Differential Internet Behavior's of Students from Gender Groups

Rozita Jamili Oskouei
Computer Science & Engineering Department
Motilal Nehru National Institute of Technology
Allahabad, UP, India
rcs0702@mnnit.ac.in, rosejamili2009@gmail.com

ABSTRACT

World Wide Web (WWW) contains huge amount documents with various types of categories. Internet usage behaviors of users in Internet are one of the main research areas. Many Website classification schemes were present in literature based on different perspectives. For categorizing Websites based on its content, we used a classification scheme [1]. This investigation is the application of Web Usage Mining (WUM) and Education Data Mining (EDM) to differentiate students Internet usage behaviors based on their gender group.

Keywords

WUM (Web Usage Mining), EDM (Education Data Mining), Internet Usage Behavior, Gender Group

1. INTRODUCTION

World Wide Web (WWW) is an important source for information in various subjects. Different usage of Internet based on each person's requirements is necessary, because on an average most of the people are connecting to the Internet from a few minutes to hours each day.

Most of the educational institutions are providing Internet facility for their students and staff members. One important aspect according to the usage is to find users behaviors on Internet and the effects on their academic performances and qualities.

Our research concern is to answer the queries about relationships between Internet usages and students' performances from different aspects by analyzing different categories of Websites usage based its contents such as academic related or social network or entertainment related Websites etc.

We believe by capturing and analyzing Websites based on their contents and a Website classification scheme, we can find the positive of negative effects of these Websites on their academic performances.

We analyzed different aspects of Internet usage behaviors of male and female students of engineering in India by using Web access log files.

Understanding different Internet behaviors of each gender group can be useful for the instructors and course coordinators with in a college to better understand each gender's requirements based on different conditions and periods and managing schedule or Curriculum based on each group's favorite behaviors for better performance of students.

As far as we know, our research concern is the first attempt in terms of applying Web mining techniques on students' access logs for identifying their behaviors related to different categories of Websites [1] and analyzing different aspects of each category of websites usage on their different activities and performance based on their undertaken program and branch. Our attempt tries to answer the questions about the effects of Internet usages without any limitation for access to different available Websites in education environment on students' performance.

Our paper consists of seven sections, second section presents related work. Section 3 contains definitions used. Section 4 presents data collection and pre-processing step. Section 5 contains methods used for analysis. In section 6 we conclude our discussions.

2. Related work

Several research efforts [2, 5] have been made for data pre-processing step in Web mining, which involves identifying specific fields such as user identification, session identification etc. There is a large literature [6 ~ 10] on Web usage mining and application of data mining techniques for pattern discovery. In recent years, many researchers attempted behavior mining of users [11 ~ 15] and predicting users future behaviors with the help of analyzing past behaviors [16, 17, 18]. In [19~23] authors applied the data mining techniques in educational area and students data. Some other authors [24, 25, 26] focused on gender based usage of Internet. These papers attempted to extract the general effects of Internet usage on students' performances. Our research is different from all those researches. Our focus is on discovering students Internet usage behaviors based on analyzing their previous access logs and relating their usage patterns with their CPI (27,28,29) and proposing a methods for identifying students at risk (the students whose might fail) before final examination and inform them. Other focus is on students' usage behaviors related to different category of Websites and analyzing different category of Websites usage behaviors and the performance effects of non-academic websites usage [30, 31].

This investigation analyzes gender based Internet usage behaviors of students from different aspects.

3. Definitions

Data Mining (DM) is the process of extracting interesting patterns from data [32].

Web Mining (WM) can be broadly defined as the discovery and analysis of useful information from WWW with the help of DM techniques [33].

WM can be divided into three different branches [34]:

- Web Content Mining (WCM): describes the discovery of useful information from the Web contents or data or documents.
- Web Structure Mining (WSM): is the process of discovering knowledge from Web pages and the links.
- Web Usage Mining (WUM): is the process of extracting interesting patterns or knowledge from various web access log records.

Web log records are analyzed during the process of WUM and various rules and patterns are explored. WUM consists of three phases:

- Data pre-processing: The main purpose is to extract useful data from raw Web logs.
- Pattern discovery : find out various rules and patterns by taking advantage of data mining techniques
- Pattern analysis: filters out the *useless* rules discovered in the period of pattern discovery and then extracts the *interesting* rules and patterns

WUM includes the data from the Web server access log, proxy server access logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls and any other data as the results of information [35].

