International Journal of Advanced Trends in Computer Applications
www.ijatca.com

Web Usage Mining: A Survey

Jaspreet Kaur[1], Rupinder Singh[2]

[1] Jaspreet Kaur
Chandigarh University
Mohali (Punjab)
Student
jass_dhaliwal60@ymail.com

[2] Rupinder Singh
Chandigarh University
Mohali (Punjab)
Assistant Professor
rupipanjgotra1989@gmail.com

Abstract: *The use of web applications is increasing enormously and so are its users. Most of the businesses these days are running through online applications. So it is very important to access the basic information and the interests of the users. Web usage mining is the application of data mining that deals with extracting useful information from the websites. For this, log files are maintained by web servers which contain information about User Name, IP Address, Time Stamp, Access Request, Number of bytes transferred, result Status, URL, and User Agent. This paper presents survey of web usage mining, its various types, phases of web usage mining, web usage mining tools, attacks in web usage mining and applications of web usage mining.*

Keywords: *Web Usage Mining, Log files, Preprocessing, Pattern discovery, Pattern analysis*

1. Introduction

Now-a-days internet has become an essential part of our daily life. In the past decade, there is an outstanding growth in number of websites & visitors. Because of this, large amount of data has been generated [5]. Data mining involves analysis of data sets to find unsuspected relationships and to summarize the data. Web mining is one of the most noteworthy fields in the area of data mining. Web Mining is the application of data mining techniques to Web data, which can be Web document content or hyperlink structure or Web log file, to discover and mine the undiscovered knowledge and useful patterns. Web mining is divided in the following three categories: Web content mining is the extraction, mining and integration of useful data, information and knowledge from Web page content. Web content mining is differentiated from two different points of view: Information Retrieval View and Database View. Web structure mining focuses on the underlying structure of the websites. It can be used to categorize web pages and is useful to generate information such as the similarity and relationship between different web sites. Web usage mining is the process of finding out what users are looking for on the Internet. Web Usage Mining is the application of data mining techniques to find out interesting usage patterns from web data in order to understand and better serve the needs of Web-based applications.

1.1 Web usage mining

Web usage mining uses data mining procedures to access the web data. This technology is mainly concentrated upon the use of web technologies. [4]The world's largest portal like yahoo, msn etc., needs a lot of insights from the activities of their users web visits. It will be difficult to structure their monetization efforts without this usage report. The main areas of research in this domain are Web log data preprocessing and identification of useful patterns from this preprocessed data using mining techniques. Web Usage mining has straight impact on businesses. For this purpose, log files are maintained by web servers. Log file is a text file which keeps record of the requests that are submitted by the user to the server while accessing a website. [3]. Log files are used to list the actions that have been occurred and contain information about User Name, IP Address, Time Stamp, Access Request, number of Bytes

Transferred, Result Status, URL that Referred and User Agent. Analysis of these files gives a neat idea about the user. Web usage mining involve following steps:

1.1.1 Data Collection

There are three main sources from where data can be collected i.e. a) Server log files b) Proxy log files c) Client Log fie.

a) Server log file

The most widely used source for web usage mining is web server log data [11]. This web log data is generated automatically by web server whenever a user request arrives, which contains all information about visitor's activity. The frequent server log file types are access log, agent log, error log and referrer log.

b) Proxy log fie

A Proxy server is an intermediate server that exists between the client and the Web server. Therefore if the Web server gets a request of the client via the proxy server then the entries to the log file will be the information of the proxy server and not of the original user. To gather the information of the user, these web proxy servers maintain a separate log file.

(c) Client log file

This type of log fie is present at the client side but still, all the entries are made at the server side. Client log files include activities and events that happen within the premises of client machine.

