EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

# Knowledge Extraction Using Web Usage Mining

Muhammad Waqas[1], Maria Iram[2], Sara Shahzad, Sidra Arshad, and Tahir Nawaz

1 University of Lahore, Sargodha, Pakistan, hwr44ever@gmail.com
2 University of Lahore, Sargodha, Pakistan, merum2@gmail.com

## Abstract

Web log files are the greatest source of knowledge now days, which keeps all the information about users interaction to web. This interaction provides us the usage patterns of the user by using web usage mining. These files contain all the information about visitors of the web which is used as input for analysis. These files are converted to required formats after completing the preprocessing so Web Usage Mining (WUM) techniques can apply on these logs. Web usage mining gives us the details of user patterns. In this study we are going discover different behaviors patterns from the web proxy server log file of an educational organization with web usage mining technique. Results are based on the interest of users towards educational websites.

**Keywords:** Data Mining, Web Usage Mining, Web Mining, Knowledge Extraction, KD, WWW

Received on 22 December 2017, accepted on 10 January 2018, published on 13 April 2018

doi: 10.4108/eai.13-4-2018.154551

## 1. Introduction

Data mining purpose is to predict unknown, useful and understandable patterns from huge data. Data mining general steps are displayed in figure 1. Web mining is an application of predicting knowledge from web log files. Because of the complex infrastructure and scalability of web it had led to numerous quality data issues like identification of page view, user and filtering robot activity[1].

World Wide Web (WWW) possess huge amount of data and growing exponentially with respect to time and usage. Web has become complex for end user to browser effectively. Maintaining is as important as to building. To improve and update we need to know our user interests so we update our website according to user web surfing [2]. Maintaining a web site may include improving in design which can be known user patterns. These patterns help us identify the user's interests on these websites. By visiting a website users accomplished different tasks such as viewing, buying of product as well as user can register for online courses and can attend classes online. By analyzing of web logs interaction of user with web can provide different kind of useful information which can help in enhancing web in

means of efficiency and effectiveness. By browsing through a website, users complete different tasks, such as buying products, registering for classes, and attending classes online. Analysis of an interaction log file can provide useful information that helps a website engineer in enhancing the website structure in a way that will make the website usage easier and faster in the future.

## 2. Web Mining

Data mining has different application and web mining is one of the most common technique which extract knowledge from web log data. Because of the complex infrastructure and scalability of web it had led to numerous quality data issues like identification of page view, user and filtering robot activity [1]. The extracted knowledge quality results are based on the selected algorithm criteria. Web mining is further divided into following three categories:

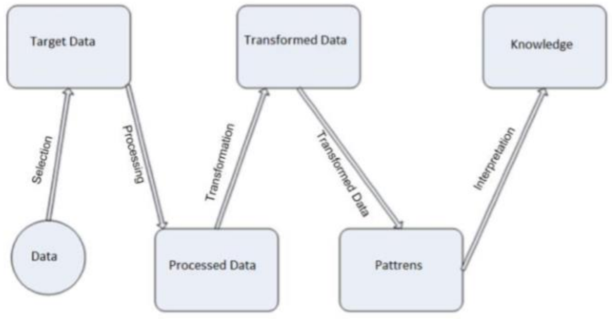


Fig. 1. Data Mining Process [3]

1. Web Usage Mining
2. Web Content Mining
3. Web Structure Mining

## 3. Web Usage Mining

WUM is an application of data mining technology to mining the data of the Web server log files [4]. WUM is defined as applying data mining techniques to log interactions between users and a website[5]. WUM also known as web log mining is the application of data mining technique on web log repositoriesto discover useful knowledge about users behavioral pattern. Data source of WUM are textual log files gathered at web servers. Log records possess a lot of useful information like IP address, URL and Time [6]. WUM mainly consist of two major techniques, statistical analysis, and association rule, clustering, and sequential patterns which is advance form of web mining [7]. Both techniques require huge data gathered from different sources such as proxy servers, web clients, and web servers [8]. Other sources like web application data can also be used [9] first statistical approach gives common and consolidated estimated statistical usage, While in the second technique provide help to identify the user patterns. WUM has four stages as to data mining which are discussed above.

## 4.Web Content Mining

This web mining technique refers to discovery of knowledge about collections of main traditional multimedia documents objects such as audio, image, text, and video, embedded in our web page or linked to our page [7]. WCM has two approaches, Agent and database based. Agent based approach comprises of personalization web agent, filtering information and its categorization and intelligent search-agent. Database approach consists of web Query System and Multilevel database [10].
There are number of existing techniques to extract knowledge through web mining.

## 5. Web Structure Mining

In this technique we extract knowledge from the links on the web and from organization. Its basically works on the web hyperlink structure, And with technique help graph structure is made which usually provide authoritativeness, or ranking, and improve page search results by filtering [11].

