Stability Analysis for Users' Web Preference

Bo Zhang, Wenli Zhou, Hao Yan, Yinan Dou, Jing Ma
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications, Beijing
zhangbo0207@gmail.com

***Abstract*—Advanced personalized e-applications require comprehensive knowledge about their user's preference in order to provide individual product recommendations and custom-tailored product offers. Much research has been conducted using Web logs to infer user preferences and predict users' behavior. However, little research measures the stability of user preferences over time. In this paper, the index named Web stability coefficient (*ST*) based on a comprehensive analysis of Web access records is proposed to represent the stability of user's preference. Then we used k-means clustering algorithm to assign users into different groups, which is computed based on the values of vector representation of user's *ST*. By analyzing user patterns, we present some interesting conclusions that facilitate us to better understand behavioral characteristics of Web user.**

Keywords-Web usage mining; stability of preference; clustering; k-means

I. INTRODUCTION

With the explosive growth of Web content, it has become increasingly difficult to let users to find their own interested information. Many systems were designed to retain on-line customers using pop-up window or other methods. However, the needs of different users are not the same, a category of target Web pages to some one may compels others onerously. We can make use of Web usage mining techniques to capture and model the behavioral patterns that can be used to create a personalized customized for users by providing dynamic recommendations.

One way to obtain user preference is to analyze characteristics of content used by each user. Authors in [1] introduce the three categories of Web mining and present a typical analysis process associated with electronic business applications. Some information systems identify frequently appearing terms as user interest keywords using the TF-IDF [2]. In [3], a novel approach of Web page classification using Naive Bayes classifier based on Independent Component Analysis is proposed. In particular, clustering algorithms are frequently used in Web mining [4] [5] [6]. The temporal properties of user preferences are seldom studied, in [7] a metric is proposed to measure the longitudinal changes of user preferences, and in [8], the authors investigate changes in users' preference of services at different time scales.

In the application process of data mining algorithms, feature extraction often plays a decisive role in the formation and interpretation of the results. Blind application of data mining algorithms on the data will ignore a lot of hidden features. Extracting feature of the raw data appropriately according to the analysis purpose, could not only simplify the analysis process and improve the data's applicability of the algorithm, but also reflect the will of data miner more directly in analysis results. In this paper, based on the understanding of overall statistical properties of the experimental data, we put forward the concept of the user's Web access preference stability coefficient (*ST*), with in the domain knowledge of experimental data. We analyzed the access sequences' stable characteristic of users' Web preference through this indicator and studied the characteristics of similarities and differences between different categories of Webs.

The rest of the paper is organized as follows: After a description of the collection of experimental data in Section II, we analyze the statistical characteristics of users' access sequences in Section III. In Section IV, we propose the *ST* index of users' Web preference. Section V present clustering algorithms for mining preference models and analyze the experimental results from different angles. We conclude our paper with a summary and outlook in Section VI.

II. DATA COLLECTION

The data collected by network traffic monitoring equipment arranged on the portal of a typical WAN owned by an ISP in China, which include 283618 ADSL users' Web access logs in two weeks in Feb. 2009. In order to investigate the stability of Web users, the users who logged in all the 14 days are selected and 272567 users' Web access data were retained.

We classified a large number of pages which hold most access times launched by users into 10 different content categories, including: Portal, Television, Shopping, Blog, Forum, News, Finance, Music, Games and Search. Then, Web access logs are transformed into average access times (AAT) of the 10 pages categories for each user in one day according to the classification. After the process of 14 days' logs, we got every user's Web access times record which is a 10-dimensional sequence.

III. DATA STATISTICS ANALYSIS

To grasp the data characteristics and the access trend, we firstly analysis the statistical properties of overall data set.

This work was supported by National Natural Science Foundation of China (61072061), 111 Project of China under Grant No. B08004 and Chinese Universities Scientific Fund (2009RC0123)

