iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** XII **Month of publication:** December 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue XII, December 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Model Survey on Web Usage Mining and Web Log Mining

Vikas Tiwari[1], Dhruvesh Chudasama[2], Prof. Avinash Ingole[3] (Guide)
[1]*Department of Computer Engineering, BVCOEL, Pune-412115, India*
[2] *Department of Computer Engineering, BVCOEL, Pune-412115, India*
[3] *Professor Department of Computer Engineering, BVCOEL, Pune-412115, India*

Abstract - At present in our day to day life internet plays a very important role. It has become a very vital part of human life. As internet is growing day by day, so the users are also expanding at much greater rate. Users spend lot of time on internet depending on the behavior of different user. Internet provides huge amount of information and from this information knowledge is extracted for the users. This extraction of information demands for the new logics and method. The data mining techniques and applications can be used in web based applications for performing this job which is also known as web mining. Web based mining or web usage mining is one of the trending topics nowadays. When user uses internet or visits some web pages, the associated information are stored in the server log files. Using these log files of server the human nature or behavior can be predicted. This paper focus on the web based mining and how it can be can be used to predict the human behavior using the server log files. The paper contains some of the techniques and methods associated with web mining.
Keywords: Web Based Mining, Log Files, Data Cleaning, Session Identification, User Identification.

I. INTRODUCTION

The Internet or World Wide Web is growing or widening at a very high rate. As it is growing in terms of user and size, the problems associated with it are also expanding. One of the problems is the extraction of knowledge and to capture the taste of the user or to analyze his or her behavior. The data mining techniques can be used on the web based application to overcome this problem. The web based mining or web mining is one of the applications of the data mining which is used to extract knowledge from web data such as web documents, hyperlinks between documents, usage logs pf websites, etc.
Web mining can be basically divided into three categories, depending on the data to be mined.

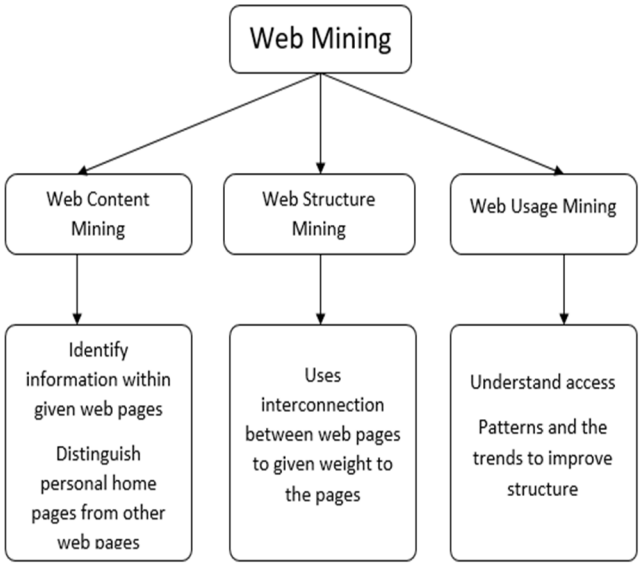


Fig: 1 Web Mining Taxonomy

A. Web Content Mining

Web Content Mining is the process in which the useful information is extracted from the content of web documents. Content data is

Volume 4 Issue XII, December 2016
ISSN: 2321-9653

the collection of facts a web page is designed to contain such as text, images, audios, videos, structured records such as lists and tables. Issues addressed in Web content mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages.

B. Web Structure Mining

Web structure mining is the process which is used to discover structure information from the web. The Structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Hyperlinks are a structure that connects a location to a different location within a page or out of the page with different one. Documents Structure the content within a Web page can also be organized in a tree structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

C. Web Usage Mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications (Srivastava, Cooley, Deshpande, and Tan 2000). Usage data captures the identity or origin of web users along with their browsing behaviour at a web site.

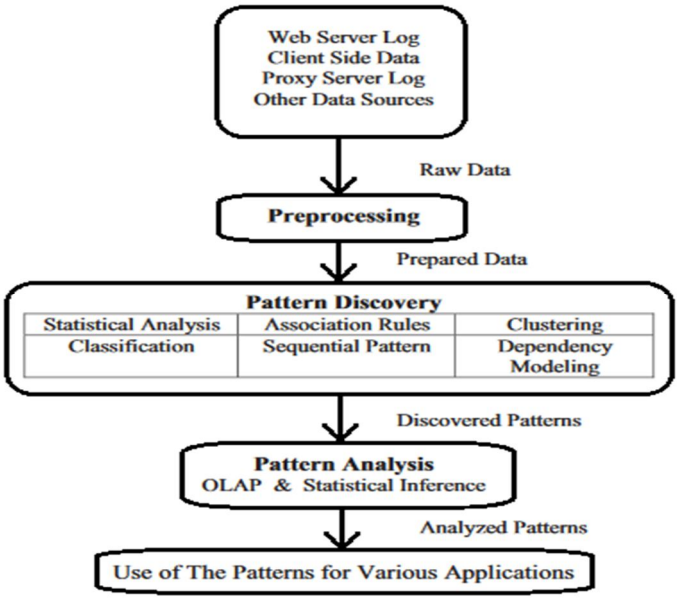


Fig: 2 Phases of Web Mining

Web usage mining can be classified depending on the kind of data usage [13]:

