Web usage mining based on WAN users' behaviors

Hao Yan, Bo Zhang, Yibo Zhang, Fang Liu, Zhenming Lei
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications, Beijing
yanhao71@gmail.com

Abstract—**Web mining focuses on extracting useful information from large volumes of Web data. Web usage mining (WUM) is one of important application which applies Web mining techniques to discovery usage patterns from Web accessing data. Meanwhile clustering performs a key role in distinguishing different kinds of usage patterns from raw data. Considering usage features of activities, information scope and preference, we propose a two-step K-means clustering algorithm to search user groups in realistic data collected from WAN. In the paper, some useful practical conclusions are also presented to facilitate design of targeting and recommending applications.**

Key words: Web usage mining; usage pattern; clustering; targeting and recommending applications

I. INTRODUCTION

Web mining is an important work to extract useful knowledge from web data. With the extracted knowledge, researchers can recognize patterns about web traffic, network users' behaviors and interaction between web pages and users. Furthermore, these patterns are used to support strategies in business management such as website maintaining, page personalization, directional marketing and so on [9].

In terms of data source, Web mining is divided into three types, namely Web content mining, Web structure mining and Web usage mining (WUM) [11]. Web content mining focuses on the real data in the web pages such as text and graphics. Web structure mining analyzes the hyper-links to find web pages structure information. WUM analyzes web logs including IP addresses, page references and timestamps and so on.

This paper deals with WUM, which faces challenges of data availability, mining efficiency, diversity of users' behaviors, features selection, and evolving usage patterns [4] [8] [9].

Traditionally, web usage mining technologies always analyze data collected from a single web site [1] [2] [4], when apply those approaches in to multiple web sites, the entire users' behaviors can be described and the discovered patterns can be referred by overall web sites. In this paper we collect data from a WAN to achieve comprehensive analysis.

Data mining technologies include statistical analysis, clustering, classification, association rules and dependency modeling are applied to Web usage data [2] [4]. Among these technologies, clustering is used frequently to capture different user behavior classes. An efficient clustering algorithm should be used when it comes to huge data set. In [10], the time and space complexities of clustering algorithms are estimated and K-means is proved to be efficient enough in huge data set.

Recent year, targeting customers, personalization and recommending systems are become hot topics for their valuable applications in commerce [9]. These applications are mainly achieved on the foundation of understanding the requirement and preference of users by WUM, and they can help the providers to design services to attract more customers and meanwhile maintain scale of existing customers.

In this paper, we introduce two-step K-means cluster algorithm to mining web usage patterns with four attributes extract from collected data. The four attribute, namely page access times, category number, relative entropy and element of categories can reflect users' activities, information scope and preference which can embody users' requirements in different aspects. The conclusions based on the analysis of result clusters provide some heuristic ideas to design targeting or recommending applications.

The rest part of paper are organized as follows: SectionⅡ presents the related work, SectionⅢ shows the data collection, the methodology is describe in SectionⅣ, SectionⅤ presents the clustering results and analysis, conclusions and future works are posed in SectionⅥ.

II. RELATED WORK

Several cluster algorithms are applied in web usage mining. In [1], rough set based BLEM2 algorithm is used to classify and predict web usage patterns. Then they show the improvement of predicting accuracy by comparing predicting results between BLEM2 algorithm and centroid based algorithm in their data set. In [2], Kobra applies Kohonen map for clustering phase and detecting user's navigation behaviors. The advantage of Kohonen map is that the cluster number does not need to be given. Fuzzy clustering algorithm is also introduced in [3] and [4]. In [3], page-click number and web browsing time are used to cluster similar web users. At the end of [3], authors point out that how to apply cluster algorithm to internet is a direction to study. In [4], Jianxi Zhang lands the partition matrix using gradient-based scheme and apply the improved fuzzy clustering algorithm for marketing in a bank. In [5], Marcelo proposes a Customer Behavior Model Graph based Workload Characterization Algorithm to classify users. By analyzing the classified user groups, Daniel finds browsing time in a web store has the negative relationship with the

This work is supported by 111 Project of China under Grant No. B08004 and the Fundamental Research Funds for the Central Universities (BUPT2009RC0123).

