ISSN 2395-1621

A Novel Approach for User Navigation Pattern Discovery and Analysis for Web Usage Mining

[#1]Prof.Priya U.Thakare , [#2]Pravin Varpe , [#3]Dipa Shedage , [#4]Jayshri Raykar , [#5]Archana Pawar

[1]thakarepriya1@gmail.com
[2]pravinvarpe97@gmail.com
[3]shedagedeepa@gmail.com
[4]jayshriraykar1@gmail.com
[5]apawar6262@gmail.com

[#12345]Department of Computer Engineering,
SITS, Narhe

ABSTRACT

In the last few decades increase in popularity of web, web mining has attracted lots of attention. Web usage mining is an important area in web mining, the discovery of patterns in the browsing and navigation data of Web users. In order to better serve the needs of web based applications, web log mining is the technique to discover usage patterns from web data. In user access log files very significant information about log servers is present. This paper describes about preprocessing of Web Log Data, analysis of Web Log Data to find information about web site, users navigational patterns generation of the particular web site other information which will help system administrator and Web designer to improve their system by determining the patterns and the usage of web pages. The results which obtained will represented in the graphical and tabular format. These results will be used in further development of web site in order to increase its effectiveness.

Keywords- **Web Mining, Pattern Analysis, Navigation Pattern, WUM, FCM.**

ARTICLE INFO

Article History

Received :28th September 2015

Received in revised form :

1st October 2015

Accepted : 5th October , 2015

Published online :

6th October 2015

I. INTRODUCTION

The web is an important source of information retrieval and interaction produces a huge quantity of data stored in web log files. Web Usage Mining (WUM) is the technique of data mining which is used to know the user access to websites. In WUM, data can be collected from server logs, proxy logs, and browser logs or can be obtained from an organization's database. These data collections are different in terms of the location of the data source, the different types of data available, the segment of population from which the data was collected. For pattern extraction the web logs are analysed by using the technique of Web Usage Mining (WUM). The web log mining process initially starts with data pre-processing where the cleaning is done o the data sets and the clustering techniques are used to extracts hidden knowledge. Pre-processing is the process which extract the useful information and also used for data formatting. After pre-processing of log file or data the analysis is done which is used to extract information. The pattern generation is done after extraction of useful data according to the user's needs. In our project we have implemented the same phases of web usage mining i.e. pre-processing of the web logs and on the knowledge gained Fuzzy C - Mean algorithm is applied and then outcome is analysed for user's navigation results.

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from web data in order to understand and better serve the needs of web based applications. Usage data captures the identity or origin of web users along with their browsing behaviour at a web site. Fuzzy C-Mean (FCM) is a data clustering technique in which a dataset is grouped into n clusters with every data point in the dataset belonging every cluster to a certain degree. For example certain data point that lies close to data point that lies far away from the canter of cluster will have a low degree of belonging or membership to that cluster.

Data pre-processing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data pre-processing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. Web Usage mining can be used to

uncover patterns in server logs but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data. The following are the pattern discovery methods. This is the final step in the Web Usage Mining process. After the pre-processing and pattern discovery, the obtained usage patterns are analysed to filter uninteresting information and extract the useful information. The methods like SQL (Structured Query Language) processing and OLAP (Online Analytical Processing) can be used.

In the proposed methodology first web log data of "E-Commerce site" is taken as input and then pre-processing is done on input dataset. Apply data mining clustering on the pre-processed dataset and then from the clustered values patterns are generated. Lastly useful knowledge is extracted from the generated patterns. The methodology implemented for the generation of online navigational patterns is proposed in which pre-processing is done and later applying clustering algorithm to cluster the patterns generated. The proposed methodology implemented here provides the patterns about user's navigation on various parameters which are beneficial for the company owner or organizations for tuning up of their websites according to requirements of the users. The web log data is pre-processed by using the Fuzzy- C-Mean algorithm and clustering methodologies. The results are generated after analysis and generation of pattern s in the form of graphs and the tabular list format.

