A Novel Approach for Extracting Students Temporal and Periodic Internet Usage Behavior

Rozita Jamili Oskouei

Computer Science & Engineering Department

Motilal Nehru National Institute of Technology

Allahabad, Up, India

ABSTRACT

Understanding the students' temporal and periodic Internet usage behaviors are essential for administrators, instructors and course coordinators for evaluating Internet usage behaviors related to those students' undertaken courses or programs. In this investigation we propose a new Web usage mining approach for automatic extraction and visualization of periodic and temporal based Internet usage behaviors of individual users. Our model is able to extract each individual user's changes on periodic behaviors during different periods of a semester.

There is lot of research papers in education and Web usage mining exists. We believe our work is totally different because our concern is extracting individual users' behavior and we used Web site classification scheme for extracting each individual student's behavior's.

Keywords

Web usage mining, individual users, Internet usage behavior, periodic and temporal-based behavior, Website classification

1. INTRODUCTION

World Wide Web (WWW) contains a large set of data including different topics and contents. Every day different users visit different Websites regarding their requirements. Most colleges and institutes around the world provides Internet infrastructure for their students and teachers. An important research concern in Internet usage aspects has to answer the following questions:

- How the students in different colleges and under different programs navigate different Websites based on those Websites' contents?
- What are the preferred periodic navigation patterns of these students?
- Can periodic or temporal Internet usage behavior of students be recognized?
- What are the relationships between students on different programs and semesters and their Internet usage behaviors?

The first problem focuses on the Internet usage behavioral patterns of students related to specific Website categories based on contents of those Websites. The remaining problems capture the navigational periodic and temporal usage behaviors of students and relate them with Websites categorization and students' undertaken programs or semesters.

Web usage mining [1] aims at discovering interesting and frequent Internet usage behavior by extracting access patterns of users from log files. The discovered knowledge is useful for modeling users' Web access behavior and decision making.

We developed a matrix based our model for extracting each individual students periodic (during a semester, a week before examination, during examination) Internet usage behaviors. Our proposed model is able to distinguish each student's changes on behaviors and report them. This model can be helpful for understanding excellent and weak students Internet usage behavior and identifying all at- risk students and inform them before examination.

As far as we know, our investigation is the first study of periodic or temporal-based students' behaviors related to different categories of Websites [2] usages based on their undertaken program and branch.

The content of the paper is ordered as follows: Section 2 includes related works and section 3 presents data collection and pre-processing step. In section 4 we discuss our model of extracting each individual user's periodic and temporal based Internet usage behaviors. Section 5 presents our experimental results and section 6 contains conclusion.

2. Related Works

Many efforts related to extracting users' usage patterns in Internet are available for the purpose of clustering or classification[3 ~ 6] . But very little research about students Internet behaviors [7]. Authors on [8,9] proposed an approach to construct a user behavioral model from Web usage logs using the fuzzy formal concept analysis technique. This approach enables to costly personalized resources preparation process to be done in advance rather than in real time. In [10] authors presents a matrix model based approach for learning individual behavior models for Web users with using maximum entropy and markov mixture models for generating probabilistic behavior models.

This paper is different of all of those researches, because we made an approach for extracting each user's individual behavior related to his/her undertaken program or branch and finally we applied our own classification scheme based on academic related contents of visited Websites for capturing students periodic and temporal behaviors. We hope this approach will be helping us for better understanding students' behaviors and activates related to their academic performances.

3. Data Collection and Pre-Processing

We collected proxy server's access logs from Computer Center, Motilal Nehru National Institute of Technology, Allahabad for two years including four semesters. Students' information such as their undertaken program, branch, and semester etc, is taken from Dean Academic office.

I was permitted to use the pre-processing and filtered contents of the log files for the research purposes. The pre-processing hides the actual identity of students and replaces virtual_Id instead of real User_Id.

Data pre-processing in a Web usage mining model aims at reformatting the original access logs to identify all Web access sessions. The proxy server usually records all users' access activities of the Websites as access logs. Due to different proxy server parameters, there are many types of access logs, but typically the log files record some basic information such as user_Id, IP_Address, Time of connection, HTTP-Status code etc.

We developed a tool for analyzing Web access behaviors of individual users from proxy server log files. Since proxy server logs consist of 15 fields for each record (each user's interaction with Internet), we selected four fields including (IP-Address, User_Id, Time of connection, Website URL).

