Volume-2, Issue-5, May-2015 ISSN: 2349-7637 (Online)

RESEARCH HUB – International Multidisciplinary Research Journal (RHIMRJ)

Research Paper

Available online at: www.rhimrj.com

Web Usage Mining: Survey on Process and Methods

Payal Sagar[1st]
Post Graduate Student,
Department of Computer engineering,
C.U.Shah College of Engineering &Technology,
Surendranagar, Gujarat (India)

Prof.A.V.Nimavat[2nd]
Assistant Professor,
Department of Computer engineering,
C.U.Shah College of Engineering &Technology,
Surendranagar, Gujarat (India)

Abstract**: *In today's era, the internet playing an essential role in our day-to-day life. The internet has influenced every area of users. The tremendous growth of an internet raises the complexity to browse efficiently by the users. To understand and increase the performance of the website, a website is changed as per user's needs. Web log files analyses to fulfil the needs of auser; these log files are stored on server side. Through which the behaviour of users is capture. Web usage mining is an application to analyse interesting usage patterns from web log data. Web usage mining process is to discover patterns in three phase processes: data pre-processing, pattern discovery, pattern analysis. Log data contain an impurity and noise that why this data cannot directly used for pattern discovery. Pre-processing step is important in the mining process to get a right pattern from log data. Data pre-processing includes three steps data cleaning, user identification and session identification. In this paper, provides a detailed review on web usage mining process and methods.***

Keywords: web usage mining, web log mining, pattern discovery, pattern analysis.

I. INTRODUCTION

The internet is growing rapidly in terms of size and usage with respect to time. Knowledge present at the web data is gained by using the web data mining. Web data mining can be categorized into three areas [1]. 1. Web content mining: is the process of discovering useful knowledge from text, image, audio or video etc. 2. Web structure mining: it operates on the webs hyperlink structure can provide information about ranking or authoritativeness and enhance search results of a page through filtering. [5]. 3.Web usage mining: is the process of extracting useful knowledge from the web log data. This is used to understand and better serve the needs of users. This knowledge can be applied for anefficient reorganization of website, better personalization, recommendation, navigation and attracting more advertisement. As a result, more user attracts toward website [1].

In this paper, we discuss a collection of web log data, are the sources for the weblog data in section II. In section III methods for data pre-processing section IV discuss various techniques used for pattern discovery section V discuss on various pattern analysis ways. These results are represented by using interesting measure section VI conclusion.

II. SOURCE OF DATA

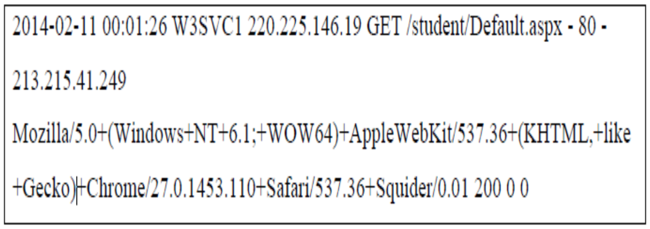
There are three main sources to get raw log data. Which are namely *1) web server log file 2) proxy log file 3) client log file.*

A. web server log file:

The most used source for web usage mining is web server log data. This log data is generated authentically at server side. Log data contain the information about the user's activity. The common server log file types are access log, agent log, error log and referrer log [3].

1. Access log file contains all the information that provides to the client by the server.2. Agent log file contains the information about the user's browser and os name with version etc. 3. Error log file contains the error details while processing the user's request. 4. Referrer log file [1] is used to allow websites and web servers to identify where users are visiting them from, for promotional or security purposes.

Depending on a web server, web log file data differs from a number, type of attributes and format of web log file [3]. Many log file formats are available 1.Common log format, 2.Extended common log format, 3.centralized log format, 4.NCSA common log format, 5.ODBC logging, and 6.centralized binary logging [3]. Among this common or extended log file format is mainly used by a web server.



```
2014-02-11 00:01:26 W3SVC1 220.225.146.19 GET /student/Default.aspx - 80 -

213.215.41.249

Mozilla/5.0+(Windows+NT+6.1;+WOW64)+AppleWebKit/537.36+(KHTML,+like

+Gecko)|+Chrome/27.0.1453.110+Safari/537.36+Squider/0.01 200 0 0
```

Fig 1: Sample Log FileFormat

W3C extended log file format is very important in web usage mining, it can be reform. It contains some additional attributes than Common Log File.

Above sample log file format contains fields like date, time, s-sitename, s-ip, cs-method, cs-uri-stem, cs-uri-query, s-port, cs-username, c-ip, cs(User-Agent), sc-status, sc-substatus, sc-win32-status, where s-server, c-client, cs-client to server, sc-server to client.

```
Date: 2014-02-11
Time: 00:01:26
S-sitename: W3SVC1
S-ip: 220.225.146.19
cs-method: GET
cs-uri-stem :  /student/Default.aspx
cs-uri-query: s-port:80
cs-username: c-ip : 213.215.41.249
cs(User-Agent):
        Mozilla/5.0+(Windows+NT+6.1;+WOW64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/27.0.1453.110+Safari/537.36+Squider/0.01
sc-status: 200
sc-substatus: 0
sc-win32-status: 0
```

B. Proxy Server Log File:

The proxy server plays an intermediate role between the user and the server. All the user request and services are passed through this proxy server. Proxy server log files, whose format is same as of web log file may reveal the actual HTTP request coming from multiple clients to multiple web servers and characterizes, reveals the browsing behaviour for a group of anonymous users sharing a common proxy server [12].

