IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Machine Learning Based Approach for Web User Behavior Identification

Gaurav Vyas
M.Tech. Research Scholar,
Dept. of CSE, SIRTS, Bhopal

Prof Chetan Gupta
Assistant Professor,
Dept. of CSE, SIRTS, Bhopal

Dr Kapil Chaturvedi
Associate Professor,
Dept. of CSE, SIRT, Bhopal

Abstract — As we know that data size is huge and complex and it is very difficult to analyze the data manually. So for that we need a mechanism to discover patterns of user behavior for the web apps optimization which automatically give a result and according to that result a decision will be taken to create web pages which are associated with each other according to user's interest for that we apply web usage mining techniques by which we can extract the information from the web logs which is useful to us.

Keywords: Web Mining, Web Usage Mining, Association Rules mining.

I. INTRODUCTION

Web mining [1] is an application of data mining. In web mining we can extract meaningful patterns or meaningful information from huge or complex data sets by using web documents or web related activities. According to researchers. Web Mining is divided in to three categories.They are

A. Web Content Mining

Web content mining [2] is to get the useful information by using the mining techniques on text, images, video etc. Web content mining is also known as Web Text Mining.

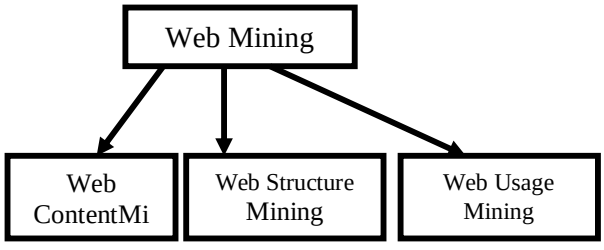

Fig.1- Web Content Mining

B. Web Structure Mining

Web structure mining means mining the data from the hyperlinks or the structure of the website. In Web structure mining graph theory is use to analyze the nodes and the interlink structure.

C. Web Usage Mining

Web usage mining [3] is the application of data mining. Web Usage mining means mining the web log data or the browsers log data. It is an activity which automatically discovers the patterns of one or more web servers.

Introduction to Web Log files.

When the user submits a request to a web server one file is created which are having an information related to web activity and this file is known as web log file.[4]

Location of Web Log Files

The three places of Web log files are:

(i) Web Server
(ii) Proxy Server
(iii) Client Browser.

(i) Web Server log files:

In the web server log files, the accurate and complete information will be stored.

(ii) Proxy server-side log files:

Proxy server take the HTTP request from the client side and send to the web server then result passed to

web server and return to user. But the difficult task is the construction of proxy server and the request interception is also limited.

(iii) Client Browser:
Log file can reside in client's browser HTTP cookies used for client browser. HTTP cookies needs for future access.

II. WEB LOG STRUCTURE

Web server logs [5] are plain text files. This text file may be a comma delimited, space delimited or tab delimited. In web log each line (record) have some fields via Remote Host Field, Date/Time field, HTTP Request field, Status code field, Transfer volume (Bytes) field.

(i) Remote Host Field
The Remote Host field consists of IP address or Domain name of the remote host making the request.
(ii) Date/Time field
Date/Time field are in [DD: HH:MM: SS]
Format. Where DD is the Date: HH is the Hour represent in 24 hours: MM is the minute: SS is the second.
(iii) HTTP Request Field
The HTTP Request field contains the information that client Browser has requested from the web server.
(iv) Status Code Field
Status code field gives three-digit responses from the web server to the client browser about the success or failure of the requested. If there was an error. Following series help us to identify the success or failure.
(v) Transfer Volume (Bytes) Field
Transfer Volume indicates the size of the file sent by the web server to the client Browser.

Successful Transmission	(200 Series)
Redirection	(300 Series)
Client Error	(400 Series)
Server Error	(500 Series)

III. PHASES OF WEB USAGE MINING

The phases of web usage mining [6] are:
(i) Pre-processing
(ii) Pattern Discovery
(iii) Pattern Analysis