4. Data collection and pre-processing

For our analysis, we collected proxy server access log files from Motilal Nehru National Institute of Technology, Allahabad (India) for a period of two year including four academic semesters. This period includes four final examination weeks' data in addition to four mid-term test exam weeks.

I was permitted to use the pre-processed and filtered contents of the log files for research purpose. The pre-processing step hides the actual identity of students and replaces *virtual_id* in place of real *user_id*.

Computer center provided all users information including User*_id, Full Name, and Department* in a text file.

Other data including students' academic information were collected by Dean (Academic Affairs) office for two semesters of an academic year, separately, as excel files. Fields included in a record are:

Registration No, Full Name, Program, Branch, Semester, Gender, CPI, Email, etc.

After collecting data, we pre-processed and saved necessary data for our analysis. Each access log file contains 15 fields per record. We selected 3 fields for our analysis including:

User_ID, Visited_Website's_URL, Time-of-Connection

After pre-processing step, we have a table consist of following fields per each user:

[Virtual_Id, Program, Semester, Department, No. of sessions per a day with time spent on each session and respective Website category, Total number of sessions per a day, Average_Time_Spent (per a day), Average_Hits_No (per a day), Number of Unique Visited Websites (per a day), CPI (Cumulative Performance Index), Gender].

We have developed and implemented algorithms to compute the number of sessions that individual user connected to Internet per a day in different periods of a semester along with total time spent in each session on different categories of Websites and total time spent per day with the time stamp of the access log. We used minute as a unit of time duration to measure the time spent by users on a Website during a day. This unit of measurement implies that all durations less than a minute has been rounded to one minute. Our algorithm computes the average total time spent by each user per day and also average time spent by him/her on different category of visited Websites.

A session refers to the unit of interaction between a user and a web server. It consists of pages accessed by a user in a certain amount of time. If the time between page requests exceeds a certain limit (our threshold was 20 minutes), it is assumed that there is another user-session. Users visit different Websites from time to time and spend arbitrary amount of time between consecutive visits. These users may have more than one session.

For computing the total number of hits (number of visited pages) per day we computed all opened Web pages by a user with respect to different categories of websites.

We used ***Rapid Miner,*** an Open Source Software for statistical analysis, for analyzing all statistics related data. All data are transferred in excel files and all various inconsistencies are removed. These inconsistencies include missing visited URL's, missing user id, etc.

Our analysis is based on the collected access logs and the questionnaire that is conducted on all students of four different semesters.

5. Methods for Analyzing Data

In first step, we used Website classification scheme [1] to categorize the Websites visited by students based on its contents. This classification is based on academic (AC) or non-academic (NAC) related Websites, shown in Figure 1, and is more useful than other classification schemes [36] & [37].

AC Websites includes university or institution, e-learning, eBooks (free), journals and conferences related, professional such as science related, free or non-free downloading software or tools and non-free eBooks Websites etc.

NAC Websites are divided into:

- Social Networks (SN) Websites that includes:
  - Blog
  - Advanced SN: includes all Websites that designed for the purpose of social communication goals such as orkut.
  - Special SN Websites: includes websites for sharing files or documents with others, download/upload, special community of users such as music community etc.

