2004－ICS－138　(18)
2004／12／5

Structural Analysis of Web User Communities

Tsuyoshi Murata[†,††]

There are two kinds of communities in the Web; communities of related Web pages (Web communities) and communities of users who watch such related pages (user communities). Discovery of the former communities has been attempted by many researchers such as Kumar's trawling and Flake's method. Discovery of the latter communities is also important for clarifying the behaviors of Web users. Moreover, it is expected that the characteristics of user communities in the Web correspond to those in real human communities. The author proposed a method for discovering user communities based on client-level log data. Web audience measurement data are used as the description of users' Web watching behaviors. Maximal complete bipartite graphs are searched from the graph obtained from the log data without analyzing the contents of Web pages. Since there are many user communities discovered in the above method, choosing a small number of "interesting" ones is required. As the criteria for judging interestingness of user communities, discrepancies of distance among community members are proposed in this paper.

1. Introduction

World Wide Web is an important media that allows 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 often called Web communities, and the latter (groups of users 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 watching Web pages. Both communities are mutually related: 1) if many users have interests to some specific topic, the number of Web pages about the topic is increased, and 2) if the structure of Web pages about specific topic has been changed, users' Web watching behaviors will be affected. Discovering the structure of both communities and the interactions between them is important for predicting the development of the Web. Several methods have been proposed for the discovery of Web communities, such as Kumar's trawling based on graph search[5] and Flake's method based on network-flow theory[2]. Compared with the research for Web communities, however, little attention has been given to the research for user communities in the Web. There are many practical applications for discovering user communities, such as recommendation of suitable Web pages and adaptation of Web sites for users of similar tastes.

The author previously proposed a method for discovering user communities from the data of users' Web watching behaviors. Discovery of user communities are important for understanding users' information needs, and for recommending suitable Web pages by social Web filtering. Web audience measurement data are client-level log data that record users and their visited URLs. The data are transformed into a graph, and complete bipartite graphs that correspond to user communities are searched from the graph. Since there are many user communities discovered in the above method, choosing a small number of "interesting" ones is required. As the criteria for judging interestingness of user communities, discrepancy of distance among community member is proposed in this paper. There are many aspects in entities (terms and users), and many distance functions between entities can be defined accordingly. If entitites grouped together based on one distance function are mutually dissimilar based on another function, the group is worth manual inspection since it identifies discrepancies between distance functions. On the assumption that distance functions between entities are defined, interestingness of user communities defined in this way is simple and powerful for post-processing of Web mining.

† National Institute of Informatics
†† Japan Science and Technology Agency

time	userID	sec	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/.../clean.htm
00:01	9701	4	auctions.yahoo.co.jp/jp/2...-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

userID	sex	b/year	b/month	job	area
0016	M	1971	9	22	3
0017	M	1981	9	74	3
0019	M	1939	12	94	3
0020	M	1950	11	21	3
0021	F	1980	3	75	3
0022	F	1976	12	95	3
0023	F	1975	7	96	3
0024	M	1945	5	41	3
0025	M	1963	12	13	5
0026	M	1960	11	41	3
0027	M	1971	4	11	3
0028	F	1946	8	81	3
0029	M	1944	9	42	3
0030	M	1975	9	75	3
0031	F	1976	4	82	3

Fig. 2 Example of Users' Personal Data

2. Web Audience Measurement Data

Web audience measurement data are just the same as audience data of TV programs. Randomly selected users are asked to use special modified Web browsers that record users' behaviors (such as visited URL, time of the visit, and elapsed time at the URL) at users' PC. Example of Web audience measurement data are shown in Fig. 1. Each row of Web audience measurement data represents a user's visit of a URL. Attributes of the data include time, user ID, visited URL, elapsed time, and other miscellaneous information about users' actions such as click of hyperlinks, back to previous pages, or manual enter of URLs.

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/). Web audience measurement data are used mainly for statistical analysis, and for detailed investigation of the visitor of specific site. For example, 1) Investigation of the situations of internet usage (users' age or sex, access time, environment of users' computers, and so on), 2) Investigation of the relation between campaigns of sales promotion of a company and the number of visitors to the company's Web site, and 3) Investigation of the relation between behaviors of buyers at online shops and the results of their questionnaire. Fig. 2 shows the example of users' personal data. The data contain attributes such as user ID, sex, age, birth year, birth month, occupation, address, and so on.