1.1.2 Data Pre-processing

The raw data collected from the previous steps may contain noise, impurities or may be unformatted. So, data pre-processing is performed. The various data pre-processing methods are:

a) Data cleaning and feature selection: This method is used to remove unnecessary/irrelevant fields from the raw data. For example- removing JPEG, GIF, JPG files and audio/video files because they are not executed on the basis of user's request. Even the entries which are made to mark the unavailability of any resource or page are also removed. The entries occurred from crawlers or spiders also need to be removed because they do not show the way how users navigate through the web sites.

b) User Identification: User Identification means identifying a unique user. In most cases log files provides IP addresses. In other cases, log files also contain user login [7]. When user login is not available then IP addresses are used along with type of OS and browsing software.

c) Session Identification: A user session is defined as a set of pages visited by the same user inside the duration of one particular visit to a website [4]. A user may have a single or multiple sessions during a period. When a user has been identified, the click stream of each user is divided into logical clusters. The method of portioning into sessions is called as Sessionization or Session Reconstruction. There are mainly three methods in session reconstruction among which two methods depend on time and one depends on navigation. The time oriented methods are simple. The set of pages visited by a specific user at a specific time is called page viewing time. It varies from 25.5[8] minutes to 24 hours [9] while 30 minutes is the default timeout by Cooley. The second method depends on page stay time which is calculated with the difference between two timestamps. If it exceeds 10 minutes then the second entry is assumed as a new session. Time based methods are not reliable as the users may involve in some other activities after opening the web page and factors such as busy communication line, loading time of components in web page, content size of web pages are not considered. Third method based on navigation uses web topology in graph format. It considers web page connectivity; however it is not necessary to have a hyperlink between two consecutive page requests. If a web page is not connected with previously visited page in a session, then it is considered as a different session.

1.1.3 Pattern discovery

It is the decisive stage where some useful knowledge will be derived by applying various statistical and/or data mining techniques at hand from various research areas like data mining, machine learning, statistical method and pattern recognition. Following methods are used for pattern discovery process:

a) Statistical Analysis

It is the most general method to take out knowledge about visitors to a web site. We can perform different kinds of expressive statistical analyses like mean, median, mode, frequency etc on variables such as page visit, the time of visit and navigational path length. There are various web traffic analysis tools produce which generate an intervallic report containing statistical information such as the most commonly accessed pages, average view time of a page or length of navigational path.

b) Association Rules

It is a procedure for finding frequent patterns, correlations and associations among sets of stuffs and it is used to relate pages that are most frequently located together in a single server session. Association rules are used in order to disclose correlations among pages accessed together throughout a server session. Those types of rules point out the possible relationship between pages that are often viewed together even if they are not

directly connected, and can disclose associations between groups of users with specific interests.

c) Clustering

Clustering is used to group together a set of items that have similar characteristics. In the Web Usage Mining, there are two kinds of interesting clusters to be discovered user clusters and page clusters. User clustering results in groups of users that seem to behave similarly when navigating through a Web site and Page clustering identifies groups of pages that appear to be conceptually related according to the user's perception.

d) Classification

Classification is the process of mapping a data into one of several predefined classes. In the Web area, one is interested in developing a users profile belonging to a particular category or class. These necessitate selection and extraction of features that best explain the properties of a known class or category. Classification can be done by using supervised inductive learning algorithms such as k-nearest neighbor classifiers, Vector Machines, decision tree classifiers, naive Bayesian classifiers etc.

e) Sequential Patterns

Sequential patterns indicate the correlation between transactions. The method of sequential pattern discovery challenged to find inter-session patterns such that the presence of a set of objects is followed by another object in a time-ordered set of episodes or session. With the help of this approach, Web marketers can forecast future visit patterns which will be helpful in placing advertisements intended at certain user groups.

f) Dependency Modeling

Sequential patterns indicate the correlation between transactions. The method of sequential pattern discovery challenged to find inter-session patterns such that the presence of a set of objects is followed by another object in a time-ordered set of episodes or session. With *the* help of this approach, Web marketers can forecast future visit patterns which will be helpful in placing advertisements intended at certain user groups.

1.1.4 Pattern analysis

Results of pattern discovery might not be in the form for real interpretation. Pattern analysis provides ways to contrast the results and to extract interesting rule or pattern from output of Pattern discovery. For this purpose, [11] various visualization and presentation tools are used which represent data in 2D, 3D pictorial representation. These tools compare and characterize result in terms of charts, graphs, tables, Wien diagram and so many others visual presentations. Most of the times result generated or data itself are stored in data cubes or in data ware house.

2. Literature Survey

The focus of literature review is to study the various steps to be performed in web usage mining and the various threats to web usage mining. Websites attacks are also very common. These attacks are implemented using information available from the web log files.
In 2000, Jaideep Srivastava et al. gave brief overview of web usage mining with the help of webSIFT which performs web usage mining in the standard NSCA format. [1] The webSIFT system provides the option of converting server session into episodes. The authors also explained the possible ways to protect the data and for maintaining individual identity.