978-1-4244-3709-2/10/$25.00 ©2010 IEEE

probability of a customer buying an item. K-means clustering algorithm is used in [6] to divide over 1 million users into 5 groups and each of the group are defined by their social behaviors.

Most of data used in aforementioned works come from web servers, and they are handled by a clustering process in one time. Here we propose two-step K-means clustering algorithm which is introduced in SectionⅣ to mine the realistic data of multiple web sites, and group users with attributes extracted from two concept levels to reveal hierarchical user clusters. Finally, we have a further analysis in result clusters.

III. Data Collection

The data collected by network traffic monitoring equipment arranged on the portal of a typical WAN owned by an ISP in China include 39165 ADSL users' web access logs in a week in Sep. 2009. Each record in web access logs is formed by a user account and a page URL. The process of collection is as follows: while a user is browsing a web page, the client PC sends HTTP GET request packet to web server. The packet is copied by the monitoring equipment when it reaches the network portal. Then the equipment analyzes payloads of the GET request packet and captures the page URL from URL domain. Finally, a software in the equipment writes user account and URL into the log.

We classified 2700 pages which hold more than 95% access times launched by users into 16 different content categories, including: IT, economy, e-commerce, service, software, international, portal, living, search, sports, wireless business, news, leisure, games, video and knowledge. Next, web access logs are transformed into average access times of the 16 pages categories for each user in one day according to the classification.

In the following analysis, we replace the original account with corresponding unique meaningless random string for privacy concern while maintaining consistency of data for each user, which does not affect the results of the analysis.

IV. Methodology

A. Attribute selection

Attributes extracted from data play important roles in result interpretation. In this paper we try to explore in the following aspects: first, the users' activities in web usage; second, the scope of information obtained from web pages; third, the significance of preference.

For the consideration above, we select four attributes as follows:

- **Page access times** (PAT): This is total web pages access times of a user, which reflects the users' activities in web usage in quantity. It is easy to see that PAT equal to the sum of average access times of the 16 pages categories.
- **Category number** (CN): This is the number of different web page categories accessed by users. It can present the scope of information users absorb from the internet.
- **Relative entropy** (RE): In information theory, entropy measures the "observational variety" of variation. With the increasing of the entropy, the frequencies of observed values tend to indistinguishable [7]. Based on the above consideration, entropy can describe users' signification of preference in quantity. The higher entropy is, the less significant preference users have. Since different number of page categories lead to different value ranges, we use RE instead.

  The RE is defined as:

$$RE = \begin{cases} \dfrac{E}{E_{\max}} = \dfrac{-\sum_{i}^{n} p_i \log_2 p_i}{\log_2 n} & n \geq 2 \\ 1 & n = 1 \end{cases}$$

  where pi is the percentage of PAT of category i, and n is the number of categories. When n=1, we define RE=1, which means the user only visit one page category.
- **Element of categories** (EC): This attribute has 16 dimensions which are in terms of the access times percentage of the 16 categories. It reflects the constitution of PAT and can reveal users' concrete interests.

B. Two-step K-means clustering

In this paper we use K-means algorithm and separate the clustering process in to two steps: In step one, we cluster PAT, CN and RE, and in step two we cluster EC.

We use two-step K-means algorithm on the following considerations:

K-means algorithm is easy to implement and efficient enough to deal with large data set. The drawbacks exit in K-means algorithm is "*How many clusters?*" [6]. A general criterion is: after clustering, data within a cluster have a high similarity and data between clusters have low similarity. The authors in [5] presented a method to measure the similarity with coefficient of variation in which C_{intra} is defined as coefficient of variation of intra cluster distance and C_{inter} is defined as coefficient of variation of inter cluster distance. The β_{cv} which is donated as the ratio between C_{intra} and C_{inter} helps us to find the value of K.

The four attributes can be divided into two concept levels that PAT, CN and RE describe users' web usage profiles in macroscopic while EC describes users' web usage profiles in microscopic. Meanwhile, the former attributes have one dimension in each and the later attribute has 16 dimensions. That is to say if we give equivalent weight to each dimension when clustering, EC may have a lager effect on the results.