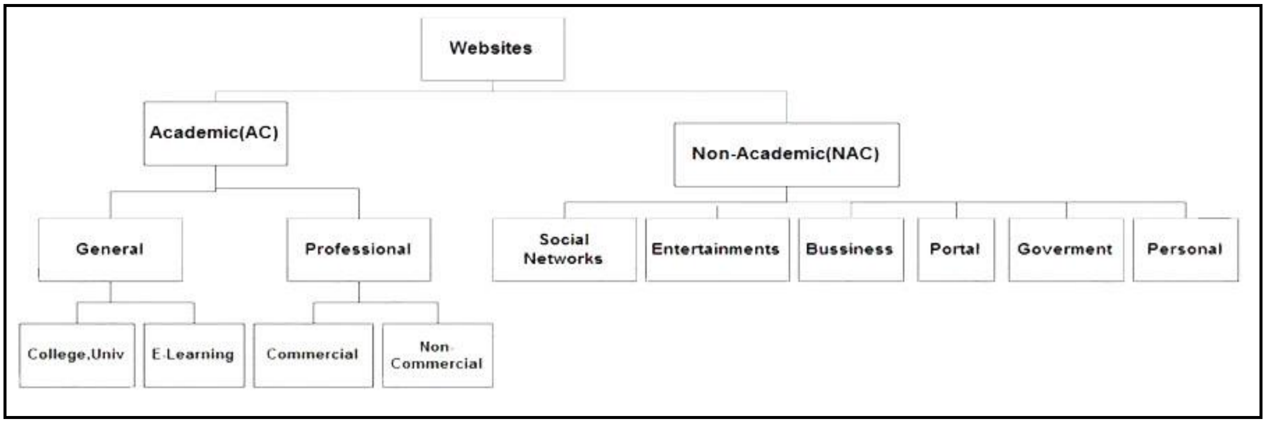

Figure 1: Websites Classification [1]

- Entertainment Websites (EN): including
  - Undesirable Websites: contains all adult or drug etc.
  - Media: includes all TV, Radio, Movie, Video, Hollywood actors and photos etc.
  - Online Game
  - News
  - Greetings cards
  - Sports
  - Fun
  - Emotion Websites
  - etc
- Business
  - Undesirable related Websites (not free)
  - Job Agency
  - Marriage Websites such as shadi.com
  - Travel Agency
  - E-trading
  - E-Shopping
  - Advertisements Websites
  - etc
- Portals such as Google or Yahoo
- Government Websites such as embassy or passport office
- Personal

In second step, the access log files are pre-processed and selected all necessary data for our analysis.

One main aspect of this research concern is to identify the average number of session (each user connecting to Internet per day), average time spent (on Internet per a day), most visited websites (names and category). We then compared these results in terms of different periods of a semester such as examination weeks, a day before final examination, during a semester separately for holidays and working days.

Female students are in minority in most of the engineering colleges in all around the world. This paper attempts to compare all Internet behaviors of female students of engineering with male students' usage behaviors of finding the differences on their usage.

For defining each Website's category, we analyzed 4000 Websites manually by visiting their homepage and notifying their important special keywords. The remaining Websites are categorized automatically based on those special keywords.

6. Experimental Results

Overall female students' population in our college was 4300 which is approximately 14% of all enrolled students and remaining were male students. All students are provided a user ID for accessing Internet and E- mail.

Total students gender based populations of different programs are presented in Figure2.

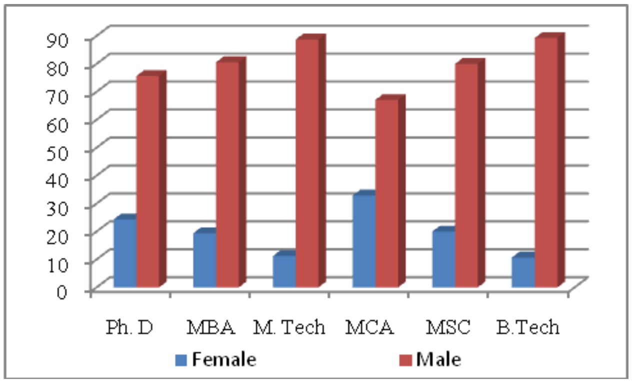

Figure 2: Female / Male Students % in Each Program

The horizontal axis in Figure 2 represents different programs in our college and vertical axis represents percentage of students in each program.

From our analysis on 2 years data, we observed that 11% of continues users (users having at least one session per day) belongs to female and the remaining 89% were male students.

From our analysis, we also observed that the numbers of unique Websites visited are approximately 2000. Figure 3 shows the unique Websites visited by female and male per day.

Boxplots are useful for computing distributions of multiple attributes or the same attribute for different groups. *Boxplot* displays the median, the quartiles, the range of values covered by the data and any outliers which may be present. The *box* is a rectangle with edges defined by the lower and upper quartiles; so it indicates where the 'middle 50%' of the data can be found. The vertical line inside the box is located at the median.

Figure 3 shows the upper and lower quartiles and also median of unique visited Websites by each female and male students.

The vertical axis shows number of unique visited Websites by users per a day, horizontal axis declared usage is belongs to female or male.

As a result of Figure 3, it can be observed that on an average, the number of male users that having number of visited Websites are exceptional than female users. This seems the relative median for female is more than male users. The reason for this situation is during examination days number of unique visited Websites by female students per a day is increased.

