Indian Journal of Science and Technology

Perceived Internet Usage Behaviours as Predictor to Outlier Detection in Students' Communities in Academic Environments

Rozita Jamili Oskouei[1]*, Mohsen Askari[2] and Phani Rajendra Prasad Sajja[3]

[1,3]Computer Science & Engineering Department, Motilal Nehru National Institute of Technology, Allahabad, UP, India; Rozita2010r@gmail.com, sajja.phani@gmail.com
[2]Computer Engineering Department, Islamic Azad University, Ramsar, Iran; mohsen.askari65@yahoo.com

Abstract

It is important to provide perspectives about the effects of Internet usage on students' personal and social behaviours along with the impacts of these usages on their academic performances. To explore students' Internet usage behaviors and predicting outliers in student's community, we have developed Web based data mining tool named Education Data Miner (EDMiner), which provides user friendly interface for different stockholders of the system including professors and deans. This research study was conducted with a sample of 5210 students from one engineering college in India during 36 months continually. The primary focus of this study is to extract Internet usage pattern of students by exploring proxy server access log files. These patterns were then used for identifying outliers in students' community. We have applied centroid and density based clustering methods to identify outliers. Further, the relationship between Internet usage behaviours and various Academic and Non-academic activities were explored. Based on our results the majority of visited Websites, 35 percent, belongs to Websites under Extra-Curricular category whereas for curricular Websites it is 24 percent. Further, our results also contradict the perception that the Internet usage adversary affects the academic performance. Moreover, our analysis results show higher average time spent on Internet did result into nonparticipation in other activities, which are very essential for the growth of these students. This nonparticipation in other activities may prove to be an indicator for loneliness of these individuals.

Keywords: Educational Data Mining, Internet Usage Behaviours, Academic Performance, Curricular and Co-curricular Activities, Web Usage Mining.

1. Introduction

Internet and World Wide Web (WWW) technologies have significantly influenced daily activities and quality of life of individuals and organizations including academic institutions. Most of the academic institutions in India have made significant investment in Internet and computing infrastructure with an objective to make quantum jump in academic productivity and quality. Academic institutions have also opted for these technologies in a big way in their library services, classrooms, research labs and residential complexes. These computing and Internet infrastructure have come at significant cost and with the following expectations. (1) Easy access to learning resources through Internet will result in improvement of quality of teaching and learning. (2) Computer assisted class rooms will increase productivity of both students and teachers. (3) Virtual class rooms may be able to partially address the shortage of faculties in institutions. (4) Internet access in residential complexes will lead to self-learning. However,

*Corresponding author:
Rozita Jamili Oskouei (Rozita2010r@gmail.com)

there is a growing perception in academic community that these objectives are either not achieved or achieved only marginally. Many people believe that students use Internet as a stress buster and are increasingly becoming addictive to it, to the extent that it hurts them.

Further, students in academic communities in India in particular are heterogeneous group of individuals. These heterogeneities are due to economic, social and cultural environments from which they have come. Therefore their behavioural patterns and preferences are likely to be different. Classification of their behavioural pattern and preferences based on their Internet usage pattern may help to identify the dominant patterns and outliers of the community. Identification of the outliers of the community may help to proactively initiate measures to improve academic environments. We also report elsewhere a significant gender gap exists in terms of Internet usage.

This research work is motivated by the desire to extract the Internet usage behaviours of students and determine the effects of Internet attitudes and behaviour on their academic performance, extra-curricular and co-curricular activities. Secondly, identifying the outliers in the students' community, who are unable to cope up with academic and environmental stress and strain, which can enable institutions to initiate proactive measures, if required? Finally, to design and implement a Web based tool named EDMiner, which can be used by different stockholders including academic administrators and professors to control students' Internet usage patterns along with their various activities in academic environments.

This paper is organized in five sections. Section 2 explores the related works and the tools (commercial, free and open source) which are available in market for analysis of data. In Section 3, we describe our proposed Website Classification scheme and its mapping to Open Directory Project (ODP) [11]. Section 4 describes the background of the study and details its methodological approach (sampling, data collection and analysis and outlier detection) along with the architectural design and the main components of EDMiner tool, and with the help of some GHU windows, we describe some of the functionalities of this tool. Section 5 includes discussions about this study and its limitations and concludes our paper.

2. Related Works

In this section, we present the related works in the area of data mining and educational data mining in particular. It begins by reviewing the usage pattern discovery methods followed by a short discussion on different data/Web mining tools which are accessible in both free or open source and commercial forms.

2.1 Usage Pattern Discovering

Several research studies [3–7, 9, 22–27, 29, 30, 32, 33, 36, 37, 39, 42, 45 and 48] have been made to model individual and group behaviours and to evaluate usage patterns of different services. These models have used different sources of input data for modelling. These input data includes access log files [3, 4, 25, 27, 29, 30, 33 and 48], click trace [23], questionnaires [36 and 37], interviews [7] and other relevant documents [6, 26 and 32]. In [25, 42], Web access log files and clicking patterns of visitors to the Website have been used to evaluate the usage patterns of contents of the visited Websites and to cluster the users based on their preferences for the pages from the Websites. These studies have been used to improve the Website contents and to eliminate those contents which are not being used. In [23], a similar research study has been made to predict Website visitors' genders, age and their ethnicity.