In general, data source for Web usage mining can be divided into two classes: server-level data and client-level data[6]. Web audience measurement data is the latter. As is often pointed out, usage data at client-level reflect users' true behaviors since the usage of cached data cannot be recorded on server-level data.

3. A Method for Discovering Web User Communities

As the method for discovering user communities, graph mining approach is employed. This is because of the following reasons: 1) in order to analyze huge Web usage data, graph mining approach is generally faster than content based approaches, and 2) it is expected that methods for Web structure mining are applicable for graph representation of Web usage data.

Several approaches have been attempted for the discovery of Web communities, which is one of the important research topics of Web structure mining. Kumar[5] claimed that Web pages whose hyperlinks constitute a bipartite graph structure are regarded as a community sharing common interests. The idea is simple and powerful for finding related nodes from a graph, and it is expected to be applicable to the graph structure of Web usage data. The initial idea of our approach is to search complete bipartite graphs from the pairs of user and visited URL. Let us suppose group of users (P, Q, R, ...) visit URLs (A, B, C, ...). Users' Web watching behaviors can be represented as a graph when we regard each user and URL as a node and each visit of URL as an edge. Bipartite graphs such as (P, Q, R, A, B) and (S, T, B, C) can be regarded as user communities sharing common interests since the users in the graph visit the same URLs.

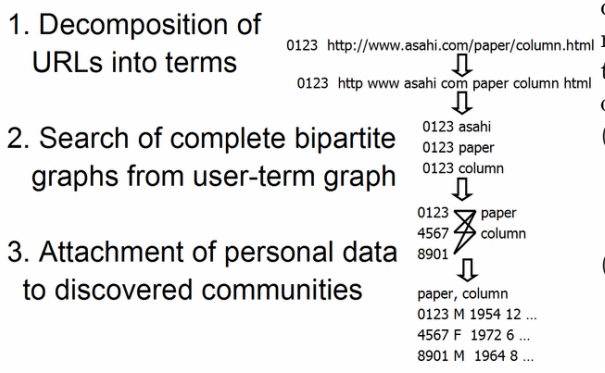

Fig. 3 A Method for Discovering User Communities

However, this naive idea is not applicable to real Web audience measurement data since most of the URLs are visited by only one user. The graph structure of users and URLs (user-URL graph) is sparse and most of the complete bipartite graphs contained in the graph are composed of only few nodes. As a method for discovering user communities from the structure of Web audience measurement data, the following procedures are applied in our discovery method:

(1) Decomposition of URLs into terms
A URL contains information about the page it points to. Each URL string is decomposed into terms that are used as labels of the URL.

(2) Search of maximal complete bipartite graphs from user-term graph
As claimed by Kumar, Web pages whose hyperlinks compose a complete bipartite graph are mutually related. The same idea is applied to the user-term graph.

(3) Attachment of personal data to discovered user communities
In order to assist the understanding of discovered user communities, personal data (such as sex and birth year) of all the members are attached to user communities.

Overall procedures are shown in Fig. 3.

3.1 Decomposition of URLs into terms

As mentioned above, each URL contains useful information about the page it points to. Another reason for using URL strings for discovery is that many Web pages are dynamically generated at the time of visit. Contents of such pages cannot be obtained at the time of discovery because of privacy reasons and storage capacity reasons. It is expedient to use URL strings as the labels of huge Web log data. Decomposition of URLs is performed as follows:

(1) Terms are extracted from URL strings by chopping URL at the point of punctuations, such as periods (.), slashes (/), dashes (-), and question marks (?).

(2) Overly frequent terms (such as "www" and "html") and numbers are removed from the extracted terms because they do not characterize the contents of original URLs. In our experiment, the following terms are removed: "www", "html", "htm", "cgi-bin", "cgi", "co", "ne", "or", "jp", "net", and "com".

(3) By replacing URLs with the terms generated in the above procedure, a graph of userIDs and URLs (user-URL graph) are expanded to the graph of userIDs and terms (user-term graph). In general, obtained user-term graph is denser than original user-URL graph since more than one term are usually extracted from each URL.