978-1-4244-6252-0/11/$26.00 ©2011 IEEE

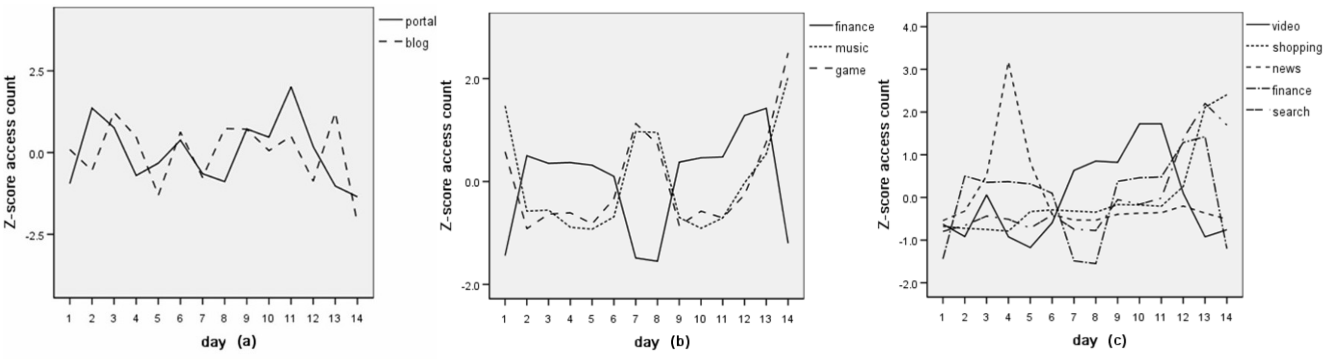

Figure 1. The distribution of total amount of AATs over date in 10 pages categories.

TABLE I. STATISTICAL FEATURES OF WEB PAGES CATEGORIES

Statistics	Web Pages Categories									
	Portal	*Video*	*Shopping*	*Blog*	*Forum*	*News*	*Finance*	*Music*	*Game*	*Search*
number of AU	20851	55892	49225	121724	162376	96530	58420	107083	70266	168447
average AD of AU	2.30	2.65	2.53	3.47	5.44	3.27	3.71	3.35	2.98	5.02
average access times of AU	13.8	46.9	115.4	173.2	354.4	128.1	1351.1	125.6	78.8	205.7

A. Access Sequence Trend

We begin by calculating the sum of access times of every category. Taking into account the difference of sequence baseline among Web pages categories, the z-score normalization transformation was applied to observe different curves' trends. Ten pages categories were divided into three groups intuitively according to their trends.

In fig.1, the horizontal axis shows the date, from a Sunday to the next Saturday. Fig. 1 (a) describes the first main class of Web pages. The two curves both fluctuate around baseline which means that the users' demands of Portal and Blog are relatively stable. In fig. 1 (b), the three types of Webs' trends all have periodic distribution with time. Specifically, the distribution curves of Finance and Game/Music have a negative correlation. During weekends, users' preference of Finance decrease significantly, while keeping high level during weekdays, which is exactly consistent with the domestic stock market period. On the contrary, users' demands of Music and Game show higher during weekends. The curves presented in fig. 1 (c) are relatively stable overall, but appear significant trends in some days. For example, the curve of News rises in the third day, and reaches the peak in the fourth day. There are special events, such as a sharp drop in the stock market, occurred in these days, but the specific factors which led the distribution of the curve needs further research.

B. Access Sequence Characteristic

In this section, we continue to analyze the properties of the access sequences based on the terms defined as following ：

- Active Users (AU): If a user visits a type of Web, then this user is an active user of this Web.
- Access Days (AD): its value represents the number of days in which the AU visits a type of Web.

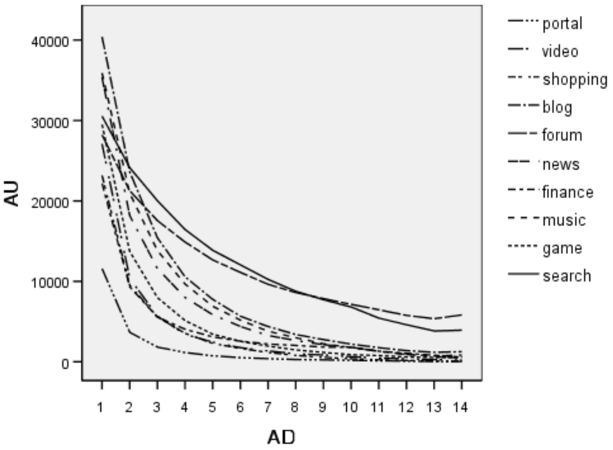

Figure 2. Distribution of AU with different AD.

We have calculated the value of AD to visit a certain type of Web for each AU within two weeks. In fig. 2, each curve represents a type of Web. The value of AU for various categories of Web pages vary greatly, but overall shows a similar trend. The number of users monotonically decreased with the increasing of AD. In Table I, the AD' averages of different Webs are between 2.30 and 5.44. This shows that for a certain category of Web, most of its AU do not need to visit it frequently. By further analysis combining with the raw data, we found that only a few users meet the trendy characteristics shown by the overall, and stochastic volatility is obvious. For most users, time series data of accessing different pages doesn't have determined movements, nor can be described with determined time function. Only probabilistic methods can reflect its changes.