Several other studies [1, 2, 21, 28, 31, 32, 34, 36, 38 and 40, 47] have focused on studying the usage of information technology at different levels of education. These studies have been made either to assess the impact of technology on extent of learning or to predict the performance of students in one or more courses. Some of the researchers [1, 2, 21, 31, 34, 38 and 40] have studied the impacts of the Internet usages on students' academic performances. In [38], Q.A. Al-Radaideh has proposed a model to predict the performance of students in the final examination of C++ course based on their performance in the mid-term examination, their attendance and their activities on C++ course Website including questions and answers and the assignments. Similarly, in [2, 21 and 31], a model has been proposed to predict the performance in a Web based course based on users' on line activities. In [28], Liccardi has addressed the role of social networks in computer science education. This study concluded that social networking plays a positive role in students' learning experiences.

2.2 Data Mining Tools

Several data mining tools have been developed to analyze different kinds of data. These tools vary in terms of the input data which they use to analyze. The contents used by these tools include textual data, multi media contents or unstructured Web data, etc. These tools are available in

both free or open source (Weka [19], Orange[20], R-Miner [46], Rattle [16], Rapid Miner [15], Weblizer [18], etc.) and commercial (Statistica 8 [44], SAS Enterprise Miner [8, 17], IBM Intelligent Miner [12], Microsoft SQL Server 2005 [35], Angoss Knowlwdge STUDIO [10], KXEN [13], Website parser tool [14], Website Scraper [14], Web extractor software [14], etc.) forms. Most of these tools have very interesting and useful features and are generally targeted to the enterprise level data sets for specific domains and applications such as business intelligence.

None of these tools including free and open source and commercial has features to help in analyzing data related to students 'Internet activities. A prime motivation to design a data mining tool was to help academic administrators and teachers to extract the Internet usage behaviours of students, identify outliers in students communities based on their Internet usage behaviours, academic performance and to study the relationships between different groups of outliers.

3. Website Classification Scheme

Web page classification is the process of assigning a Web page to one or more predefined category labels. In Website classification, categorization can be done based on Website's content or structure. Most of the general purpose search engines and portals use the Website classification scheme of Open Directory Project (ODP) [11]. These search engines and portals include Google, Netscafe Search, AOL Search, Lycos, DirectHit, etc. ODP is a multilingual open content directory of WWW links and is constructed and maintained by a community of volunteer editors. ODP defines 16 top level categories, which are 1: Arts, 2: Business, 3: Computers, 4: Games, 5: Health, 6: Home, 7: Kids and Teens, 8: News, 9: Recreation, 10: Reference, 11: Regional, 12: Science, 13: Shopping, 14: Society, 15: Sports and 16: World.

Since, ODP categories to which Websites visited by students, do not related to activities of academic environments. We need to have the concepts in the classification scheme which explicitly related to the activities of students in a residential academic institution. In an academic institution, students' activities are generally classified as curricular (includes course works, which may have lectures, tutorial or practical classes and other activities directly related to courses), co-curricular (include technical paper writing, paper presentations, seminars and conferences and etc.), extra-curricular (include sports, cultural activities (dance, theatre, etc.) and literary activity, hobbies (photography, robotic, etc) and non-curricular activities (include business, shopping and career related activities).

It is pre-requisite to classify the Websites visited by students according to curricular, co-curricular, extra-curricular and non-curricular categories. We have augmented the ODP classification scheme with the following concept.

- Curricular, Co-Curricular, Extra-Curricular, Non-Curricular, Media, General, Professional, Undesirable, Adult, Webcam, Free SMS, Sharing Websites, Special Communities, Terrorism and Criminals.

Among the categories given above, Curricular, Co-Curricular, Extra-Curricular, Non-Curricular are the higher or top level categories.

These concepts have been introduced as either generalization or specialization of ODP concepts. Super and Sub categories for the above concepts are given below.

- *Curricular category* is generalization of *Science, Computer and Reference* categories.
- *Media* is a specialization of Art and generalization of *TV, Radio, Music, Movie, Video* and *Animation*.
- *Social and communication networking* is generalization of *Social Networking*. *Webcam, free SMS, Special Community, Resource Sharing Websites* are specialization of *Social Networking (SN)*.
- *Undesirable* is specialization of *Recreation with Adult*, *Terrorism and Criminals* its generalization along with *Drugs*.

These categories have been introduced based on analysis of contents of the popular Websites visited by students. From our analysis, Website contents, which are related to academic courses, were mostly under Science, Computers and Reference categories.

All the above categories, we have considered either based on related activity or usage. The Undesirable category is introduced to capture social acceptable norms. Terrorism, Criminal and Drugs related Websites are undesirable in most of the societies. However, the Adult Websites may not be in the same category. In the context of Asian and Middle-East societies including Indian society, we considered these to include under Undesirable category. It is evident from our analysis, that majority of visited Websites (35%) belong to Extra-Curricular and Subcategory of Social Network Websites (27%) and not to curricular Websites

(24%). It is heartening to note that, there is only small percentage (lesser than 1%) who visited drug related Websites. Further, there is a small percentage (10%) of students, both male and female, who visit Adult Websites.

4. Methodology

In this section, we start our approach first by mapping the categorization scheme according to ODP. We then give the outlier detection followed by the design and implementation details of EDMiner tool. Finally, we provide the results with the help of GUI windows.

4.1 Mapping Categorization of Websites

In order to meet our requirements of Website classification according to Curricular, Co-Curricular, Extra-Curricular and Non-Curricular categories, we have mapped our scheme to ODP classification scheme. For example, in Curricular category of our classification scheme, we have mapped the 'Top/Computers', 'Top/Reference' and 'Top/Science' from ODP classification as Computers, Reference and Science under Curricular category. We have further classified Reference as General and Professional sub-categories. Professional has a sub-category of other, which maps ODP's Top/Reference category except Education sub-category. A complete list of these mappings is given in Appendix A.

