社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2003－ＩＣＳ－133 (22)
2003／9／15

Discovery of Web user communities from Web audience measurement data

Tsuyoshi Murata† ‡

† National Institute of Informatics　2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan
‡ Japan Science and Technology Corporation　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 user community, Web audience measurement data, discovery

Web 視聴率データからの Web ユーザコミュニティ発見

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 ユーザコミュニティ, Web 視聴率データ, 発見

1. Introduction

World Wide Web is an important media that allow us to distribute messages with Web pages, and to refer others by hyperlinks. The Web has abilities of uniting related Web pages as well as humans of similar tastes. The former (groups of related Web pages) is called Web communities, and the latter (groups of humans of similar tastes) is called user communities in this paper. Web communities are based on the connection of pages with hyperlinks, and user communities are based on the users' behaviors of viewing Web pages. Both communities are mutually related: 1) if many users have interests to some specific topic, the number of pages about the topic is increased, and 2) if the structure of Web pages about specific topic has been changed, users' page viewing behaviors will be affected.

Discovering the structure of both communities and the interactions between them is important for predicting the development and the phenomena in the Web. Practical applications for discovering such communities are as follow: development of efficient algorithms for Web page crawling, improvement of ranking algorithms, recommendation of suitable Web pages, and personalization of Web sites for group of users of similar tastes.

The ultimate goal of Web mining is to discover useful knowledge from the Web. There are three main approaches of Web mining: Web content mining, Web structure mining, and Web usage mining [5]. Web content mining is based on contents of Web pages, Web structure

mining is based on graph structure of hyperlinks, and Web usage mining is base on data generated by users such as log data and bookmark files. These three approaches are mutually related; combination of these approaches is expected to achieve discovery about the Web and its users.

The author has been working on Web structure mining based on hyperlink graph structure, and has developed systems that discover Web communities from a few given URLs [7][10]. These systems are based on an assumption that Web pages whose hyperlinks form complete bipartite graph can be regarded as a Web community sharing common interests. Data acquired from a search engine are used for searching bipartite graphs in the systems.

Bipartite graphs are simple and powerful structure for detecting relations among nodes. The strategy of searching bipartite graph employed in our Web structure mining system can be applied to Web usage mining as well for discovering relations among users.

There are several approaches for Web usage mining, and Web log mining is one of the main research topics. As the source for Web log mining, Web audience measurement data are discussed in this paper. The data are acquired by recording page views of pre-specified user group. Data of this sort are used as the basis for detecting trends of popularity among Web sites of same industry. The data are composed of access time, user ID, viewed URL, elapsed time at the URL, and so on. Among these attributes, user ID and their viewed URLs are focused. Connections of user ID and viewed URLs can be regarded as a (huge) bipartite graph.

Connection among Web pages can be represented as graph structure as well. The methods for discovering Web communities that is applied to hyperlink graph can be applicable to the above graph of Web audience measurement data in order to find groups of user ID and viewed URLs, which can be regarded as user community of similar tastes.

This paper proposes a method for discovering Web user communities based on Web audience measurement data. Discovery of such communities will be useful for Web page recommendations that match users' tastes. Discovery of Web user communities is important for Web site builders in that it help better understanding of users' viewing behaviors and their trends. User communities are important for market segmentation also.

Kumar's trawling [6] discovers Web communities by searching bipartite graph structures from Web snapshot data. Moreover, hubs and authorities are regarded as important pages in Kleinberg's HITS algorithms [4] since they are regarded as crucial components of bipartite graphs. This paper also focuses on bipartite graph structures; users and Web pages that compose a bipartite graph are regarded as a closely related group.

2. Web audience measurement data

2.1. Investigation of users' behaviors

Web audience measurement data are just like audience data of TV programs. Behaviors of certain group of users are recorded at client sides. In Japan, there are four major companies for this sort of investigation: Nielsen//NetRatings (http://www.netratings.co.jp/), VideoResearch Netcom (http://www.vrnetcom.co.jp/), Nikkei BP (http://www.nikkeibp.co.jp/), and Nihon Research Center (http://www.nrc.co.jp/).

Major ways for utilizing Web audience measurement data is statistical investigations of users. For example, the following investigations are shown in their Web sites:

- Investigation of the situations of internet usage (users' age or gender, access time, environment of users' computers, and so on)
- Investigation of the relation between campaigns of sales promotion of a company and the number of visitors to the company's Web site
- Investigation of the relation between behaviors of buyers at online shops and the results of their questionnaire