Several pre-processing tasks need to be done before performing Web mining on the access logs. For our work, these include data cleaning, user differentiation and session identification.

4. Our Proposed Algorithm for Extracting Individual's Usage Behavior

In this section, we examine the problem of identifying frequent navigational behaviors of students related to Website categories. For example, it may be interesting to know students having different behaviors in Internet based on their undertaken program or branch or semester. Identifying these differentiations can be helpful for academic affairs and teachers for taking better decisions for improving their quality of knowledge by looking at different groups of students' behaviors on visiting different categories of Websites. Administrators can identify program requirements of students under different programs based on most visited Websites and Internet behaviors.

Here data mining problem is to determine all navigational behaviors, related to contents of Websites, which are frequently exhibited by students under different programs from various colleges. Given the proxy server log files, we have populated users sessions and their used Websites' within that session are categorized. The next step is to capture the students' navigational behaviors. The extracted behaviors may include the following instances:

- Average number of sessions per day
- Most visited Websites
- Sequence of visited Websites during the semester
- Average time spent on each Website
- Average number of pages visited for each session per day

Our proposed algorithm is able to capture various behaviors of students in different periods during a semester.

All the users' access logs are pre-processed into independent user sessions. Each user session details includes visited Website, total time spent on that Website in each session, number of hits (visited or opened pages), category of that Website in terms of content and regarding our Website categorization [2].

By examining each user's sessions against the category of visited Website, it is possible to extract the category of each visited Website by each user and the behavioral navigational patterns of the users during each and every session of a day or for a period.

In our proposed approach, for each user we build a matrix that shows different sessions per day and the visited Websites in the following order:

- Each row represents different sessions per day
- Each column represents the visited Websites in different sessions

Each row is symbolic as S and each column as W, therefore we have:

$$\begin{cases} 1 & \text{if Website j visited by user i} \\ 0 & \text{if Website j not visited by user i} \end{cases}$$

for each user we have matrix similar to the following matrix

$$\text{User (i)} = \begin{array}{c|ccccccc} & W(1) & W(2) & W(3) & \ldots\ldots & \ldots\ldots & & Wm \\ \hline S(1) & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ S(2) & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ \ldots. & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ \ldots. & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ \ldots. & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ \ldots. & 1 & 0 & 1 & 1 & 1 & 0 & 0 \end{array}$$

In this matrix rows S1, S2, …, Sn shows different session during a day for each user, we used session timeout of 20 minutes, this means whenever a user do not have any hits for a 20 minutes duration program automatically creates a new session and the page is to be considered for new session. Each column W1, W2, ..., Wm indicates Websites visited by user. Value 1 indicates that the user visited that Website in that session and 0 otherwise.

Separate matrices are building for each user for different sessions during a semester.

Support for each Website with respect to number of sessions is computed as:

$$\text{Support (SjWi)} = \frac{\sum W_i}{\sum S_n}$$

$\sum W_{(1,..,m)}$ is the number of times that Website W_i visited by user in session S_j and $\sum S_n$ is the number of total sessions per a day or per specific period.

So our second matrix will be formed as following:

$$\text{User (i)} = \begin{array}{c|cccccc} & W(1) & W(2) & --- & -- \; -- & ---- & W(m) \\ \hline D(1) & 0 & 8/10 & 0 & 9/10 & 10/10 & 0 \\ D(2) & 3/8 & 7/8 & 0 & 8/8 & 7/8 & 0 \\ D(3) & 0 & 9/10 & 0 & 10/10 & 9/10 & 0 \\ .. & 2/10 & 7/10 & 0 & 9/10 & 10/10 & 0 \\ ;;; & 1/20 & 16/20 & 0 & 20/20 & 19/20 & 0 \\ ;;; & 0 & 7/8 & 0 & 6/8 & 8/8 & 0 \\ ;; & & & & & & \end{array}$$

In this matrix each row represents different days from day 1 to our specified period (day (n)) and each column represent different visited Websites. And the numbers displayed in cells are support for each Website during each day. For example, support count of D (1, 2) is 11/15, shows user (i) in day (1) visited Website (2) with the support amount 11/15. That is, the user had 15 sessions during D1 and within that 15 sessions s/he visited Website (2) in 11 sessions.

