International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 5, Issue 2, March - April 2016 **ISSN 2278-6856**

CLUSTERING OF USERS BASED ON BROWSING BEHAVIOR

M. Sree Vani

Associate Professor, Dept of CSE, MGIT , Hyderabad-500075

Abstract: ***Web mining is the intelligent analysis of Web data. With Web mining techniques, business organization can gain a better understanding of both the web and web users' preferences to help them run their business more efficiently. By analyzing the characteristics of the user clusters, web users can be understood better and thus can be provided with more suitable and customized services. To identify subgroups of users, we used clustering to categorize users on the basis of their behaviors. Each clustering method uses different criteria to group data objects. Our solution proves that group together clients or data items that have similar characteristics gives best results.***

Keywords: web mining, user behavior, browsing data

1. INTRODUCTION

Web mining is the intelligent analysis of Web data. With Web mining techniques, business organization can gain a better understanding of both the web and web users' preferences to help them run their business more efficiently. Recently, Web Usage Mining (WUM) is an active area of research and commercialization. The goal of WUM is to leverage the data collected as a result of user interactions with the web to learn user models which are beneficial for web personalization. An important topic in Web Usage Mining is clustering web users - discovering clusters of users that exhibit similar information needs, e.g., users that access similar pages. By analyzing the characteristics of the clusters, web users can be understood better and thus can be provided with more suitable and customized services. Nowadays, various data mining techniques have been successfully applied to Web access logs to extract useful information. Among them, clustering allows us to group together clients or data items that have similar characteristics.

2. RELATED WORK

Tasawar et al., [1] proposed a hierarchical cluster based preprocessing methodology for Web Usage Mining. Web usage mining based on fuzzy clustering in identifying target group is suggested by Jianxi et al., [3]. The author enhanced the fuzzy clustering technique to identify groups that share common interests and behaviors by examining the data collected in Web servers.Houqun et al., [4] proposed an approach of multi-path segmentation clustering based on web usage mining. They deals with examining and researching methods of web log mining and bringing forward a multi-path segmentation cluster technique, that segments and clusters based on the user access path to enhance efficiency.

3. METHODOLOGY

We present an efficient mechanism for clustering of users based on user browsing behavior. A user can be browse number of pages based on user's interests and it generates a more user response. This mechanism provides an opportunity to generate a higher user response and satisfaction increasing business value. The steps in the algorithms are as follows:

Step 1: Web usage Data Preprocessing.
Step 2: Finding Similarity between user's
Step 3: Clustering of users based on Browsing Behavior

3.1 Web Usage Data Preprocessing

Web server logs are the primary sources for usage data in which the activities of web users are registered. These log files can be stored in various formats such as Common or Extended log formats. Basically, an entry in Common log format consists of the following fields as shown in the following Figure 1.This original format of log file needs be modified according to the following reasons as shown in Table 1.

```
141.243.1.172[29:23:53:25]”GET/Software.html HTTP/1.0” 200 1497
Quer2.lycos.cs.cmu.edu[29:23:53:36]”GET/Consumer.html     HTTP/1.0”
200 1325
Tanuki.twics.com[29:23:53:53]”GET/News.html HTTP/1.0”200 1014
Wpbfl2-45.gate.net[29:23:54:15]”GET/HTTP/1.0”200 4889
Wpbfl2-
45.gate.net[29:23:54:16]”GET/icons/circle_logo_small.gifHTTP/1.0”200
2624
Wpbfl2-
45.gate.net[29:23:54:18]”GET/logos/small_gopher.gifHTTp/1.0”200 935
```

Figure 1: Web log file - sample template

For IP address, some values starting with same prefix can be interpreted requests from the same user or group of users. In order to understand the user behavior the following information could be extracted from server logs. In order to extract a particular user behavior from remaining logged users, each record in the log file should be written in such a way to uniquely identify users who performed it to study browsing behavior. A user here typically represents a person, computer, domain or company. It is an easy task if the log file records a person ID such as login user or computer name. However, it is a non trivial task in case of multiple users logging from a single computer especially when web sites do not require users to log in with a user name. Most web servers do not assist providing consistent user login identity to take help out. Thus, the information available according to the HTTP standard is not adequate to distinguish a user among all other users when browsing on same host and proxy.