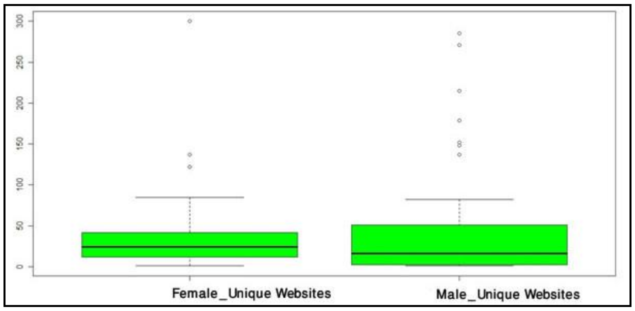

Figure 3 : Number of Unique Websites visited by each per a Day

Other aspect of our analysis is comparing average time spent by students based on their gender. Figure 4 shows female and male students time spent per day.

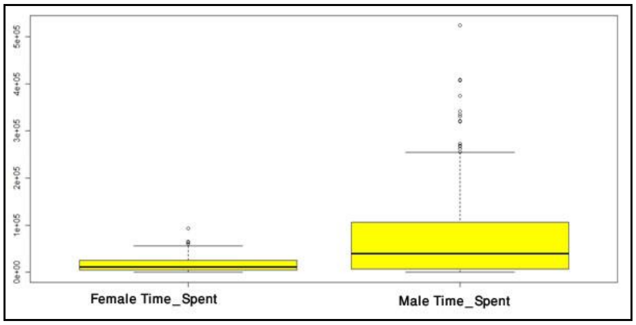

Figure 4: Average Time Spent per a Day

Figure 4 shows that maximum time spent on Internet belongs to male and female users' median (average) time spent is less than male students.

From Figure 3 and Figure 4, we can see that the female users are spending less than male during a day on Internet and the average number of unique visited Websites are equal to or greater than male users. The speed of visiting a page by female is more than male students.

Figure 5 shows the relationship between CPI and average number of visited websites per a day by users.

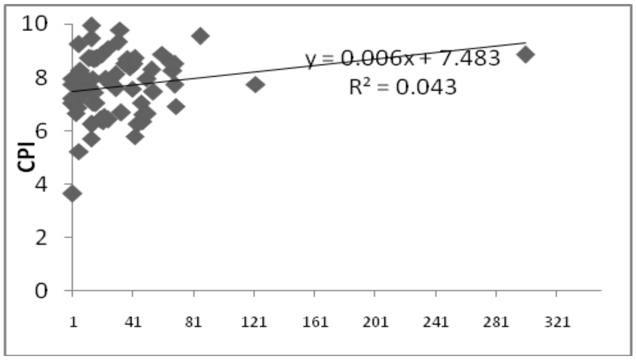

Figure 5: Average Number of Unique visited websites Vs. CPI

In Figure 5, horizontal axis shows Number of unique Websites visited by users and vertical axis shows CPI a scale of 0-10. Figure 5 shows that maximum number of unique websites are visited by students whose CPI was (6.5<CPI<8.5).

Gender based comparison of CPI and average time spent per a day by female and male students are shown in Figure 6.

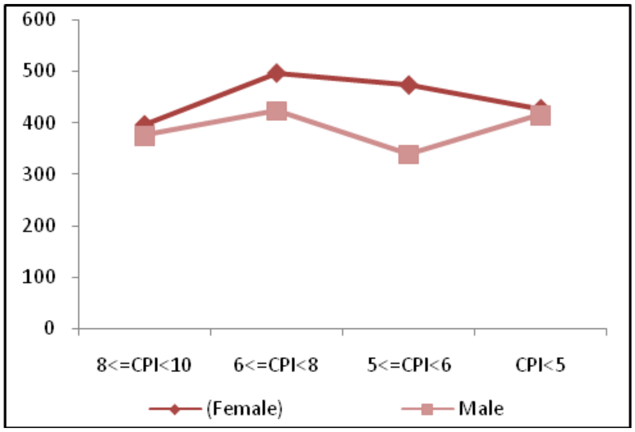

Figure 6: Female & Male Time Spent Vs. CPI

Figure 6, horizontal axis shows CPI and vertical axis shows average time spent on Internet per a day. As a result of Figure 6, it can be observed that both male and female whose are having CPI between 6 and8 had the highest time spent on Internet and those having CPI<5 spent less time on Internet.

One important point related to female and male is related to their different behaviors during examination weeks. For deducing these differences, we compared the average time spent during examination weeks by female and male students.

