Effective Web Log Mining and Online Navigational Pattern Prediction: A Survey

Omkar Chandrakar 1st, Toran Verma 2nd

Computer Science & Enginneering, Rungta College of Engineering & Technolog[1], Bhilai, C.G., India
Computer Science & Engineerin, Rungta College of Engineering & Technology[2], Bhilai, C.G., India

***Abstract*— The web has become the world's largest repository of knowledge. Web usage mining is the process of discovering knowledge from the interactions generated by the user in the form of access logs, cookies, and user sessions data. Web Mining consists of three different categories, namely Web Content Mining, Web Structure Mining, and Web Usage Mining (is the process of discovering knowledge from the interaction generated by the users in the form of access logs, browser logs, proxy-server logs, user session data, cookies). Accurate web log mining results and efficient online navigational pattern prediction are undeniably crucial for tuning up websites and consequently helping in visitors' retention. Like any other data mining task, web log mining starts with data cleaning and preparation and it ends up discovering some hidden knowledge which cannot be extracted using conventional methods. After applying web mining on web sessions we will get navigation patterns which are important for web users such that appropriate actions can be adopted. Due to huge data in web, discovery of patterns and there analysis for further improvement in website becomes a real time necessity. The main focus of this paper is using of hybrid prediction engine to classify users on the basis of discovered patterns from web logs. Our proposed framework is to overcome the problem arise due to using of any single algorithm, we will give results based on comparison of two different algorithms like Longest Common Sequence (LCS) algorithm and Frequent Pattern (Growth) algorithm.**

***Keywords*— *Web Usage Mining, N*avigation Pattern, Frequent Pattern (Growth) Algorithm.**

I. INTRODUCTION

As daily use of World Wide Web is increasing, mining of the database is having more demanding. The database of web is in the form of web sessions with session id or session owner. So for this type of mining is called Web Mining. And in the Web Mining when we use to find path traversal pattern for decision management in website design, then it comes under the web usage Mining [12]. Web usage mining is also called web log mining. It is a web mining technique which is based upon the discovery and analysis of web usage patterns from web logs. These web logs include web server logs, proxy server logs, web browser logs, etc., and are created when users communicate with the web server [13]. [1]Web usage mining is the process of finding out what users are looking for on the internet. Few users might be looking at only documented data, whereas some others might be interested in multimedia data. It is the submission of facts and figures mining techniques to find out interesting usage patterns from World Wide Web facts and figures in alignment to realize and better serve the desires of Web based applications.[2]A new session in LCS is classified into any one of the clusters and prediction list is generated based on the navigation patterns of corresponding cluster. A prediction model by considering order information of pages and time spent on them in a session.[3] Data used for Web usage mining, can be collected at one of these three parts: Server level, Client Level, and proxy level collection. The input for the Web Usage Mining procedure is a client session document, which is fundamentally a pre-processed record and comprises of information, for example, who visited the site and what pages were visited and for to what extent, with their respective order. However it may contain unnecessary information. This unnecessary information can be minimized or reduced by data pre-processing of the web log data. After data cleaning the unique number of user and session identify and at last stage of data preprocessing we get the frequent user's access pattern from the web server access log file data.[4]A particular user or a set of user taking advantage of knowledge gained from user's navigational pattern and interest of the individual user with the conjunction of content and structure of Web Sites.[5] Web usage mining involves the automatic detection of user access patterns on one or more web servers. It is an application of data mining algorithms to web access logs to find the trends and regularities in web users' navigation patterns. There are many kinds of data that can be used in web mining and can be classified into following five types:

- Content of Web Page.
- Intra-Page Structure of Web Page.
- Inter-Page Structure of Web Page.
- Usage Data.
- User Profile.

In [7] a method to predict the user's navigation patterns is proposed using clustering and classification from Web log data. First phase of this method focuses on separating users in Web log data, and in the second phase clustering process is used to group the users with similar preferences. Finally in the third phase the results of classification and clustering are used to predict the users‟ next requests.

[6] To overcome the drawbacks of the current recommender system such as intelligence, adaptability, flexibility, limitation of accuracy, we present architecture for integrating semantic information about the products with web log data and generate a list of recommended products by using LCS and frequent pattern Algorithm. [8] A data cleaning approach should satisfy several requirements. Firstly, it should detect and remove all major errors and inconsistency in the database. The approach should be

supported by tools to limit manual inspection and programming effort and be extensible to easily cover additional source. Furthermore, data cleaning should perform mapping function and merging function. Mapping functions for data cleaning and other data transformations should be specified in a declarative way and be reusable for other data sources as well as for query processing. [10] The main aim of using association rule mining technique is to find the associations between items in a certain transactional record. Finding association rules is totally based on the support and confidence models, where a minimum support has to be specified to start the search .