Fig. 2 shows that the curves of different types of Web have some difference, indicating that the users' demands of different Web have different characteristics. Combining with the average AD of AU in Table II, it can be found that the two types of Web, Search and Forum, have most users with frequent demands. But the users of Portal and Shopping only visited the pages in less than three days on average within 14 days. The average number of access times of Finance is significantly higher than other types', which indicates that this category of users usually have very significant preference, resulting in a large number of access times.

IV. SST Representation of Time-Series Stability

Through the analysis above, we confirmed the randomness of Web access sequence and concluded that users seldom show stable or cyclical behavior in short-term. A common method for determining trend is to calculate a moving average of order n as the following sequence of arithmetic means [9]:

$$\frac{c_1+c_2+\cdots+c_T}{T},\frac{c_2+c_3+\cdots+c_{T+1}}{T},\frac{c_3+c_4+\cdots+c_{T+2}}{T},\cdots$$

To describe the changes in trend, the time series is replaced by its moving average sequence. The process of the smoothing of time series tends to eliminate unwanted fluctuations.

Take order as *T*, then the *j*-th moving average $\bar{d}_j = \sum_{i=j}^{T+j-1} C_i$. Before the calculation we must first determine the value of T, the general method is to choose the value that minimize the sequence *RSE* or according with experience. Since the process of moving average will lose the data in begin and end of sequence, considering that we have only 14 days' data, it'll be cause serious distortion if the order is too big. On the premise of reduce information loss, we choose 3 as T to achieve good smooth effect.

As for each user, the sequence means have large gap between different categories, we introduce coefficient of variation (CV) to describe the dispersion of the smoothed time series. Coefficient of variation is the ratio of standard deviation to means. It can eliminate the units and the average of different data on the comparison of variation degree.

$$CV=\sigma/\bar{d}=\frac{1}{\bar{d}}\sqrt{\frac{1}{D-T}\sum_{j=1}^{D-T+1}(\bar{d}_j-\bar{d})^2}$$

in which, the mean of sequence is defined as:

$$\bar{d}=\frac{1}{D-T+1}\sum_{j=1}^{D-T+1}\bar{d}_j=\frac{1}{D-T+1}\sum_{j=1}^{D-T+1}\left(\sum_{i=j}^{T+j-1}C_i\right)$$

and the standard deviation of sequence is defined as:

$$\sigma^2=\frac{1}{D-T}\sum_{j=1}^{D-T+1}(\bar{d}_j-\bar{d})^2\cdot$$

In order to facilitate the clustering analysis, we defined a stability coefficient *ST=1/CV*. For a certain type of Web, the higher the stability coefficient is, the more stable its users' demands are. It can be obtained from Table II that for each type of Web, there are a lot of users who do not have access logs, and their sequence means are 0. We set their *ST* as 0 which means that they own no preference for these Webs. In addition, there are still a small amount of users, whose have access logs but its σ is 0. For this part of the users, the stability coefficients are set to infinity. The *ST* is defined as:

$$ST=\begin{cases}0 & \bar{d}=0\\ 1/CV & other\\ \infty & \sigma=0\end{cases}$$

V. Clustering Analysis

A. Pattern Discovery From Web

In this section we will make use of the stability coefficients obtained earlier to assign users into different groups, and then mine the behavior patterns of the users. Since the K-means algorithm is simple and efficient, and has good scalability for large data sets, we choose this algorithm to cluster the users. Before the application of this algorithm, several existing problems should be solved first.

First, the algorithm needs to determine an initial division based on the initial cluster center, and the final result depends strongly on the initial value. In [10], the authors identified one of the distance optimization, namely KKZ method, as complements of the k-means learning characteristics towards a better cluster separation.

Second, since a small amount of these data would have a huge impact on the mean, it is sensitive to noise and outlier data. And we used the characteristic that K-means algorithm is sensitive to outliers to remove noise. First we selected a large value of *K* for the initial cluster, then removed the clusters contain very few users from the cluster results in.

Finally, before execution of the algorithm, we should determine an exact value of *K*, which required us to find the relatively appropriate number of the clusters. In [11], the index of β_{cv} is proposed to help us to determine the value of K. β_{cv} is the ratio between the intra-cluster *CV* and the inter-cluster *CV*. It represents the difference of the similarity in a class and the similarity between classes. The best indication for *K* would be when β_{cv} becomes relatively stable.