ODP data is available as RDF dumps in compressed format and is encoded in UTF-8. These latest RDF dumps can be downloaded from http://dmoz.org/about.html. One dump file is structure.rdf.u8.gz, which provides the category hierarchy information. In this classification scheme, each category and the sub-categories within that category are available as a hierarchical structure. For example, category Top/Computers has sub-categories of Internet, Software, hardware, Security, etc. Another one file is content.rdf.u8.gz, which provides the links within each category. Each category of Website is defined by the community editors. For example, one of the categories of google.com is Computers: Internet: Searching: Search Engines: Google.

We have downloaded the contents dump latest by September 6, 2012. We have used Apache Jena [49] framework to load this RDF [50] dump in to TDB database and queried (SPARQL) [51] the RDF data for finding the category. This dump file has been pre-processed to correct the errors or warnings that are occurred while loading the dumps into the database. The errors or warnings handled while pre-processing the dump includes correcting the RDF namespace declaration, removing empty namespace, adding namespace prefix and changing id to about.

After making the above changes to the content RDF rump, we have converted the dump from RDF/XML into n3 format using Jena's command line tool 'rdfcat' to load into TDB database.

We have developed a two step process to find the category of each visited Webpage by querying the RDF data in TDB using SPARQL query language. This two step process has been described below.

Step 1:

- For each URL visited, get the top level domain name
  - First the name of Websites will be extracted by trimming the resource accessed. For example, the URL http://www.google.com/resources/xyz.abc.html will become http://www.google.com.
  - Extract the top level domain name from the URL extracted. For example, the formatted URL from the above step http://www.google.com will be become google.com

    Note: For all sub-domains, this process produces the same top level domain. For example, for different Websites visited including http://www.google.com, https://mail.google.com/mail/, https://plus.google.com/, google.com will be the resultant extracted top level domain.
- After retrieving the top level domain name, this domain name will be searched in 'Content Database' for finding the different categories that the extracted top level domain has been listed.

For example, the search results obtained for Websites google.com and facebook.com are 2417 category matches for *google.com* and 249 category matches for *facebook.com.*

Step 2:

- In this step, for each unique top level domain names extracted from step 1, we have mapped the categories to our proposed Website's classification scheme.
- For example, for twitter.com, one category we got from ODP is Computers: Internet: On the Web: Online Communities: Social Networking: Twitter. From our mapping it can be categorized under Extra-Curricular → Society → Social Network → Social Communication category.

Note: Not all Websites have been listed in ODP. For those that are not listed in ODP, we have visited the Websites manually and identified the categories by looking at the contents. For example, for tvunetworks.com, we mapped the category for this Website under Non-Curricular → Other.

4.2 Outliers

Several definitions or descriptions of outliers [39, 41, 43 and 45] are defined in the literature. In the context of our study, we defined outliers as individuals whose Internet activities, academic performance, features, extent of engagement to academic activities are different from the majority of the members of the community to which s/he may be belong to. Outliers may belong to bad or good categories. We used centroid and density based clustering methods to identify outliers based on average time spent on Internet per day, academic performance (CPI), co-curricular and extra-curricular activities and relationship between different groups of outliers. We used Rapid-Miner and Weka for examining the results of clustering made by our tool.

We have identified the following outliers:

- Students having CPI<= 2.7 or CPI >= 9.7
- Students daily average time spent on Internet >= 457 minutes or less than 5 minutes

4.2.1 Outliers vs. Various activities

Internet usage behaviors of students who participated in *co-curricular activities* (IEEE conferences) shows:

- 20% of female and 14% of male participants of co-curricular activities were in outliers based on academic performance with CPI > 9.7.
- None of the female or male participants were outliers based on average time spent on Internet.

Exploring the category of visited Websites by students who participated in co-curricular activities (IEEE conferences, etc.) shows:

- There is no gender based differences either in the category of visited Websites or percentage of average time spent on these category of Websites. Further, both female and male students spent relatively more time on Websites belonging to curricular activities.

Examples of *Extra-Curricular activities* are cultural activities and sports. It is important to note that, there is no single female or male students, who is outliers with CPI < 2.5 and participated in these activities, whereas 2 female (out of 60) and 3 male (out of 160) regular students had CPI >= 9.7.

One significant difference between students who participated in co-curricular activities are these group of students spent more time on Websites belonging to co-curricular and extra-curricular activities. By analyzing gender based academic performance of all participants of extra-curricular activities including sport and cultural events along with their average time spent, we can conclude that outliers with respect to the average time spent on Internet more than 457 minutes (both genders) participate minimally in these activities. In the other words, higher average time spent on Internet did result into nonparticipation in other activities which are very essential for the growth of these participant students. This nonparticipation in these group activities may prove to be an indicator for loneliness of these individuals. Further, it is interesting to note that 70% of these participants belong to first year students.

4.3 EDMiner Tool

This section describes some of the functionalities of EDMiner tool. EDMiner has the following major objectives:

- To discover the distribution of visitors during different hours of a day in a semester continually including examination periods.
- To identify the category of visited Websites and average time spent.
- To identify outliers based on academic performance and average time spent.
- To establish relationships between different groups of outliers.

The tool provides user friendly interface for the following stakeholders:

- System and Network Administrators
- Course coordinators and Professors
- Dean (Academic Affairs/ Students Welfare)

The input for this tool:

- Proxy server access log files.
- A text file containing students' data with fields: Registration-Number, Full-Name, Program, Branch, Semester, Gender, CPI.
- A text file containing User-Id, Full-Name and department name.
- A text file, which includes Registration-Number, Full-Name, Program, Activity-ID.

