International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 9 Issue 3 ǁ 2021 ǁ PP. 47-51

Survey On An Integrated Approach for the Analysis of users' future request by Web Mining

[1]Hinal Rathod, [2]Anita Anand
[1] PG Scholar, Hinal J. Rathod, LDRP ITR, Gandhinagar, Gujarat, India
[2] Professor, Mrs. Anita Anand, LDRP ITR, Gandhinagar, Gujarat, India

Abstract - *In the web servers, log repositories plays a key role as it keeps record of user pattern for different users and thus it is great source of knowledge. Web Usage Mining is an area, where the navigational access behaviour of users' over the web is tracked and analyzed. So that websites owner can easily identify the access patterns of its users'. By collecting and analyzing this behaviour of user activities, websites owner can enhance the quality and performance of services to catch the attention of existing as well as new customers. An approach is proposed to get more accurate prediction results by using both kmean clustering and logistic regression. The predictions of users' future access requests by this manner can improve accuracy of results and will helps in order to reduce the search time.*

Date of Submission: 27-02-2021 Date of acceptance: 12-03-2021

I. INTRODUCTION

Information dominates the world more than any time before. As the volume of data increases, it becomes a very tedious and tough task to comprehend it. Information technology has made it possible to analyze as well as manage large amount of data electronically and to be able to search for probably very interesting knowledge inside this deep ocean of data. Data mining seems the only solution to this ever growing problem.

Data mining is primarily a conception of as the process of extracting implicit, previously unknown and potentially useful information from the large set of databases. Exercising large amount of data for superior decision making by looking for interesting patterns in the data has become prime task in today's environment. Hence, the significance of data mining is arising conspicuously. The techniques such as association rules mining, classification, clustering, genetic algorithms and other statistical patterns etc. are often used to discover useful patterns and knowledge that are previously unknown to the system and users. Also data mining has been used in various applications such as marketing, engineering, customer relationship management, decision making, expert prediction, web mining, crime analysis, medicine, and cloud computing, among others.

With the technological advancement as well as by the raised popularity of World Wide Web (WWW), many websites typically experience thousands of visitors and its users' daily. So there has been huge interest towards web mining.

The web mining is intended to discover useful patterns from large sets of web data, where at least one of structure or usage data is used in the mining process. Web based data includes different kinds of information like web structure based data, document based data, log based data and data of user profiles. For small sites, an individual Web designer's intuition along with some straightforward usage statistics may be adequate for predicting and verifying the users' browsing behaviour. However, as the size and complexity of a website increases, the simple statistics provided by existing Web log analysis tools are inadequate for providing meaningful insight into how a website is being used.[12]

The web server automatically generates the usage information of the websites and it is stored in the web server as log file. Information of each page requested and accessed by web users' stored in log file commonly referred to as web access log. So web service providers only need a tool to analyze these logs. Web Usage Mining (WUM), a sub part of web mining is used for this purpose.

II. WEB USAGE MINING

2.1 Web Mining

To mine interesting information from large datasets, data mining techniques can be applied. Web mining can be referred as the transformation of the data mining techniques to web data[12]. It is an application of data mining techniques to retrieve, discover and evaluate interesting information from the WWW. There are four categories of web mining subtasks [10]:

A) Finding and selecting information: requested information is extracted and retrieved from textual documents existing on the web.
B) Pre-processing: preliminary processing is performed automatically on the retrieved information.
C) Generalization: General patterns of personal websites are discovered.
D) Analyzing: Finally the estimation of the validity of extracted patterns is performed and the patterns whose accuracy and validity are confirmed are accepted and offered.
Based on the interest and/or final objective of what kind of knowledge to mine from web data the web data mining can be categorized into three categories [9].
A) Web Content Mining: Discovers the useful information or knowledge from web page contents or documents or services i.e. text, image, audio, videos etc.
B) Web Structure Mining: Deals with analyzing, discovering, and modeling, link structure of web pages and/or website to generate structural summary on which various techniques are applied and outcomes of these techniques can be utilized to recreate, redesign the website to improve the structural quality of website.
C) Web Usage Mining: Deals with understanding of user behavior, while interacting with website, by using various log files to extract knowledge from it.

