザコミュニティを見出す手法について検討する。

キーワード　Web コミュニティ, ユーザコミュニティ, Web 視聴率データ

Toward the Discovery of Web User Communities from Web Audience Measurement Data

Tsuyoshi MURATA† ‡

† National Institute of Informatics (NII)　2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan
‡ Japan Science and Technology Corporation(JST)　3rd Floor, Yoyogi Community Building, 1-11-2 Yoyogi, Shibuya-ku, Tokyo, 151-0053 Japan
E-mail: † tmurata@nii.ac.jp

Abstract　The author has been working on the methods for discovering sets of related Web pages (Web communities). It is expected that groups of people who watch such pages (user communities) also exist. Discovering such user communities and analyzing them are important for clarifying the behaviors of Web audiences of similar tastes, and detecting dynamic changes of the communities is expected to discover trends of real human society. The ultimate goal of this paper is to discover such human communities. In order to achieve this, we focus on log data of Web audiences' behaviors (Web audience measurement data) and discuss methods for discovering user communities using the data.

Keyword　Web community, user community, Web audience measurement data

1. はじめに

Web の膨大なネットワークにおけるコミュニティとして、同一のトピックについての関連 Web ページの集合(Web コミュニティ)とともに、興味を共有しているユーザの集合(ユーザコミュニティ)が存在すると考えられる。前者は Web ページ間を結ぶハイパーリンクのグラフ構造によって特徴づけられ、後者はユーザの Web 閲覧行為などの振る舞いによって特徴づけられる。多数のユーザが関心を持つトピックの Web ページは増大し、逆に Web ページの構造の変化がユーザの閲覧行為に影響を及ぼすなど、両者は互いに影響を及ぼしあっていると考えられる。

これらの両コミュニティの構造を明らかにし、両者の間の相互作用を解明することは、Web 構造の発展やネットワーク上における現象の予測をする上で重要であるとともに、Web ページ収集アルゴリズムの効率化や Web ページのランキングアルゴリズムの改良などの実際的な応用を考える上でも有用であると言える。

Web mining の研究は、大まかに分けて Web content mining, Web structure mining, Web usage mining の三つに分類できる[5]。Web content mining は Web ページの内容に基づいたマイニングであり、Web structure mining は Web ページ間を結ぶハイパーリンクで構成されるグラフ構造に基づくマイニングであり、Web usage mining は Web 閲覧者のログデータ等に基づくマイニングである。これらの 3 つのアプローチは完全に分けられるものではなく、相互に関連性のあるものである。

著者は、ハイパーリンクのグラフ構造に注目して、関連ページ集合からなる Web コミュニティを発見するシステムの構築を行なっている[7][10]。ハイパーリンクが完全2部グラフを構成するようなページ集合は、興味を共有する Web コミュニティであるという仮定に基づき、関連ページ集合を発見する手法についての実験を行なっている。2部グラフに注目したこの Web コミュニティ発見手法は、Web structure mining だけでなく、Web mining の他のアプローチにも適用できると考えられる。

Web オーディエンス・メジャメントデータ(Web 視聴率データ)は、特定のユーザ集合の Web ページ閲覧行動を時間を追って記録することで得られるデータであり、その Raw data はアクセス時刻、ユーザ ID、閲覧 URL などの属性からなるデータ集合である。このようなユーザの振る舞いに関するデータは、同業種のサイト間の人気の比較や、ページ閲覧における順序の傾向などを見出すための基礎となるものである。Web オーディエンス・メジャメントデータにおける時系列データの様々な属性の中で、ユーザ ID と閲覧 URL に注目することによって、2部グラフの構造としてとらえることができる。

一方、ハイパーリンクによって結合されている Web ページ集合のグラフ構造も、リンクで結合している URL の組の集合として表現することができる。従って、上述の Web コミュニティ発見の手法を Web 視聴率データに適用することが可能であり、それによってユーザコミュニティを発見することができると考えられる。