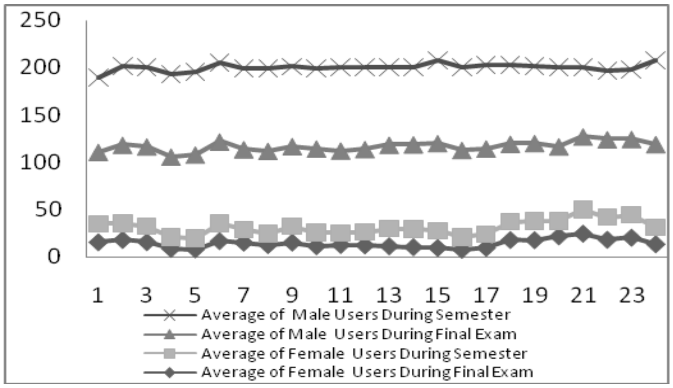

Figure7: Female & Male Users during Mid/Final Exam

In Figure 7, the horizontal axis shows the 24 hours in a day and vertical axis shows the average time spent per a day by students. This figure shows the average time spent by female and male students during final examination weeks and during semester. This figure shows that the average time spent on Internet by male students in both periods were more than that of female students and for both the average time spent is decreased during final examination weeks.

One questionnaire is made for collecting students observations related to our research. In this questionnaire one question was related to students' semester and the average time that they spent every day on Internet. The results of the students' answers and log files analysis shows that, students on different semesters having different behaviors in terms on average time spent on Internet per day (shown in Figure 8).

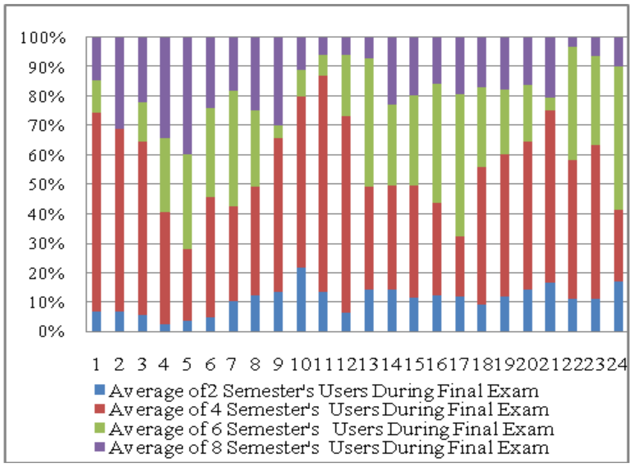

Figure 8: Different Semester's Users

In Figure 8, horizontal axis shows 24 hours of a day and vertical axis shows the percentage of users with their semester's in each hour of a day during final examination week. Figure 8 shows that minimum time spent users in every hour of a day belongs to 2nd semester and maximum time spent belongs to 4th semester. Figure 8 also shows that the maximum and minimum usage percentages with respect to each hour. For example, 2nd semester students' maximum usage is at 10am.

One more important and interesting aspects of Internet behaviors of students are related to AC and NAC websites' usage and effects of these behaviors on students' academic behaviors. Figure 9 shows the students usage of AC and NAC websites per a day during 24 hour. Figure 8 compares different days of a semester's AC Websites usages. The results of shows that maximum AC usages are at 1Pm during final examination weeks and maximum usage are at 4Pm on a day before final examination weeks.

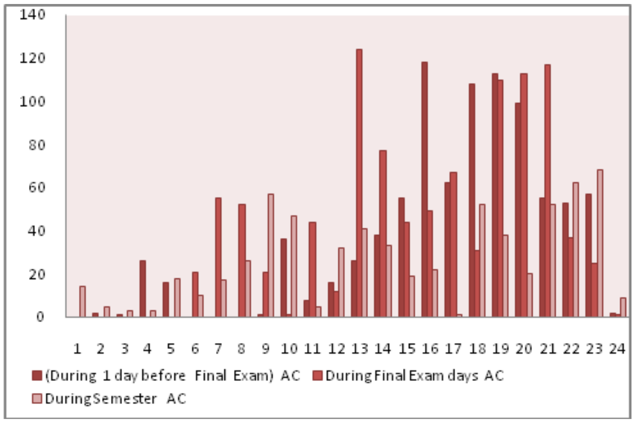

Figure9: Comparing No. Of Unique AC Visited Websites

Horizontal axis shows 24 hours a day and vertical axis shows the average time spent by students per a day. This is interesting that the average time spent is minimum on a day before examination except at (9, 10) Am and (10, 11) Pm.