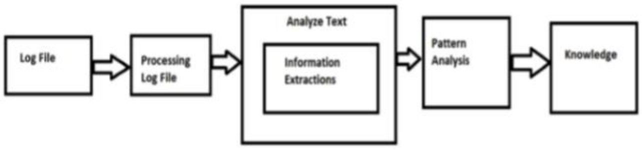
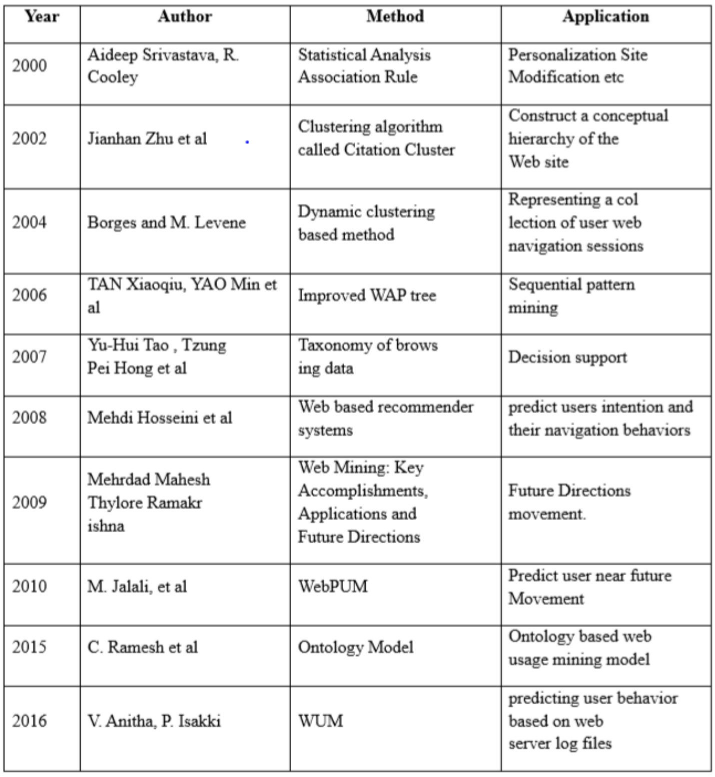


Fig. 2. Web Usage Mining Process [12]

## 6. Background Study

In this section you can find different proposed techniques and objectives which can be achieved by using web mining techniques. Wu et. al, proposed a technique for WUM to find out the grid computing environment by clicking pattern [13]. Aghabozorgi et. al, presented the idea of uzzy clustering incremental clustering of WUM in [2]. Inbarani et. al, proposed Rough set based on feature selection for WUM in [5]. Ladekar A. Pawar A. et al. [14] gave details of a widely used algorithm in web mining, which amends output’s draft of association rule mining.

Table 1. Web Using Mining Works



| Year | Author | Method | Application |
|---|---|---|---|
| 2000 | Aideep Srivastava, R. Cooley | Statistical Analysis Association Rule | Personalization Site Modification etc |
| 2002 | Jianhan Zhu et al | Clustering algorithm called Citation Cluster | Construct a conceptual hierarchy of the Web site |
| 2004 | Borges and M. Levene | Dynamic clustering based method | Representing a col lection of user web navigation sessions |
| 2006 | TAN Xiaoqiu, YAO Min et al | Improved WAP tree | Sequential pattern mining |
| 2007 | Yu-Hui Tao , Tzung Pei Hong et al | Taxonomy of brows ing data | Decision support |
| 2008 | Mehdi Hosseini et al | Web based recommender systems | predict users intention and their navigation behaviors |
| 2009 | Mehrdad Mahesh Thylore Ramakr ishna | Web Mining: Key Accomplishments, Applications and Future Directions | Future Directions movement. |
| 2010 | M. Jalali, et al | WebPUM | Predict user near future Movement |
| 2015 | C. Ramesh et al | Ontology Model | Ontology based web usage mining model |
| 2016 | V. Anitha, P. Isakki | WUM | predicting user behavior based on web server log files |

Parvatikar S. and Joshi B. [3] this paper concentrated on Web Usage Mining is the client route designs and their utilization of web assets. The distinctive stages engaged with this mining procedure and with the relative examination between the example disclosure calculations Apriori and FP-development calculation. Information Preprocessing is one of the essential undertakings previously applying mining calculations. It changes over the crude log record into client session. In this work, we have quickly presented log document preprocessing and executed it in a CTI log record. Likewise, we create the rundown of the client session document. We have utilized separating system to expel minimum asked for assets.

Deepa and Raajan [10] implemented the preprocessing techniques to convert the log file into user sessions which are suitable for mining and reduce the size of the session file by filtering the least requested pages using the preprocessing technique. Data Preprocessing is one of the important tasks before applying mining algorithms. It converts the raw log file into user session. In this work, we have briefly introduced log file preprocessing and implemented it in a CTI log file. Also, we produce the summary of the user session file. We have used filtering technique to remove least requested resources. Researches whose only focus is to create a personalized website misses the effects of the web pages content. Adding this content to the knowledge of users patterns gives broader view for personalizing web. Author explores users searching web usage patterns relation with queries [15]. A site-keyword graph is formed based on these two attributes based on which recommendations are generated for the new users. Improving personalization of web usage is mining is also aim of [16].

## 7. Proposed Work

In this system our aim is to find the usage patterns of users of the University of Lahore, Sargodha campus. Client request the web pages which are stored on the proxy server in log files. By exploring these log files by applying web usage mining algorithms and techniques we will get user interest over the world wide web. In this system we have divided our users into two groups Students, and Faculty. Result of this study will help this organization to provide better facilities to the users.

## 8. Methodology

Web Usage Mining requires huge data gathered from different sources such as proxy servers, web clients, and web servers [8]. Our study focuses on the web usage of educational organization, so we got our log file from proxy server of the organization. Figure 4 shows the implemented idea to get interesting results.

### Data Collection and Attribute Selection

In Web Usage Mining we need data of the browsed web pages from the institute web log server. Web proxy server is the best source because all the web request are logged there in a log file. There are as many as 65 different variables to work on. Five variables were selected to achieve desired result of this study. These attributes are shown in figure 4.

### Nave Bayes Classifier

This algorithm is Bayes Theorem based. It is a collection of algorithms rather than a single algorithm. But these algorithms share independent classified of any other feature. In this study this algorithm first finds the identity of the users which are faculty/staff or students. Further it classifies the accessed URLs and classifies them in term of their percentage of categories wise.

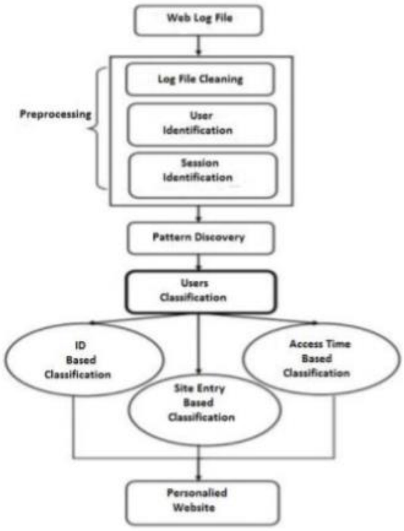


**Fig. 3.** Users Classification through WUM [17]

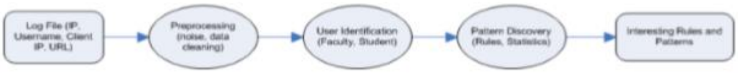

Log File (IP, Username, Client IP, URL)
Preprocessing (noise, data cleaning)
User Identification (Faculty, Student)
Pattern Discovery (Rules, Statistics)
Interesting Rules and Patterns

**Fig. 4.** Proposed Idea

### JRip

Its one of the most popular and basic algorithm. It has set of rules which examine classes in growing size and generate incremental reduce error class proceeded by all particular decision by treating all example and also find new set of rule which cover all class members. Then it proceeds to next class and repeat it until all classes are discovered.

### Results

As this chart in figure 5 clearly shows 65.98% are the students who Uses University provided internet facility to get connect to web. Due to this factor institute has to take some decision to provide facility where they can find their course related materials easily in less time.

**Table 2.** User Ratio of Data

| Web Type | Faculty/Staff | Student |
|---|---|---|
| **Education** | 87 | 66 |
| **Entertainment** | 7 | 4 |
| **Facebook** | 31 | 59 |
| **Google** | 38 | 68 |
| **MSN** | 13 | 52 |
| **Youtube** | 16 | 64 |
| **Other** | 229 | 631. |
| **Rate** | 34% | 66% |

**Table 3.** Web Accessed Pages

| **Education** | **Entertainment** | **Google** | **Facebook** | **Yahoo** | **MSN** | **Other** |
|---|---|---|---|---|---|---|
| 11.77% | 0.84% | 8.19% | 6.98% | 6.28% | 5.03% | 44.67% |

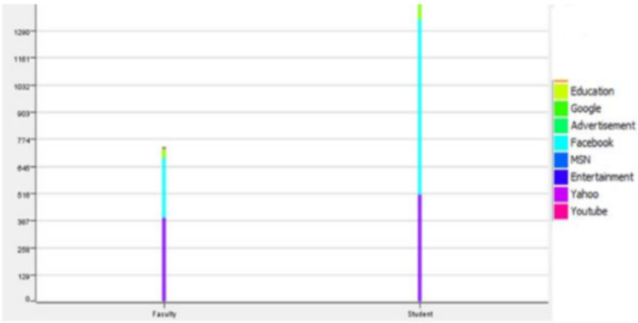


**Fig. 5.** Users Classification through WUM

In table 2 a category wise comparison is shown of our two types of users. It is clearly shown that faculty has accessed educational websites more than student by a margin of 31%. While in all other categories student usage is greater than Faculty/Staff users. It shows the trend of our users in our case Faculty and Student interests of this institute. While in all other categories student usage is greater than Faculty/Staff users. It shows the trend of our users in our case Faculty and Student interests of this institute.