本稿では Web 視聴率データを基に、興味を共有するユーザ集合である Web ユーザコミュニティを発見するための手法について検討する。そのようなユーザコミュニティを発見することができれば、個々のユーザにとっては嗜好に合った情報を Web から獲得するのが容易になるなどのメリットがある。それと同時に、Web ページを作成する側にとっても、ユーザの興味の動向を把握し、今後の変化の方向性を考える上で有効であると言える。

Kumar らの trawling[6]においては Web のスナップショットデータからハイパーリンクによる二部グラフ構造を探索することで Web コミュニティの発見を行なっている。また、Kleinberg らの HITS アルゴリズム[4]において hub や authority などを導入しているのも、基本的にはリンクによる二部グラフ構造を関連性の指標として重視しているためである。本稿では、視聴行動データにおけるグラフ構造においても、ユーザ集合と Web ページ集合が二部グラフを構成するようなものを、ユーザ間の関連性を示すものとして注目する。

2. Web オーディエンスメジャメントデータ

2.1. Web オーディエンスメジャメントデータとは

Web オーディエンスメジャメントデータ(Web 視聴率データ)とは、テレビの視聴率と同じように、予め定めたユーザ集合の閲覧パターンについて、クライアント側でデータを取得するものである。日本では、ネットレイティングス社(http://www.netratings.co.jp/)、ビデオリサーチネットコム(http://www.vrnetcom.co.jp/)、日経 BP 社(http://www.nikkeibp.co.jp/)、日本リサーチセンター(http://www.nrc.co.jp/)などがデータ取得の調査を行なっている。

これらの視聴データの利用方法としては、Web 視聴者の傾向に関する統計的な調査が主である。具体的には、以下のようなものがある。

・インターネット利用状況の調査(利用年齢層、時間帯、接続環境など)

・販売促進キャンペーンと Web ページ訪問者との関係の調査

・オンラインショップでの購入者の視聴行動とアンケート結果とを組み合わせた追跡調査

これらの調査のベースとなるのは各ユーザの視聴行動に関する Raw data である。具体的には、以下のような各属性に関する大量のデータであり、データ収集のために改変したブラウザ等をユーザに使ってもらうことによってデータ収集を行なう。Raw data の例を以下に示す。各属性はそれぞれ時刻、ユーザ ID、閲覧 URL、経過時間を表す。属性として、ブラウザのバージョンやリンクアクション（リンクをたどる、戻る、URL を入力するなど）が含まれる場合もある。

```
time userID elapsed time    URL
00:00 9601 10 www.jpncm.com/cgi-lib/cmbbs/wforrum.cgi
00:00 9701 27 www.dion.ne.jp
00:00 3502 19 search.auctions.yahoo.co.jp/search
00:00 5201 14 eee.eplus.co.jp/shock/shock03.html
00:01 5502 10 user.auctions.yahoo.co.jp/jp/show/mystatus
00:01 0501 6 user.auctions.yahoo.co.jp/show/mystatus
00:01 3301 36 www.pimp-sex.com/amateur/raimi/01/clean.htm
00:01 9701 4 auctions.yahoo.co.jp/jp/2....-category-leaf.html
00:02 8501 3 www.uicupid.org/chat/csp_room.php
00:02 8001 3 page.auctions.yahoo.co.jp/jp/show/qanda
00:02 1501 11 www.nn.iij4u.or.jp/~movie/pm/main.html
00:02 9002 12 www.umai-mon.com/user/p_category.php
```

図 1:Raw data の例

また、ユーザ ID で示された各ユーザについては、性別、年齢、生年月、居住エリア等の属性値が対応している。

UserID	gender	year	month	occupation	area
16	M	1971	9	22	3
17	M	1981	9	74	3
19	M	1939	12	94	3
20	M	1950	11	21	3
21	F	1980	3	75	3
22	F	1976	12	95	3
23	F	1975	7	96	3
24	M	1945	5	41	3
25	M	1963	12	13	5
26	M	1960	11	41	3
27	M	1971	4	11	3
28	F	1946	8	81	3
29	M	1944	9	42	3
30	M	1975	9	75	3
31	F	1976	4	82	3