3.2 Search of Maximal Complete Bipartite Graphs

Generated user-term graph is a fairly large bipartite graph. From the graph, maximal complete bipartite graphs are searched. Such graphs indicate that a group of users visit URLs that include a group of terms. Search of bipartite graph is performed by the algorithm proposed by Uno[8]. The algorithm has abilities of searching each complete bipartite graph with the time complexity $O(D^3)$, where D represent the maximum number of degree in a graph. For the search of user-term graph, the algorithm is much faster than conventional search methods whose time complexity are $O(|V||E|)$ at best. In our experiment, maximal complete bipartite graphs which are equal or larger than (3, 3) graph are searched since there are many small isolated complete bipartite graphs (such as one-to-one connection of one user and one term).

3.3 Attachment of Personal Data to UserID

Many complete bipartite graphs are obtained by the above search. Each complete bipartite graph is shown as a list of userID and terms. As a way to assist the understanding of a user

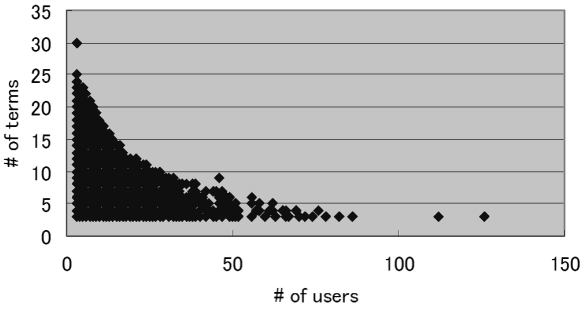

Fig. 4 Distribution of the Size of User Communities

community, personal data that correspond to each userID are attached to resultant graphs. Therefore, each user community is finally represented as terms and personal data (such as sex and birth year) as shown in Fig. 2.

4. Preliminary Experiments

Based on the method described above, preliminary experiment for discovering user communities is performed.

Fig. 4 shows the distribution of the size of discovered user communities. You can see from the figure that the more users included in a user community, the less the number of shared terms is. Most of the discovered user communities have less than 50 users. Reasonable size of user communities is controversial and it depends on how the communities are used. At least, we can claim that the method has abilities of discovering several sizes of unobvious user communities.

The followings are the examples of discovered user communities. As explained above, each user community is represented as a collection of terms and personal data of users (sex, birth year, occupation, and address).

User Community about Soccer

Terms:

- "news"
- "soccer"
- "nikkansports" (name of Japanese sports newspaper company)

Users:

- Male, 1980, college student, kanto
- Male, 1976, college student, kanto
- Male, 1965, service job, kanto
- Male, 1972, college student, kanto

5. Interestingness of user communities

As shown in Fig. 4, many bipartite graphs are searched from user-term graph. Manual inspection of these graphs inposes heavy burden to humans. In order to automate (or semi-automate) the processes of selecting interesting user communities, the concept of interestingness of communities have to be analyzed.

The user-term graph extraced from Web audience measurement data shows relations between users and terms. There are similarities among users and among terms irrespective of the user-term graph. If the similarities among users are defined in the form of distance function, clusters of similar user are obtained based on the distance function. If the members of a user community discovered from user-term graph are dissimilar based on the distance function, the user community reveals unexpected relations among users.

The above explanation can be rephrased in the following way: groups of users can be obtained by 1) clustering based on a distance function between users, and 2) discovery of user communities based on user-term graph. If the users from the latter communities are dissimilar to the former clusters, the user-term graph clarifies novel relations among users that cannot be found from the former clustering. It is expected that such discrepancies between different distance functions can be used for finding interesting user communities.

The same argument is true with terms. If you can define appropriate distance functions among terms, a group of dissimilar terms contained in the same user community will be unexpected and interesting. In order to employ the above idea of interestingness, distance functions of terms and users have to be defined.

5.1 Distance between users

Users' personal data are declared by themselves. As described above, there are several attributes contained in users' personal data, such as birth year/month/day, distinction of sex, income, educational background, occupation, and so on. Some of the attributes that cannot be used for defining distance functions are omitted, and the remaining ones (such as age and sex) are used in our method.

5.2 Distance between terms

Defining appropriate distance between terms or words are not a simple task. For the sake of convenience, the number of co-occurrences of terms obtained from a search engine can be used as the distance of the terms. Since the terms in our method are extracted as substrings of URLs, some of the terms are meaningless strings. Such meaningless strings can be deteced by performing search on a search engine.

6. Discussion

6.1 Related work of user community

There are several attempts for the research of Web log mining such as discovering frequent patterns of log data, or clustering users of specific Web site. Our approach is quite different in that the goal is to discover user communities of similar tastes using Web audience measurement data that record users' visits of several Web servers.