II. BACKGROUND STUDY

Data mining efforts associated with the Web, called Web mining, can be broadly categorized into three areas of interest based on which part of the Web to mine; Web Content mining, Web Structure mining, and Web Usage Mining (Kosala and Blockeel, 2000). In Web mining, data can be collected at the server-side, client-side, proxy servers or a consolidated Web/business database (Srivastava et al., 2000). The information provided by the data sources can be used to construct several data abstractions, namely users, page-views, click-streams and server sessions.

.

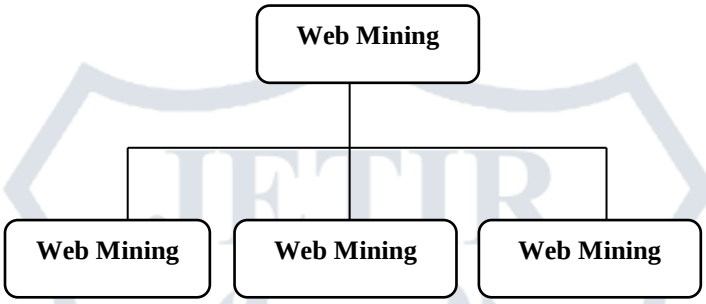

Figure 1: Web Mining Taxonomy

Web Usage Mining is defined as the process of applying data mining techniques to the discovery of usage patterns from Web logs data which to identify Web user's behavior (Srivastava et al., 2000). Web Usage Mining is the type of Web mining activity that involves an automatic discovery of user access patterns from one or more Web servers [3].

Process of Web Usage Mining: As shown in Fig. 2, three main tasks are performed in Web Usage Mining; Pre-processing, Pattern Discovery and Pattern Analysis. Fig. 2 represents a brief description about the main task of Web Usage Mining process. Web Usage Mining involves determining the frequency of the page access by the clients and then finding the common traversal paths of the users. First task is the data is collected from web server log file.

A. Data Pre-Processing

Web prediction models are generally used to identify the most probable future action for sequence of requests for the following class of users:

a. Specific users (client based models are more suitable)

b. Similar minded users (e.g. group of students in the same research team)

c. General users (e.g. for wide range of users in an internet cafe) Information about the above mentioned class of users and their browsing patterns have to be extracted from web server log files. These web log files have thousands of log entries and do not readily come with the information that we need to classify users. Some of the issues associated with log files and the preprocessing steps to make them more useable are discussed below.

The input for the Web Usage Mining process is a user session file, which is basically a pre-processed file and consists of information such as who accessed the web site and what pages were accessed and for how long with their respective order. However it may contain unnecessary information. This unnecessary information can be minimized or reduced by data pre-processing of the web log data. After data cleaning the unique number of user and session identify and at last stage of data preprocessing we get the frequent user's access pattern from the web server access log file data [12] [4].