図 2:ユーザの属性の例

2.2. データ獲得の場所とその性質

一般に、Web usage mining におけるデータ源としては、以下の 3 つが考えられる[9]。

1. サーバレベルのデータ収集
2. クライアントレベルのデータ収集
3. プロキシレベルのデータ収集

これらは、データ源が異なるだけでなく、利用可能なデータの種類や、データの粒度等においても異なっている。

1 のサーバレベルでのデータ収集は、Web サーバのログをデータ源とした不特定多数の閲覧者に関するものである。具体的にはアクセス元の IP アドレス、アクセス時間、閲覧ページ URL、プロトコル、そのページの参照元などの属性からなる。このようなデータは Web サーバの管理者が他からのアクセス状況を知る上で最も入手しやすいデータであるが、Web においてはさまざまなレベルのキャッシュが存在しており、キャッシュからの読み出しがログに反映されないなどの欠点もある。

2 のクライアントレベルでのデータ収集は、Java スクリプトや Java アプレットなどによるリモートエージェントを使うか、データ収集ができるよう改変したブラウザによって、クライアント側でデータ収集を行なうものである。これにはユーザ側の協力を必要とするが、サーバ側でデータを収集する場合と比較して、キャッシュやセッション同定等の問題を改善する上で有効である。

3 のプロキシレベルでのデータ収集は、共通のプロキシサーバを使うユーザ集合の閲覧行動についてのデータを得るものである。これについては上の二つとは性質が異なるので除外して考える。

Web オーディエンスメジャメントデータは、2 のクライアントレベルで収集されたデータである。従来の Web usage mining 研究では、1 のサーバレベルで収集されたデータを使うことが多かったが、上記の理由から Web 閲覧者の詳細な行動を反映したデータになっていない可能性がある。

Web 視聴率データにおいては、改変したブラウザをユーザに使用してもらうことによって、サーバレベルでは収集できないデータを得ることができる。例えば、ブラウザにおける back ボタンを押して以前見たページを再度見る場合、キャッシュのデータを利用しているため、サーバレベルでは記録されないがクライアントレベルでは記録される。また、ユーザが各ページでとるリンクアクションや、他のアプリケーションソフトウエアの起動など、Web ページの閲覧以外の行動と組み合わせることによって、より現実に沿った法則が見出せると期待できる。

3. ユーザコミュニティ発見に向けて

3.1. 視聴データの二部グラフ構造

ユーザコミュニティ発見における最終的な目標は、類似したページ集合を閲覧するような、興味を共有するユーザ集合を得ることである。Web 視聴率データからユーザのコミュニティを見出す方針として、視聴行動における二部グラフ構造に注目する。

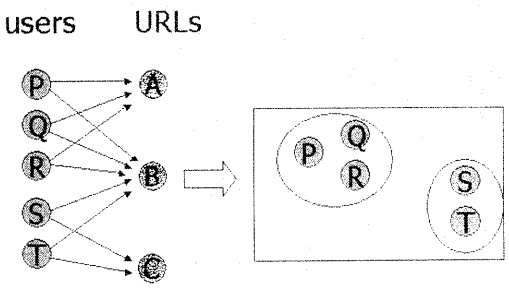

図 3: Web 視聴行動における構造

Raw data においてはユーザが閲覧した URL 等が時間順に列挙されている。このデータから興味を共有するユーザ集合を見出すにあたり、各々の閲覧 URL は粒度が非常に細かく、同一の URL を閲覧するユーザが複数存在する可能性は低い。従って、図 3 に示すように、Web ページとユーザの両方でまとまりを見出すことを考える。

- URL よりも粒度の荒い Web ページ集合の検出
- それらを閲覧するユーザ集合の検出

Web のグラフ構造においても、上述の trawling や HITS に代表されるように、二部グラフ構造が関連性を示す重要な指標となっている。関連するユーザの集合においても同様に、二部グラフ構造がコミュニティに対応していると考えられる。

