Web User Clustering Analysis based on KMeans Algorithm

JinHuaXu
Computer and Information Engineering College
Zhejiang Gongshang University
HangZhou, China
e-mail:xjh@mail.hzic.edu.cn

HongLiu
Computer and Information Engineering College
Zhejiang Gongshang University
HangZhou, China
e-mail:LLH@mail.hzic.edu.cn

***Abstract*—As one of the most important tasks of Web Usage Mining (WUM), web user clustering, which establishes groups of users exhibiting similar browsing patterns, provides useful knowledge to personalized web services. In this paper, we cluster web users with KMeans algorithm based on web user log data. Given a set of web users and their associated historical web usage data, we study their behavior characteristic and cluster them. Experiment results show the feasibility and efficiency of such algorithm application. Web user clusters generated in this way can provide novel and useful knowledge for various personalized web applications.**

Keywords- web user;clustering; KMeans; vector matrix; similarity

I. INTRODUCTION

Internet has become increasingly important as a medium for life, work and study as well as for dissemination of information. Web mining is the mining of data related to the World Wide Web. It is categorized into three active research areas according to what part of web data is mined, of which Usage mining, also known as web-log mining, which studies user access information from logged server data in order to extract interesting usage patterns.

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. One kind of outcome of Web mining is Web browsing patterns. By the use of Web browsing patterns, business organizations can perform mass customization and personalization, adapt their Web sites, and further improve their marketing strategies, product offerings, and promotional campaigns. Therefore, Web browsing pattern mining has special meaning for business organizations. Thus, it has attracted much attention from data mining, machine learning, and other research communities for many years. Of the existed methods, some are non-sequential, such as association rule mining and clustering; and some are sequential, such as sequential or navigational pattern mining. Both approaches ignore the site topology and need to identify user sessions.

In this paper, we explore the problem of user clustering based on vector matrix and KMeans algorithm. 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 rest of this paper is organized as follows: in Section 2, we introduce related work. Section 3 we explicate the approaches and algorithms applied. The experiment and discuss are introduced in Section 4. Finally, Section 5 summarizes the paper.

II. RELATED WORK

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. Existing web usage data mining techniques include statistical analysis [1], association rules [2], sequential patterns [3], classification [4], and clustering [5]. 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. What' more, a project aiming at extracting navigation behavior models of a site's visitors was introduced in [7]. Two classification-type experiments were implemented to predict visitors' sex and if visitors would be interested in some section of the website. The results of both experiments were not very good with classification accuracy all under 56%.

Clustering analysis to mine the Web is quite different from traditional clustering due to the inherent difference between Web usage data clustering and classic clustering. Therefore, there is a need to develop specialized techniques for clustering analysis based on Web usage data. Some approaches to clustering analysis have been developed for mining the Web access logs.

Perkowitz and Etzioni [8] discuss adaptive Web sites that learn from user access patterns. The PageGather algorithm uses the page co-occurrence frequencies to find clusters of related but unlinked pages. Mobasher, Cooley and Srivastava [6] propose a technique for capturing common user profiles based on association-rule discovery and usage-based

978-1-4244-8106-4/$26.00 © 2010 IEEE

clustering. Cooley [10] introduces an algorithm that classifies users using a hypergraph partitioning technique. Cooley's method is used to identify particularly interesting and similar path histories, but it cannot be used to gain an overall picture of all usage of a Web site. Nasraoui and Krishnapuram [11] use unsupervised robust multi-resolution clustering techniques to discover Web user groups. Xie and Phoha [12] use belief functions to cluster Web site users. They separate users into different groups and find a common access pattern for each group of users. Unfortunately the approach still needs to identify sessions.

III. TECHNOLOGY OVERVIEW

In this section, we firstly give the formal definitions of the problem and then introduce some general information on interrelated technology.

A. Problem statement

For a Web site, we use hits(user, url) to denote the frequency of a user who browses Web page url of the Web site during a period of time. Suppose a.html is a Web page which has been visited by user User1 10-times and by user User2 20-times during a given period of time, the state of the Web page a.html is State(a.html)={a.html,{User1, 10}, {User2, 20}}. And a user cluster CoU is a group of users that seem to behave similarly when navigating through a Web site, specifically, they access conceptually related Web pages of a Web site during a given period of time.

We suppose that:

(1)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.