4.3.1 System Architecture

In this section, we give a high level architecture overview of EDMiner system and describe the major components of the tool. Figure 1 depicts the system architecture of EDMiner. This figure depicts the system architecture in terms of high level view of components.

The three major components of the EDMiner are:

- Data resources are responsible for collecting data from various data sources for inputting to the Pre-processing component.
- Pre-processing component is responsible for data selection, cleaning and integration. Data selection extracts user id, time of access and URL of visited Website fields from the log files. Data cleaning removes records with inconsistent or missing values and data integration combines data.
- Data Mining & Pattern Discovery component is responsible for transforming the data into knowledge. This component extracts each user's daily average time spent (minutes) and the total number of visited WebPages. Moreover, with the help of data mining techniques, Internet usage pattern for each user is extracted and with the help of k-means and density based algorithms, outliers based on time-spent per day and number of visited WebPages and CPI has been extracted. With the help of visualization techniques, we have visualized our analysis results in tabular and pictorial representation for making more understandable reports for academic people.

4.3.2 EDMiner Modules

From the usage point of view, EDMiner has four functional modules including pre-processing, database, Internet usage patterns extractor and outlier detection modules.

4.3.2.1 Pre-processing Module

Data pre-processing is a data mining process deals with the preparation and transformation of the initial data. This module involves extracting the fields and records from log files by cleaning data. Data cleaning removing records with inconsistent or missing values. The fields which are extracted for the analysis is *User_ID*: id to access the Internet through proxy server authentication, *Website-URL*: URL of the visited Webpage and *Time-of-Connection*: Timestamp of the Webpage visit. The other data files are students academic details from Dean Academics which includes registration, CPI related information, data from Computer Centre (CC), which includes Internet access details, students other activities reports. These data files are also cleaned and loaded into the database for further analysis.

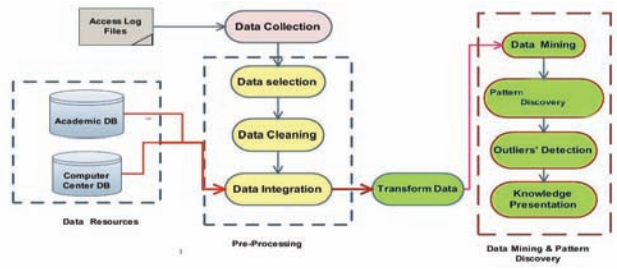


Figure 1. EDMiner System Architecture.

4.3.2.2 Database Module

During the analysis various database tables are populated with the analysis data. The database module is responsible for managing the populated data, processing query, etc. The identified database tables are users, day-history, session, website, category, etc. Some of the major tables are given below with the field information.

User: This table is populated with the data obtained from Dean (Academic Affairs) which includes the following fields for further analysis: *User-ID*, *Program*, *Branch, Semester, Gender and CPI*.

Session: This table contains the analysis data which includes session details. Session time is the amount of time which a user spent on Internet continuously. If the amount of time between two consecutive hits for a user is greater than the pre-defined (here 15 minutes) session time then a new session is created. This table includes the fields: *Session_Id*, *User_Id*, *Day_History_Id*: reference to Day-history-table, *Start_Time*, *End_Time* and *Session_Duration*.

Day-History: This table contains the records of day history for each user. The fields included for this table are: *Day_History_Id*, *User_Id*, *No_of_Sessions*, *Average_Session_Duration*, *Minimum_Session_Duration*, *Maximum_Session_Duration*.

Further, this module also populates and handles data for *Website, Category, Session-Website, and Website-Category* tables.

4.3.2.3 Internet Usage Patterns Extractor Module

Identifying Internet usage patterns can be useful for studying the behaviours of students. These patterns include the amount of time spent on Internet, the frequency of visiting Websites, the category of visited Websites, etc.

During analysis, the Session and Day-History tables are populated with the records of visited Websites. This module uses these tables to compute daily average time spent on Internet and also number of visited WebPages (hits count) per day by each individual user. It also computes the support value for each visited Website and prepares a database for the Website having support value greater than or equal to 50. These support values are computed on the basis of different duration in terms of continues days such as 5, 10, 15, 20, etc., days. These support values along with their duration of computation are stored in the database for further observation. Further, this module classifies the Internet users into *Regular* (users who connect to the Internet frequently), *Non-Regular* (users who do not connect to the Internet frequently) and *Internet-Absentees* (users who never connected to the Internet).

4.3.2.4 Outliers Detector Module

This module clusters students based on their CPI, daily average time spent on Internet. This module uses *k-means, Density-Based* Spatial *Clustering* of Applications with Noise (*DBSCAN*) clustering methods to create these clusters. This module is also responsible for identifying outliers based on the thresholds provided by users and relationship between these outliers. It also analyzes the relationships of outliers with curricular, co-curricular, extra-curricular activities.

4.4 EDMiner User Interface

EDMiner provides functionalities for different stakeholders of the system through various user interface windows. Each user of the system has given one or more roles. These roles and responsibilities of each user have been described in below section.

4.4.1 The Academic and System Administrators

The Academic and System Administrators are authorized to upload log files, files containing personal information and academic records of students. In addition to uploading, the user interface also allows to add or edit the individual records.

These Administrators can also query the system to get the following information:

- Distribution of Internet users on hourly basis for a day.
- Hourly distribution of number of WebPages visited during a day.
- For a specific period, total number of users who connected to Internet, list of popular visited Websites and their categories.
- Further, they can query daily based Internet usages of students and compare usages on different days or for a selected period.

