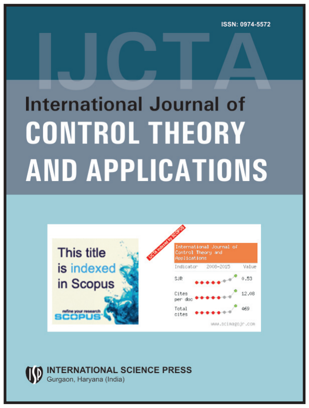

International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 25 • 2017

Behavior Mining of Male Students by Analyzing Log Files

S. Kamalakkannan[a] and S. Prasanna[b]

[a]Research Scholar, Vels University, Chennai – 600117, Tamil Nadu, India. Email: kamalindia81@yahoo.com
[b]Associate Professor, Vels University, Chennai - 600117, Tamil Nadu, India. Email: prasanna@velsuniv.org

Abstract: Internet usage behaviors of users on the Internet are one of the major study areas. Various Website organization schemes were present in the text based on diverse perspectives. Web mining is used to extort information based on the client query from the huge collected works of data accessible on the net. It is disturbed generally with its content, structure and usage. The web data include web links, web logs, objects on the web and web pages. In this paper, we describe web mining and nearby a method to make use of web mining in an enhanced way to know the users and website activities which in turn develop the web site information to draw more students by analyzing log files.

Keywords: Web usage mining, Behavior mining, Internet usage Behavior.

1. INTRODUCTION

Web Mining is used to extort information from the raw unstructured data. The rising field of web mining aims at discovering and extracting linked information that is secreted in Web linked data, in particular in content documents published on the Web. Web mining is performed in three behaviors they are: (1) web usage mining (2) web content mining (3) web structure mining. Web usage mining provides the support for the web site intend, providing a personalization server and other industry, making decision, etc. Web content mining is the development of extracting information from the content of credentials or their images. Web document text mining, store finding based on concepts indexing or manager and based knowledge may also drop in this category. Web structure mining is the method of inferring data from the World Wide Web association and links between references in the Web. Finally, web usage mining, also recognized as Web Log Mining, is the development of extracting attractive patterns in web access logs. In organize to better serve for the users, web mining applies the data mining, the artificial intelligence and the chart technology and so on to the web data and traces users visiting individuality, and then extracts the users' using the model.

We propose decayed Web mining into these sub tasks, namely

1. Resource finding: the task of retrieving proposed Web credentials.
2. Information selection and pre-processing:automatically selecting and pre-processing detailed Information for retrieving Web resources.

3. Generalization: automatically discovers universal pattern at character Web sites as well as transversely multiple sites.
4. Analysis: Validations and/or reading of the mined patterns.

1.1. Web Content Mining

Web content mining describes the mechanical search of information belongings available online, and involves mining web data contents. In the web mining field, web content mining essentially is an analog of data mining techniques for relational databases, since it is possible to discoverrelated types of information from the unstructured information residing in web credentials. The web document typically contains numerous types of data, such as text, image, audio, video, metadata and hyperlinks. Some of them are semi-structured, such as HTML documents or a more ordered data like the data in the tables or record generated HTML pages, but the majority of the information is unstructured manuscript data.

1.2. Web Structure Mining

It is that division of Web Mining which focuses on the arrangement of the Web sites and source data mainly consists of the structural information nearby in Webpages (e.g., links to other pages); typical applications are linked-based classification of Web pages, position of Web pages throughout a grouping of content and structure, and turn around engineering of Web site models.

1.3. Web Usage Mining

It is that division of Web Mining which deals with the withdrawal of data from server log files; source data generally consist of the (textual) logs that are together when the users' way in Web servers and force be represented in typical formats (e.g., Common Log Format, extensive Log Format, Log ML), distinctive applications are those based on user modeling techniques, such as Web personalization, adaptive Web sites, and user modeling.

2. LITERATURE REVIEW

The Behavior of female **students** by **analyzing log files,** here they to obtain outlook regarding the web exploitation and the behavior of female students. It determines the sample of dissimilar separation employees, individuals and standby, to decide the presentation, admission relating to web employ and the sexual orientation division[1] bythe technique for minute spent on the Internet can't provide points of attention on the difference between the wonderful understudy and Weak student. Required some more constituent as for supportive went to sites linked to educational. It utilizes federation to differentiate student.

Customer behavior using WUM in ecommerce E-business is characteristically utilized for internet shopping so need, to know the client captivity, what they have to obtain. The endeavor is an effort server log manuscript, inspecting the field like user recruitment, and some additional data missing from the customer through the program and the web. For reviewing in sequence utilized some expression like association parameter to get to the next page foreseeing for the client requires to stay. A priori algorithm is executed for typical page removal that help the site designed to show signs of development site organization. To stare at the most tremendous ahead method and occasionally went to method by the customer they used to go for hierarchy replica. Bunching utilized K-mean calculation[2].