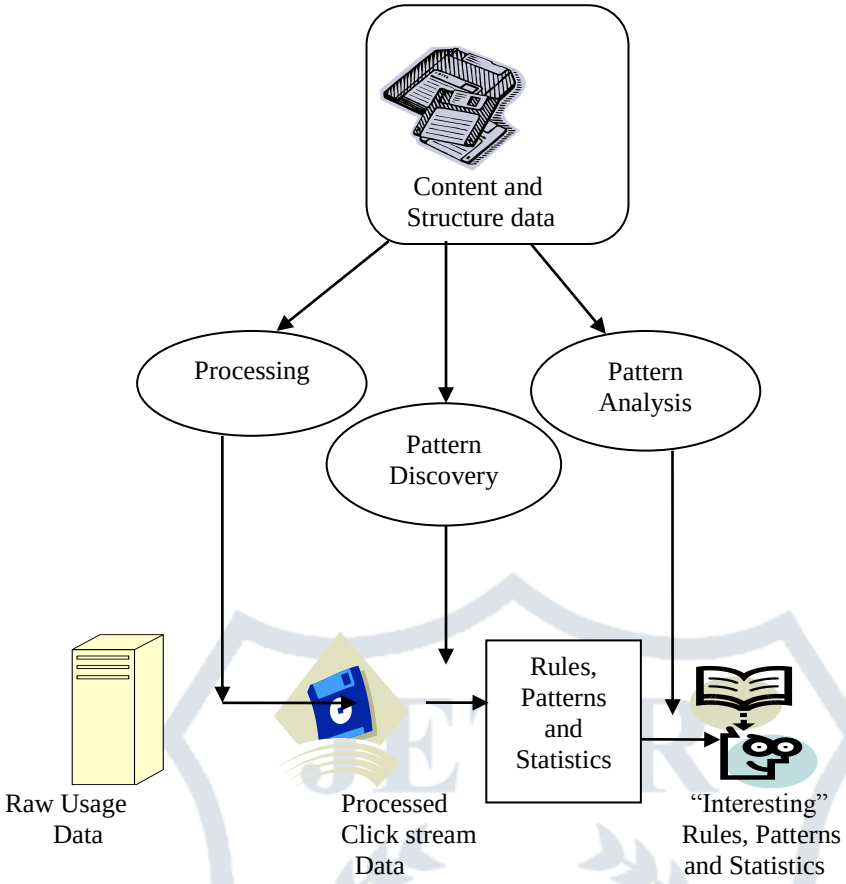

Figure 2: Web Usage Mining Process

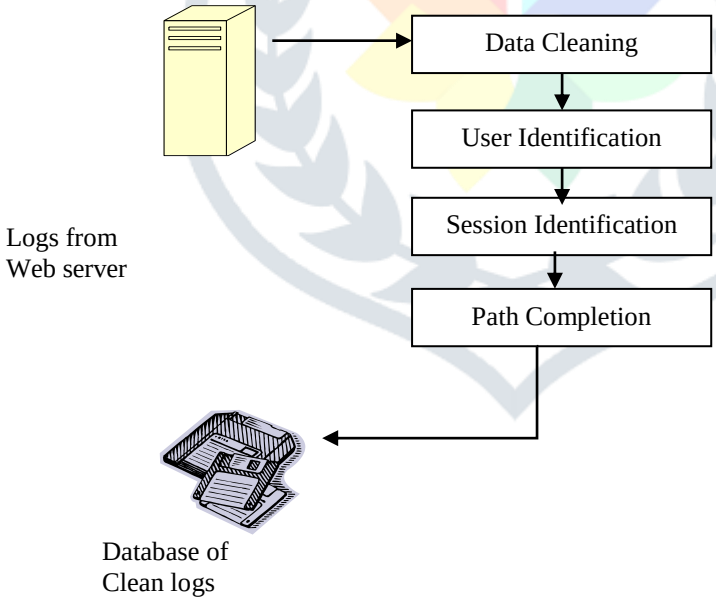

Figure 3: Data Pre-processing

B. Pattern Discovery

Pattern Discovery Tools implement techniques from data mining, psychology, and information theory on the Web traffic data collected. To discover the novel, potentially useful and interesting information, several methods and data mining algorithms are applied such as Path Analysis, Association Rules, Sequential Patterns, Clustering, and Classification etc. [3]

C. Pattern Analysis

The analysis of the pre-processed data is very beneficial to all the organizations performing different businesses over the web. The administrators of web site are interested in frequent usage pattern of web site like "How people are using web site?" "Which pages are being accessed most frequently by user's?" etc. Therefore analyses the structure of usage access log file are required to find the answers of above questions.