These search results can be used to identify the popular visited Websites by different users and the time in which the users connect to the Internet. These results can be used improve the quality of Internet or for a particular category of Website during a period.

Figure 2 shows the distribution of Internet users on hourly basis for a selected day (September 06, 2011). From this window, we can observe that 483 different users are connected to Internet during 12:00 to 01:00Am Also observe that the number of users connected to the Internet during the period 01:00 to 02:00PM is 54.

4.4.2 Course Coordinators and Professors

Course Coordinators and Professors can query the system for:

- Total number of registered students for a selected course and their gender-wise distribution.
- Classification of students registered for a course based on group (regular, non-regular and Internet-absenters)
- The Internet absentees from the course along with their CPI and whether they are outliers with respect to their CPI
- Outliers within regular users based on CPI, and average time spent.
- All *Regular users*' Internet usage patterns and the category of visited Websites.

One of the functionalities provided by this system to the course coordinators is finding the outliers. Figure 3 shows the population of outliers of the selected course (for example #134). Here average time spent (>457min) is the measure to find the outliers.

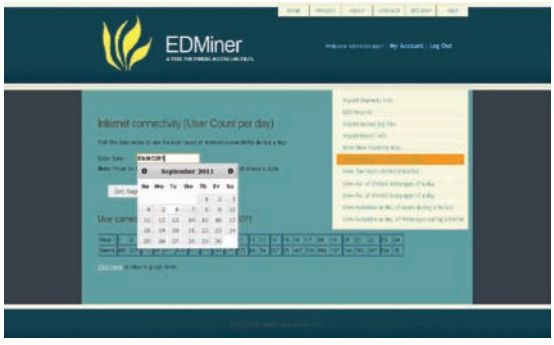


Figure 2. Distribution of Internet users on hourly basis for a selected day.

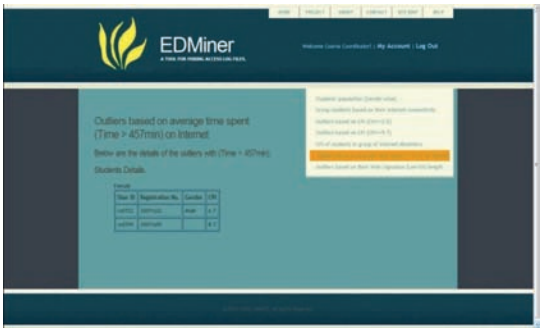


Figure 3. Outliers based on average time spent (>457min) for the selected Course #134.

4.4.3 Dean (Academic Affairs and Student Welfare)

These are the highest privileged users and can query the following details regarding an individual student:

- Group to which a user belongs (Regular, Non-Regular or Internet-Absentees).
- Daily Average time spent on Internet
- Daily Average number of visited Web pages
- The category of visited Websites
- CPI
- Is s/he belongs to group of outliers (time-based or CPI)

Following details can be queried on the basis of branch, semester and course.

- Gender wise and course wise total number of students
- Number of regular, non-regular users of Internet and Internet absenters' based on gender and program.
- Internet usage patterns based average time spent on Internet, average number of visited Web pages and the category of visited Websites of each regular user.
- CPI basis Internet usage pattern
- Distribution of Internet users per hour for a favorite day
- Hourly distribution of number of Web pages visited during a favorite day
- Internet usage pattern of outliers based on CPI (CPI<=2.5 or CPI<=9.7) including the category of visited Websites these outliers.
- Internet usage pattern of participants on co-curricular or extra-curricular activities

One of the features which will be available for the Deans is to find the users who connected to the Internet for a particular day based on their CPI. The window in Figure 4 shows users' connectivity report for a day (September 6, 2011) for the selected CPI (< 5.0). The query results include

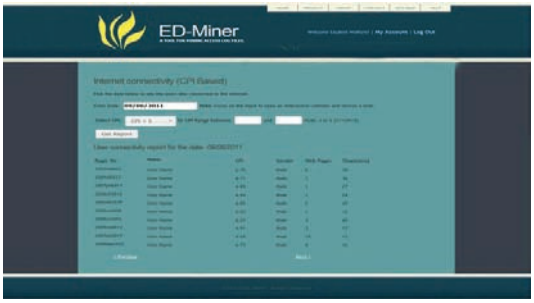


Figure 4. Internet Usage Pattern of Users with CPI<5.

registration number, name, CPI, gender, Web pages visited and the time spent.

5. Discussions and Conclusion

In main goal of this investigation was to explore positive or negative effects of Internet usage on students' academic and nonacademic activities. We attempted to extract Internet usage behaviours of students during a semester by analyzing log files for a period of 36 continues months. Our main focus is to extract the outliers of students' communities based on their academic performance and the relationship between their Internet usage behaviours and comparison of behaviours of students of outliers with others. To visualize the Internet usage behaviours and outliers, we have implemented a Web based tool named EDMiner using Java, which can be accessed by different stakeholders including administrators and professors. This tool is able to extract individual students' usage pattern or history of connection along with the name of visited Website and the category of the Website for a selected period. We have used k-means and DBSCAN methods for clustering students and for identifying the outliers.