TABLE I. RESULT OF STEP ONE

Cluster	1	2	3	4	5	6	7	8	9	10
PU	24.67%	16.04%	13.92%	12.84%	9.29%	6.62%	5.41%	5.42%	3.70%	2.15%
PAT	35.64	44.93	90.14	50.11	45.38	143.27	69.42	158.67	267.13	198.77
CN	4.6	6.4	4.1	3.9	2.7	3.4	2.7	5.7	4.1	2.8
RE	0.90	0.90	0.88	0.74	0.90	0.84	0.54	0.89	0.86	0.53

V. RESULTS AND ANALYSIS

A. *Result of step one*

In order to give equivalent weigh to PAT, CN and RE, we should normalize PAT and CN before clustering, so that the three attributes have same effect on results.

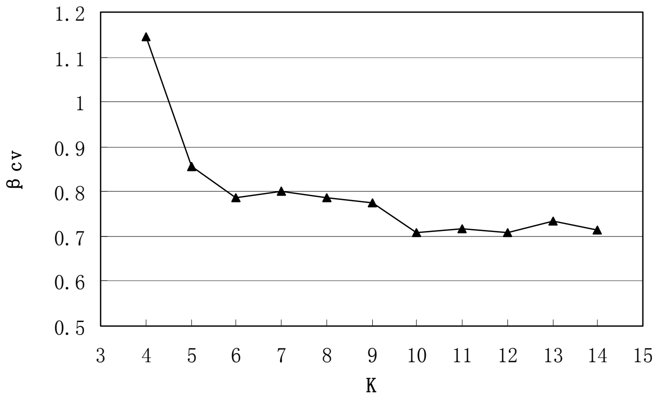

Figure 1. β_{cv} varies with K

Figure 1 shows the value of β_{cv} varies with the increase of K. As it can be seen from the figure, β_{cv} falls fast when K changes from 3 to 6. From 6 to 10, K declines slowly, only appearing a slight fluctuation. The change between 10 and 15 tends to smooth. This is an indication that 10 is an appropriate value of K.

Table I presents the Percentage of Users (PU) in the first row and centroids of attributes in the following rows. Clusters are in descending order by PU.

B. *Analysis one*

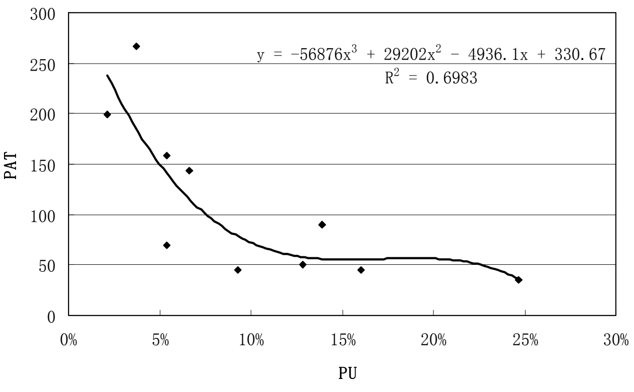

Figure 2. PAT varies with PU in clusters

It can be seen from Table I that the centroids of PAT tend to decline with the increase of PU. To show this pattern clearly, we plot the centroids of PAT vary with PU in Figure 2. The line regressed by the points which present the clusters manifests that there is negative relationship exiting between PAT and PU. That is to say the fewer users clusters have, the more active they use web service. This pattern enlightens us that small size clusters should be considered first when we define active user groups.

In Table I we can see that although we have classified the pages into 16 categories, the centroid of CN in each cluster does not catch up the half number of 16. This indicates a pattern that most users require limited scope of information when they are surfing. Based on above analysis, web sites can provide personalized service with knowing users' requirements to enhance users' loyalties and maintain existing users.

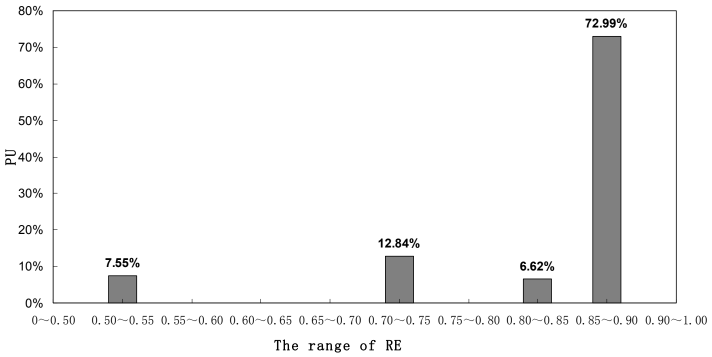

Figure 3. The PU distributes in the range of RE

In Figure 3, the range of RE are distributed in x axis, and the sum PU of the clusters whose RE centroids fall in the same range are presented in y axis. In [7], when the RE is around 0.9, the observed values of variance are close to be "uniformly distributed". From Figure 3, we can see that 72.99% users fall into the span of 0.85 to 0.90, which indicates most of users visit page categories approximately evenly. Only a few users have significant preference. The analysis above illustrates that there are huge number of users whose preferences are ambiguous, and this part of users should be filtered or compressed to ensure targeting rate in directional marketing.