C. Client Log File:

The log file can reside in client's browser windows itself and recording the activities within the client machine. In some case it is advantageous.

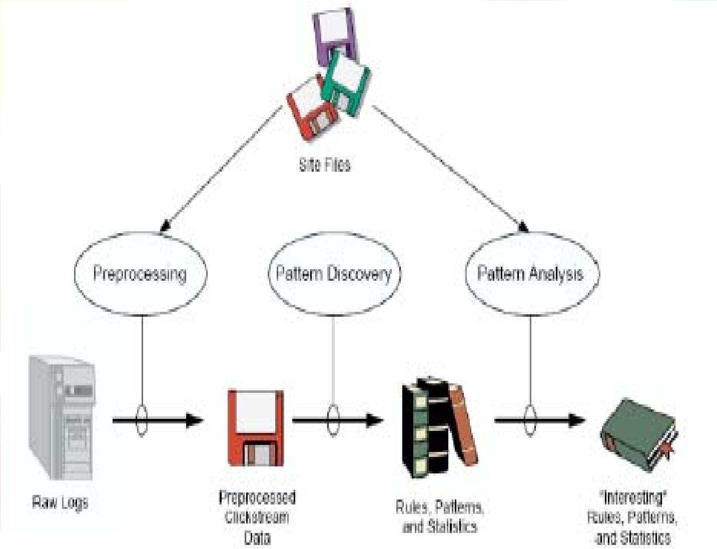

Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Fig 2: Web Usage Mining Process

Web usage mining process is a three step co-related process. As shown in figure-2, which are pre-processing web log data, pattern discovery and pattern analysis.

III. DATA PREPROCESSING

Variety of sources is individual or combined raw log data. This raw log data may contain noise and impurities. Therefor raw log data undergoes a data pre-processingphase, which consisting a series of steps called data pre-processing. By which we can removes such impurities and convert data into the format on which data mining techniques can be applied to extract the knowledge. Data pre-processing is the time consuming task because as qualitative the data better the results. Data pre-processing includes data cleaning, user identification and session identification. Algorithms and techniques are developed for data pre-processing.

A. Data Cleaning:

In this process unnecessary and irrelevant fields from the raw log file are removed. Extensions like gif, jpg, css in target URL are removed because this type of file actually not requested by users but automatically downloaded by HTML tags. For data cleaning HTTP status code is also considered, HTTP status code under 200 and upper 299 are removed. User request method is also considered, other than the GET method like POST, HEAD is removed. Data cleaning removes the records containing robots.txt in the requested source name. Because web robot itself follows all the hyperlinks from web pages and like Google periodically use web robots to gather all the pages from a website to update their search indexes. These techniques are used to remove the irrelevant data from log data.

B. User Identification:

A unique user can be identified by using the user identification process. By using client IP address, the unique user can identify.

1. ***Base on Client Information:*** Is one of the heuristic techniques used for user identification. Agent field of a log file contains operating system and browser name with version. If same IP address having different agent field it shows a different user. E.g. if a user is visiting two pages of a same website by using two different browser simultaneously on a single device then this technique consider two records by the different user even they are from a single user.
2. ***Base on Topology:*** Topology of a website used to identify a user. If a user request a page that is not accessible from its previously requested pages is considered as a new user, this can be done by using referrer attribute of log format and link information from site topology [3]. E.g. if a user make a request by using bookmarked pages which are not concerned via links. Like this approach confusion accurse.
3. ***By Using Cookies:*** Cookie is a small variable which stores some parameter value at client side. A Cookie created at server side and send to the client side. This cookie contains some useful information regarding user so it is possible to identify a unique user. But this technique may not support, in which some browsers disable cookies.

C. Session Identification

A user session is referring to a number of pages visited by the single user during a certain time period. We can differentiate entries into different user sessions through a timeout.

1. ***Session identification by time oriented heuristic:*** time gap is used between two entities, if time exceeds certain threshold new session is created.

   If s.tn+1-s.tn>= timethreshold then new session.

   Mostly threshold value is 30 minutes. This value depends on application, site topology and on many parameters. Therefore, the fix threshold is not suitable for all applications. A Dynamic threshold is suggested according to the type of application.
2. ***Session Identification by time spent on observing page:*** pages are categorized into navigational pages and information pages based on time spent by users. Information pages are a goal of user's, more time is spent by users on information pages to study the content compared to navigational pages. This information is used to define the session. If we know the percentage of navigational pages in log data, the maximum length of such page can be identified by formula.

   $$Q=-\ln(1-\gamma)/\lambda$$

   Where q is athreshold of a navigational page,γ is the percentage of the navigational page, λ is observed duration time mean of all pages in log data.
3. ***Session identification by referrer:*** W3C Extended log format have referrer URL attribute. This attribute is existed in the same session. If no referrer is fount then it is a first page of a new session. If two consecutive request a and b, where p and S (p is a page and s is a session) if referrer(r) for a page b was invoked within session S: r and S, then n is added to S, otherwise a new session.