Third matrix contains aggregation of results during a period of days. During that period the support of each visited Website are calculated.

$$\begin{array}{cc} & \begin{array}{cccccc} W(1) & W(2) & .. & .. & .. & W(m) \end{array} \\ \text{Support} & \begin{bmatrix} 12\% & 78\% & 2\% & 93\% & 95\% & 3\% \end{bmatrix} \end{array}$$

After computing support (W_i), those Websites that are having support less than the threshold value are eliminated.

The threshold value used is 65%. Those Websites having support equal to or more than the threshold are selected as periodic visited Websites and remaining are considered as temporal-based visited Websites.

Finally Web classification is applied for the selected Websites. For instance, academic related Websites, social network and entertainment Websites usage behaviors by students and study those Websites usage affects on students academic behaviors and performances are identified.

In data mining with different purposes it is possible to make a various analysis and get different results. For example, in our proposed approach by applying average time spent by each individual user or average number of hits (number of visited pages of each Website) can deduce different results of individual students Internet usage behaviors and relate these results to category of Websites and get a very important points for extracting behaviors and relationships with academic activities and performances.

5. Experimental Results

Our investigation concerning analyzing periodic behavior of individual students in terms of average number of sessions connected to Internet per day and number of unique Websites visited per day and average time spent per day with respect to those users (students) undertaken programs and branches'.

Because different programs have different courses and different professors, our main goal is to detect special cases in behaviors of users and relate them with their branch and program, and predict their future behaviors for helping and guiding them for improving their perform.

Proxy server's access log files of Motilal Nehru National Institute of Technology Allahabad are collected for two years including four final examinations and midterm examinations data.

Our analysis is based on female students on different programs. For example, analysis includes students Internet usage behavior from different programs such as P.hd, M.tech, B.tech.

The queries to be addressed are:

- Can we define special Internet behaviors for Students in different periods (periodic-based behaviors distinguishing)?
- Is it possible to differentiate Internet behavior of students related to their undertaken programs or branches or semesters?
- Can we predict future (new) students' behaviors by applying periodic-based analyzed results?

We applied our proposed matrix model on all log files during 4 semester in different periods including normal days of a semester, the weeks before examination and during examination weeks, and displayed interesting changes on students behaviors in different periods and relationships of these changes with their undertaken semesters and programs and branches of study.

Table 1, shows average, minimum and maximum number of sessions and time spent and number of visited Websites in different periods of a semester.

Table 1: Session & Time Spent & Total visited Websites in Different Periods

	Sessions(per a day)			**Time Spent(Minutes)**			**Number of Visited Websites**		
	Average	Minimum	Maximum	Average	Minimum	Maximum	Average	Minimum	Maximum
During Semester	6	1	39	300	1	900	68	1	440
A Weak Before Examination	4	1	28	89	1	330	49	1	230
Examination Weeks	9	1	58	150	1	400	72	1	510

Table 1 shows maximum number of sessions per day belongs to examination weeks and minimum number of sessions per day belongs to the day or a week before final examination.

Other results of Table 1, is related to time spent on Internet, average time spent on Internet during a semester is maximum and during a week before examination is minimum, number of unique Websites visited by users per a day. Table 1 shows the maximum visited Websites are in examination weeks.

From Table 1, during examination weeks average time spent on Internet is decreased but number of sessions and unique visited Websites is increased. This seems students visiting more number of unique Websites during exam weeks and the average time spent is less compared to normal weeks.

Results of applying our proposed approach for extracting percentage of temporal and periodic or frequently visited Websites are shown in Table 2.

Table 2, shows maximum percentage of periodic Websites usage belongs to a week before examination weeks.

Table 2: Temporal and Periodic Percentage of Internet Usage Behavior

	Periodic Websites%	Temporal Websites%
During Semester	14%	86%
A Weak Before Examination	40%	60%
Examination Weeks	30%	70%

Finally we analyzed the results based on category of visited Websites. Results shows that during examination weeks and a week before examination most users are visiting news, email, Academic Websites related to their examinations or online games. So the varieties of visited Websites during examination are not similar so percentage of periodic Websites visiting is increased.