Chintan R. Varnagar et al. [2] also discussed the work done so far on data collection and pre-processing stage of web usage mining. The authors presented the main sources of data i.e. Client Log File, Proxy Log File, Web Server Log File. The various data pre-processing techniques such as data cleaning/feature extraction user identification and session identification were also by the authors.

Sanjeev Dhawan et al. in their research work concluded web usage mining as an important research domain of web mining [6]. The author discussed various fields of web log files such as date, time, server IP, server method, client IP, number of bytes transferred, session, etc. Log files usually contain noisy and irrelevant data. Therefore preprocessing is done to remove unnecessary data from log file. User and session identification is also included in preprocessing step which includes identifying users and sessions. Pattern discovery techniques such as association rule mining (such as Apriori), clustering and classification are applied on the reduced log file. As a final point, these are presented to the user in an appropriate format.

Doru Tanasa et al. [8] proposed a new methodology for Intersites WUM that combines the classical preprocessing steps of data fusion, data cleaning, and data structuration with a new step called data summarization. To support their methodology, the authors designed and implemented AxIS LogMiner. This software tool takes as input the different log files and outputs a MySQL database. The authors used Perl scripts to implement data fusion, data cleaning, and data structuration. Java and SQL were used for user interface and data summarization. The whole experiment was done on INRIA (the French National Research Institute for Research in Computer Science and Control) web log files.

Olfa Nasraoui et al. presented a structure mining, tracking and validating evolving multi faceted user profiles on websites that have all the challenging aspects of real-life Web usage mining, including evolving user profiles and access patterns, dynamic Web pages, and external data describing an ontology of the Web content [3].

V.Chitraa et al. proposed a new method of session identification for extraction of user patterns. The authors focused on preprocessing of web log files. They used matrix method for session identification in which the number of rows represented the users and columns represented web pages. [4] The author also discussed other methods of session identification such as graph method. In experimental work the authors proved the matrix method to be more effective in session identification.

Ankita Kusmakar et al. [5] described three types of web mining i.e. web content mining web structure mining and web usage mining. The author has discussed that increasing demand of the Web has greatly evolved the Web mining technology. Web mining techniques such as sequential pattern mining, association rules mining, clustering are also discussed by the author. Applications of web usage mining include personalizing the delivery of web content, improving user navigation through pre-fetching and caching, improving web design and customer satisfaction.

Roger Meyer in his work [9] detected various types of attacks on web log files in 2008. According to the author, there are two attack detection methods – rule based detection (static) and anomaly based detection (dynamic). The author discussed 9 different attacks on web log files such as: Cross site scripting, malicious file execution, insecure direct object reference, cross-site request forgery, etc. To detect these attacks, attack vectors have to be known in order to make detection rules and various standards and encoding variants needs to be known.

In 2012, an anomaly detection scheme was proposed by Yi Xie and Shensheng Tang to detect the Web-based distributed denial of service attack [10]. For this, a new dynamic hidden semi-Markov model was proposed to model the time-varying user-behavior which was a major challenge in the existing approaches. The authors used entropy as filter policy for HTTP request of anomalous browsing behavior. The model and algorithms were used to implement DDoS attacks detection from sets of empirical workload data from a real traffic data.

Dusan Stevanovic et al. worked on the methods for separating malicious and non malicious users. Detecting malicious web crawlers is one of the most dynamic research areas in the field of network security [11]. The authors analyzed this problem using unsupervised neural network learning. SOM and ART2 were used by the authors.

B.Naveena Devi et al. [12] also discussed the importance of web mining in the field of e-commerce. The authors introduced a web usage mining intelligent system to provide classification on user information based on transactional data by applying agglomerative clustering algorithm, and also offer a public service which enables direct access of website functionalities to the third party.

In 2012, K Sudheer Reddy et al. implemented preprocessing techniques on especially designed Web Sift (WebIS) tool on an IIS web server [13]. The authors presented two proficient algorithms, one for the retrieval of web log files and other foe joining of web log files. Moreover, these two algorithms were included in the functionality of the already developed tool WebIS. P. Sampath in [14] presented a new technique for pattern mining operations named as Systolic trees. This architecture has been proved to have high throughput and faster execution. This system also includes automatic weight estimation. It minimizes the mining time by partitioning the tree into intense and sparse parts and then sending the dense tree to the hardware.
In 2014, Javed Ashraf analyzed various machine learning techniques that can be used to handle DDoS attacks and intrusions in software defined networks [15]. The authors also discussed the pros and cons of various

3. Applications of web usage mining

3.1 Personalization of web content

This is the most important application of web usage mining as the website owners can personalize their web services based on the user interests. Recommendation Systems are the most common application in this area as their aim is to recommend interesting links to the users.