Bases of these investigations are raw data of users' Web viewing behaviors. The following rows are the examples the data. These data are recorded at users' computers by using modified Web browsers. Attributes of raw data include time, user ID, elapsed time, and viewed URLs. Attributes of raw data are not the same among the above investigation companies: some raw data include versions of Web browsers and link actions (click hyperlinks, back to previous pages, or entering URLs manually).

```
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

Fig 1: Example of Web audience measurement data

Data about users are composed of the following attributes: user ID, gender, age, year and month of birth, address, and so on.

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

Fig 2: Example of users' attributes

2.2. Characteristics of Web log data for discovery

In general, data source for Web log mining can be classified to the following three cases [9].

1. server-level data collection
2. client-level data collection
3. proxy-level data collection

Source of Web log data affects the characteristics and granularity of available data. In case 1 (server-level data collection), available data are about multiple unspecified users who visit specific Web server. Attributes of log data include IP address, time of access, accessed URL, protocols, referrer of the URL, and so on. Although this sort of log data is easily available for system administrators of Web server, the data do not always reflect users' true behaviors since there are several levels of caches on the Web.

In case 2 (client-level data collection), data are collected at client sides by using remote agents such as Javascript, Java applets, or modified Web browsers. Although users' cooperation is required for this method, client-side collection has an advantage over server-side collection because it does not suffer both the caching and session identification problems.

In case 3 (proxy-level data collection), proxy caching are collected and analyzed for the purpose of reducing loading time of Web pages. Abilities to predict future page request will improve the performance of proxy servers. The data obtained at proxy-level (multiple clients – multiple Web servers) are quite different from other Web log data.

Web audience measurement data are collected at client-level (case 2). Although most of the previous researches of Web usage mining use data collected at server-level, the data is not suitable for detailed analysis for the reasons mentioned above. Web audience measurement data are expected to contain information of users behaviors that cannot be obtained from server-side data collection. For example, when viewing previous Web page, users often click back button. Although server-level data cannot capture such behaviors since cached data are usually used, they are recorded on data collected at client-side. In addition to the behaviors of viewing Web pages, other behaviors of operating their PC, such as activating other application software, can be recorded by client-level data collection. By using this sort of rich usage data, more realistic rules are expected to be discovered.

3. Idea for discovering Web user communities

3.1. Structure of Web pages and users

The ultimate goal of discovering Web user communities is to find groups of users who view similar Web pages sharing common interests. In general, discovery from Web usage data can be classified into statistical analysis, clustering, classification, sequential patterns, and dependency modeling [9]. Our goal of discovering Web user communities can be regarded as a kind of clustering.

As the strategies for discovering user communities from Web audience measurement data, search of bipartite graphs on Web viewing behaviors are discussed in this paper. Figure 3 shows the outline of the search. Let us suppose that users (P, Q, ...) view Web pages (A, B,...). A group of users and URLs that compose a bipartite graph can be searched from the graph of Web audience measurement data.

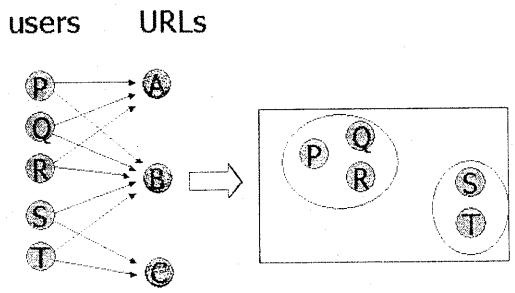

Fig 3: Bipartite graph of Web viewing behaviors

As mentioned above, Web audience measurement data is a list of URLs that are viewed by pre-specified users in the order of time sequence. In order to search bipartite graphs from the data, Web pages have to be grouped in advance because of fine granularity of viewed URLs. As a way for grouping Web pages, truncation of URLs into sites is the most naive. It is true that evaluating the contents of Web pages by some similarity measure is better than simple truncation. However, the amount of log data is huge (about 180MB for one month). As the first step for grouping related Web pages, truncation of URLs is employed in our approach.

3.2. Evaluation of user communities

As an attempt for extracting humans communities from given graph structure, Girvan et al. [2] focus on betweenness of each edge in a graph. Betweenness is defined as the number of shortest path of arbitrary pairs of vertices. Since an edge of highest betweenness is regarded as the edge that "bridge" densely connected subgraphs, iterative removal of edges of highest betweeness is performed in order to extract such subgraphs. For evaluating discovered communities, their method is applied to human network of Karate club and analyzes the correspondence between discovered communities and real factions in the club. However, evaluating the validity of user communities is not an easy task in general.

In the case of user community discovery from Web audience measurement data, attributes of each user such as age and gender can be used for judging the similarities of users to a certain extent. Another way for evaluating user communities is to use data sets of same users obtained in different time. Since Web audience measurement data is collected consecutively, data of viewing behaviors can be partitioned into subsequences that can be used for both discovery and the evaluation of the results, just the same as cross validation of machine learning.

3.3. Web communities and user communities

Web communities and user communities are closely related, and some changes of one side may affect the other. Clarifying the interactions between both communities is important for retrieving information from the Web as well as for supporting human relationship through the Web.

As the bridge for connecting both communities, search engines often play an important role. For example, when a worldwide accident such as 9.11 terror occurred, many search of related words such as "CNN" or "world trade center" were performed immediately after [3]. As the next step, information exchange on bulletin boards becomes active, and Web pages about the accident are newly built and linked together. Detailed analysis of the process of this kind of community generation is expected to assist humans' smooth communications through the Web.

4. Concluding remarks

In the field of Web structure mining, several attempts have been made for discovering "Web communities". The phrase is mainly used for groups of related Web pages. This paper claims the importance of discovering user communities that are composed of users sharing common interests. An idea for discovering user communities from Web audience measurement data is described. As the next step, systems for discovering user communities have to be developed in order to verify this idea and evaluate the validity of discovered user communities.

There are several attributes in Web audience measurement data. It is expected that importance of each attribute for discovering user communities becomes clear in the process of actual Web usage mining. In order to judge user communities as valid, we have to discuss more about the attributes that are useful for evaluating user communities, and also about the strategies for actively collecting needed attributes from users.

References

[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 and 9.11 terror affect Web accesses - Investingation of Netratings (in Japanase) 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] T. Murata: Discovery of Web communities based on co-occurrence of references (in Japanese), Journal of JSAI, Vol.16, No.3, pp.316-323, (2001).