1) *Web Server Data:* user logs are stored in the web server and basically include IP address, page references and access time.
2) *Application Server Data:* Commercial application servers such as Web logic, 1, 2 Story Server, 3 have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.
3) *Application Level Data:* Commercial application servers such as Web logic, 1, 2 Story Server, 3 have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.
4) *Web Log Files:* Web log files are the files which contain complete information about the users browse activities on the web server by [2] [11] [4].

D. Web Log Files

Web log files are the files which contain complete information about the users browse activities on the web server by [2] [11] [4]. These web log files are created automatically by every user click to the corresponding web servers. These log files is in text format, most of the times and the size varies from 1KB to 100 MB.

There are three types of log files which are as follows:

1) *Web Server Logs:* History of web page requests is maintained as a log file. Web servers are the costly and the most common data source. They collect large volume of information in their log files. These logs contain name, IP, date, and time of the request, the request line exactly came from the client, etc. These data can be bound together as a single text file, or divided into different logs, like access log, referrer log, or error log. However, user specific information is not stored in the server logs [15]
2) *Proxy Server Logs:* It acts as an intervening level of catching lies between client browser and web servers. Proxy caching is used to decrease the loading time of a web page as well as the reduce network traffic at the server and client side. The actual HTTP request from multiple clients to multiple web servers are tracked by the proxy server [9]. The proxy server log is used as a data source for browsing behaviour characterization of a group of unauthorized users sharing a common proxy server.
3) *Browser Logs:* On client side using JavaScript or Java applets the browsing history is collected. To implement client side data collection, user cooperation is needed. Here pre-processing discussed using Web Server Logs [11]. Web server logs are used in the web page recommendation to improve the E-Commerce usability.

E. Types Of Log Files Format

Log file is a simple text file recorded from each user. This log files can be displayed in three different formats:

1) W3C Extended Log File Format
2) NCSA Common Log File Format
3) IIS Log File Format

W3C format allows user to choose properties, the user wants to log for each request. It is the default log file format in the IIS server. It can be customized, that means administrator can add or remove fields depending on what information want to record. Omitting unwanted attribute fields when log file size is limited.[11][15][14][2]

```
#Version: 1.0 #Date: 12-Jan-1996 00:00:00
#Fields: time cs-method cs-uri
00:34:23 GET /foo/bar.html
12:21:16 GET /foo/bar.html
12:45:52 GET /foo/bar.html
12:57:34 GET /foo/bar.html
```

NCSA Common Log File Format is a fixed ASCII text-based format, so the customization cannot be done. It is available for websites and for SMTP and NNTP services but not for the FTP services. Because HTTP.sys handles the NCSA Common log file format, this format records HTTP.sys kernel-mode cache hits.

```
216.67.1.91 - - [01/Jul/2002:12:11:52 +0000] "GET /
index.html HTTP/ 1.1" 200 431
```

IIS Log File Format is also a fixed ASCII text-based format, so this cannot be customized. Because HTTP.sys handles the IIS log file format, this format records HTTP.sys kernel-mode cache hits.[9][11]

```
172.16.255.255, anonymous, 03/20/01, 23:58:11,
MSFTPSVC, SALES1, 172.16.255.255, 60, 275, 0, 0, 0,
PASS./Intro.html
```

II. METHODOLOGY

In the data preprocessing, it takes web log data as input and then process the web log data and gives the reliable data. The goal of preprocessing is to choose primary features, then remove unwanted information and finally transform raw data into sessions. So to do this Data preprocessing is divided into sub processes which are known as Data Cleaning, user identification, and Session Identification [12] [2].