C. *Result of step two*

In this step, we cluster users using the EC, i.e. the percentages of access times of the 16 categories, based on the result clusters in step one. The 10 values of K are selected with the same method described in Section Ⅳ. The result clusters in step two are list in TableⅡ.

TABLE II. RESULT OF STEP TWO

Cluster	PU	Main page categories	Cluster	PU	Main page categories
$C_{1,1}$	9.84%	leisure/36%, portal/26%, search/12%	$C_{6,3}$	0.62%	service/48%,leisure/18%, search/12%
$C_{1,2}$	6.55%	search/36%, portal/18%, leisure/17%	$C_{6,4}$	0.22%	economy/67%, leisure/14%
$C_{1,3}$	4.16%	service/32%, portal/17%, search/13%	$C_{6,5}$	0.10%	news/72%, leisure/10%
$C_{1,4}$	4.10%	video/34%, portal/18%, leisure/17%, search/13%	$C_{7,1}$	2.93%	leisure/83%
$C_{2,1}$	4.47%	portal/25%, service/17%, leisure/15%, search/12%	$C_{7,2}$	2.00%	search/37%, portal/20%, service/14%
$C_{2,2}$	4.20%	leisure/36%, portal/13%, search/13%	$C_{7,3}$	0.48%	video/83%
$C_{2,3}$	3.83%	search/33%, leisure/14%,portal/12%	$C_{8,1}$	1.48%	leisure/44%,search/16%,portal/12%
$C_{2,4}$	2.59%	video/29%, search/14%, leisure/14%,portal/13%, service/10%	$C_{8,2}$	1.29%	leisure/22%,portal/20%,search/20%
$C_{2,5}$	0.95%	games/25%, portal/15%, search/14%, leisure/13%, service/11%	$C_{8,3}$	1.13%	search/41%,leisure/18%,portal/11%
$C_{3,1}$	5.24%	leisure/43%, portal/17%, search/15%	$C_{8,4}$	0.68%	video/32%,leisure/19%,search/16%,portal/12%
$C_{3,2}$	3.94%	search/38%, leisure/20%, portal/16%	$C_{8,5}$	0.49%	service/32%,leisure/18%,search/17%,portal/11%
$C_{3,3}$	2.36%	video/43%, leisure/19%, portal/16%, search/15%	$C_{8,6}$	0.18%	economy/40%,leisure/15%,search/12%
$C_{3,4}$	2.03%	service/33%, leisure/20%, search/14%, portal/13%	$C_{8,7}$	0.12%	living/31%,leisure/21%,search/18%,service/10%
$C_{3,5}$	0.35%	economy/39%, leisure/16%, search/15%, portal/11%	$C_{9,1}$	1.40%	leisure/30%,search/25%,portal/16%
$C_{4,1}$	6.18%	leisure/63%, portal/12%	$C_{9,2}$	1.06%	leisure/61%,search/15%,portal/11%
$C_{4,2}$	2.45%	search/58%, leisure/12%	$C_{9,3}$	0.66%	search/57%,leisure/18%
$C_{4,3}$	1.82%	portal/60%, leisure/16%	$C_{9,4}$	0.26%	service/47%,leisure/20%,search/15%
$C_{4,4}$	1.24%	video/62%, portal/10%	$C_{9,5}$	0.20%	economy/72%
$C_{4,5}$	0.91%	service/59%, leisure/10%	$C_{9,6}$	0.10%	news/65%,search/10%,leisure/10%
$C_{4,6}$	0.22%	economy/61%, leisure/10%	$C_{9,7}$	0.02%	software/49%,leisure/22%,search/13%
$C_{5,1}$	4.79%	leisure/48%, search/22%, portal/13%	$C_{10,1}$	1.07%	leisure/83%
$C_{5,2}$	3.36%	portal/46%, leisure/25%, search/13%	$C_{10,2}$	0.48%	search/55%,service/32%
$C_{5,3}$	1.14%	video/49%, search/15%, leisure/13%, portal/10%	$C_{10,3}$	0.32%	economy/85%
$C_{6,1}$	3.24%	leisure/58%, search/14%, portal/13%	$C_{10,4}$	0.18%	e-commerce /26%,living/16%,portal/15%
$C_{6,2}$	2.25%	search/45%, leisure/23%, portal/12%	$C_{10,5}$	0.11%	news/87%