II. LITERATURE SURVEY

Now a day's, multiple Web Usage Mining algorithms have been proposed to mining user navigation behaviour. In the following we examine some of the most significant navigation pattern mining systems and algorithm in web usage mining area that can be compared with our system.In this paper [1] they have discussed about precise web mining of logs and an efficient online navigational pattern prediction which is a vital for alteration up websites and subsequently helps for the visitors' preservation. Like any other data mining task, it starts with cleaning of data and its preparation and at the end determining some hidden knowledge which cannot be pull out using conventional methods. In order for this process to obtain good results it has to depend on some good quality input data. For that reason, additional focus in this process is on cleaning of data and then pre-processing. While on the other side, the scalability is the confront facing for online prediction. As a result of any improvement in the efficiency of online prediction solutions is more than necessary. As reply to the above mentioned concerns they proposed is an upgrading the web mining process of log and to the online navigation of pattern prediction. Their contribution contains 3 different components. First, they prefer a sophisticated time-out based heuristic for session identification. Second, they suggested the practice of precise an algorithm depends on the density for navigational pattern detection. Finally, they recommended a new loom for efficient online prediction. The bearing experiments reveal the applicability and efficiency of the projected approach. In this paper [2] they have discussed about log file analysis of web which began with the reason to offer to the Web site administrators a method to ensure the adequate bandwidth and ability of server for the organization. This analysis field made huge advances with the fleeting of time, and now companies look for ways to use Web log files to get information related to visitor profiles and activities of buyer. The investigation of Web log may tender advices regarding to advance the offer, information relating problems occurred to the users, and even about effort for the welfare of the site. Traces about heavy use in exacting intervals of the time or hacker attacks may be actually useful to systematize the server and regulate the Website. From the users point, web is a growing collection of bulky amount of information, usually a great portion of time is essential to look for and find the suitable information. Personalization is a option for the achievement of the developing of a web infrastructure. A client or a visitor who finds effortlessly what he was probing for is a client or a visitor that will return. For this reason, Web sites are shaped and modified to made contents more easily reachable, using profiles found to make recommendations or to aim users with ad hoc advertising. In this paper [3] they discussed about WUM an application of data mining techniques for gaining knowledge for serving better the requirements of web based applications. WUM analysis is done by applying techniques of pattern reorganization on the logs of web data. Pattern recognition is defined as the act of captivating in raw data and making an action based on the category of the pattern. WUM is divided into 3 parts: Pre-processing, Pattern discovery and Pattern analysis. Further, this paper intended with experimental work in which web log data is used. These log data of the "NASA" from NASA web server which is analysed by "Web Log Explorer". Web Log Explorer is a web usage mining tool which plays the vital role to carry out this work. In this paper[4] they converse the most topical loom to log of web data illustration aspire to incarcerate the navigational patterns of users with revere to the on the whole structure of the website. One such depiction is tree structured log files which is the main focus of this entire work. Most accessible technique for analysis such kind of data are based on the use of frequent mining sub tree techniques to pull out frequent user activity and navigational paths. In this paper they evaluated the use of extra typical data mining techniques facilitating by a recently proposed structure preserving flat data illustration for ordered data. The originally proposed agenda was attuned to better suit the task of log mining. Experimental estimate is executed on 2 data sets of real world web log and assessments are prepared with an existing status of the art classifier for tree structured data. The results show the huge potential of the process in facilitating the application of a broad range of data mining or analysis techniques to tree structured web log data. In this paper [5], they confer about the speedy expansion of www in its quantity of passage and the range and complication of web sites. In this paper, a new loom is offered based on hybrid clustering methods for WUM. The WUM process has 3 steps: pre-processing of data, data mining and analysis of result. Firstly, they gave a concise depiction of the WUM process and data on web, then the appearance of the pre-processing step and the data warehouse were engaged. The amalgam clustering technique based on FCM clustering are worn for analysing and the Web logs taken from the real world servers of web. The outcome obtained after applying these technique and the equivalent interpretation are also

offered. Furthermore, this paper also described WUM with regards to cloud which cloud is mining.

III. FIGURES GRAPHS AND TABLES

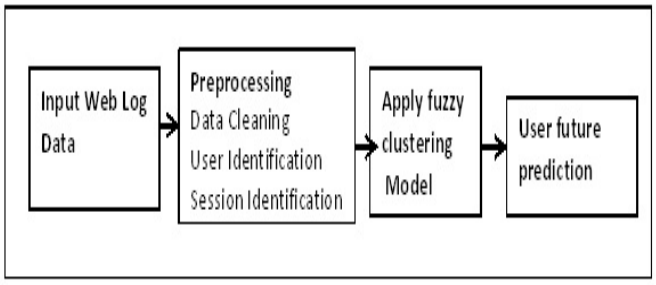

Fig. 1 Block diagram of the system

1. Read the web log files.
2. Pre-processing
   i) Select required attribute from log file like IP Address/ URL, Date and Time, Request Type of User request, Protocol, Port Number and Page Number & remove other attributes if present.
   ii) Remove irrelevant or invalid entries like robot request.
   iii) From cleaned log files identify unique users according to IP address and unique web pages.
   iv) Session identification.
3. Clustering is done by using Fuzzy-C-Mean.
4. Prediction: On various criteria the pattern prediction is done by using pattern analysis and pattern discovery for pattern generation using E-Commerce dataset.