A. Data Cleaning

Data cleaning is the process of removing the unwanted data from the huge amount of data. So the use of data cleaning procedure is necessary to remove all the unwanted data used in data analysis and mining. The efficiency of data mining can be increased by using

data cleaning process. The cleaned data include removal of local and global noise, elimination of videos, graphic records and the format efficiency, elimination of HTTP status code records, robots cleaning.[10]

B. The Records Of- Graphics, Video And The Format Information

In every record of URI field, JPEG, GIF, CSS filename extension is found, these extensions are going to be eliminated from the web log file. But the files with these extensions are the documents embedded in the web pages. So it's not necessary to include these files while identifying the user interested web pages. This process support to identify user interested sessions.[10]

C. Failed HTTP- Status Code

In this cleaning process it will find the user's interested sessions by reducing the evaluation time. In this process the status field of every record in the web access log is checked and the status code over 299 or below 200 are removed.

D. Robots- Cleaning

It is a software tool which is used to scan a website periodically to extract all the content. This software tool is known as spider. All the hyperlinks from a web page are automatically followed by WR. The uninterested session from the log files are automatically removed by the removal of WR.[10]

E. User Identification

User identification is the process of identifying different user who are accessing the website. The main aim of this process is to gather information related to every user's access characteristics, then make user clustering and provide recommendation service for the users. Different users are identified by their different IP addresses.

F. Session Identification

The process of session identification involves the identification of different sessions of the user. A sequence of pages viewed by a user during one visit is known as the Session. The session is recorded and stored in the log file. In pre-processing it is necessary to find session of each user. It defines the number of times the user has accessed a web page. It takes all the page reference of a given user in a log and divides them into user sessions. These sessions can be used as an input data vector in classification, clustering, prediction and other tasks. Based on a uniform fixed timeout a traditional session identification algorithm is used. A new session is identified when the interval between two sequential requests exceeds the one hour.