Figure 10 compares the NAC Websites usage of students in different periods of a semester.

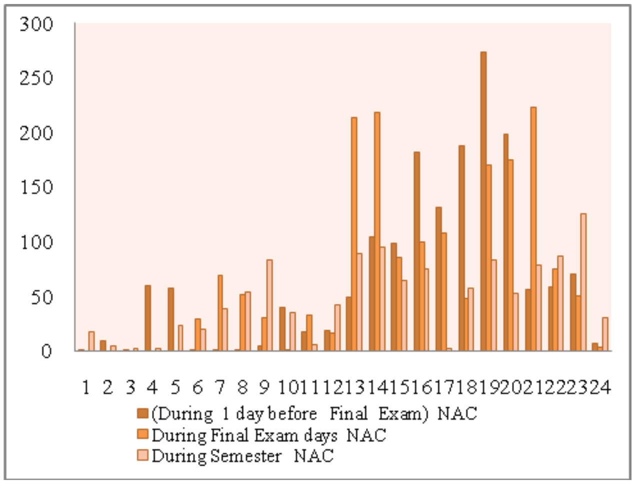

Figure10: Comparing No. Of Unique NAC Websites visited

In Figure 10 horizontal axis shows 24 hours of a day and vertical axis shows the average time spent per a day. As a results of this figure, total minimum time spent in a semester belongs to a day before examination except at (10, 11) Pm and 9Am.

7. Conclusions

We have used WUM for analyzing proxy server log files for extracting students Internet usage behaviors of an engineering college in India during a period of two years including 4 mid-term exams and four final examinations.

We have analyzed gender based usage behaviors of students of an engineering college in India. We analyzed gender based behavior of students based on total time spent on Internet per a day during different periods of a semester including a day before examination day, final examination weeks and during a semester. As for our results, behaviors of students are changing during examination weeks and also based on their semester and their undertaken program. We analyzed the category of visited websites visited by students during that specified periods and observed that a day before examination students' time spent is less than other periods.

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] 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

[2] Z.Huiying, L.Wei "An Intelligent Algorithm of Data Pre-Processing in Web Usage Mining" , Proceedings of the 5th World Congress on Intelligent Control and Automation, June 15-19, 2004 Hangzhou, P.R.China

[3] D.Tanasa et.al "Advanced data preprocessing for inter sites Web Usage mining", IEEE computer society 2004

[4] R.Cooley,J.Srivastava,B.Mobasher "Web Mining: Information and Pattern Discovery on the World Wide Web. In Proc. of the Ninth IEEE International Conf. on Tools with Artificial Intelligence (ICTAI'97), 1997

[5] R.Cooley,J.Srivastava,B.Mobasher "Data prepration for Mining: Discovery and applications of usage pattern from Web data, SIGKDD Explorations, Vol.1(2),12-23,2000

[6] X.Wang, Y.Ouyang, X.Hu, Y.Zhang "Discovery of User Frequent Access Patterns on Web Usage Mining" IEEE 8th International Conference on Computer Supported Cooperative

[7] S.Bai, Q.Han, Q.Liu, X.Gao "Research of an Algorithm Based on Web Usage Mining" IEEE 2009, 978-1-4244-3894-5/09

[8]S.P.Nina,M.Rahaman,K.I.Bhuiyan, K.E.U.Ahmed "Pattern Discovery of Web Usage Mining" IEEE ,2009 International Conference on Computer Technology and Development

[9] M.Hogo, M.Snorek, P.Lingras "Temporal Web Usage Mining" Proceedings of the IEEE/WIC International Conference on Web Intelligence(WI'03)

[10] P.Geczy,N.Izumi, S.Akaho, K.Hasida "Human Web Behavior Mining" IADIS International Conference WWW/Internet 2007

[11] E.Manavoglu, D.Pavlov, C.L.Giles "Probabilitic User Behavior Models" Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)

[12]B.Zhou,S.C.Hui, A.C.M.Fong "Discovering and Visualizing Temporal-based Web Access Behavior" Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence(WI'05)

[13] B.Zhou,S.C.Hui, A.C.M.Fong " An Effective Approach for Periodic Web Personalization" Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence(WI'06)

[14] K.R.Suneetha, K.R.Krishnamoorthy "Identifying User Behavior by Analyzing Web Server Access Log File" IJCSNS International Journal of Computer Science and Network security, Vol.9 No 4, April 2009