During analysis, we found that 22% of the students never connected to the Internet. We further tried to identify the reasons for this large percentage of Internet absentee. One of the reasons is that the students who do not reside in the institute campus may not have access to institute Internet infrastructure from outside the campus and they may not be able to use during regular working hours due to lectures or practical classes. We found that this is specifically true for MBA, MCA and M.Sc degree programs. However, it does not explain the Internet absentees from B.Tech students as most of them resides with in campus and have access to Internet from hostels. Our further investigation of IP addresses revealed that a small number of students impersonated their teacher and used

their Ids to access Internet. Further, 41% of total female students falling under category of Internet absentees and should be a matter of concern for administrators. One of the dominant reasons for these absentees might be that the Internet infrastructure and accessibility to female students in the hostels is limited. Further, there might be a restriction on the hours of accessibility in the institute. Another reason might be due to personal inhibitions or resistance to adapt to technology due to cultural, social or religious beliefs and traditions. We have found that the CPI of students from Internet absentees is very low and is tempting to hypothesis that this is one of the reasons for the poor academic performance.

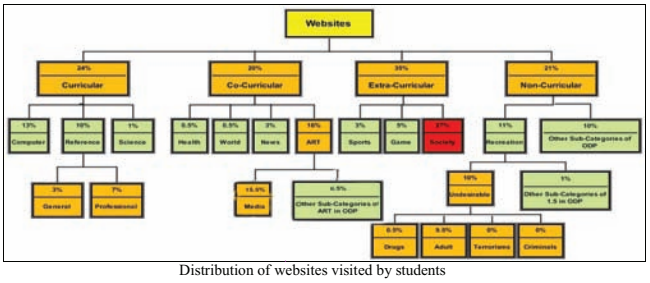
A large number of students from some of the post graduate programs such as MBA and M.Sc. either did not use Internet or used very infrequently. This may be due to different orientation of their academic programs or evaluation scheme. The curriculum contents may not be compelling these students to use the Internet. All the observations made in this research work are based on analysis of data from one of the technical institutions in India. It may be inappropriate to generalize these observations to other students' communities in different institutions. At best, these results may be applicable to other institutions, which have similar infrastructure and student population. Therefore, it is important to validate these observations for other data from other heterogeneous group of institutions with different population mix in terms of gender and different academic programs. Despite these limitations, this study addressed a major gender based gap in Internet usage behaviors of students which mostly confirm to the results of study done by Butakov [42].

One of the prime motives for undertaking this study was to identify outliers in the students' community, who are unable to cope up with academic and environmental stress and strain. For this purpose, one needs to identify more input about students' activities in addition to their Internet related activities. The additional input could be network of friends and their Internet and mobile usage patterns or visits of outside the campus for different personal needs, etc. It is desirable to have a comprehensive catalog of activities and to collect data about them so that proactive detection of outliers can be performed. Table 1 presents the distribution of students who are falling in the outliers groups.

Students who are in outliers based on CPI and average time spent have used limited use of Internet and the popular category of Websites visited by those belongs to curricular activities. Further, it is evident from our results that outliers

Table 1. Distribution of Outliers (from Different Aspects)

Dimension		Threshold	Female	Male	Total
CPI	*Total Users*	*CPI<=2.5*	10	51	61
		CPI<=9.5	10	20	30
	Regular Users	*CPI<=2.5*	2	10	12(20%)
		CPI<=9.5	9	15	24(80%)
Time-Spent	*Regular Users*	*Time<5 min*	8	42	50
		Time>=457 min	31	125	156



Distribution of websites visited by students

from different groups used Internet dominantly for extra-curricular activities and most popular Websites visited by them are Sports and Social Networking. It is interesting to note that some of them in the outliers have visited adult Websites, which are not permitted socially as well as legally.

It is interesting to note that majority of outliers based on CPI are those who either do not connect to Internet or connect very in-frequently. Further, there is no single student who never connected to Internet in the group of outliers having CPI >= 9.7. From these details, it is tempting to conclude that those who either do not connect to Internet or connect infrequently were not performing well in their academic programs. This also contradicts the perception that the Internet usage adversary affects the academic performance.

Further, exploring the relationships between outliers based on CPI and average time spent revealed:

- 58% of outliers with CPI >=9.7 are also outliers in terms of average time spent on Internet with threshold of 457 minutes per day. In other words, majority of academic outliers with excellent performance use Internet extensively.
- Only 22% of outliers with average time spent on Internet more than 457 minutes have CPI>=9.7. It implies that

we cannot generalize that more time spent on Internet lead to better academic performance.

- 70% of female and 50% of male outliers with CPI>=9.7 were outliers with time-spent>=475.

In the other words, higher average time spent on Internet did result into nonparticipation in other activities, which are very essential for the growth of the students. This non-participation in other activities may prove to be an indicator for loneliness.