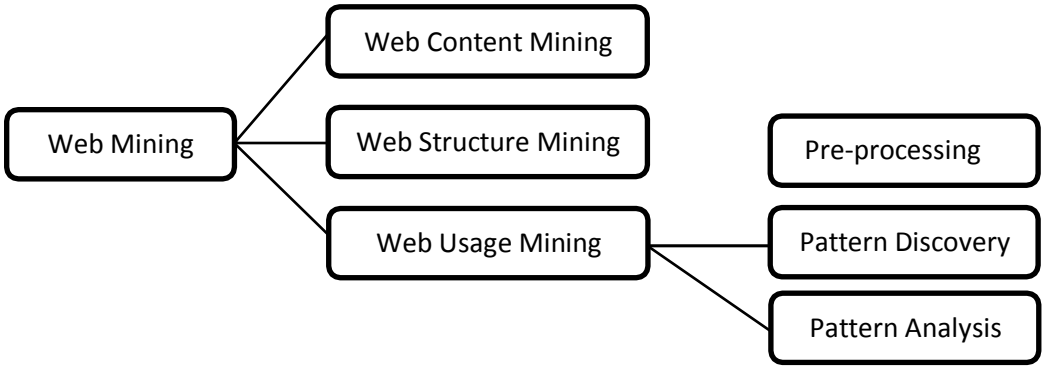

Fig 2.1: Classification of Web Mining

2.2 Web Usage Mining

Web usage mining used to discover the useful and interesting information [10] or usage statistics from the secondary data derived from the interaction of the users while surfing on the web, in order to understand and better serve the needs of web based applications. It mainly focuses on the techniques that could forecast users' behaviour while the user interacts with WWW. Web usage data captures the identity or origin of the users of web along with their browsing behaviour and navigation patterns at a website. The significant source of data for web usage mining is textual logs collected by web servers all around the world.

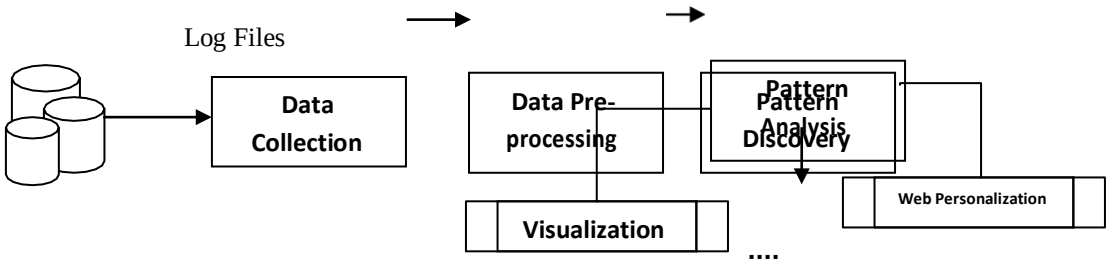

Fig 2.2: Web Usage Mining Process [12]

The Web usage mining process mainly consists of four steps and they are: Data Collection, Data Pre-processing, Pattern Discovery and Pattern Analysis.

2.2.1 Data Collection [10]

Data collection is the first step in the WUM. Usage mining applications are based on data gathered from the main sources: Web servers and Proxy servers. Web servers are evidently the richest and most common source of data which collects vast amount of data in their log files. Data collection is the process of obtaining the information from the log files. Web log file is automatically created and maintained by a web server. Every hit

to website including each view of a HTML document, image, or other object is logged. Irrespective of source of collection, the web log file has the following general characteristics:

A) It is simple plain textual file and maintains records in identical format.

B) Each record in the log file represents a single HTTP request.

C) Each record contains significant information about a request: the client host name or IP address, the date and time of the request, the requested file name, the HTTP response status and size, the referring URL, and the browser information [10].

2.2.2 Data Pre-processing

The raw data that collected from the various sources are usually diverse, voluminous and unstructured. Therefore, data pre-processing phase is predominantly significant and prerequisite for discovering access patterns in usage mining. The collected log data may contain the number of raw, irrelevant, noisy and unformatted entries. Such collected raw log files cannot be used directly for the process of web usage mining. So the collected raw log data must be undergoes through a complex processes, consisting of series of steps called Data pre-processing. At this stage, impurities, noisy and irrelevant data are removed and/or data may be converted into some required format on which data mining techniques can be applied [10]. Good data sources not only discover quality patterns but also improve the WUM algorithm. Hence, data pre-processing is an important activity for the complete web usage mining process and vital in deciding the quality of patterns.

It is aimed to transform the raw collected click stream data into a set of user profiles. Data pre-processing is the most challenging, complicated and time consuming task. Generally, data pre-processing can be summarized into different sub-tasks are: Data Cleaning & Feature Selection, User Identification, Session Identification, Path Completion and Formatting, as shown in figure.