Table 1: Web log file Re-formatting

Field	Original format	Re-formatted
IP address	129.173.66.192	129.73
Date/Day	18/Oct/2009	{Mon,Tue,Wed,THU,Fri,Sat,Sun}
Time	13:04:56	{morning,afternoon,evening,night}
URL	/~user/index.html	/~user/index.html

The most widespread remedy for this problem is the use of cookies and session variables. Another way to identify unique users is using a heuristic method in which unique IP address as a user will be identified with IP address when IP addresses resolve into domain names registered to a person, domain or company as it is possible to gather more specific information from Domain name servers. Once the users are identified, server log data passes through a session reconstruction step in which we process reconstructing the user's original sessions by using server log data. Reconstructing user sessions from server logs is a challenging task since the access logs protocol is stateless and connectionless. If neither the cookies nor the user-login information are available, the reconstruction of original session is based on two basic heuristics: 1. Time oriented (h1,h2) and 2. Navigation oriented (h3). Time oriented heuristic considers the browsing time patterns and past browsing analysis. Navigation oriented heuristic considers the site topology and trace route information. Accesses to cached pages are not recorded in the web log due to the browser or proxy cache does not send request to web servers. Therefore, references to cached pages will not be logged. However, the missing references in the log file can be found using a set of assumptions. The referrer field of the web log or the web site structure can be used to infer cached pages when requests are analyzed with other cookie and session information. If a requested web page P_i is not reachable from previously visited pages in a session, then a new session is constructed starting with page P_i. The irrelevant page requests which comprise of URLs of embedded objects with filename suffixes like .gif, .jpeg. png, .pdf etc., can be considered to remove from logging records unless they help to evaluate users behavior. Eventually, this step produces a set of user sessions $S=\{S_1,\ldots.S_m\}$ with necessary parameters to uniquely identify a logged in user.

3.2 Similarity between user's

We use the cosine similarity as the similarity measure. The similarity between user vector Ul = {Y1,1, Y1,2, … Yl,k}, and user vector, U2 = {Y2,1, Y2,2, …Y2,k}, is defined as a cosine similarity measure:

$$sim(u_i,u_j)=\frac{\sum_{l=1}^{m}u_l^i * u_l^j}{\sqrt{\sum_{l=1}^{m}(u_l^i)^2}\sqrt{\sum_{l=1}^{m}(u_l^j)^2}}$$

With the user-page_view matrix , we can easily discover the user clusters by measuring the similarities among row/column vectors, respectively. Specifically, we first compute the similarities among different vectors (row vectors for user clustering) and obtain the similarity matrix Simk*k which is shown in Eq. (3). It means that we evaluate the similarity values row by row: if the similarity value is great than given threshold, the corresponding row number which represent corresponding users fall into one class.

$$sim_{kxk}=\begin{pmatrix} 1 & sim(1,2) & sim(1,3)\cdots & sim(1,f)\cdots & sim(1,m) \\ & 1 & sim(2,3)\cdots & sim(2,f)\cdots & sim(2,m) \\ & & 1 & \ddots\ \vdots & \ddots\ \vdots \\ & & & 1\,sim(i,j)\cdots & sim(i,m) \\ & & & & \ddots\ \vdots \\ & & & & 1 \end{pmatrix}$$

3.3 Clustering of users based on Browsing Behavior

A user cluster is a group of users that seem to behave similarly when navigating through a Web site, they access conceptually related Web pages of a Web site during a given period of time. We suppose that:

(l)The users with similar interests should have the similar browsing patterns.

(2) Associated Web pages should be browsed by the users with similar interests.

(3) The general browsing patterns are not changeable during a given period of time for a given user, although different users' browsing patterns maybe different during the specified period of time.

Based on the above assumptions, we can draw user clusters from Web logs by the analysis of users' browsing information during the period of time. To identify subgroups of users, we used clustering to categorize users on the basis of their behaviors so that appropriate advertisements could be served to them. We used K-Means clustering because ,our solution does not require the identification of user sessions from Web logs and a user can be assigned to not more than one cluster. Furthermore, the approach is not based on sequential pattern mining, so it avoids the difficulties of performance and scalability. The inputs to clustering are user impression data stored in web log files [4]. Next, we picked k1 number of user clusters using K-Means

Websites

clustering. User clustering groups Internet customers on the basis of similar website visitation behavior (Figure 2).

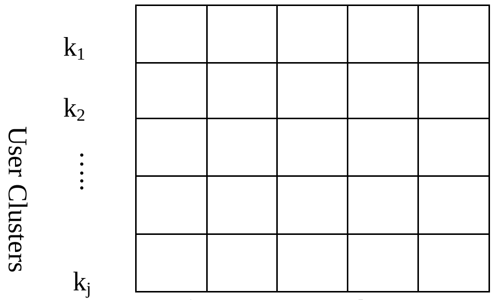


Figure 2: User Clustering

4. Experiments and Results

We used Data sets consisting of web log records for 5446 users are collected from De Paul University website. There are mainly two data source: user table and log table, user table saves web user information, which includes some attribute as: id, machine name, user name, true name, class name, start time, total online time. And log table saves web user browsering log information, which includes some attribute as: logonname, machine name, neturl, logtime, etc. For Web user clustering, by scanning the test data sets, we obtain a user i and the corresponding Web pages URLsri.Then we decide which user cluster the user i belongs to according to the discovered user clusters, we obtain the predicted URLspi. By comparing URLspi and URLsri, we obtain the precision metric of Web user clustering as follows:

$$\mathrm{Pr}ecision(UserClustering) = \frac{1}{m}\sum_{i=1}^{m}\frac{\|URLs_i^p\ \theta URLs_i^p\,|}{\|URLs_i^p\,\|}$$