Improving methods of preprocessing in WLM it gives abetter filtration organization of edge page, the sorting out computation is scheming every one of the pages are Frame or Sub Frame and on the off chance that it is Sub Frame, erases it one by one[3]. Filtration it gets restored in way development. The Decision tree is complete by utilizing algorithm and distinct and decide policy that enhances the ability and filtration technique.

Web logs, cleaning for mining of the Web Usage purpose of integration is on data preprocessing by charitable that field clearing and Data Cleaning figuring. Field Extraction is used to parcel fields from the log evidence that empties unnecessary data, data cleaning add up is furthermore used. Once a person has joined Facebook, he or she can examine for anyone and viewpoint the other customer profiles[4].

Web Mining – A Catalyst for E-Business In this paper, we describe web mining and present a technique to employ web mining in an improved way to recognize the users and website behavior which in turn improve the web site information to draw more users[5]. This paper also presents an impression of the variety of researches complete on pattern extraction, web content mining and how it can be taken as a method for E-business.

Data Preprocessing Method of Web Usage Mining for Data Cleaning and Identifying User navigational Pattern Web log file is a server log file which is a basic data sources in Web usage mining, in which it contain - access logs of the web server. The important task in the WUM is Data Preprocessing phase. It consists of data cleaning, user identification, session identification, path completion[6]. Data preprocessing is used to clean the irrelevant data from the log file so it can be provided for the pattern discovery to identify the user pattern.

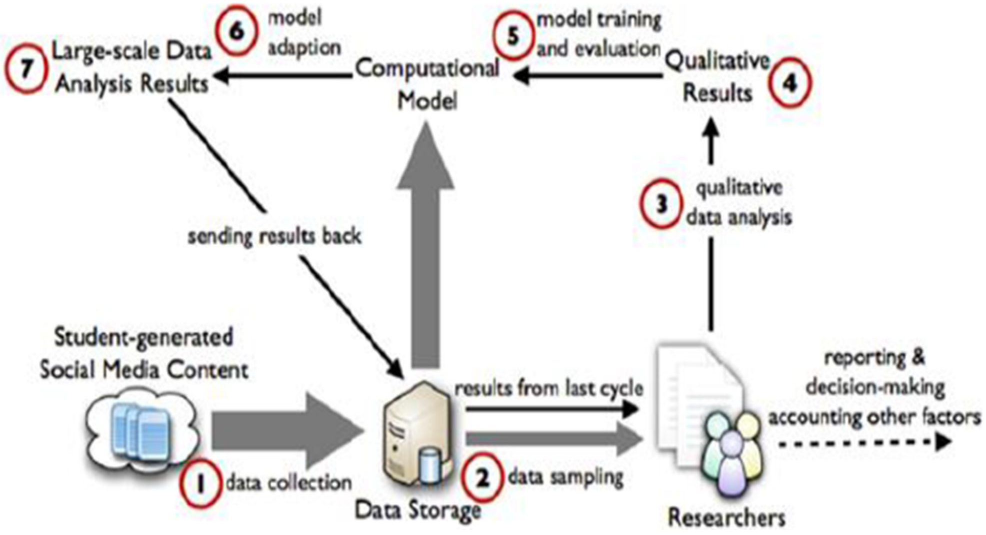
7
Large-scale Data Analysis Results
6
model adaption
Computational Model
5
model training and evaluation
Qualitative Results
4
3
qualitative data analysis
sending results back
Student-generated Social Media Content
results from last cycle
reporting & decision-making accounting other factors
1
data collection
Data Storage
2
data sampling
Researchers

Figure 1: Architecture

Mining Educational Data to Analyze Student s' Performance applied the categorization as a DM technique to review the students' presentation, create use of the decision tree method for organizing. The purpose of their study is to take out associate that characterizes students' performance in the end semester assessment. They used students' data from the student preceding file, Class test, including Attendance, Assignment marks and Seminar[7].

Clustering web transactions using a rough approximation In this paper, we present an irregular approximation-based clustering to cluster web dealings from web access logs. Using this move toward, users can successfully mine web log records to find out web page access patterns [8].

A Survey on Web Mining and Its Techniques This paper also reports the review of a variety of techniques of web mining approached from the subsequent angle like Feature Extraction, Transformation and symbol and Data Mining Techniques in various application domains [9].