6. References

1. Merceron A, and Yacef K (2005). TADA-Ed for educational data mining, Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, vol 7, No. 1.
2. Minaei-Bidgoli B, Kashy D A et al. (2003). Predicting student performance: an application of data mining methods with an educational web-based system, Proceedings of The 33rd ASEE/IEEE Frontiers in Education Conference,(T2A), vol 1, T2A-13–T2A-18, DOI: 10.1109/FIE.2003. 1263284.
3. Zhou B, Hui S C et al. (2006). An affective approach for periodic web personalization, Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence IEEE Computer Society Washington, USA, 284–292, doi:10.1109/WI.2006.36.
4. Zhou B, Hui S C et al. (2004). An intelligent recommender system using sequential web access patterns, Proceeding of the IEEE , International Conference on Cybernetics and Intelligent systems, Singapore, vol 1, 393–398, doi: 10.1109/ICCIS.2004.1460447.
5. Lindemann C, and Littig L (2007). Classifying web sites, (WWW 2007), Banff, Alberta, Canada. ACM 978-1-59593-654-7/07/0005.
6. Chen Q, Hsu M et al. (2000). A data-warehouse/OLAP framework for scalable telecommunication tandem traffic analysis, 201–210, Data Engineering, Proceedings 16th International Conference, DOI: 10.1109/ICDE.2000.839413.
7. Madge C, and Connor H (2002). On-line with e-mums: exploring the Internet as a medium for research, Royal Geographical Society (with The Institute of British Geographers), vol 34(1), 92–102, DOI: 10.1111/1475-4762.00060.
8. Wedding D K, Extending the data mining software packages SAS enterprise miner and SPSS clementine to handle fuzzy cluster membership: implementation with examples, Thesis paper, Available from: http://content.library.ccsu.edu/cdm/ref/collection/ccsutheses/id/946
9. Manavoglu E, Pavlov D et al. (2003). Probabilistic user behavior models, 3rd IEEE International Conference on Data Mining,(ICDM03), 203–210, DOI: 10.1109/ICDM.2003.1250921.
10. Angoss predictive analytics software and solutions, Available from: http://www.angoss.com/about-angoss/why-angoss
11. Open Directory Project (ODP) (2012). Available From: http://www.dmoz.org/
12. Technical resources for IBM Information Management software, Available from: http://www.ibm.com/developerworks/data/downloads/uima
13. Predictive Analytics / Data Mining, Available from: http://www.kxen.com
14. Statistica, Data Mining Software, Available from: http://www.mozenda.com/web-mining-software
15. Open source software for big data analytics, No programming required, Available from: http://www.rapid-i.com
16. Open source solutions and data mining over 20 years of research and development, Available from: http://www.rattle.togaware.com
17. SAS providing software solutions since 1976, Available from: http://www.support.sas.com
18. The Webalizer , free web server log file analysis program, Available from: http://www.webalizer.org
19. New Zealand's weka website is for the use of disabled people, Available from: http://www.weka.net.nz
20. Open source data visualization and analysis for novice and experts, Available from: http://www.orange.biolab.si
21. Wu H, Zhang J et al. (2009). The explore of the web based learning environment based on web sequential pattern mining, Proceedings of International Conference on Computational Intelligence and Software Engineering, (CiSE 2009), 1–6, DOI: 10.1109/CISE.2009.5364267.
22. Stepanikova I, Nie N H et al. (2010). Time on the Internet at home, loneliness, and life satisfaction: Evidence from panel time-diary data, Computers in Human Behavior, vol 26(3), 329–338, DOI: 10.1016/j.chb.2009.11.002.
23. Tullio J, Goecks J et al. (2002). Augmenting shared personal calendars, Proceeding of the 15th annual ACM symposium on user interface software and technology, ACM New York, NY, USA, (UIST '02), 11–20, DOI: 10.1145/571985.571988.
24. Begole J B, Tang J C et al. (2003). Rhythm modeling, visualizations, and applications, Proceedings of the ACM Symposium on User Interface Software and Technology, Vancouver Canada , 11–20, Doi: 10.1145/964696.964698.
25. Pierre J M (2001). On the automated classification of web sites, University Electronic Press, Computer and Information Science, vol 6.
26. Kang J H, Welbourne W et al. (2004). Extracting places from traces of locations, the ACM International Workshop on Wireless mobile applications and services on (WLAN) Hotspots, 110–118, New York, NY, USA.
27. Figl K, Kabicher S et al. (2008). Promoting social networks among computer science students, Proceedings of the 38th ASEE/IEEE Frontiers in Education Conference, Saratoga Springer, NY, S1C, 15–20, DOI: 10.1109/FIE.2008.4720676.

28. Liccardi I, Ounnas A et al. (2007). The role of social networks in students learning experiences, Proceeding of the Working group reports on Innovation and technology in computer science education, (ITiCSE-WGR '07), ACM New York, NY, USA, 224–237, DOI: 10.1145/1345375.1345442.
29. Terveen L, Hill W et al. (1999). Constructing, organizing, and visualizing collections of topically related web resources, ACM Transaction on Computer-Human Interaction, vol 6(1), 67–94.
30. Halvey M, Keane M et al. (2005). Predicting navigation patterns on the mobile-internet using time of the week, (WWW 2005), 958–959, ACM Press.
31. Leiba M, and Nachmias R (2006). Web usage patterns and learning styles in an academic course in Engineering, International Conference on Information Communication Technologies in Education, (ICICTE 2006), Rhodes.
32. Huynh M Q, Lee J et al. (2006). The insiders perspectives: a focus group study on gender issues in a computer-supported collaborative learning environment, Journal of Information Education, vol 4, 237–255.
33. Ester M, Kriegel H et al. (2004). Accurate and efficient crawling for relevant websites, VLDB '04 Proceedings of the Thirtieth international conference on Very large data bases, vol 30.
34. Jalali M, Mustapha N et al. (2008). A new classification model for online predicting users future movements, Proceeding on the International Symposium on Information Technology, (ITSim (2008), 1–7, DOI: 10.1109/ITSIM.2008.4631852.
35. Andronie M, and Crisan D (2010). Commercially available data mining tools used in the economic environments, Database Systems Journal, vol I, No. 2/2010.
36. Eagle N, and Pentland A (2009). Eigenbehaviors: Identifying structure in routine, Behavioral Ecology and Sociobiology, vol 63(7), 1057–1066, DOI: 10.1109/ISWC.2002.1167224.
37. Kwon O, and Lee J (2000). Web page classification based on k-nearest neighbor approach, ACM 5th International Workshop on Information Retrieval with Asian Languages (IRAL), New York, NY, 9–15.
38. AI-Radaideh Q A, AI-Shawakfa E M et al. (2006). Mining students data using decision tress, International Arab Conference on Information Technology, (ACIT2006), 1–5.
39. Rasmussen J L (1988). Evaluating outlier identification tests: Mahalanobis "D" Squared and Comrey "Dk", Multivariate Behavioral Research, vol 23(2), 189–202.
40. Kim S, and Chang M (2007). The differential efforts of computer use on academic performance of students from immigrant and gender groups: implications on multimedia enabled education, Proceeding of the Ninth IEEE International Symposium on Multimedia Workshops IEEE Computer Society Washington, (ISMW '07), DC, USA, DOI: 10.1109/ISMW.2007.83.
41. Schwager S J, and Margolin B H (1982). Detection of multivariate outliers, The annals of statistics, 10, 943–954.
42. Butakov S, Odinma A et al. (2009). Web content usage behaviour: a case study of a university in Sub-Saharan Africa, AMCIS Association for Information Systems, (AMCIS 2009).
43. Stevens J P (1984). Outliers and influential data points in regression analysis, Psychological Bulletin, vol 95, 334–344.
44. U. S. Headquarters: StatSoft, (2007). New Features in Statictica 8, Available From: http://www.statsoft.com/Portals/0/pdf/statistica8features:pdf
45. Hodge V J, and Austin J (2004). A survey of outlier detection methodologies, Artificial Intelligence Review, vol 22, 85–126.
46. Williams and Graham, (2011), Data Mining with Rattle and R, 1st Edn., 374.
47. Yao-Guo G., Lin-Yan S et al. (2006). A research on emotion and personality characteristics in junior high school students with internet addiction disorders, Chinese Journal of Clinical Psychology, vol 14(2), 153–155.
48. Tian Y H, Huang T et al. (2003). A web site mining algorithm using the multi-scale tree representation model, WEBKDD'03 workshop of the ACM KDD-2003 conference, Washington DC, U.S.A.
49. Apache Jena project! Apache Jena™ is a Java framework for building Semantic Web applications, Available from: http://jena.apache.org/
50. Klyne G, and Carroll J J (2004). Resource Description Framework (RDF): Concepts and Abstract SyntaxW3C Recommendation, Available from: http://www.w3.org/TR/2002/WD-rdf-concepts-20021108/
51. Prud'hommeaux E, and Seaborne A (2008). SPARQL Query Language for RDF, , W3C Recommendation, Available from: www.w3.org/TR/rdf-sparql-query/

Appendix

ODP Hierarchy

Top

- (01) Arts
- (02) Business
- (03) Computers
- (04) Games
- (05) Health
- (06) Home
- (07) Kids and Teens
- (08) News
- (09) Recreation
- (10) Reference
- (11) Regional
- (12) Science
- (13) Shopping
- (14) Society
- (15) Sports
- (16) World

Mapping ODP to our Websites classification scheme

Website

- Curricular
  - Computers (03) {'Top/Computers'}
  - Reference (10) {'Top/Reference'}
    - General
      - E-learning {'Top/Reference/Education/Distance Learning/Online Teaching and Learning'}
      - Distance Learning {'Top/Reference/Education/Distance Learning/' except 'Online Teaching and Learning'}
      - Education {'Top/Reference/Education' except 'Distance Learning'}
    - Professional
      - Other ('Top/Reference/' except 'Education')
  - Science (12) {'Top/Science'}
- Co-Curricular
  - Health (05) {'Top/Health'}
  - World (16) {'Top/World'}
  - News (08) {'Top/News'}

```
|    |-- Arts (01) {'Top/Arts'}
|    |    |
|    |    |-- Media
|    |    |    |
|    |    |    |-- Television {'Top/Arts/Television'}
|    |    |    |-- Radio {'Top/Arts/Radio'}
|    |    |    |-- Music {'Top/Arts/Music'}
|    |    |    |-- Movies {'Top/Arts/Movies'}
|    |    |    |-- Video {'Top/Arts/Video'}
|    |    |    |-- Theatre {'Top/Arts/Performing Arts/Theatre'}
|    |    |
|    |    |-- Other {All other under Arts except the above 6}
|
|-- Extra-Curricular
|    |
|    |-- Sports (15) {'Top/Sports'}
|    |-- Games (04) {'Top/Games'}
|    |-- Society (14) {'Top/Society'}
|    |    |
|    |    |-- Social Network
|    |    |   |
|    |    |   |-- Social Communication {'Top/Computers/Internet/On the Web/Online Communities/Social
                  Networking'}
|    |    |        |-- Professional Communication {'Top/Computers/Internet/On the Web/Online Communities/'
                       except Social Networking}
|    |    |        |-- Webcam {'Top/Computers/Internet/On the Web/Webcams'}
|    |    |        |-- file Sharing Computers/Internet/File Sharing'}
|    |    |        |-- Blog {'Top/Computers/Internet/On the Web/Weblogs'}
|    |    |
|    |    |-- Other (14)
|
|-- Non-Curricular
|    |
|    |-- Recreation (09) {'Top/Recreation'}
|    |    |
|    |    |-- Undesirable
|    |    |    |
|    |    |    |-- Drugs {'Top/Recreation/Drugs'}
|    |    |    |-- Adult {'Top/Society/Sexuality'}
|    |    |    |-- Terrorism {'Top/Society/Issues/Terrorism'}
|    |    |    |-- Crime {'Top/Society/Crime'}
|    |    |
|    |    |-- Other (Except Drugs in Recreation)
|    |
|    |-- Other (02, 06, 07, 11, 13)
|
```