As the research for discovering humans' social network, ReferralWeb[4] is a famous one. Co-citation of technical papers is used for finding relations among AI researchers. Although Kautz's approach focuses on the network among humans, our method utilizes both users and terms, which enables the discovery of user communities that are labeled with terms.

Another related approach is social network analysis from graph structure[3][7]. In this approach, given graphs are decomposed into subgraphs by removing unnecessary edges based on a criterion called edge betweenness. Although the approach succeeds in discovering fractions in a sports club and informal groups in a company, it is not applicable to huge Web data because of its scalability.

In general, discovery of user communities are important for fostering communication among users (such as I2I project[1]), and for performing collaborative filtering (such as Grouplens project). User communities discovered by our new method is rather unsubstantial; each member does not know which communities he or she belongs. Such user communities may not be appropriate for fostering explicit communication among users; they can be used for implicit collaborative filtering (for users) and for site modification (for Web site administrators).

6.2 Evaluation of User Communities

One of the most crucial problems for the research of community discovery is to evaluate its outputs. As mentioned above, evaluation of user communities is not an easy task. In Girvan's paper[3], several examples of communities are shown. Tyler[7] attempts interviews to human subjects in order to evaluate discovered communities.

Evaluation of user communities depends on the way they are used. Our ultimate goals for user community discovery are to build a map of communities that allows us to command a bird's-eye view, and to clarify dynamic changes of communities. As well as real human communities, user communities in the Web are expected to born, grow, and decay over time. As the first step for analyzing user communites, the criteria for judging interesting ones are proposed. Discrepancies of distance among users are simple and applicable to other domains when appropreate distance functions can be defined.

6.3 Web Communities and User Communities

Discovery of user community proposed in this paper is the first step for clarifying dynamic changes of user communities. It should be mentioned that Web communities and user communities are closely related, and that 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 the 9.11 terror occurred, many search of related words such as "CNN" or "world trade center" were performed immediately after. 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.

7. Concluding Remark

The author previously proposed a method for

discovering user communities from Web audience measurement based on the search of maximal complete bipartite graphs. Choosing interesting user communities discovered by the method is crutial for post-processing of the discovery. This paper propose a criterion of interestingness based on discrepancies of distance functions. The author will perform experiments for choosing interesting user communities in order to show the validity of the interestingness.

As the next steps for this research, appropriate definion of distance functions for user communities and Web communities should be investigated. The criteria of interestingness proposed in this paper is based on the assumption that appropriate distance functions are defined in advance. Although the definition of distance between entities is domain-dependent, the idea of focusing on discrepancies between distance functions is expected to be applicable to domains other than user communities. Another direction for further research is to focusing on the static / dynamic relations among user communitites, such as detecting relations among different user communities, and analyzing dynamic changes of user communities.

Acknowledgments

The author would like to express his thanks to Prof. Takeaki Uno for permitting the use of his fast bipartite search algorithm.

References

1) Budzik, J., Bradshaw, S., Fu, X., and Hammond, K. J., (2002). Clustering for Opportunistic Communication, in Proc. of WWW 2002.
2) Flake, G. W., Lawrence, S., Giles, C. L., Coetzee, F. M. (2002). Self-Organization and Identification of Web Communities, IEEE Computer, Vol. 35, No. 3, pages 66–71.
3) Girvan, M., Newman, M. E. J. (2001). Community structure in social and biological networks, online manuscript, http://arxiv.org/abs/cond-mat/0112110/ .
4) Kautz, H., Selman, B., Shah, M. (1997). The Hidden Web, AI Magazine, Vol.18, No.2, pages 27–36.
5) Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A. (1999). Trawling the Web for Emerging Cyber-Communities, Proc. of the 8th WWW conference.
6) Srivastava, J., Cooley, R., Deshpande, M., Tan, P.-N. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, ACM SIGKDD Explorations, Vol.1, No.2, pages 12–23.
7) Tyler, J. R., Wilkinson, D. M., Huberman, B. A. (2003). Email as spectroscopy: automated discovery of community structure within organizations, http://xxx.lanl.gov/arXiv:cond-mat/0303264 .
8) Uno, T. (2003). Fast Algorithms for Enumerating Cliques in Huge Graphs, submitted.