B. Vector matrix representation

Before clustering web user based on web logs, we firstly construct vector matrix about url and user. A URL—User associated matrix R is used to describe the relationship between Web pages and users who access these Web pages. Let n be the number of Web pages and let m be the number of users, the matrix can be denoted as

$$R_{m\times n}=\begin{pmatrix} hits(1,1)\ hits(1,2)\cdots hits(1,j)\cdots hits(1,n) \\ hits(2,1)\ hits(2,2)\cdots hits(2,j)\cdots hits(2,n) \\ \vdots \quad \vdots \quad \ddots \quad \vdots \quad \ddots \quad \vdots \\ hits(i,1)\ hits(i,2)\cdots hits(i,j)\cdots hits(i,n) \\ \vdots \quad \vdots \quad \ddots \quad \vdots \quad \ddots \quad \vdots \\ hits(m,1)\ hits(m,2)\cdots hits(m,j)\cdots hits(m,n) \end{pmatrix} \tag{1}$$

Where, hits mean one kind of user browsing information. We can directly extract the hits of all users who access the Web pages of a Web site during a given period of time. In the matrix, we view users as rows, Web pages as columns, and the count of hits as the values of the elements of this matrix, that is hits(i,j) as the count of user i accesses Web page j during a defined period of time. The ith row vector R[i,] records the counts of the ith user accesses of all the Web pages during the specified period of time, and the jth column vector R[, j] records the counts of all users who access the jth Web page during the same period of time.

C. Similarity measure

We use the cosine similarity as the similarity measure. The similarity between vector, $u_1=\{y_{1,1}, y_{1,2},..., y_{1,n}\}$, and vector, $u_2=\{y_{2,1}, y_{2,2}, ..., y_{2,n}\}$, is defined as a cosine similarity measure:

$$Sim(u_i,u_j)=\frac{\sum_{k=1}^{m} u_k^i * u_k^j}{\sqrt{\sum_{k=1}^{m}(u_k^i)^2}\sqrt{\sum_{k=1}^{m}(u_k^j)^2}} \tag{2}$$

With the matrix R, 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 Sim_{n*n}, 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.

$$sim_{n\times n}=\begin{pmatrix} 1\ sim(1,2)\ sim(1,3)\cdots sim(1,j)\cdots sim(1,m) \\ 1 \quad sim(2,3)\cdots sim(2,j)\cdots sim(2,m) \\ 1 \quad \ddots \quad \vdots \quad \ddots \quad \vdots \\ 1\ sim(i,j)\cdots sim(i,m) \\ \ddots \quad \vdots \\ 1 \end{pmatrix} \tag{3}$$

D. KMeans algorithm

Kmeans is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this

algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (4)$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j is an indicator of the distance of the n data points from their respective cluster centres.

The flow of algorithm is shown as the following steps:

(1) Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

(2) Assign each object to the group that has the closest centroid.

(3) When all objects have been assigned, recalculate the positions of the K centroids.

(4) Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

E. *Experimental evaluation*

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:

$$\Pr ecision(userclustering) = \frac{1}{m} \sum_{i=1}^{m} \frac{\| URLs_i^p \theta URLs_i^r \|}{\| URLs_i^r \|} \quad (5)$$

IV. EXPERIMENT AND DISCUSS

In this paper, experiment data is from web log data of school lab. 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. The log data are shown as Table I.

TABLE I. CLUSTERING RESULT

Logon Name	Machine Name	NetUrl	LogTime
09054 00212	236-68	sns.qzone.qq.com/cgi-bin/friendsale/frisale_cgi_pkspace?otype=js	2010330 12:55
09054 00212	236-68	m101.mail.qq.com/cgi-bin/mail_list?sid=…..	2010331 19:43
09054 00212	236-68	acookie.taobao.com/1.gif?acookie_load_id=…	2010418 11:39
….	….	…..	….

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 vector matrix, which is shown as follows:

21	16	0	0	0	5	0	0
23	44	0	0	0	94	0	0
312	46	3	0	0	35	0	0
69	311	0	0	78	8	13	37
0	44	0	0	67	8	0	0
576	44	0	0	0	102	0	0
0	143	0	0	73	7	0	0
164	121	34	0	125	9	0	0
71	110	0	0	18	110	0	0
99	63	269	0	0	7	0	0
142	140	0	15	339	37	7	60
115	809	0	4	82	12	0	3
62	5	0	0	0	7	0	1
557	114	0	0	0	0	0	19
0	0	0	0	0	0	0	0
71	37	0	0	0	2	0	0
199	20	0	1	0	18	39	63
32	90	0	0	0	3	38	12
0	55	0	0	0	0	127	59
123	31	0	0	0	10	8	6
137	10	53	1	0	0	19	28
233	38	0	0	158	16	0	0
192	11	0	0	0	16	59	49
20	2	0	1	0	0	0	0
25	3	0	0	0	0	0	0
321	0	0	0	0	0	0	0
336	0	239	0	0	2	0	0
349	17	110	0	27	12	17	87
925	85	47	13	0	7	184	150
92	15	503	0	0	0	0	1
695	154	140	54	17	21	0	0
8024	5	726	2	0	4	0	0
1863	168	2	18	123	51	35	184
2232	388	2913	1	551	100	0	4
3318	154	599	0	574	60	32	115
4448	59	0	0	136	28	0	0