## Conclusion

Web usage mining is web mining technique which is useful to find out the trends of user towards web. By which we can explore and meet our end user needs. This technique is also equally beneficial to find out any organizational needs. In this study we choose an educational organization to find out our stakeholders. In this experiment we found that for this specific organization Faculty is more attracted towards educational related websites, whereas student used more internet services as much as 66%. Our results shows educational websites accessed by faculty and student usage is 69% and 31% respectively. Students used organizational provided facilities on other stuff rather than the educational materials.

## Future Work

This study is the initial study towards improvement for providing better facilities to student, but it requires much more deep study in following areas. Preprocessing the web logs is not easily available so we need more compact and precise preprocessing technique in order to make our data much more meaningful. This data can be used to identify the usages in different departments of university campus so each department can improve their facility according to their students need. Data is useful to discover hyperlink structure.

## References

1. M. Hogo, M. Snorek, and P. Lingras. Temporal web usage mining. In *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, pages 450 453, Oct 2003.
2. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorer. Newsl.*, 1(2):1223, January 2000.
3. S. R. Aghabozsssorgi and T. Y. Wah. Using incremental fuzzy clustering to web usage mining. In *2009 International Conference of Soft Computing and Pattern Recognition*, pages 653658, Dec 2009.
4. H. H. Inbarani, K. Thangavel, and A. Pethalakshmi. Rough set based feature selection for web usage mining. In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, volume 1, pages 3338, Dec 2007.
5. M. Jalali, N. Mustapha, N. B. Sulaiman, and A. Mamat. A web usage mining approach based on lcs algorithm in online predicting recommendation systems. In *2008 12th International Conference Information Visualisation*, pages 302307, July 2008.
6. Zhang Huiying and Liang Wei. An intelligent algorithm of data pre-processing in web usage mining. In *Fifth World Congress on Intelligent Control and Automation (IEEE Cat. No.04EX788)*, volume 4, pages 31193123 Vol.4, June 2004.
7. Federico Michele Facca and Pier Luca Lanzi. Mining

interesting knowledge from weblogs: a survey. *Data & Knowledge Engineering*, 53(3):225  241, 2005.

8. D. Dong. Exploration on web usage mining and its application. In *2009 International Workshop on Intelligent Systems and Applications*, pages 14, May 2009.
9. K. Etminani, A. R. Delui, N. R. Yanehsari, and M. Rouhani. Web usage mining: Discovery of the users' navigational patterns using som. In *2009 First International Conference on Networked Digital Technologies*, pages 224249, July 2009.
10. Wang Bin and Liu Zhijing. Web mining research. In *Proceedings Fifth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2003*, pages 8489, Sept 2003.
11. C. H. Wu, Y. L. Wu, Y. M. Chang, and M. H. Hung. Web usage mining on the sequences of clicking patterns in a grid computing environment. In *2010 International Conference on Machine Learning and Cybernetics*, volume 6, pages 29092914, July 2010.
12. Faustina Johnson and Santosh Kumar Gupta. Article: Web content mining techniques: A survey. *International Journal of Computer Applications*, 47(11):4450, June 2012. Full text available.
13. S. P. Nina, M. Rahman, K. I. Bhuiyan, and K. E. U. Ahmed. Pattern discovery of web usage mining. In *2009 International Conference on Computer Technology and Development*, volume 1, pages 499503, Nov 2009.
14. Tzung-Pei Hong, Ming-Jer Chiang, and Shyue-Liang Wang. Mining weighted browsing patterns with linguistic minimum supports. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 5 pp. vol.4, Oct 2002.
15. T. Murata and K. Saito. Extracting users' interests from web log data. In *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*, pages 343346, Dec 2006.
16. Massimiliano Albanese, Antonio Picariello, Carlo Sansone, and Lucio Sansone. A web personalization system based on web usage mining  techniques. In *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers &Amp; Posters*, WWW Alt. '04, pages 288289, New York, NY, USA, 2004. ACM.
17. A. Bhargav and M. Bhargav. Pattern discovery and users classification through web usage mining. In *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, pages 632636, July 2014.