[15] E.adar, D.S.Weld, B.N.Bershad, S.D.Gribble "Why Web Search: Visualization and Predicting User Behavior" International WWW Conference committee(IW3C2),WWW2007, May 8-12,2007,Banff, Alberta, Canada

[16]Y.Q.Peng, G.X.Xiaq, T.Lin "Predicting of User's Behavior Based on Matrix Clustering" Proceedings of the fifth IEEE Conference on Machine learning and Cybernetics, Dalian,13-16 August 2006

[17] M.Jalali, N.Mustapha, A.Mamat, N.B.Sulaiman "A New Classification Model for Online Predicting Users' Future Movements" 2008 IEEE, 978-1-4244,6/08

[18] A.EL-Halees "Mining Students Data to Analyze Learning Behavior: A Case Study" International Journal of Computer Applications (IJCA Journal) , Number 22, Article 4

[19] Q.A.AI-Radaideh,E.M.A.Shawakfa,M.I.A.Najjar "Mining Student Data Using Decision Trees" The 2006 International arab Conference on Information Technology (ACIT'2006)

[20] A.Merceron, K.Yacef "Educational Data Mining: a Case Study" Proceeding of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology Pages: 467-474

[21] P.Cortez, Alice Silva "Using Data mining to Predicting Secondary School Student Performance" www.dsi.uminho.pt/pcortez/student.pdf

[22]B.M.Bidgoli, D.A.Kashy, G.Kortemeyer, W.F.Punch "Predicting Student Performance: An Application of Data Mining Methods with an

Educational Web-Based System" 33rd ASEE/IEEE Frontiers in Education Conference, Nov 5-8, 2005

[23] N.V.Anand Kumar "Improving Academic Performance of Students by Applying Data Mining Technique" European Journal of Scientific Research ,ISSN 1450-216X Vol.34 No.4(2009),pp.526-534

[24]J.Sander "Gender and Technology in Education : A Research Review" www.jo**sanders**.com/pdf/gendertech0705.pdf

[25]S.Kim, M.Chang "The Differential Effects of Computer Use on Academic Performance of Students from Immigrant and Gender: Implications on Multimedia Enabled Education" Ninth IEEE International Symposium on Multimedia 2007-Workshop

[26] Rozita Jamili Oskouei "Identifying Students Behaviors Related to Internet Usage Patterns (A Case Study)" In proceedings of IEEE 2nd International Conference on Technology for Education , Bombay 1-3 July 2010

[27] 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

[28] Rozita Jamili Oskouei "Behavior Mining of Female Students by analyzing log files" , In Proceeding of IEEE fifth International Conference on Digital Information Management ICDIM 2010, Canada July 5-8

[29] Rozita Jamili Oskouei "Analyzing Different Aspects of Social Network Usage on Students Behaviors and Academic Performances" In proceedings of IEEE 2nd International Conference on Technology for Education , Bombay 1-3 July 2010

[30] Rozita Jamili Oskouei, B.D. Chaudhary "Impact of Non academic Websites Usage on Students Academic Performance(A Case Study) In proceeding of IEEE , International Conference on Education and Network Technology, China 25-27 Jun 2010

[31] Jamili Oskouei, B.D. Chaudhary "Relationship Between Academic Performance and Usage Pattern of Non-Academic Websites Usage by Gender Group" In proceeding of IEEE , International Conference on Education and Network Technology, China 25-27 Jun 2010

[32] S.L.Tanimoto "Improving the Prospects for Educational Data Mining" http://www.educationaldatamining.org/UM2007/Tanimoto.pdf

[33] J.Srivastava "Web mining: Accomplishments and future Directions", http://www.ieee.org.ar/downloads/srivastava-tutpaper.pdf

[34] X.Wang, Y.Ouyang, X.HU, Y.Zhang "Discovery of User Frequent Access Patterns on Web Usage Mining", The 8th International Conference on computer supported corporation work in design proceedings

[35] S.P.Nina, M.Rahaman, K.I.Bhuiyan, K.E.U.Ahmed "Pattern Discovery of Web Usage Mining" 2009 International Conference on Computer Technology and Development, 978-0-7695- 3892-1/09

[36] Open Directory Project (ODP). http://www.dmoz.org/

[37] Wikipedia, http://en.wikipedia.org/wiki/