Figure 1. Vector 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 number of user3 hitting web page1.

After getting hits information vector matrix, we apply KMeans algorithm into clustering web user into some clustering with different k values, whose result is shown in Table II.

TABLE II. CLUSTERING RESULT

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

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 Fig.2 based on k=6.

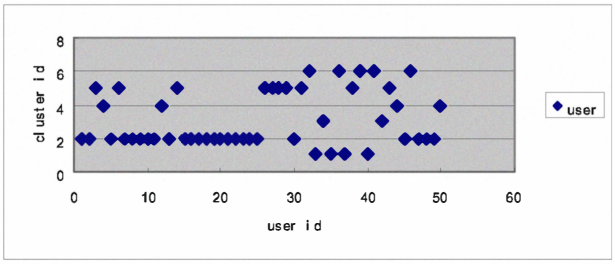

Figure 2. Clustering result of web user.

And the cluster center value is shown in Table III.

TABLE III. CLUSTER CENTER

id	1	2	3	4	5	6	7	8
1	2629.75	117.5	150.25	4.5	214.25	52.25	16.75	84.75
2	91.08	40.08	48.72	1.04	31.2	13.96	12.4	12.2
3	1446.5	250	2848	0.5	555.5	72.5	0	12
4	69	438.75	0	1	40	11	3.25	10
5	485.8	47.3	102.8	6.7	4.4	23.2	20.1	32.9
6	5452.8	75.4	317.8	11.4	157.6	18.6	12.4	32.4

V. CONCLUSION

In this paper, we have presented vector analysis and KMeans based algorithms for mining user clusters. We have also applied the proposed algorithms to the real world data and our experimental results show the proposed algorithm is feasible, and have scalability.

REFERENCES

[1] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan, Web usage mining: discovery and applications of usage patterns from web data, in: SIGKDD Explorations, vol. 1 (2), 2000, pp. 12–23.

[2] B. Mobasher, H. Dai, T. Luo, M. Nakagawa, Effective personalization based on association rule discovery from web usage data, in: Proc. of wIDM, 2001.

[3] Q. Yang, H.H. Zhang, T. Li, Mining web logs for prediction models in WWW caching and prefetching, in: Proc. of ACM SIGKDD, 2001.

[4] T. Li, Q. Yang, K. wang, Classification pruning for web-request prediction, in: Proc. of www, 2001.

[5] B. Mobasher, R. Cooley, J. Srivastava, Creating adaptive web sites through usage-based clustering of URLs, in: Proc. of IEEE KDEX workshop, 1999.

[6] Mobasher, B., Cooley R., and Srivastava, J. Creating Adaptive Web Sites Through Usage-based Clustering of URLs, Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, 1999.

[7] Lalmas, M. Dempster-Shafer's Theory of Evidence applied to Structured Documents: modeling Uncertainty, SIGIR97, Philadelphia, USA, 1997.

[8] Perkowitz, M., Etzioni, O. Adaptive Web sites: automatically synthesizing Web pages in Proceedings of Fifteenth National Conference on Artificial Intelligence, Madison, WI, 1998.

[9] D.S. Phatak, R.Mulvaney, Clustering for personalized mobile web usage, in: Proceedings of the IEEE FUZZ'02, Hawaii, USA, 2002, pp. 705–710.

[10] R. Cooley, Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data, Ph.D. Thesis, University of Minnesota, May 2000.

[11] O. Nasraoui, R. Krishnapuram, A new evolutionary approach to Web usage and context sensitive associations mining, International Journal on Computational Intelligence and Applications—Special Issue on Internet Intelligent Systems 2 (3) (September 2002) 339–348.

[12] Y. Xie, V. Phoha, Web user clustering from access log using belief function, in: Proceedings of the ACM K-CAP'01, First International Conference on Knowledge Capture, Victoria, British Columbia, Canada, (2001), pp. 202–208.