There are 50 clusters named $C_{m,n}$ listed in Table Ⅱ, where m presents the serial number of clusters in step one and n presents the serial number of clusters in step two. Main page categories, whose percentages of access times more or equal to 10%, are listed to observe user patterns easily and clearly.

D. Analysis two

The Maximum percentage of access times (MPAT) corresponds to the most preference page category in a cluster, and it is meaningful to find users who have significant preference.

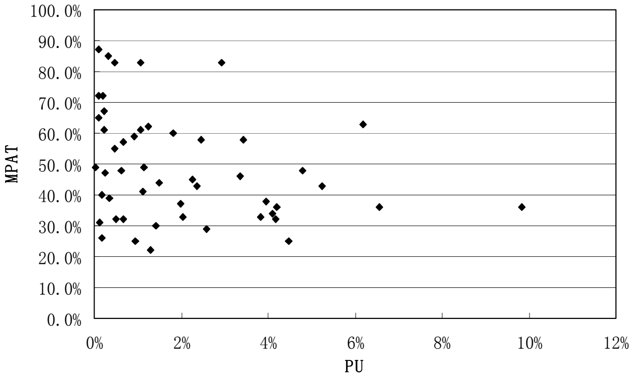

Figure 4. The distribution of MPAT over PU

In Figure 4, each point presents a cluster. The x axis presents the PU and y axis presents MPAT. From figure 4, we can see the values of MPAT distribute from 0.2 to 0.9 and demonstrate a large distinction among clusters. This indicates that there are notable diversities exit among the preferences of clusters. Moreover the clusters with higher MPAT distribute in the lower range of the PU shows that the clusters which have significant preferences are in smaller size.

TABLE III. CLUSTER GROUPS WITH DIFFERENT S

	S	Cluster number
Group1	0.2%	8
Group2	0.5%	16
Group3	2.0%	31
Group4	10%	50

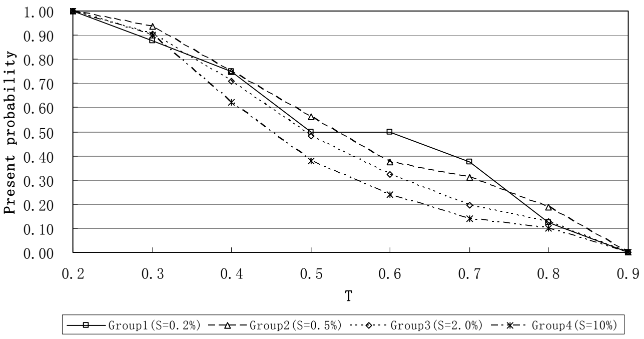

Figure 5. Present probabilities vary with T in the four groups

In order to examine the probability of eligible clusters varies with the PU and MPAT, we plot Figure 5 where S is defined as the threshold of the PU and T is defined as the threshold of MPAT. In TableⅢ we select 4 values of S in ascending order. Clusters are grouped when their PU less than S. The numbers of clusters in each group are list in TableⅢ. Then we examine the present probability of clusters whose MPAT are more than T in each group. From Figure 5, we can see that the present probability declines with the increase of T, and when T is equal to a certain value the present probability is higher in the group with smaller S. From the above analysis, we can conclude that: first, the more significant preference a cluster appears, the less present probability it has; second, clusters with significant preference are likely to have smaller size.