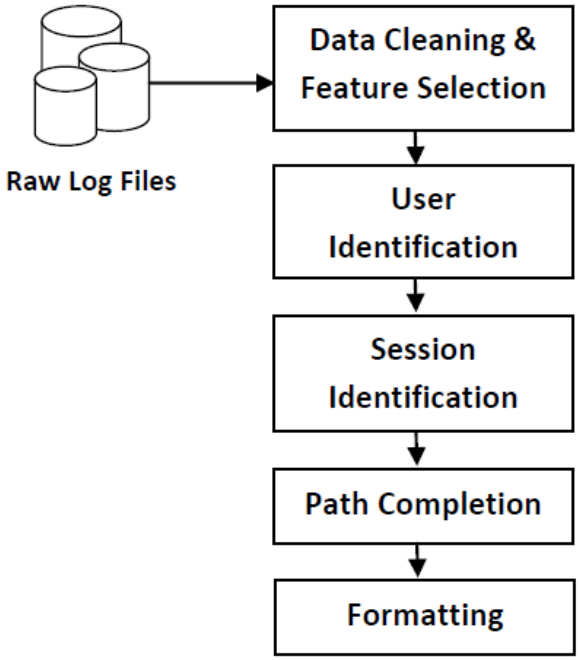

Fig 2.5: Sub-steps in data pre-processing for WUM

Data Cleaning &Feature Selection :

The first step in data pre-processing is to clean the collected raw weblog data. The request processed by auto search engines (e.g. Crawler, Spider and Robot) and requests for graphical page content such as jpg, JPEG, and gif images or uncompleted queries are removed [10]. Hence elimination is required for files which have suffixes as: jpg, jpeg, gif, css, cgi, etc. were found in cs_uri_stem field. The following table shows the summary of tasks performed in data cleaning stage.

User Identification :

The identification of individual users who access a websites is one of the most significant issues for successful personalization of website and also it is vital step in the usage mining. It is intended to find out the

unique individual users from the web log files. It analyses the access log file and clusters the users based on their similar characteristics. There are several methods to be followed for the users' identification. The simplest method is to assign a different user to each different IP-address identified in the access log file so that different users are being distinguished by using their IP-addresses.

The method used for this process is referred as referrer-based method. User identification may complex due to the presence of resident caches, firewalls and proxy serves. In the proxy servers many users' shares the same address and same user uses many browsers. An Extended Log Format may overcome this problem by referrer information and a user agent. To deal with this problem, the usage mining methods were employed that depends on user co-operation.

Session Identification :

Once a user was identified, click-stream is divided into clusters; this method of division is referred as Sessionization or Session Reconstruction. A user session is defined as a series of web pages visited by the same user within the duration of one particular visit to a website. A user may have single or multiple sessions during a time period. It is aimed to find out the different user sessions from the web access log file and it involves – dividing the page accesses of every user into separate sessions. The rules that used to identify a users' session are as follows :

A) For any new IP-address in web access log file, a new user as well as a new session will be created.

B) If the refer page in an entry of web access log file is null, in one user session; a new session will be created.

C) It is assumed that the user is starting new session, if the time between page requests exceeds more than 25.5 to 30 minutes.

Path Completion :

There is another critical issue that arises and required to be resolved is path completion. The path completion process recognize unique user sessions by adding significant accesses that are not recorded in the access log files. Sometimes some users' access does not being recorded in the access log due to the existence of proxy server and local caching problems.

Therefore, it is aimed to obtain complete users' access path by filling up the missing page references and the incomplete access path is recognized based on user session identification. For successfully performing this process the analyst must know the topology, network of hierarchy and the relationship among WebPages. The missing pages are added as: The page request is checked whether it is directly linked to the last visited page or not. If there is no link exists with last visited page then we can check the recent history. If the log record is available in the recent history then we can use 'back' button for caching until the particular page has been reached. Else if the referrer log is not clear, the site topology can be used for the same.

Formatting :

It is the last step in the data pre-processing. Here, the structured file containing sessions and visits are transformed in to a relational database model. The data generalization method is applied at request level and aggregated for visits and user sessions to completely fill in the database.

2.2.3 Pattern Discovery

Once a user's transactions have been identified, in this ultimate stage a various types of data mining techniques are performed to discover the useful patterns from the weblogs in net usage mining. There are various methods to discover useful pattern (i.e. knowledge) form weblogs as: Statistical analysis, Association Rule mining, Classification, Clustering, Sequential patterns and Dependency modeling.