Before experiment, a critical step in effective Web mining is data preprocess, whose aim is to transform log data to an appropriate format for analysis, according to the need of the mining analysis. Generally, data preparation needs to meet the requirements of the particular mining task. For our clustering analysis, data preprocessing contains two steps: data cleaning, data statistics. Data cleaning means removing redundant data, leaving useful data for analysis, and data statistics means getting hits information of web user access different pages. And then based on such information, we construct user-page_view matrix, which is shown in figure 3.5 as follows:

	P_1	P_2	P_3	P_4	…	P_k
U^1	28	12	33	10	…	45
U^2	32	76	23	133	…	54
U^3	312	333	223	33	…	113
U^4	231	13	23	323	… ..	233
… .	…	…	…	… .	…	…
U^m	23	234	343	34		333

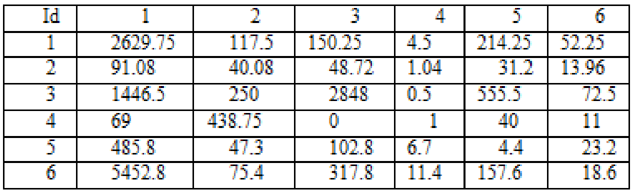
Figure 3 : User-Page_view matrix

In Fig.1, the number of Web pages is 8, the number of users is 50, and the value 312 in the third row and the first column means frequency of user3 visited web page l. After getting hits information from matrix, we apply K-Means algorithm into clustering web user into some clustering with different k values, whose result is shown in Table 3.2.

Table 2 : Clustering Result

K value	Precision(user clustering)
4	0.7582
5	0.8283
6	0.8912
7	0.8523

Table 3: Cluster center



Id	1	2	3	4	5	6
1	2629.75	117.5	150.25	4.5	214.25	52.25
2	91.08	40.08	48.72	1.04	31.2	13.96
3	1446.5	250	2848	0.5	555.5	72.5
4	69	438.75	0	1	40	11
5	485.8	47.3	102.8	6.7	4.4	23.2
6	5452.8	75.4	317.8	11.4	157.6	18.6

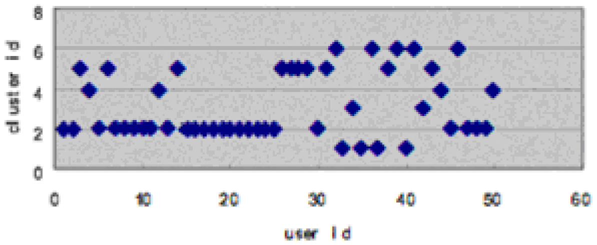


Figure 4: Clustering result of web user

From Table II, we can see that when the value k is set 6, the similarity preference of algorithm is best, so the clustering result is shown as Figure3.6 based on k=6. And the cluster center value is shown in Table 3.3.

5. Conclusion

By analyzing the characteristics of the user clusters, web users can be understood better and thus can be provided with more suitable and customized services. To identify subgroups of users, we used clustering to categorize users on the basis of their behaviors. Each clustering method uses different criteria to group data objects. Our solution proves that group together clients or data items that have similar characteristics gives best results.

References

1. O. Etzioni, "The World-Wide Web: Quagmire or gold mine?," Commun. ACM, vol. 39, no. 11, pp. 65–68, 996.
2. H. Chen and M. Chau, "Web mining: Machine learning for Web applications," Anu. Rev. Inf. Sci., vol. 38, pp. 289–329,2004.
3. X. Fang, M. Chau, P.J. Hu, Z. Yang, and O.R.L. Sheng, "Web mining based objective metrics for measuring Website navigability," in Proc. Int. Conf. Inf. Syst., Milwaukee, WI, 2006.

4. R. Kosala, “Web mining research: A survey,” ACM SIGKDD vol. 2, no. 1, pp. 1–15, 2000.

5. J. Cho, H. Garcia-Molina, and L. Page, “Efficient crawling through URL ordering,” in Proc. 1998 WWW Conf., Brisbane, Australia, 1998.

6. S. Chakrabarti, M.V.D. Berg, and B. Dom, “Focused crawling: A new approach to topic-specific Web resource discovery,” in Proc. 1999 WWW Conf., Toronto, Canada, 1999.

7. W.Y. Chung, G. Lai, A. Bonillas, W. Xi, and H. Chen, “Organizing domain- specific information on the Web: An experiment on the Spanish business Web directory,” Int. J. Hum.-Comput. St., vol. 66, no. 2, pp. 51–66, 2008.

8. H. M. Chen and M. D. Cooper, “Using clustering techniques to detect usage patterns in a Web-based information system,” J. Amer. Soc. Inf. Sci. Tech., vol. 52, no. 11, pp. 888–904, 2001