IV. Pattern Discovery

Pattern discovery is a process in which various data mining techniques are applied to find frequent patterns. These techniques are data mining, machine learning, statistical methods and pattern recognition. Mostly used techniques are classification, clustering, association rule, sequential pattern etc.

1. Classification is a supervised learning method. It is an automated process of assigning a class. Decision tree induction, Bayesian Classifier, K-nearest neighbour classifier and support vector machines etc. 2. Clustering is an unsupervised learning method in which clusters are build, users or pages are grouped who have similar characteristics, is known as user clustering and page clustering. Pages which are conceptually connected are identified. Formation of a cluster is done by using similarity measures between two entities. Commonly used similarity measure techniques are Euclidian Distance, SPO and Fuzzy C-Mean etc [12].clustering is useful for inferring user demographics in order to perform Market Segmentation in E-Commerce applications and in Personalization. 3. Association Rules are used to discover the related pages or items together in a same transaction. By applying different association rule mining techniques, we can identify which pages or items are frequently accessed together by the users. Support and Confidence are two measures used for the rules importance and quality. Like A-Priori, Eclate and FP-growth etc techniques are used to generate rules. 4. Sequential Patterns used to discover frequent sub sequences among a large amount of sequential navigation patterns that appear in user's sessions frequently. Techniques used for association rule mining are also used for sequential pattern mining.

V. PATTERN ANALYSIS

After the pattern discovery generated results are not suitable for interpretation. Therefore on that results pattern analysis is needed. In this process, uninteresting pattern and rules filter out from the generated results. To represent the data in 2D, 3D and pictorial representation many visualization tools are used. These tools provide a way for comparing and characterizing result in the form of charts, graphs, tables, wein diagram and in so many other visual presentations [2].

VI. CONCLUSION

In web, usage mining log data are mined to understand and better serve the user's needs. By mining,these log data website owners can provide better functionality to users. To achieve the knowledge from web log data various heuristic processes, methods and techniques are used.

REFERENCES

1. Ashika Gupta, Rakhiarora, Ranjanasikarwar, NehaSaxena, "Web Usage Mining Using Improved Frequent Pattern Tree Algorithms" 978-1-4799-2900-9/14/$31.00 ©2014 IEEE
2. ShailyG.Langhnoja, Mehul P. Barot, Darshak B. Mehta, "Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery" International Journal of Data Mining Techniques and Applications, June 2013.
3. Chintan R. Varnagar, Nirali N. Madhak, Trupti M. Kodinariya, Jayesh N. Rathod, "Web Usage Mining: A Review on Process, Methods and Techniques" Department of Computer Engineering, Atmiya Institute of Technology and Science, Rajkot, Gujarat, India.
4. Nanhay Singh, Achin Jain, Ram Shringar Raw," COMPARISON ANALYSIS OF WEB USAGE MINING USING PATTERN RECOGNITION TECHNIQUES" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013.
5. TheintTheint Aye, "Web Log Cleaning for Mining of Web Usage Patterns" IEEE, 2011
6. K. Sudheer Reddy M. Kantha Reddy V. Sitaramulu, "An effective Data Preprocessing method for Web Usage Mining"
7. Mirghani. A. Eltahir, Anour F.A. Dafa-Alla, " Extracting Knowledge from Web Server Logs Using Web Usage Mining", INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE), 2013.
8. BinaKotiyalt, Ankit Kumar, BhaskarPant , R.H. Goudar, Shiv aliChauhan, SonamJunee, "User Behavior Analysis in Web Log through Comparative Study of Eelat and Apriori", Proceedings of7'h International Conference on Intelligent Systems and Control(ISCO 2013).
9. Omer Adel Nassar,Dr.Nedhal A. Al Saiyd" The Integrating Between Web Usage Mining andData Mining Techniques", International Conference on Computer Science and Information Technology IEEE, 2013.
10. Mr. Rahul Mishra, Ms.AbhaChoubey," Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining" International Journal of Advanced Research in Computer Science and Software Engineering,September 2012.
11. P.Nithya, Dr.P.Sumathi," Novel Pre-Processing Technique for Web LogMining by Removing Global Noise and Web Robots", National Conference on Computing and Communication Systems IEEE, 2012.
12. L.K Joshila Grace, V. Maheswari, DhinaharanNagamalai, "Analysis of Weblogs and Web User in Web Mining," International Journal of Network Security & Its Applications (IJNSA), Vol. 3, No. 1, January 2011.
13. MitaliSrivastava, RakhiGarg, P. K. Mishra, "Preprocessing techniques in Web Usage Mining:A Survey, "International journal of Computer Applications",2014.