3.2. ユーザコミュニティの評価

Girvan らの研究[2]はグラフの頂点や辺における中心性(betweenness)に注目して、それが一番小さい辺をグラフから除去していくことによって密なグラフ構造を見出すものである。その評価方法として、空手クラブにおけるメンバー間のネットワークに適用し、派閥などを見出すことによって妥当性を示している。しかし、得られたユーザコミュニティが妥当なものであるかどうかを評価することは一般には困難なことである。

上記の Raw data に基づいて Web コミュニティを発見した場合、ユーザ集合に関する年齢や性別などの属性値によって、ユーザの類似性をある程度示すことが可能である。また、視聴行動のデータは継続して収集されており、時間間隔をおいて収集されたデータで発見されるコミュニティを比較することによって、ある時点で発見されたコミュニティの妥当性や、その時間変化等を見出すことが可能になると考えられる。

3.3. Web コミュニティとユーザコミュニティ

Web ページ集合からなる Web コミュニティと、ユーザ集合からなるユーザコミュニティは、相互に密接に関わりあいながら存在していると考えられる。その相互作用のメカニズムを解明することは、Web からの情報獲得のみならず、Web を介した人間間のつながりを支援していく上でも重要である。

その二つのコミュニティをつなぐ存在として、サーチエンジンが重要な役割を果たすと言える。例えば、飛行機墜落などの大きな事故が起こると、サーチエンジンの検索キーワードとして"CNN"や"world trade center"などの関連語が頻出することが多い[3]。そして、主要なサーチエンジンにおける掲示板等を出発点として、必要な情報を発信するための Web ページや、情報交換のための掲示板等が作られ、それらが互いにリンクを張り合うことによって Web ページのネットワークが形成されていくと考えられる。

4. おわりに

本稿では、Web におけるコミュニティとして、関連するページからなる Web コミュニティと、興味を共有するユーザ集合からなるユーザコミュニティを発見することの重要性と、そのための方向性について示した。

Raw data を用いた実際のプロセスを進めていくにつれて、どのようなデータ属性がコミュニティ発見に必要であるかが明らかになっていくと期待される。コミュニティ発見において指標となる属性について明らかにし、必要に応じてそのようなデータを積極的に収集して検証していくことも重要であると言える。

文　献

[1] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener: Graph structure in the Web, Proc. of the 9th WWW conference, (2000).

[2] M. Girvan, M. E. J. Newman: Community structure in social and biological networks, online manuscript, http://arxiv.org/abs/cond-mat/0112110/, (2001).

[3] Internet Watch:「Nimda」と米テロ事件は Web アクセスにも影響～ネットレイティングス調査, http://www.watch.impress.co.jp/internet/www/article/2001/1019/netra.htm (2001).

[4] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins: The Web as a Graph: Measurements, Models, and Methods,Proc. of the 5th Annual International Conf. on Computing and Combinatorics (COCOON '99), LNCS 1627, pp.1-17, Springer, (1999).

[5] R. Kosala, H. Blockeel: Web Mining Research: A Survey,ACM SIGKDD Explorations, Vol.2, No.1, pp.1-15,(2000).

[6] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins: Trawling the Web for Emerging Cyber-Communities, Proc. of the 8th WWW conference,(1999).

[7] T. Murata: Finding Related Web Pages Based on Connectivity Information from a Search Engine, Poster Proc. of the 10th WWW conference,pp.18-19, (2001)

[8] L. Page, S. Brin, R. Motwani, T. Winograd: The PageRank Citation Ranking: Bringing Order to the Web, Online manuscript, http://www-db.stanford.edu/~backrub/pageranksub.ps,(1998).

[9] Srivastava, R. Cooley, M. Deshpande, P.-N. Tan:Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data,ACM SIGKDD Explorations, Vol.1, No.2, pp.12-23 (2000).

[10] 村田剛志:参照の共起性に基づく Web コミュニティの発見,人工知能学会誌 Vol.16, No.3, pp.316-323, (2001).