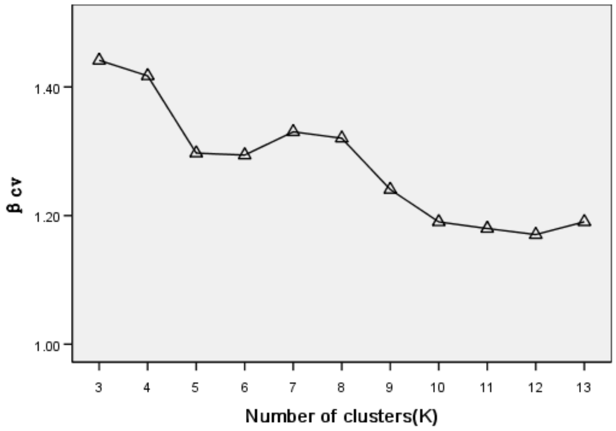

Figure 3. βcv for varying values of K.

Fig.3 shows the β_{cv} variation for executions of the *K*-means with different values of K. One can observe that β_{cv} falls fast when K changes from 3 to 0 with some fluctuation, and becomes relatively stable between 10 and 13. It indicates that we can choose 10 as a local optimum.

B. Result of clustering

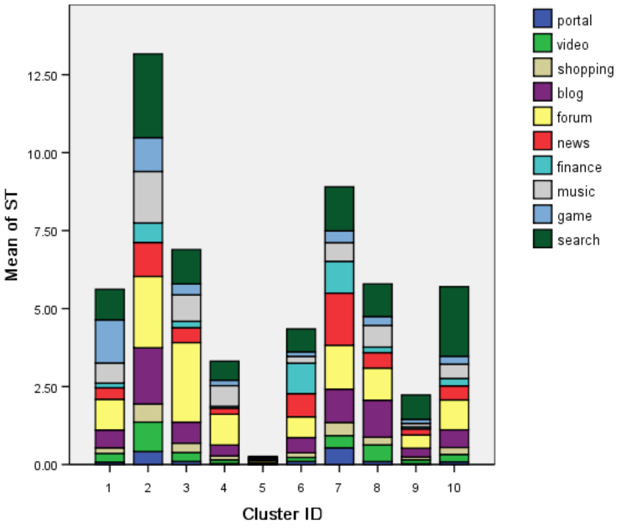

Figure 4. Ralative feature values for the 10 clusters created.

We classify all users into different models by clustering. There is the mean value distribution of each cluster in Fig.4. The mean *ST* of each feature in C5 is small. The reason is that there is a great deal of users with no access records in 14 days, which constitute C5 cluster. On the whole, Search and Forum are relatively more stable amongst clusters except C5, which means those two Web pages appeal to users generally and people use them regularly. *ST* of Music distributes evenly in every cluster except C5 and C6, however, with a relative low *ST* value, which shows that many users are demanding for Music, rather than taking it as a fixed interest.

According to the figure, we can find that the similar distribution happens in modes of C3, C4, C9, and C10. Apart from Search, Forum, and Music, there is no other steady interest in those users. Besides, all features in C2 have the highest *ST* value. Compared to other clusters, users in C2 are steadily interested in Blog, Music, Video and Game. There are lots of preferences to C2, which mainly consist of entertainment Web pages. C6 is similar to C7, having a steady demand on News, Finance and Blog, as well as slim difference on other operations. Additionally, users in C1 prefer Game Websites, while those in C8 prefer Blog.

C. Analysis Of Discovered Patterns

The *ST* value of every feature shows that the distribution of steady preferences of the operations in one cluster, while the sum of an operation *ST* values in all clusters shows the stability of Web demands for all users. We define the sum of an operation *ST* value as *SST*, and analysis the changing rules of *SST* along with the statistical characteristics of a cluster.

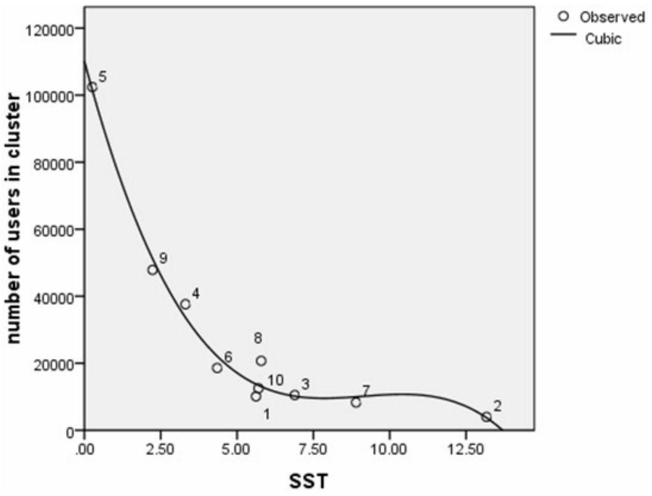

Figure 5. Number of users varies with SST in clusters.

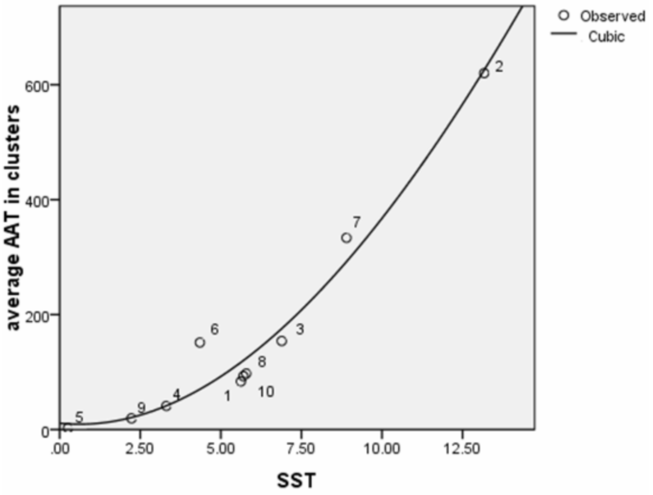

Figure 6. The distribution of average AAT over SST in clusters.

Each dot in Fig.5 and Fig.6 presents a cluster. Those two scatter graphs show the distribution of the number of users and average AAT going with *SST* value by fitting up three curve modes. As showed in the graph, in a cluster, the steadier Web accessing, the larger average AAT, nevertheless, and the fewer users it contains. A small amount of users require Web access largely and regularly, mainly concerning on entertainment part; meanwhile, users with rarely Web access times take up a great number in a cluster.

Most *SST* values fall into the range of [3, 7], where there are not many users but more behavior patterns. According to the Fig.4, users whose *SST* values are between 3 and 7 normally have 2-4 preferences for Websites. Also, the stability of a cluster is not related strongly to the number of users and average AAT. For instance, it has a large number of users in C8, while C6 has a larger average AAT.

VI. Conclusion and Future Work

To sum up, there are three change laws in time serial of all categories of Web pages' sum of AAT, and most regular users do not click any kind of Website frequently. We present the index of *ST* and research the stability of users' AAT serial based on this indicator. Through the cluster analysis we find that rare users prefer regularly such 10 Webs; and a certain number of clients, who have an explicit intent to acquire knowledge from the Webs, show up their steady demand on 2-4 of those with diverse acting modes. A few of users possess a relatively high access times and stability on Webs, and they pay more attention on entertainment contents.

The analysis in this paper is based on data obtained from Webs accessing logs in two weeks. We can study users` behavior on Webs in an extending time range in the future. Since *ST* index is referred to as the stability of access serial, it is considered to employ other index like user access rate or access frequency to go for further analysis in modeling process.

References

[1] Li Haigang,Yin wanling. Study of Application of Web Mining Techniques in E-Business. 1-4244-0451-7/06/$20.00 ©2006 IEEE

[2] Robertson, S.: Understanding Inverse Document Frequency: on theoretical arguments for IDF. Journal of documentation 60(5), 503-520 (2005)

[3] Zhongli He, Zhijing Liu. A Novel Approach to Naïve Bayes Web Page Automatic Classification. page:361-365 ISBN: 978-0-7695-3305-6.2008

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

[5] 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.

[6] Hao Yan, Bo Zhang, Yibo Zhang, Fang Liu, Zhenming Lei. Web usage mining based on WAN users' behaviors. Wireless Communications Networking and Mobile Computing (WiCOM), 2010.

[7] Yang Sun, Huajing Li, Isaac G. Councill, Wang-Chien Lee and C. Lee Giles. Measuring User Preference Changes in Digital Libraries. CIKM'08, October 26–30, 2008.

[8] Minjie Guo, Peng Liu, Fang Liu, Yuanyuan Qiao. Changes in Users' Prefernce of Services at Different Time Scales. Communication Technology (ICCT), 2010

[9] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques, Second Edition. ISBN 978-7-111-20538-8, 2006

[10] i He, Man Lan, Chew-Lim Tan, Sam-Yuan Sung, Hwee-Boon Low. Initialization of cluster refinement algorithms: a review and comparative study. 17 Jan. 2005

[11] 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