



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

EDMiner (Education Data Miner)

Rozita Jamili Oskouei¹, Satyendra Sharma², Mohsen Askari³, Phani Rajendra Prasad Sajja⁴
Computer Science & Engineering Department^{1,2}, Sindhuja Building⁴
Motilal Nehru National Institute of Technology^{1,2},
Department of computer, islamic azad university, ramsar Branch, ramsar, iran³,
Expert Software Consultants Ltd.⁴,
Allahabad, UP, India^{1,2,4}

Abstract: This paper attempts to develop a tool for exploring the Internet behavioral patterns of students from gender and cultural groups and mining those behaviors' relationships with users' academic performance. Modeling these behaviors may be helpful for different stockholders in formulating academic policies and guidelines. EDMINER (Educational Data Miner) is a Web based educational data mining which provides functionalities to understand usage patterns with respect to the category of visited Websites, and various usage statistics and detecting outliers between users based on their academic performance and Internet usage behaviors. Our experimental results show EDMINER has capability to detect academically at risk students before examination during semester and inform to professors for taking appropriate decisions.

Keywords: Data mining tool, Educational Data Miner, Internet Usage Behaviours, ODP (Open Directory Project), CPI (Cumulative Performance Index)

I. Introduction

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 following expectations.

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

Further, academic community, in general and of India in particular, is a heterogeneous group of individuals. These heterogeneities are due to economic, social and cultural environments from which they have come. Therefore their behavioural pattern 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 pro-actively initiate measures to improve academic environments.

This research work is motivated by the desire to design and implement a Web based tool named as EDMiner (Education Data Miner) for exploring the Internet usage behaviours of students and extracting relationships between their Internet usage patterns and academic performance. One of the primary motivations 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 and enable institutions to initiate proactive measures, if required.

This paper is organized in seven sections. Section 2 presents related works and tools which are available in market both commercial and free or open source tools for analysis of data. In Section 3, describes our proposed Website Classification scheme and its mapping to Open Directory Project (ODP) [11]. Section 4, presents the architectural design and the main components of EDMiner tool. Section 5 provides definitions of outliers, their detection and co-relation between Internet usage behaviours and academic performance. Section 6, shows some of the functionalities of EDMiner tool. Section 7, concludes the paper.

II. Related Works

Several research studies [1 ~12] 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 [1~5], click trace [6], questionnaires [7 and 8], interviews [9] and other relevant documents [10~12].

Several data mining tools have been developed to analyse the data. These tools vary in terms of the input data which they use to analyse. The contents used by these tools include textual data, multimedia contents or unstructured Web data, etc. These tools are available in both free [13~15] or open source and commercial forms [16]. 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 analysing data related 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 or medical problems and to study the relationships between different groups of outliers.

III. Website Classification Scheme

Web page classification or categorization is the process of assigning a Webpage to one or more predefined category labels. In Website classification, categorization can be done based on Website's content.

Most of the general purpose search engines and portals use the Website classification scheme of Open Directory Project (ODP) [17], also known as DMOZ. These search engines and portals include Google, Netscape 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.

To evaluate the suitability of this classification scheme to model the behaviour of students' community, we analyzed Web access log files of Motilal Nehru National Institute of Technology (MNNIT) Allahabad, India for a period of three months. During that period, 3496 students, who were authorized to use Internet, visited approximately fifty thousand unique Websites. The distribution of these Websites in terms of first level categories of ODP shows that, the dominant categories are Society (26%), References (22%), Business (15%) and Computer (10%). Combining these categories amounts to 73% of the Websites visited by the students.

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. 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. The contents which are related to courses were mostly under Science, Computers and Reference categories. For example, 75% of the visitors of the popular Websites have visited the Websites for downloading different tools such as Web browsers, anti-virus or software packages, etc. Accordingly we defined a generalization of curricular category for these three categories.

IV. Tool Design

A. System Architecture

In this section, we give a high level architecture overview of EDMiner system and also describe the major components of the tool with presenting the system in terms of modules. Figure 1 depicts the overall 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: 1. Data resources (responsible for handling various input resources for the tool), 2. Pre-processing component (responsible for data selection, cleaning and integrating data from different sources), 3: Data Mining & Pattern Discovery component (in this component, the data is transformed to knowledge).