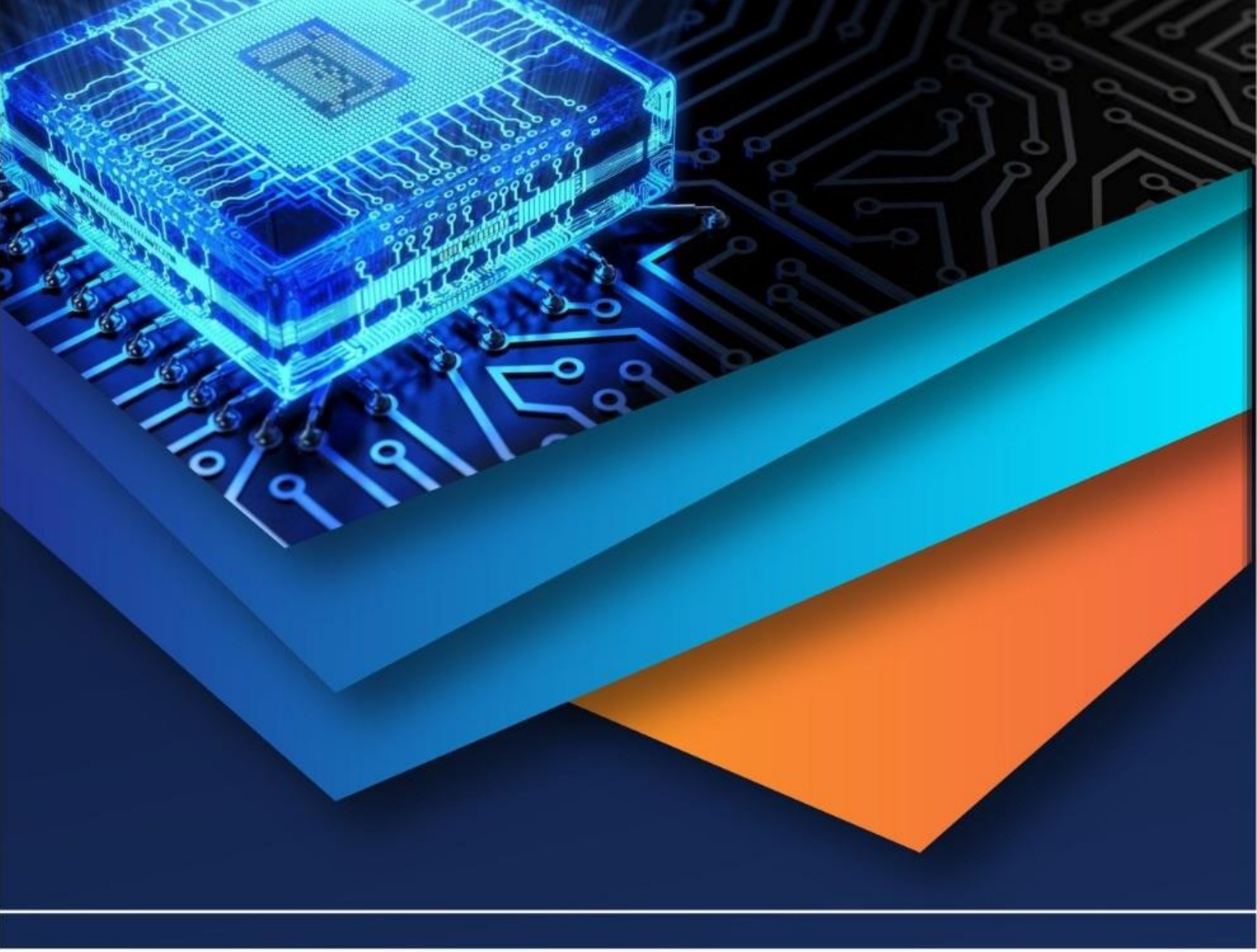
III. CONCLUSION

Web usage mining or web log mining are one the emerging fields or areas where there are many things that are yet to be discovered. In this paper we have tried to discuss few of the aspects that come under web usage mining, web log mining and how they are correlated with each other. There are many improvements that are required within near future. Various search engines are using these technologies to implement and they are still finding solutions for improving in a much better way.

REFERENCES

[1] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2nd ed., New York: Springer, 2011.
[2] P. Nithya and P. Sumathi, "A survey on Web usage mining: theory and applications," International Journal Computer Technology and Applications, Vol 3, 2012, pp. 1625-1629.
[3] Kitchenham B., S. Charter, "Guidelines for performing systematic literature reviews in software sngineering," Version 2.3, EBSE Technical Report EBSE-2007-01, Keele University and University of Durham, UK, 2007.
[4] S. Langhnoja, M. Barot, and D. Mehta. "Pre-processing: procedure on web log file for web usage mining", International Journal of Emerging Technology and Advanced Engineering, Vol 2, 2012.
[5] J. Srivastava, R. Cooley, M. Deshpande, P. Tan. "Web usage mining: discovery and applications of usage patterns from web data", ACM SIGKDD, Vol 1, 2000.
[6] S. Dhawan and M. Lathwal, "Study of preprocessing methods in web server logs", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3, 2013.
[7] Chitraa and A.S. Davamani, "A study on preprocesssing methods for web usage data", International Journal of Computer Science and Information Security, Vol 7, 2010.
[8] R. Mahajan, J.S. Sodhi, V. Mahajan, "Web usage mining for building an adaptive e-learning site: a case study", International Journal of eEducation, e-Business, e-Management and e-Learning, 2014.
[9] C.J. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M.J. del Jesus, S. García, "Web usage mining to improve the design of an e-commerce website:

OrOliveSur.com", International Journal of Expert System with Applications, 2012.
[10] G. Neelima, Dr. Sireesha Rodda, "Predicting user behavior through Sessions using the Web log mining" International Conference on Advances in Human Machine Interaction (HMI - 2016)
[11] J.D. Vela´squez, "Combining eye-tracking technologies with web usage mining for identifying Website Keyobjects", International Journal of Engineering Applications of Artificial Intelligence, 2013.
[12] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann Publisher, 2011
[13] Lya Hulliyyatus Suadaa, "A Survey on Web Usage Mining Techniques and Applications" International Conference on Information Technology Systems and Innovation (ICITSI) 2014
[14] D. Stevanovic, N. Vlajic, A. An, "Detection of malicious and nonmalicious website visitors using unsupervised neural network learning", International Journal of Applied Soft Computing, 2013.
[15] B. Verma, K. Gupta, S. Panchal, R. Nigam, "Single Level Algorithm: An Improved Approach for Extracting User Navigational Patterns To Improve Website Effectiveness", International Conf. on Computer & Communication Technology, 2010.
[16] V.V.R. M. Rao and V. V. Kumari, "An Efficient Hybrid Predictive Model to Analyze the Visiting Characteristics of Web User using Web Usage Mining", International Conference on Advances in Recent Technologies in Communication and Computing, 2010.
[17] Paliouras, "Discovery of Web user communities and their role in personalization", User Model User-Adap Inter, 2012.
[18] R. Tamimi, M. E. Mohammadpourzarandi, "The Application of Web Usage Mining In E-commerce Security", International Conference on ecommerce in developing countries with focus on e-security, 2013.
[19] Y. Slimani, A. Moussaoui,Y. Lechevallier, A. Drif, "A community detection algorithm for Web Usage Mining Systems", International Symposium on Innovation in Information & Communication Technology, 2011.
[20] B.N. Devi, Y.R. Devi, B.P. Rani, R.R. Rao, "Design and Implementation of Web Usage Mining Intelligent System in the Field of e-commerce", International Conference on Communication Technology and System Design, 2012.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY