ISSN:0975-9646

Construction of Developing User Profiles for Mining Structure

Davuluri Suneetha, Kumar Vasantha
Department of CSE, Avanthi Institute of Engg & Tech, Tamaram, Visakhapatnam, A.P., India.

***Abstract*— Now-a-days business transactions are carried out over the web with E-Commerce (Electronic Commerce) activity is undergoing a significant revolution.CRM (Customer-Relationship Management) is the ability to track user's browsing behaviour, down to individual mouse clicks, has brought the vendor and end customer to very closer than ever. This paper provides a phenomenon "Mass Customization" for vendors, to personalize their product for individual customers at a massive scale. This phenomenon is one of the applications of web usage mining, which is the process of applying data mining techniques to the discovery of usage patterns from web data. Data mining techniques when associated with the web, called web mining can be broadly divided into three classes: i) Content mining ii) Usage mining iii) Structure mining .In this paper we present an up-to-date framework for web usage mining, which will help both profit organizations and non-profit organizations, to understand "who" the users were, "what" they looked at and "how" their interests changed with time.**

***Keywords*— Individual mouse clicks, User's behaviour, Usage patterns, Web usage mining.**

1. INTRODUCTION

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases, namely pre-processing, pattern discovery, and pattern analysis. Given its application potential, Web usage mining has seen a rapid increase in interest, from both the research and practice communities. Web usage mining has recently attracted attention as a viable framework for extracting useful access pattern information, such as user profiles, from massive amounts of Web log data for the purpose of Web site personalization and organization. These efforts have relied mainly on clustering or association rule discovery as the enabling data mining technologies. Typically, data mining has to be completely reapplied periodically and offline on newly generated Web server logs in order to keep the discovered knowledge up to date. In addition to difficulty to scale and adapt in the face of large data and continuously evolving Patterns, most clustering techniques, such as the majority of K-Means variants, also suffer from one or more of the following limitations:

1.Requirement of the specification of the correct number of clusters/profiles in advance, sensitivity to initialization,

2.Sensitivity to the presence of noise and outliers in the data, and Unsuitability for sparse data sets.

Hence, there is a crucial need for scalable, noise insensitive, initialization independent techniques that can continuously discover possibly evolving Web user profiles without any stoppages or reconfigurations.

2. SYNOPSIS OF WEB USAGE MINING

Recently, data mining techniques have been applied to extract usage patterns from Web log data [1], [2], [3], [4], [5]. This process, known as Web usage mining, is traditionally performed in several stages [1], [3] to achieve its goals:

1. Gathering of Web data such as activities/click streams recorded in Web server logs,
2. Pre-processing of Web data such as filtering crawlers requests, requests to graphics, and identifying unique sessions,
3. Analysis of Web data, also known as Web Usage Mining, to discover interesting usage patterns or profiles,
4.Understanding/assessment of the discovered profiles.

In this paper, we further added a fifth step after a repetitive application of steps 1-4 on multiple time periods, i.e.,

5.Tracking the progression of the discovered profiles.

2.1 Managing Profile Progression

Most previous research efforts in Web usage mining have worked with the assumption that the Web usage data is static. However, the dynamic aspects of Web usage have recently become important. This is because Web access patterns on a Web site are dynamic due not only to the dynamics of Web site content and structure but also to changes in the user's interests and, thus, their navigation patterns. Thus, it is desirable to study and discover Web usage patterns at a higher level, where such dynamic tendencies and temporal events can be distinguished.

3. MINING USER PROFILES

The framework for our Web Usage Mining is summarized with the following steps which starts with the integration and pre-processing of Web server logs and server content databases, includes data cleaning and sessionization, and then continues with the data mining/ pattern discovery via clustering.

The total procedure can be summarized as the following steps:

1. Pre-process Web log file to extract user sessions.
2. Cluster the user sessions by using Unsupervised Niche Clustering (UNC) .
3. Summarize session clusters/categories into user profiles.
4. Track current profiles against existing profiles.