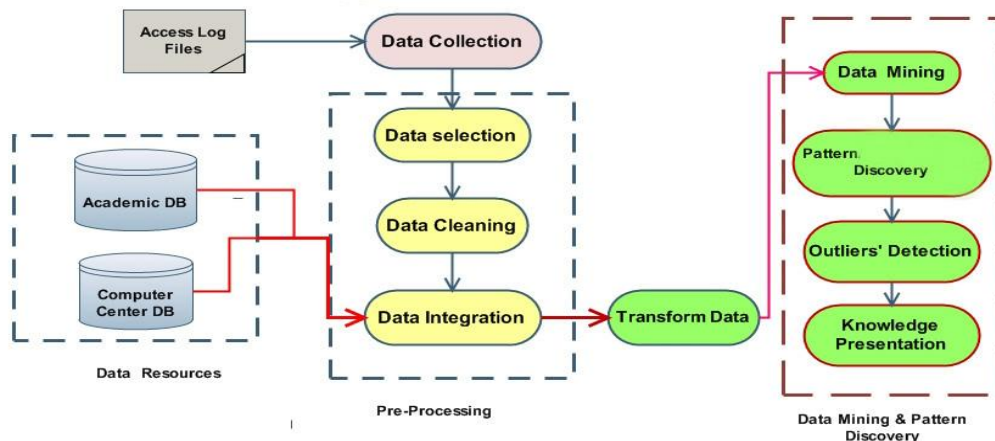


Figure 1: EDMiner System Architecture

Data resources component is responsible for collecting data from various data sources for inputting to the Pre-processing component. The input data to the EDMiner tool includes proxy server access log files, academic database information and computer center database information.

Pre-processing component is responsible for data selection, cleaning and integrating data from different sources. For example, in data selection, we extracted user id, time of access and URL of visited Website fields from the proxy server access log files for further analysis. Data cleaning and data integration removes records with inconsistent or missing values and combines data.

Data Mining & Pattern Discovery component is responsible for transforming the data into knowledge. This component extracts the following knowledge from the data.

- In data transform, each user's daily average time spent (in minutes) and the total number of visited Webpages (hit count) is extracted.
- With the help of data mining techniques, Internet usage pattern for each user is extracted.
- With the help of k-means and density based algorithms, outliers based on time-spent on Internet per day, number of visited Webpages per day and CPI has been extracted.

With the help of visualization techniques, we visualize our analysis results in tabular and pictorial representation with the help of a Web browser for making more understandable for academic people including deans, course instructors and administrators.

B. EDMiner Modules

From the usage point of view, EDMiner can be seen having four different functional modules. These are pre-processing, database, Internet usage patterns extractor and outlier detection modules.

- **Pre-Processing Module:** Data pre-processing in a data mining process deals with the preparation and transformation of the initial data. Data pre-processing further includes data cleaning, data integration, data transformation, data reduction and selection. This module involves extracting the fields and records from proxy server access log files by cleaning data. Data cleaning removing records with inconsistent or missing values. The fields which are extracted for the analysis includes user id, time of connection, URL of visited Webpage. These fields are described below. The other data files which are used were students' academic details which includes registration, CPI related information, data from Computer Center (CC), which includes Internet access details, students other activities and medical reports. These data files are also cleaned and loaded into the database for further analysis.
- **Database Module:** 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.
- **Internet Usage Patterns Extractor Module:** During the analysis part, 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.

- **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.

V. Outliers

Several definitions or descriptions of outliers [18~21] are defined. 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 he or she may belong to. In EDMiner, we use centroid and density based clustering methods to identify outliers based on average time spent on Internet per day and academic performance (CPI). We used Rapid-Miner and Weka for examining the results of clustering made by our tool. This section further continues the discussion of outliers' detection along with the results of our analysis. We have identified the following outliers:

- Students having $CPI \leq 2.7$ or $CPI \geq 9.7$
- Students daily average time spent on Internet ≥ 457 minutes or less than 5 minutes

The summary of relationships between outliers based on CPI and average time spent on Internet are given below.

- 58% of outliers with $CPI \geq 9.7$ 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 $CPI \geq 9.7$. It implies that we cannot generalize that more time spent on Internet lead to better academic performance.
- 80% of female outliers with time-Spent ≥ 457 minutes per day on Internet, had $8 \leq CPI < 9.7$. Whereas it is true only for 23% of male outliers.
- 70% of female and 50% of male outliers with $CPI \geq 9.7$ were outliers with time-spent ≥ 457 .

VI. EDMiner Tool Implementaion

We have designed and implemented a Web based data mining tool named as Education Data Miner (EDMiner) using Java and J2EE technologies. It has the following major objectives:

- To discover the distribution of visitors during 24 hours of a day, during a semester and during examination periods.
- To identify the category of Websites visited by students and average time spent on these Websites to assist Academic and System Administrators.
- To identify outliers based on academic performance and average time spent on Internet.
- To establish relationship between these different groups of outliers.

It 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 these fields: Registration-Number, Full-Name, Program, Branch, Semester, Gender, CPI (Cumulative Performance Index).
- A text file which includes User-Id, Full-Name and department name

VI. Conclusions

We designed a tool named as Education Data Miner (EDMiner) which has an simple user interface and can be easily use by administrators, dean academics, dean students' affairs and course coordinators. This users can be use this tool with different access permissions.

EDMiner can be used as an open source tool for mining access log files of proxy servers in academic areas. This tool is programmed with Java and can process large access files up to 2 gigabytes. It provides data preparation, transformation, integration, summarization and exploration, predictive and descriptive modelling. It uses k-means and DBSCAN methods for clustering the users and for identifying outliers. For example, for a selected date, the tool can creates flowchart which shown the number of users, the most popular visited Website along with overload time in that day. It is able to extract individual users' usage pattern or history of connections along with the name of visited Websites and the category of those Websites during our favourite periods.

VII. References

- [1] B.Zhou , S.C.Hui and A.C.M.Fong ,(2006), " A An Affective Approach for Periodic Web Personalization ", In Proceedings of the 2006 IEEE/WIC/ ACM International Conference on Web Intelligence IEEE Computer Society Washington,USA, pp.284-292, doi:10.1109/WI.2006.36.
- [2] B.Zhou, S.C.Hui and K.Chang, (2004), " An Intelligent Recommender System using Sequential Web Access Patterns ", In Proceeding of The IEEE , International Conference on Cybernetics and Intelligent systems , Singapore, 1-3 December, pp.393-398, doi: 10.1109/ICCIS.2004.1460447.
- [3] John M. Pierre, (2001), " On the Automated Classification of Web Sites", Published on February 4, 2001 by University Electronic Press, Computer and Information Science, Vol. 6(2001).
- [4] K.Figl, S.Kabicher , K.Toi ,(2008), " Promoting Social Networks among Computer Science Students " , In Proceedings of the 38th ASEE/IEEE Frontiers in Education Conference, October 22-25, Saratoga Springer , NY , S1C, pp.15- 20, doi: 10.1109/FIE.2008.4720676.
- [5] Loren Terveen, Will Hill, and Brian Amento, (1999), " Constructing, organizing, and visualizing collections of topically related web resources", ACM Transaction on Computer-Human Interaction.
- [6] J. Tullio, J. Goecks, E. Mynatt and D. Nguyen, (2002), "Augmenting Shared Personal Calendars ", In Proceeding of The 15th annual ACM symposium on User interface software and technology ,ACM New York, NY, USA, (UIST '02), pp.11-20, doi: 10.1145/571985.571988.
- [7] Nathan E N. Eagle and A. Pentland,(2009), " Eigenbehaviors: Identifying Structure in Routine ", Behavioral Ecology and Sociobiology , Volume: 63, Issue: 7, Publisher: Springer, pp.1057-1066 , doi: 10.1109/ISWC.2002.1167224.
- [8] Oh-Woog Kwon and Jong-Hyeok Lee, (2000), "Web Page Classification Based on k-Nearest Neighbor Approach", ACM 5thInternationalWorkshop on Information Retrieval with Asian Languages (IRAL). New York, NY,pp.9-15.
- [9] Clare Madge and Henrietta Connor, (2002), " On-line with e-mums: exploring the Internet as a medium for research " , Article first published online: 16 DEC 2002 Royal Geographical Society (with The Institute of British Geographers), Volume 34, Issue 1,pp.92-102, DOI: 10.1111/1475-4762.00060
- [10] Chen, Q., Hsu, M., Dayal, U.(2000), " A data-warehouse/OLAP framework for scalable telecommunication tandem traffic analysis" , pp.201-210, Data Engineering, Proceedings. 16th International Conference on Digital Object Identifier: 10.1109/ICDE.2000.839413.
- [11] Jong H. Kang, William Welbourne, Benjamin Stewart and Gaetano Borriello., (2004), " Extracting places from traces of locations " , the ACM International Workshop on Wireless mobile applications and services on (WLAN) Hotspots, pp.110- 118, New York, NY, USA.
- [12] M.Q.Huynh , J-Nam.Lee and B.A.Schuldt, (2006), " The Insiders Perspectives: A Focus Group Study on Gender Issues in a Computer-Supported Collaborative Learning Environment " , Journal of Information Education volume 4, pp.237-255, ISSN-1547-9714.
- [13] <http://www.weka.net.nz>
- [14] <http://www.orange.biolab.si/>
- [15] Williams and Graham, (2011), "Data Mining with Rattle and R ", 1st Edition., 374 p. 95 illus., 80 in color, Softcover, ISBN 978-1-4419-9889-7.
- [16] <http://www.support.sas.com>
- [17] Open Directory Project (ODP). <http://www.dmoz.org/>, 2012. {Last accessed on 06-September-2012}.
- [18] Rasmussen, J. L. (1988). " Evaluating outlier identification tests", Mahalanobis D Squared and Comrey D. Multivariate Behavioral Research, 23(2), pp.189-202.
- [19] Schwager, S. J., and Margolin, B. H., (1982), " Detection of multivariate outliers ", The annals of statistics, 10, pp.943-954.
- [20] Stevens, J. P., (1984), " Outliers and influential data points in regression analysis", Psychological Bulletin, 95, pp.334-344.
- [21] V.J Hodge and J. Austin., (2004), "A survey of outlier detection methodologies ", Artificial Intelligence Review, 22: pp.85-126.

VII. Acknowledgments

We are thankful to Professor B.D.Chaudhary for helping us in preparing this paper. We also acknowledge the help rendered to us by the staff of the computer center and of the Dean (Academic Affairs) office. We were permitted to use preprocessing and filtering contents of the log files for the research purpose. The pre-processing hides the actual identity of students and replace virtual_Id instead of real User_Id.