3. PROCESS GAP REGARDING EXISTING METHODOLOGY

Educational data mining uses many techniques such as decision tree, rule induction, neural network, *k*-nearest neighbor, naïve Bayesian and many others. By using this technique, numerous kinds of information can be exposed such as association rules, classifications and clustering. In general, a web log can be regarded as a series of pairs of client identifier and occurrence. In this study, web log files are separated into pieces for each mining purpose. Pre-processing can be functional to the unique web log files, so that pieces of web log can be obtained. Each portion of web log is a progression of actions from one user or meeting in timestamp ascending order. We model pieces of web log as sequences of actions, and extract the sequential patterns over convinced support entrance.

4. PROPOSED METHODOLOGY

Before applying the data mining techniques on the statistics set, there should be a method that governs our effort. Figure 1 depicts the work method used in this paper, which is based on the proposed framework. The methodology starts from the problem description, then preprocessing which are discussed in the beginning and the data set and preprocessing sections, then we come to the data mining methods which are association, classification, clustering, and outlier detection, followed by the evaluation of results and patterns, finally the knowledge symbol procedure.

4.1. Preprocessing

Data preprocessing has a primary position on the Web Usage Mining application. Ref. [10] notices.

That even if preprocessing techniques are extensively used in Web Usage Mining, the literature on this issue is still quite imperfect, and that the most absolute position on preprocessing [11] dates back to 1999.The preprocessing of Web logs is usually complex and time demanding. It comprises four dissimilar tasks: (i) the data cleaning, (ii) the recognition and the rebuilding of users_ sessions, (iii) the retrieving of information about page content and arrangement, and (iv) the data formatting.

4.2. Data Cleaning

This step consists of removing all the information tracked in Web logs that are hopeless for mining purposes [12, 13] e.g.: needs of graphical page content (e.g., JPG and GIF images) and needs for any additional categorizer which capacity be incorporated into a web page or even steering sessions performed by robots and Web spiders. While requirements for graphical inside and files are simple to do away with, robots and Web spider's navigation patterns must be explicitly identified. This is usually done for instance by referring to the remote host name, by referring to the user representative, or by checking the right to use in the robots.txt file.

4.3. Defining user by using IP Address

In user identification, IP address helps to correspond to exclusive user. From this we recognized which IP address issues which web site or web pages are visited.

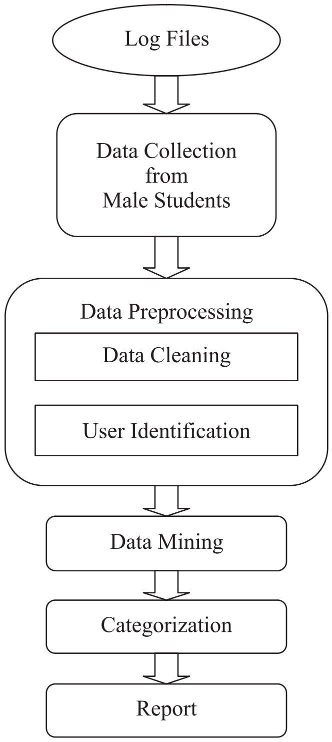

Figure 2: Methodology

4.4. Categorizing Visited URL

After preprocessing web log file the text file generated, which is used for categorization. Classification is used for categorization. Categorization equipment the substance that is grouped into categories for some specific purpose. Using Naïve Bayes classifier we defined the category of visits site.

4.4.1. Naive Bayes Multi-label Classifier

A multi-label classifier is used to categorize log files depending on the categories, residential, in prior to the simplicity examination stage. For the task, multi label classification, NB Classifier is imaginary to be customized to multi label data[7]. For multi-label classification, the conditions are somewhat more difficult, than the further classifier because every manuscript gets assigned multiple labels. Naive Bayes executes well in many complicated real-world difficulty. Even while it is often outperformed via other methods such as boosting trees, Max Entropy, Support Vector Machine, etc., NB classifier is extremely well-organized because it is less computationally and it requires a small quantity of preparation information. One accepted way to attain multi-label classification is to exchange the multi-label association trouble into multiple single-label categorization troubles.

Suppose at offer are a sum number of N words in the teaching document collection (in our case, every tweet is a manuscript) W ¼ fw_1; w_2; :::; w_{Ng}, and a total numeral of L categories C ¼ fc_1; c_2; :::; c_{Lg}. If a word when present in a category of Muncie times, and emerge in category additional than *c* for Muncie times, after that depending on the maximum probability inference, the option of this expression in a definite category *c* is

$$\mathrm{P}(wn \mid c) = \frac{m_{wnc}}{\sum_{n=1}^{\mathrm{N}} m_{wnc}} \qquad (1)$$

Similarly, the possibility of this word in various classes extra than *c*

$$P(wn \mid c') = \frac{m_{wnc}}{\sum_{n=1}^{N} m_{wnc'}} \tag{2}$$

Assume there are a total amount of M documents in the training set, and C of them is in category *c*. After that the probability of category *c* is

$$P(c) = \frac{C}{M} \tag{3}$$

and the probability of other categories *c′* is

$$P(c') = \frac{M - C}{M} \tag{4}$$

4.4.2. Naive Bayes Classifier Algorithm

This algorithm considers each sub words in the review, and accordingly classifies the reviews in different categories.