2.2.4 Pattern Analysis

This is the ultimate step in the overall usage mining process with having two fundamental goals. The first one is to retrieve the valuable, relevant and interesting rules, patterns or statistics from the results obtained in the pattern discovery stage by filtering insignificant, irrelevant information or statistics. Another aim of this analysis is to discover some information that can offer valuable insights about users' access and navigational behavior. Every time the retrieved result from pattern discovery stage might not be in the form, suitable for analysis or to determine conclusion out of it. So in this stage tools are provided which helps to transform information into useful knowledge.

Most commonly used technique in pattern analysis stage is knowledge query mechanism such as SQL (Structured Query Language), which facilitates to pose queries for information retrieval that controlled by analyst, generally kind of statistical data in the form of text. Various presentation and visualization mechanisms are also used which represent useful knowledge in 2D or 3D graphical representation such as OLAP (Online

Analytical Processing). Visualization assists an analyst to capture navigational patterns in better way as well as to forecast trends of data. So these tools provides interactive ways of comparing, characterizing, representing results in terms of charts, graphs, tables, diagrams and so many other forms.
So, Pattern analysis allows us to automatically recognize the useful patterns and knowledge in data from the same source [10] and make predictions of new data coming from the same source.

III. RELATED WORK

A) Clickstream Data Pre-processing

Whenever the users' hit search for any webpage every clickstreams they are made stored in the web server log files. The clickstream data is defined as a sequence of Uniform Resource Locators (URLs) browsed by users within a particular time period. It needed to be pre-processed to remove irrelevant entries from it before it is taken for analyze.

B) User Access Matrix

To recognize 'how many pages accessed by particular user' and 'how many users accessed particular page', there is a need to generate web access matrix of users. The retrieved data pattern from previous stage is converted into web user access matrix U_{AM} in which rows represent users and columns represent pages of website. It is used to describe relations between web users and pages who accessed by users.

C) K mean clustering

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. *k*-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids

D) .Logistic regression:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression)

IV. CONCLUSIONS

As the information of page of web site is huge and develops rapidly, increasing the user's browsing speed efficiently as well as possible and reducing the loading of web server become very important issues. As per this thesis, we will try to generate an integrated approach to recognize the frequent access patterns by analysing past users access behavior and based on those retrieve patterns the browsing behaviour of the user will be analyzed which is useful to predict the next page access requests from the user. The proposed approach will be used to improve the accuracy of predictions for users' future requests to better the web performance.

In future work, work can be extended by applying it on different kinds of websites to assess its performance. Work can also be extended by implementing the proposed approach on the parallel as-well as on the cloud technology to evaluate its effectiveness.

REFERENCES

[1]. Sowmya H.K., Dr. R.J. Anandhi, "Web Usage Mining Algorithms: A Survey", AICAAM, April 2019.
[2]. Panjawani Heena, Pooja Jardosh, "WebPage Recommendation in web usage mining using Genetic Algorithm", IJARIIE-ISSN, 2017.
[3]. Pooja Solanki, Jasmin Jha, "Web Page Recommendation System using Biclustering with Greedy Search and Genetic Algorithm", June 2015.
[4]. Kaushal Kishor Sharma, Prof. Kiran Agrawal, "A Hybrid Approach for Predicting User"s Future Request", IEEE, 2014.
[5]. Dilpreet Kaur, A.P. Sukhpreet Kaur, "User Future Request Prediction Using KFCM in Web Usage Mining", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol. 2, Issue 8, August 2013.
[6]. A. Anitha, "A New Web Usage Mining Approach for Next Page Access Prediction", International Journal of Computer Applications (IJCA), Volume 8- No. 11, October 2010.
[7]. Priyanka Makkar, Payal Gulati, Dr. A.K. Sharma, "A Novel Approach for Predicting User Behavior for Improving Web Performance", International Journal on Computer Science and Engineering (IJCSE), Vol. 2, No. 04, 2010.
[8]. V. SUJATHA, PUNITHAVALLI, "Improved User Navigation Pattern Prediction Technique from Web Log Data", Procedia Engineering 30, Elsevier, 92-99, 2012.
[9]. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, Volume 1, Issue 2, 1-12, 2000
[10]. Robert -Walker Cooley, "Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data", May 2000.
[11]. Yuhefizar, Budi Santosa, I Ketut Eddy P., Y. K. Suprapto, "Two level clustering approach for data quality improvement in web usage mining", Journal of Theoretical and Applied Information Technology (JATIT), Vol. 62, No. 2, 404-409, April-2014.
[12]. Raymond kosala, Hendrik Blockeel, "Web mining Research: A Survey", ACMSIGKDD, Volume 2. Issue 1. 1 – 15. July 2000.