3.1. CLUSTERING THE USER SESSIONS

ALGORITHM: Unsupervised Niche Clustering (UNC)

Input: Data Records

Output: Cluster Sessions

Procedure**:**

1. Randomly select an initial N candidate representative from input data
2. Update the distance for each record x relative to each cluster representative.
3. For i= 1 to N/2 do
   i) Select randomly a candidate parent Pi
   ii) Select randomly another candidate parent Pj
   iii) Obtain c1 and c2 by performing crossover and mutation between strings of Pi and Pj
   iv) Identify new clusters

a) First assign each child to closest parent
b) IF child's fitness > closest parent's fitness
THEN
Child replaces parent in the new cluster;
ELSE
Parent remains the same in the new population
4. Stop.

To cluster user sessions, we use UNC (Unsupervised Niche Clustering) that uses a Genetic Algorithm (GA) [12] to evolve a population of candidate solutions through generations of competition and reproduction. The main outline of the UNC algorithm is sketched in the following. The reason that we use UNC instead of other clustering algorithms is that unlike most other algorithms, UNC can handle noise in the data and automatically determines the number of clusters.

4. TRACKING USER PROFILES

Tracking different profile events across different time periods can generate a better understanding of the evolution of user access patterns. Note that both profiles and click streams are typically evolving. The comparison process determines which new profiles are compatible with the old profiles and which new profiles are incompatible with any previous profile. We can visualize the temporal dynamics of profiles by labeling the x-axis with the corresponding URL's. On the other hand, the y-axis is used to indicate the profile index: Finally, we generate a plot depicting the Web site user trend evolution. Moreover, this process is done offline and is only periodically done (not adding any burden on the data mining/clustering itself), since it is an offline analysis of the results of Web usage mining to help track the user profiles evolution in retrospect. The choice of the basic period length can be either arbitrary or based on the domain knowledge and intuition (like whether changes have been made to the Web site or whether new events related to the Web site domain may have occurred). In our experiments, we have chosen periods that varied from one week to one month. In general, if the periods are too small, then fewer changes will be detected, as opposed to longer periods. Thus, the right period length should be determined by trial and error. In our experiments we consider a web site which is having 20 different URL's and web log for every month. The results are shown in the following Table 1:

Table 1:URL's and their corresponding user profiles in different periods of time

URL NO	Jan 08	Feb 08	Mar 08	URL NO	Jan 08	Feb 08	Mar 08
1	3	20	25	11	5	5	5
2	1	3	5	12	18	9	10
3	12	2	5	13	24	25	25
4	4	15	15	14	5	5	5
5	8	2	2	15	8	6	15
6	1	1	1	16	17	20	25
7	4	6	4	17	30	30	30
8	9	5	5	18	15	1	8
9	10	20	30	19	2	9	10
10	20	30	40	20	8	35	35

The user profile analysis can be shown as fig 1.

From the figure 1 we can understood the comparison of user's usage pattern's at different time periods. From the above example we can make a decision that URL 1,9,10 is having more demand from users. So if the organization is profit based organization it is more profitable to manufacture the products related to URL 1, 9, 10.If the website is non-profit one we can make a decision that URL 1,9,10 is having more demand.