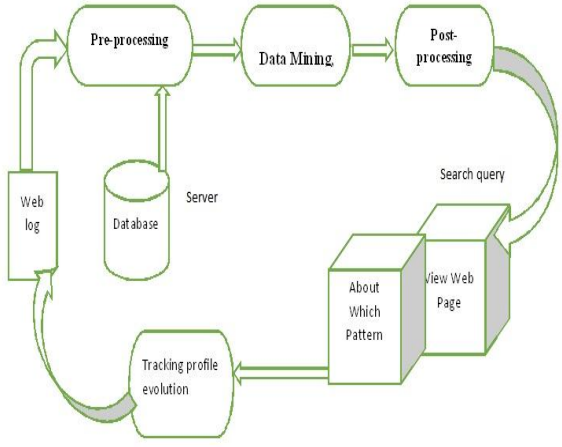

Fig. 2 Architecture Diagram for Navigation Pattern Discovery

The web logs are taken from the E-Commerce site. Then this web logs are for pre-processing where the data cleaning identification of the user's session and its reconstruction, the information retrieval of information regarding the content and structure of page and the data formatting is done. Then the pre-processed data is provided for data mining where the mining technique is used for mining. Then the post-processing is done and data is send to the search query for displaying web page and display information about the pattern. The that result is send to the profile evaluation and tracking block for evaluating and tracking the profile or patterns.

IV.EXPECTED RESULTS

In this project work we are getting the following results in the tabular and graphical form.

1. List of total number of users
   The list of uniquely identified users, total number of hits and total number of bytes consumed in the request.
2. Accessed file type
   The various types of file accessed by the users on their visit in the total duration of one week.
3. Error analysis (weekly)
   The weekly error report of request made by the users.
4. Session wise file accessed ratio
   The ratio of file accessed by users in total sessions in a day for duration of one week.
5. Product wise session accessed ratio
   The ratio of product accessed by users in total sessions in a day for duration of one week.

V. CONCLUSION

In this paper we are going to give a clear view of the web server logs, the pattern extracted from the web logs, and the results generated by extracting the patterns are represented in the graphical and tabular list format. This is used for generating online navigational patterns in which pre-processing is done and later applying clustering algorithm. In this methodology is we are providing the web log data sets and by using this web log data sets implementing the patterns about user's navigation on various parameters which are very beneficial for company owner or organizations. Here the web log data is pre-processed by using Fuzzy-C-Mean algorithm and clustering techniques.

VI. FUTURE SCOPE

Future work on this study comprises of more refined techniques for data pre-processing and recognition of access sessions, in order to assuage common problems of Web Usage Mining. Other algorithms for pattern detection shall also be incorporated in the system, so as to generate substitute methods such as Fuzzy-C-Mean algorithm and various clustering technique which can be investigated for further enhanced analysis.

REFRENCES

[1]Bowman Abedelghani Guerbas, Omar Addam "Effective Web Log Mining and Navigational Pattern

Prediction", IEEE transactions on parallel and distributed systems, vol. 23, no. 10, October 2012.

[2]Ankita Dixit, Meena Lakshmi, "Mining Web Log using Fuzzy C – Mean for Navigational Pattern Prediction" International Journal of Computer Applications (0975 – 8887) Volume 104 – No 12, October 2014.

[3]Fedja Hadzic, Michael Hecker, "Alternative Approach to Tree Structured Web Log Representation and Mining", 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent.

[4]Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem "Aggregated Search: A New Information Retrieval Paradigm", ACM Comput. Surv. 46, 3, Article 41 (January 2014).

[5] Nanhay Singh, Achin Jain and Ram Shringar Raw "Comparison analysis of web usage mining using pattern recognition techniques", International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 4, Issue 1, 2013, pp1-8 .

[6]Arvind K. Sharma and P.C. Gupta "Analysis of web server log files to increase the effectiveness of the website using web mining tool" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013.

[7]Ankita Kusmakar, Sadhna Mishra,Vineet Richhariya ,"Effective Web Log Mining: An Implementation View" International Journal of Advanced Research in Computer Science and Software Engineering 3 (4), March - 2013, pp. 304-310

[8]Sheetal Raiyani, Shailendra Jain "Effective Pre-processing technique using web log mining" International Journal of Advancement in Research & Technology, Volume 1, Issue6, November 2012.