III. RELATED WORK

In [1] paper, we propose an efficient weighted algorithm for Web Information Retrieval System. The proposed system will use frequency of a page, time spent on a page and click history of a page to assign a quantitative weight to each page for a user. The nature of this proposed approach is that the time spent on pages, visiting frequency and click history are three factors to show the interest on a page. The Proposed system can get better results of web traversal system. Here they provide a very efficient algorithm for web path traversal. Three factors frequency, time spent and click event were used to decide the Web path traversal. The experimental results show that in the proposed system when we increase the number of parameters for finding the Web path the accuracy of the system is enhanced drastically and Frequency, Time and Click based Page Weight (FTCPW) produces more accurate results than those achieved by Frequency, Time based Page Weight (FTPW) and FTPW produced better than Frequency based Page Weight (FPW). In [2] paper attempts to solve the problem of predicting the next page to be accessed by the user based on the mining of web server logs that maintains the information of users who access the web site. The prediction of next page to be visited by the user may be pre fetched by the browser which in turn reduces the latency for user. Thus analyzing user's past behavior to predict the future web pages to be navigated by the user is of great importance. Their proposed model yields good prediction accuracy compared to the existing methods like Markov model, association rule, ANN. In [3] paper the web server log files data are used of Education Web site for research work of web access prediction. Three fundamental questions that may be asked by the users while navigating the Web site are as follows: 1. Where am I now? 2. Where have I been? 3. Where can I go next? From the current browser, such as Internet Explorer (IE), Mozilla Firefox, etc. user can give a good answer to above first two questions but fail to answer third one. To know where currently the user is, he/she can check address bar field of the explorer. In [4] paper focuses on the combination of Association Rule, Frequent Pattern (Fp), Fp tree for Pattern discovery and Pattern Analysis. In [5] paper describes the methodology in three simple steps which are being discussed in the paper. The Complexity of the Steps is simple as compared to the other Algorithms or methodology that has done by the various authors. Over the decade or two there have tremendous research in this field and extensive development has been recorded. There are various challenges involved which need to look out like the user information and other activities done by the user during the visit of the web site. In this paper we focus on Weighted association-mining method, which has been successfully applied in the pages recommendation system that has not been explored in previous studies because weighted Association Rule mining assigned the different weights to each web page and improves the Association Rule Mining (ARM). They proposed a Page Weight-like algorithm for correct web page prediction. Some of the existing systems used for web traversal based on his or her navigational behavior on the web. The proposed system will use visiting frequency of a page, time spent on a page to assign a quantitative weight to each page for a user. In [6] paper them present architecture for integrating semantic information about the products with web log data and generate a list of recommended products by using LCS Algorithm. The implementation shows good performance in terms of precision, recall and F1 metrics. They have created two tier architecture for integrating semantic information with web usage mining. We have used LCS algorithm to generate a list of recommended products to the user. The results show good performance in terms of Precision, Recall and F1 metrics as compared Anexcellentstylemanualforsciencewritersis given by Young [7]. to the existing recommender system. In [9] paper, a new extension of sequential pattern, ordinal pattern, is proposed. An ordinal pattern is an ordinal sequence of attributes, whose values commonly occur in ascending order over data set. Ordinal pattern mining requests that values of different attributes must be comparable and ordinal. After each record in data set is transformed into an ordinal sequence of attributes according to their ordinal values, ordinal patterns can be mined by means of mining sequential patters. By extension to ordinal patterns, sequential pattern mining proves to be useful in identifying ordinal patterns that uncover errors in data set for data cleaning. Through transforming records to sequences of attributes, ordinal patterns can be mined by algorithms for mining sequential patterns. Ordinal pattern mining incorporates ordinal relationships among attributes. An ordinal pattern reflects a kind of ordinal relationships between attributes that commonly occur over the data. In [10] paper systematic algorithm has been proposed In this algorithm, the user is not allowed to specify any minimum support threshold values to find the frequent patterns; instead the system itself generates the minimum threshold values, thus plugging the loophole of other algorithms. Using this approach, the user is well aware of entire information aiding him to take correct informed decisions. Proposed algorithm will be very efficient in finding the association rules between the items of the provided dataset. The use of time slice also helps in avoiding the multiple scans across the database. This will leads to better performance (search space), minimization of time and useful in effective decision makings. In [11] paper, we find the throughout surfing pattern from the mining of path traversal graph. The Travelling Salesman Problem finds proves more effective to predict one step forward visit to next web page. As we get predictions of visitor, website operator can efficiently reconfigure the personalized website structure and take the help of throughout-surfing patterns for rearranging the contents of the website. First we collect the user login and surfing session from the web server. The surfing session contains the session id and consecutive browsing sessions. In [12] paper, we will do the classification of users on three aspects. Firstly, we will classify the users on the basis of their countries. Secondly, we will classify the users on the basis of their entry to the website, i.e., either direct entry or referred by the other site. Finally, we will classify the users on the basis of time of access, i.e., either accessed the site during summer season or winter season, or accessed it during different months, or accessed it during different

days. This information will then be used for efficient administration and personalization of these websites. This will ultimately result in fulfilling the specific needs of specific communities of users and in increasing the overall profit from the website, which is the main aim of this paper.

IV. PROPOSED WORK

The main objective of the proposed system, Predicting User navigation patterns using Clustering and Classification from web log data (PUCC) is to predict user navigation patterns using knowledge from (i) a Classification process that identifies potential users from web log data and (ii) a clustering process that groups potential users with similar interest and (ii) Using the results of classification and clustering, predict future user requests based on the Engine-one or Engine-two comparisons.