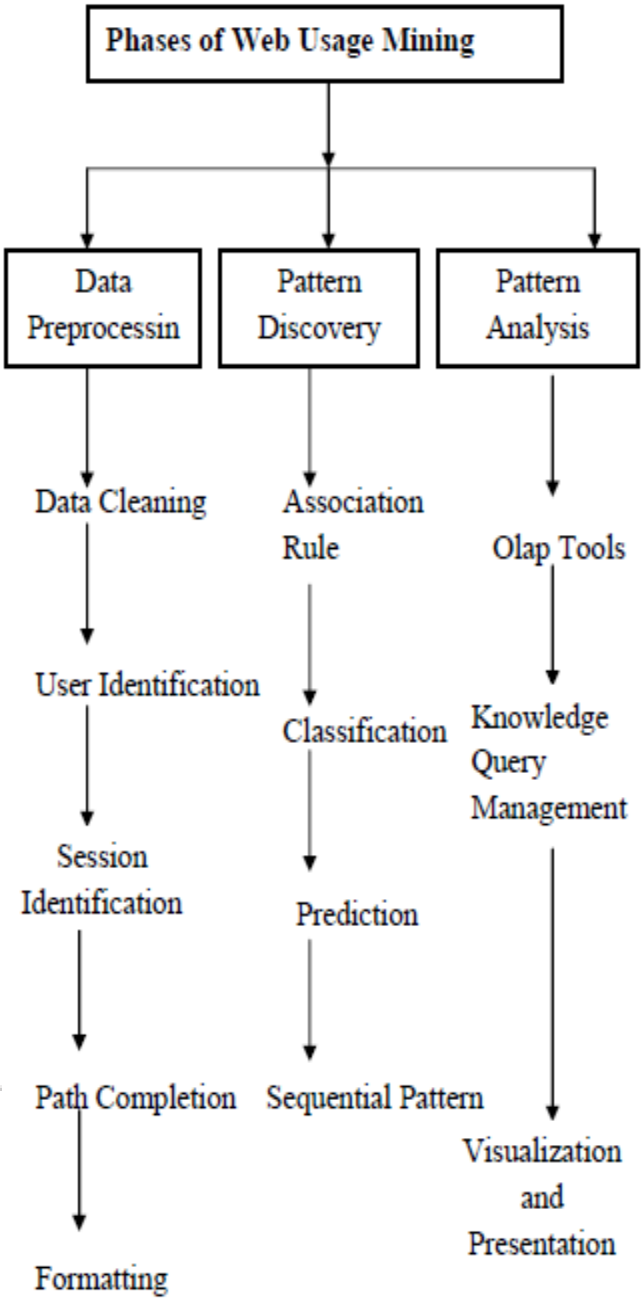

Fig.2- Phases of Web Usage Mining

(i) Data Pre processing

- The main aim of data preprocessing is the accurate and the reliable data.
- The main procedure of data preprocessing are: Data Cleaning, User Identification, Session Identification, Path Completion Formatting

- Data Cleaning

  It is the process of removing unnecessary or repetitive data. The following types of data should be removed:

  (a) Download not on the basis of the User's Request. Ex entries such as jpeg, gif,css other audio/video files which is not requested by the user should be removed.

  (b) If user request some pages and they are not available on web server. These are marked by some error codes such as Success or Failure.

- User Identification

  The following methods are there to identify the user

  (a) I.P. address: If the I.P. address is same then check the user's browser and O.S. If both are same that means they are same user otherwise not.

  (b) User Registration

  (c) Cookies

- Session Identification

  The session identification defined as the set of web pages viewed by a particular user for a particular purpose or as long as user is connected to the web site then it is called session of that particular user.

(ii) Pattern Discovery

- The types and level of analysis performed on the integrated usage data depend on the ultimate goals of the analyst and the desired outcomes [11].
- Path discovery includes Association rule, Classification, Prediction, Sequential Pattern[12].

- Association Rule

  Association Rule [6] help us to find the group of items or pages that are commonly purchases or accessed together. This helps us to make the web site content more effective.

- Classification

  Classification [6] is the task of mapping data item into one of the several predefined classes.

- Cluster Analysis

A technique which can make a group of users or a data item which are having the same characteristics. Clustering can help to make or execute the future strategies. Clustering can help to discover a group of users who are having the same navigation pattern.

(iii) Pattern Analysis

Pattern Analysis [12] is the last stage of Web Usage Analysis. In Pattern Analysis techniques are use as follows:

OLAP/Visualization Tool: For multidimensional analysis & Decision making.

Knowledge Query Management.

Intelligent Agent.

IV. ASSOCIATION RULE

Association Rule [8] mining was first proposed by Agrawal, Imielinskiand Swami. Association rule mining help us to identify the frequent item set from the transactional data base. In Association rule mining two conditions are satisfied (1) Support and (2) Confidence.

The best example of an Association rule is Market Basket[10] Analysis. Here the main objective is to identify the buying behaviour of the customer by finding the associations between the different items that customer places an order.

An Association rule [8] is defined as follows:

Let I ={i1,i2…… in} be a item sets and T ={t1,t2….tm} be a transaction that contains a set of items such that T ⊑ I, D be a database with different transaction records Ts. An association rule is then an implication in the form of X→Y, where X, Y⊆ I are the item sets and X ⋂Y = Φ.

Here X is called antecedent while Y is called consequent. The rule means X implies Y.

The two important parameters of Association mining. Support and the Confidence.

Support is the how frequent an item appears in the data set. The formulae of Support(s) are defined as follows:

$$\text{Support}(X,Y) = \frac{\text{Support count of } XY}{\text{Total number of transactions in Dataset.}}$$

Confidence:

Confidence is the ratio of transaction that contain X and Y to the number of records that contain X.

$$\text{Confidence} = \frac{\text{Support } (X \cup Y)}{\text{Support } (X)}$$

The algorithms used in Association rule mining are: Apriori, Elact, FP-Growth Algorithm.