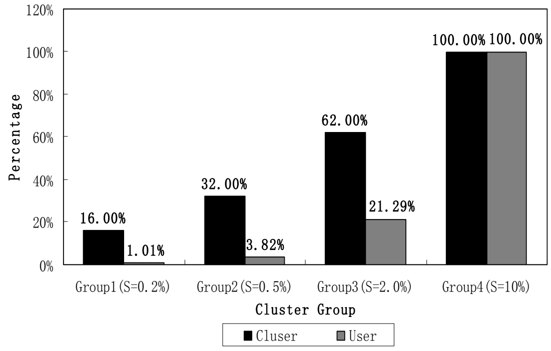

Figure 6. The percentages of clusters and users distribute in cluster groups

To compare the complexity of searching preference users in different cluster groups, we plot Figure 6 where x axis presents cluster groups and y axis presents the percentage of clusters and users contained in cluster groups. In Table 6, comparing with Group3 and Group4, Group 1 and Group 2 have much less clusters and users, which means finding preference users in smaller clusters may bring a notable improvement in complexity.

To sum up analysis on Figure4, Figure 5 and Figure 6, clusters having significant preference may be discovered accurately and efficiently by considering cluster size.

VI. CONCLUSION AND FUTURE WORK

In this paper, we use two-step K-means cluster algorithm to mine the web usage data from a WAN. By analyzing the result clusters, we get the following conclusions:

- The activity of web usage has a negative relationship with cluster size.
- Web users obtained information from web pages directionally; few users are interested in all aspects.
- Clusters with significant preference are in small number.
- Clusters with significant preference have a relative higher present probability in smaller size clusters.

The above conclusions are obtained on the point of whole network and are valuable to websites to understand entire user profiles. Also these conclusions give some enlighten in improving the targeting and recommending applications. For example, when designing targeting system using the cluster algorithm to discover user groups with special interest patterns, an appropriate threshold of cluster size can help to decline the candidate users and improve the targeting accurate especially in large data set. In the next step, we will study evolving usage patterns using temporal data and explore the migration of users in different patterns.

REFERENCES

[1] Natheer Khasawneh , and Chien-Chung Chan, “Web Usage Mining Using Rough Sets”, Annual Meeting of the North American Fuzzy Information Processing Society, 2005.

[2] Etminani, K., Delui, A.R., Yanehsari, N.R., and Rouhani, M., “Web Usage Mining: Discovery of the users’ navigational patterns using SOM”, Networked Digital Technologies, 2009.

[3] Yaxiu Yu, Xinwei Wang, “Web usage mining based on fuzzy clustering”, 2009 International Forum on Information Technology and Applications, 2009.

[4] Jianxi Zhang, Peiying Zhao, Lin Shang, and Lunsheng Wang, “Web usage mining based on fuzzy clustering in identifying target group”, Computing, Communication, Control, and Management, 2009.

[5] Daniel A.Menascé, Virgilio A.F. Almeida, Rodrigo Fonseca, and Marco A. Mendes, “A Methodology for Workload Characterization of E-commerce Sites”, Proceedings of the 1st ACM conference on electronic commerce table of contents, 1999.

[6] Marcelo Maia, Jussara Almeida and Virgílio Almeida, “Identifying User Behavior in Online Social Networks”, Proceedings of the 1st workshop on Social network systems, Apr. 2008.

[7] Kuai Xu, Zhi-Li Zhang, and Supratik Bhattacharyya, “Profiling Internet Backbone Traffic: Behavior Models and Applications”, Proceedings of the 2005 conference on applications, technologies, architectures, and protocols for computer communications, Aug. 2005.

[8] Olfa Nasraoui, Bamshad Mobasher, Brij Masand, Bing Liu. “WebKDD 2004: web mining and web usage analysis post-workshop report”, ACM SIGKDD Explorations Newsletter, Dec. 2004.

[9] Olfa Nasraoui, PMyra Spiliopoulou, PJaideep Srivastava, PBamshad Mobasher, PBrij Masand. “WebKDD 2006: web mining and web usage analysis post-workshop report”, ACM SIGKDD Explorations Newsletter, Dec. 2006.

[10] Rui Xu , Wunsch, D., II, “Survey of clustering algorithms”, IEEE Transactions on Neural Networks, Volume : 16 , Issue:3, May 2005.

[11] Robert Cooley, Pang-ning Tan, Jaideep Srivastava, “Discovery of interesting usage patterns from web data”, Proc. Workshop Web Usage Analysis and User Profiling, Aug. 1999.