3.2 Pre-fetching and caching

The results of web usage mining can be used to improve the performance of web servers and web applications. Web Usage mining can also be used to develop Prefetching and caching strategies to reduce server response time.

3.3 Design Support

Usability is very important for websites. Web Usage Mining techniques can be used to help the website owners to tackle the design and implementation issues of websites.

3.4 E-commerce

Mining business intelligence from Web usage data is very important for ecommerce. Web Usage Mining techniques help in Customer Relationship Management (CRM). In this case, the spotlight is on business specific issues such as: customer attraction, customer retention, cross sales, and customer departure

4 Conclusion

From the literature survey, we have concluded that Web usage mining plays a very important role for the web site owners. Websites are the most important way of advertisements. Web Usage Mining helps in extracting user-access pattern which can help the website owners in number of ways such as customization of web data, design support and caching. The results of Web Usage Mining depend greatly on the pre-processing stage. So, much care should be taken while performing this step. More efficient methods need to be developed to perform pre-processing. Further, to protect the web log file data from the various attacks heuristic techniques should be used such as combination of genetic algorithms with neural networks

References

[1] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", Volume 1, Issue 2 - page 12, ACM, 2000.

[2] Chintan R. Varnagar, Nirali N. Madhak, Trupti M. Kodinariya, Jayesh N. Rathod, "Web Usage Mining: A Review on Process, Methods and Techniques", IEEE, 2013.

[3] Olfa Nasraoui, Maha Soliman, Esin Saka, "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", IEEE transactions on knowledge and data engineering, vol. 20, no. 2, February 2008.

[4] V.Chitraa, Dr.Antony Selvadoss Thanamani, "A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", International Journal of Computer Applications , Volume 34– No.9, November 2011.

[5] Ankita Kusmakar, Sadhna Mishra "Web Usage Mining: A Survey on Pattern Extraction from Web Logs", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, September 2013.

[6] Sanjeev Dhawan, Swati Goel, "Web Usage Mining: Finding Usage Patterns from Web Logs", American International Journal of Research in Science, Technology, Engineering & Mathematics, 2013.

[7] G T Raju, Nandini N, "Preprocessing of Web Usage Data for Application in Prefetching to Reduce Web Latency", International Journal of Electrical& Computer Sciences IJECS-IJENS Vol: 14 No: 04, August, 2014.

[8] Doru Tanasa, Brigitte Trousse, "Advanced Data Preprocessing for Intersites Web Usage Mining", IEEE, 2004.

[9] Roger Meyer, "Detecting Attacks on Web Applications from Log files", SANS Institute, 2008.

[10] Yi Xie, Shensheng Tang, "Online Anomaly Detection Based on Web Usage mining", 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum, IEEE, 2012.

[11] Dusan Stevanovic, Natalija Vlajic, Aijun An, "Detection of malicious and non-malicious website visitors using unsupervised neural network learning", Applied Soft Computing, Elsevier, 2012.

[12] B.Naveena Devi, Y.Rama Devi, B.Padmaja Rani, R.Rajeshwar Rao, "Design and Implementation of Web Usage Mining Intelligent System in the Field of e-commerce", Elsevier, 2011.

[13] K Sudheer Reddy, Dr. G. Partha Saradhi Varma, Dr. I. Ramesh Babu, "Preprocessing the Web Server Logs – An illustrative approach for effective usage mining", Volume 37 Number 3, ACM, 2012.

[14] P. Sampatht, C. Ramesht, T. Kalaiyarasit, S. Sumaiya Banut, G. Arul Selvan, "An Efficient Weighted Rule Mining for Web Logs Using Systolic Tree", IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012) March 30, 31, 2012.

[15] Javed Ashraf, Seemab Latif, "Handling Intrusion and DDoS Attacks in Software Defined Networks Using Machine Learning Techniques", National Software Engineering Conference, IEEE, 2014.