Let S be the Sentence

Step 1: Define categories $c = \{c_1, c_2, c_3, ..., c_n\}$.

Step 2: Read data from a database.

Step 3: Divide S into sub works$\{w_1, w_2, w_3, \ldots, w_n\}$ split.

Step 4: Check sub words $\{w_1, w_2, w_3, \ldots, w_n\}$ for every categories.

Step 5: if words match with categories $\{c_1, c_2, c_3, \ldots, c_n\}$ increment the counter for that categories Else put that in “other” categories.

Step 6: Find probability of each category.

4.5. Generating Report

After the data cleaning development is completed on web log file, it shows how much memory space is will be exploiting and maintain excellence of it. After data preprocessing the manuscript file uses the organization to categorize.

5. STATISTICAL RESULTS

Facebook keeps on overpowering as the most utilized site with 93% Facebook keeps on edict as the most utilized site with 93% include social networking, maintaining up a Facebook profile although utilization seems to have fallen somewhat since 2012. In any case, it’s still a great degree well known over all ages and eras. For other social networking stages, there is a slow increment in numbers utilizing LinkedIn, Instagram and Google+ and a little decrease in Twitter clients. LinkedIn use stays higher among male, those working full-time and higher paid workers. Visual stages like Instagram, Snapchat and Tumblr keep on appealing more to the more youthful age demographics; use is much lower in more than 30s. Males use Twitter compared to females while interest is significantly all the more speaking to the last mentioned.

This table depicts the social media web sites used by the male where they use it for study purposes, business, communication, etc.

Table 1
Percentage of usage of social media sites

Facebook	*LinkedIn*	*Instagram*	*Google+*	*Twitter*	*Pinterest*	*Snapchat*	*Tumblr*	*Vine*	*Yelp*	*Foursquare*
93%	28%	26%	23%	17%	17%	15%	5%	3%	3%	1%

Table 2
Percentage of Social media sites used by male

Social networking sites used	*Male*
Facebook	92%
Linkedln	35%
Instagram	22%
Google+	21%
Twitter	23%
Pinterest	8%
Snapshot	14%
Tumblr	6%

6. CONCLUSION

In this paper, a study on the web usage of social media websites and the behavior mining of the male students by analyzing the log files are examined. This replica helps the for data cleaning and categorize the web site so recognize the user concerned web site and per user need or request. By using the Navie Bayes algorithm of classification it helps easily to categorize users visited web site and provide better efficiency and performance as compared to other algorithms.

REFERENCES

[1] Rozita Oskouei, (2010). Behavior Mining of Female Students By Analyzing Log Files.

[2] Mahendra Pratap Yadav, Mhd Feeroz, Vinod Kumar Yadav, (2012). Mining The Customer Behavior Using Web Usage Mining In E-Commerce.

[3] Huaqiang Zhou, Hongxia Gao, Han Xiao, (2010). Research On Improving Methods Of Preprocessing In Web Log Mining.

[4] Theint Theint Aye, (2011). Web Log Cleaning For Mining of Web Usage Patterns.

[5] Abdul Rahaman Wahab Sait and Dr. T. Meyappan, (2012). Web Mining – A Catalyst for E-Business

[6] Wasvand Chandrama, Prof. P.R. Devale, Prof. Ravindra Murumkar, (2014). Data Preprocessing Method of Web Usage Mining for Data Cleaning and Identifying User navigational Pattern

[7] Mohammed M. Abu Tair, Alaa M. El-Halees, (2012). Mining Educational Data to Improve Students' Performance: A Case Study

[8] Supriya Kumar De, P. Radha Krishna, (2004). Clustering web transactions using rough approximation

[9] P. Menaka MCA., M. Phil, A. Prathimadevi, (2015). A Survey on Web Mining and Its Techniques

[10] C.R. Anderson, Amachine learning approach to web personalization, Ph.D. thesis, University of Washington, 2002.

[11] R. Cooley, B. Mobasher, J. Srivastava, Data preparation for mining World Wide Web browsing patterns, Knowledge and Information Systems 1 (1) (1999) 5–32.

[12] B. Diebold, M. Kaufmann, Usage-based visualization of web localities, in: Australian symposium on information visualisation, 2001, pp. 159–164.

[13] P.-N. Tan, V. Kumar, Modeling of web robot navigational patterns, in: WEBKDD 2000—Web Mining for ECommerce—Challenges and Opportunities, Second International Workshop, 2000.