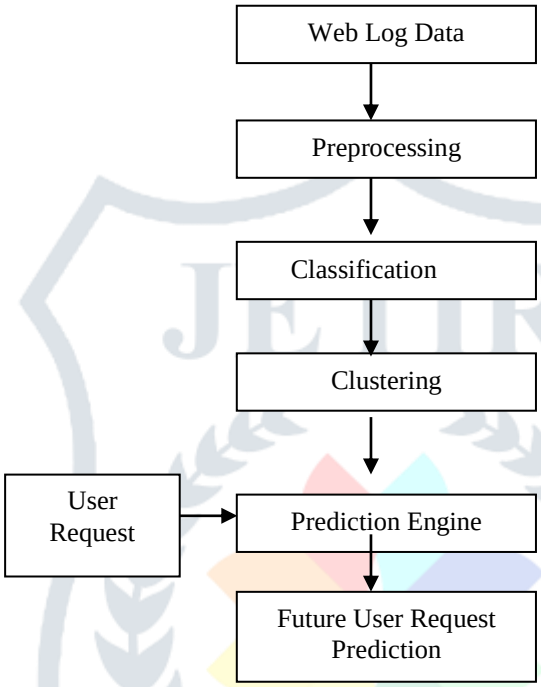

Fig 4:- Flow of work

Steps of Proposed Work:
Step 1: Original data

Step 2: Cleaning through regular expression matching algorithm

Step 3: Classification of potential user through Thresholding method

Step 4: Clustering of user navigation prediction through weighted graph algorithm

Step 5: Using association rule (Frequent pattern) for Prediction Engine1

Step 6: Using LCS algorithm for Prediction Engine1

Step 7: For decision making, compare and take out the common of Step 5 & 6.

V. CONCLUSION AND FUTURE WORK

This paper's objective is to present the Improved Enhanced User Navigation Pattern Prediction Technique from Web Log Data. This paper surveys some of the techniques in order from the year 2011 to 2014. The techniques considered in this paper are Association rule mining using Frequent Pattern Growth Mining and Longest Common Subsequence algorithm during prediction. The expected result should show combination of both of these techniques giving better accuracy of prediction and excellent result. Future work can be done in the area of better cleaning, Extraction and Clustering of data, so that this may yield the prediction in a

much refined way. Also the hybrid algorithm can be used instead of combination of these two algorithms. Using combination of two algorithms may bring a increase in time complexity, future work could be done to bring efficiency in terms of time also.

REFERENCES

[1] Rohit Agarwal, K. V. Arya, Shashi Shekhar, Rakesh Kumar, "An Efficient Weighted Algorithm for Web Information Retrieval System" 2011 International Conference on Computational Intelligence and Communication Systems, pp 126-130,2011.

[2] Poornalatha.G, Prakash S Raghavendra, "Web Page Prediction by Clustering and Integrated Distance Measure" 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp 1349-1354, 2012

[3] Priyanka S. Panchal ,Prof. Urmi D. Agravat, "Hybrid Technique for User's Web Page Access Prediction based on Markov Model" IEEE - 31661, 4th ICCCNT 2013, July 4-6, Tiruchengode, India

[4] Kaushal Kishor Sharma, Prof. Kiran Agrawal, "A Hybrid Approach for Predicting User's Future Request" 2014 Fourth International Conference on Communication Systems and Network Technologies, pp 439-443, 2014

[5] Rohit Agarwal, KarmVeer Arya, Shashi Shekhar,"An Architectural Framework for Web Information Retrieval based on User's Navigational Pattern" 2010 5th International Conference on Industrial and Information Systems, ICIIS 2010, Jul 29 - Aug 01, 2010, India

[6] Sneha Y.S, Dr G. Mahadevan, Madhura Prakash M, "An Online Recommendation System Based On Web Usage Mining and Semantic Web Using LCS Algorithm" 2011 IEEE, pp 223-226, 2011.

[7] Maryam Jafari, Farzad SoleymaniSabzchi, Shahram Jamali, "Extracting Users'Navigational Behavior from Web Log Data: a Survey" Journal of Computer Sciences and Applications, Vol. 1, No. 3, pp 39-45, 2013.

[8] Hasimah Hj Mohamed, Tee Leong Kheng, Chee Collin, Ong Siong Lee, "E-Clean: A Data Cleaning Framework for Patient Data"2011 First International Conference on Informatics and Computational Intelligence, pp 63-68.

[9] Y.B. Liu, D.Y. Liu, "Mining Ordinal Patterns for Data Cleaning" 2004 IEEE, pp 438-443.

[10] Sangeetha S, "Verdict of Association Rule Using Systematic Approach of Time Slicing for Efficient Pattern Discovery" 2012 International Conference on Computing, Electronics and Electrical Technologies [ICCEET], pp 994-999.

[11] Sagar More, "Modified Path Traversal for an Efficient Web Navigation Mining" 2014 IEEE International Conference on Advanced Communication Control and Computing Technologies (lCACCCT), pp 940-945.

[12] Anshul Bhargav, Munish Bhargav, "Pattern Discovery and Users Classification through Web Usage Mining" 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp 632-636