The Apriori Algorithm [9] proposed by R.Agarwal and R.Srikant in 1994 for mining frequent item sets. Apriori works on iterative approach known as level

wise search. Where k itemset are used to explore(k+1) itemsets. To improve the efficiency of the level wise generation of frequent item sets. An Apriori property is there to reduce the search space. The property states that if an item set less than the minimum support threshold, min_supp than that item set is not frequent. If an item X is added to the item set I, then the resulting item set i.e. X ∪ I cannot occur more frequent than I.

V. PROPOSED APPROACH FOR WEB USAGE MINING

The proposed approach for the Web Usage mining is:

Step1: Collection of Web log file and in Web log files applying preprocessing techniques. When the data is preprocessed then next task is to store in the database.

Step2: To find the discovery pattern by applying the data mining techniques such as clustering and the association rule. In our work we shall use the combined approach of clustering and association rule mining. In this step we shall use partitioning based clustering algorithm DBSCAN.

Step3: After getting the clustered group data we shall use the association rule mining technique to find the user access pattern.

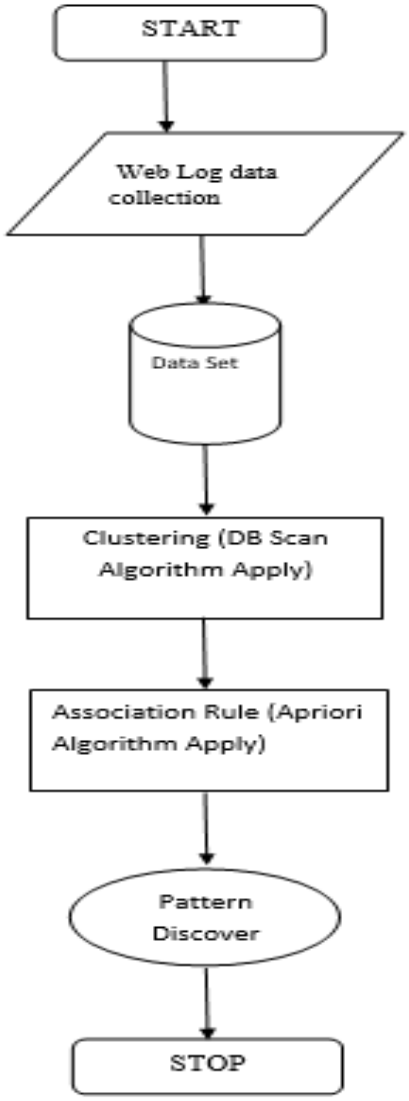

Fig. 3 – Proposed framework

VI. Conclusion

The main aim of Web Usage Mining is that to identify the behaviour of the user as how they are access the web sites. And providing the contents according to the User's Interest. In this approach this aim shall be full fill by using the association rule mining techniques in our clustered data.

REFERENCE

1. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011
2. Dunham MH. Data mining: Introductory and advanced topics. Pearson Education India; 2006.
3. Markov Z, Larose DT. Data mining the Web: uncovering patterns in Web content, structure, and usage. John Wiley & Sons; 2007.
4. Chaturvedi K, Patel R, Swami DK. Fuzzy C-Means based Inference Mechanism for Association Rule Mining: A Clinical Data Mining Approach. International Journal of Advanced Computer Science and Applications. 2015 Jun 1;6(6): pp. 103-10.
5. Jaideep Srivastava ,Robert Cooley "Web usage mining : Discovery and Applications of usage pattern from web data". SIGKDD Explorations- vol -1 ,issue -2 Jan 2000, pp. 12 – 33.
6. Liu B. Web data mining: exploring hyperlinks, contents, and usage data. Berlin: springer; 2011.
7. Naga Lakshmi, Raj Sekhara Rao , Sai Satyanarayana Reddy, "An Overview of Preprocessing on Web Log Data for Web Usage Analysis", International Journal of Innovative and Exploring Engineeering (IJITEE) ISSN:2278-3075, Volume-2 ,Issue-4,March 2013,
8. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. InProceedings of the 1993 ACM SIGMOD international conference on Management of data 1993 Jun 1 (pp. 207-216).
9. R. Agrawal and R. Srikant. Fast algorithm for mining association rules in large databases" Proceeding of the International Conference on Very Large data bases (VLDB) 1215, Sanjosh CA, pp. 487-499,1994.
10. Bakariya B, Chaturvedi K, Singh KP, Thakur GS. Efficient approach for mining top-k strong patterns in Social Network Service. In2016 Fifth International Conference on Eco-friendly Computing and Communication Systems (ICECCS) 2016 Dec 8 (pp. 104-108). IEEE.
11. Chaturvedi K, Patel R, Swami DK. An Efficient Binary to Decimal Conversion Approach for Discovering Frequent Patterns. International Journal of Computer Applications. 2013 Jan 1;75(12): pp. 29-34.
12. Chaturvedi K, Patel R, Swami DK. An Inference Mechanism Framework for Association Rule Mining. International journal of advanced research in artificial intelligence. 2014 Sep;3(9).