In other step, the students' behaviors related to their undertaken program and semester analyzed (Figure 1)

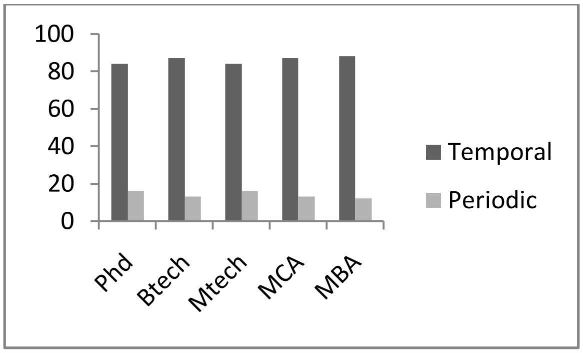

Figure 1: Different Programs and Percentage of Periodic Visits

In Figure 1, the horizontal axis displayed the number of hits (number of opened page of different Websites) and vertical axis displayed different programs. This is clear that BTech students having more percentage of periodic visited Websites (frequently visited Websites).

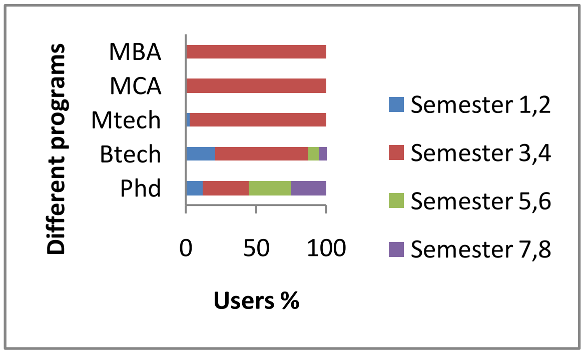

Figure 2: Relationship between Semester and Periodic Usage Percentage

In Figure 2, horizontal axis displayed the periodic usage behaviors percentage and vertical axis presented different programs. This is cleared that for students' undertaken different programs maximum periodic behaviors belongs to second year's students.

This means periodic or temporal usage behavior from Internet related to students undertaken program and also semester.

6. Conclusions

In this investigation we proposed a matrix based method for extracting temporal and periodic usage pattern of students. We believe that students every day spent time in Internet, as an course coordinators and instructors and administrators for understanding students Internet usage behaviors related to their semester and branch or programs.

Our Proposed is based on extracting students' behaviors on Internet such as number of connections (sessions) per day and number of visited unique Websites per day and recognizing the percentage of periodic (frequent) Internet behaviors' of students on different periods of a semester. Our proposed approach is able to inform the changes happening for individual users behaviors.

We plan for extending this approach and using other factors such as time spent and number of hits for future work. For classifying all students based on their behaviors on Internet and use classification for improving our Internet services in computer center of server side facilities and also for course coordinators for taking right decisions based on students behaviors on each program or branches.

ACKNOWLEDGMENTS

I am thankful to my PhD thesis supervisor, Professor B.D.Chaudhary for helping me in preparing this paper. I also acknowledge the help rendered to me by the staff of computer center and Dean (Academic Affairs) office.

REFERENCES

[1] J.Srivastava, R.Cooley, M.Deshpande and P.N.Tan "Web Usage Mining : Discover and Application of Usage Patterns from Web Data" , ACM SIGKDD Explorationss, Vol 1, No.2, 2000, pp 12-23

[2] Rozita Jamili Oskouei, B.D. Chaudhary, "Internet Usage Pattern by Female Students: A Case Study," itng, IEEE 2010 Seventh International Conference on Information Technology, 2010 pp.1247-1250,

[3]J.Xiao,Y.Zhang "Clustring of Web Using Session-based Similarity Measures" IEEE 2001

[4] M.Jalali, N.Mustapha, A.Mamat, N.B.Sulmiman "A New Classification Model for Online Predicting Users' Future Movments" IEEE 2008

[5] Y.Q.Peng, G.X.Xiaq, T.Lin "Prediction of Users' Behavioral Based on Matrix Clustring" fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 , Aguest 2006

[6] S.P.Nina, M.Rahaman, K.I.Bhuiyan, K.E.U.Ahmed, "Pattern Discovery of Web Usage Mining" IEEE 2009

[7] A.Hayashi "Two Classification Methods of Individuals for Education Data and an Application"

[8] B.Zhau, S.C.Hui, A.C.M.Fong "An Effective Approach for Periodic Web Personalization" IEEE 2006

[9] B.Zhau, S.C.Hui, A.C.M.Fong "Discovering and Visulaization Temporal-based Web access Behavior" IEEE 2005

[10] E.Manavoglu, D.Pavlov, C.Lee.Giles "Probablitistic User Behavior Models" IEEE 2003