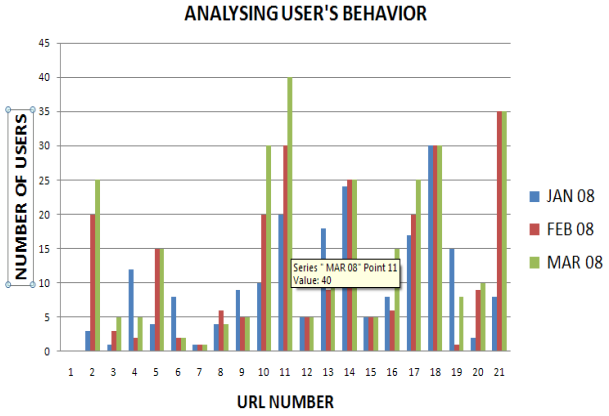

Fig 1: Analysing user profiles at different time periods

5. CONCLUSION

We proposed an approach that considers the Web usage data as a reflection of a dynamic environment which therefore requires dynamic learning of the access patterns. In this paper, we investigated a new evolutionary approach based on continuously and dynamically learning the Web access patterns from non-stationary Web usage environments. This evolutionary computation based approach can be generalized to fit the needs of mining dynamic data or huge data sets that do not fit in main memory. Preliminary experiments, performed on real click stream data, illustrated how an evolutionary algorithm can be used for mining dynamic click stream data to discover Web user profiles. Being capable to evolve with changing data is a crucial and key characteristic of an evolutionary data mining technique, if it is to be made truly scalable. More experimentation is being currently undertaken to test and extend the proposed approach with regard to scalability to large data sets, using it for clustering text documents/Web content, and incorporating Web content into Web user profiling.

REFERENCES:

[1] R. Cooley, B. Mobasher, and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web,"Proc. Ninth IEEE Int'l Conf. Tools with AI (ICTAI '97), pp. 558-567,1997.

[2] O. Nasraoui, R. Krishnapuram, and A. Joshi, "Mining Web Access Logs Using a Relational Clustering Algorithm Based on a Robust Estimator," Proc. Eighth Int'l World Wide Web Conf. (WWW '99), pp. 40-41, 1999.

[3] O. Nasraoui, R. Krishnapuram, H. Frigui, and A. Joshi, "Extracting Web User Profiles Using Relational Competitive Fuzzy Clustering," Int'l J. Artificial Intelligence Tools, vol. 9, no. 4,pp. 509-526, 2000.

[4] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," SIGKDD Explorations, vol. 1, no. 2, pp. 1-12, Jan. 2000.

[5] M. Spiliopoulou and L.C. Faulstich, "WUM: A Web Utilizatio Miner," Proc. First Int'l Workshop Web and Databases (WebDB '98),1998.

[6] O. Nasraoui, C. Cardona, C. Rojas, and F. Gonzalez, "Mining Evolving User Profiles in Noisy Web Clickstream Data with a Scalable Immune System Clustering Algorithm," Proc. Workshop Web Mining as a Premise to Effective and Intelligent Web Applications (WebKDD '03), pp. 71-81, Aug. 2003.

[7] P. Desikan and J. Srivastava, "Mining Temporally Evolving Graphs," Proc. Workshop Web Mining and Web Usage Analysis (WebKDD' 04), 2004.
[8] O. Nasraoui, C. Rojas, and C. Cardona, "A Framework for Mining Evolving Trends in Web Data Streams Using Dynamic Learning and Retrospective Validation," Computer Networks, special issue on Web dynamics, vol. 50, no. 14, Oct. 2006.
[9] M.A. Maloof and R.S. Michalski, "Learning Evolving Concepts Using Partial Memory Approach," Working Notes AAAI Fall Symp. Active Learning 1995, pp. 70-73, 1995.
[10] M.A. Maloof and R.S. Michalski, "Selecting Examples for Partial Memory Learning," Machine Learning, vol. 41, no. 11, pp. 27-52, 2000.
[11]J. Schlimmer and R. Granger, "Incremental Learning from Noisy Data," Machine Learning, vol. 1, no. 3, pp. 317-357, 1986.
[12] J.H. Holland, Adaptation in Natural and Artificial Systems. MIT Press, 1975.

Miss. Suneetha Davuluri received the B.Tech degree from the Department of Computer science and Engineering , NOVA College, JNTUniversity, Hyderabad, in 2006 and She is currently pursuing M.Tech in the Department Of Computer Science and Engineering, Avanthi Institute of Engineering and Technology, Vishakhapatnam, JNTUniversity. His research interests include Data Mining ,Web usage minig and Association Rules.

Mr. Kumar Vasantha received the M.Tech degree from the Department of Computer Science and Engineering, Avanthi Institute of Engineering and Technology, Vishakhapatnam, JNTUniversity, Kakinada in 2009 and working as a Asst. Prof in Avanthi Institute of Engineering and Technology, Vishakhapatnam His research interests include Information Security and Data Mining