An Improved User Browsing Behavior Prediction Using Web Log Analysis

Vedpriya Dongre, Jagdish Raikwal

***Abstract* - Web usage mining is widely used to discover the usage patterns from web log files. It deals with web log data which are taken from web servers, proxy server or client's cache. By analyzing user's browsing behavior, next web page prediction can be made. Various types of mining algorithms proposed over the years based on different techniques. But prediction of future request of the user mainly concern with its accuracy and efficiency. This survey describes various aspects of Web mining techniques and various issues related to Web usage mining. Along with it, a proposed work is there which addresses some of the core issues of Prediction techniques along with their solutions.**

Index Terms— Web Usage Mining, Personalization, Recommendation, Clustering.

I. INTRODUCTION

Now days, Internet has become an integral part of everyone's life. As the result of this, there is a significant increase of Web Data over the Internet. This Web Data is very huge, unstructured and disordered in nature. In addition of this, due to varying and heterogeneous nature of data, web searching has become an enormous task for the users. So prediction of user's interest or Personalization has become very essential. Web personalization can be viewed as some activity that makes individuals to access web site on their own choice.

A. Web Mining

Web mining, a data mining techniques' application, is used to discover the knowledge from the data on Web, which includes Web documents, hyperlinks of the web pages, log data of web sites, etc. On the basis of different types of web data, web mining can be categorized in three main research fields: first is web content mining, second is web structure mining and third is web usage mining. There is not a clear cut difference between these types of web mining, so different techniques are used in combination to apply on these categories.

Manuscript received May, 2015.
Vedpriya Dongre*, Computer Engineering, Institute Of Engineering and Technology Devi Ahilya University., Indore, India, 9753754889*
Jagdish Raikwal*, Information Technology, Institute Of Engineering and Technology Devi Ahilya University,Indore, India, 9893942596.*

1) Web Content Mining

The process to extract useful data from the content of a web page is called Web content mining. Basically, the Web content data contains text, image, audio, video, hyperlinks, and metadata [2].

2) Web Structure Mining

It is the process to analyze and mine the connection structure of the web sites [2]. This type of mining uses graph theory to perform mining task. A web graph has nodes, which represent web pages and edges, which represent hyperlinks that connect related pages.

3) Web Usage Mining

Web Usage Mining is a data mining application where mining methods are applied to the records of web access log files [3]. Basically here we analyze the users' browsing behavior based on their navigational pattern. There are three essential steps to perform Web Usage Mining- Data Collection and Preprocessing, Pattern Discovery and Pattern Analysis. Web Usage Mining is mainly applied to the area of Personalization, System Improvement, Business Intelligence, and User Characterization [4].

B. Web Access Log Mining

Web mining has three main categories, which involve a wide range of applications to analyze, extract and discover the useful information from the Web. Web mining also provides the means to make data access more efficient and accurate. Another important application of Web mining is to fetch data of users' activities stored as log files, analyze them and discover the information for various applications such as web prediction [5].

There are two types of data available on the web, Primary data and Secondary data. Primary data can be understood as the documents which are accessible through Web, whereas log files are the secondary data which can be used for Web usage mining.

Sources of log files are Web servers, Proxy servers and client caches. Mining process becomes more difficult when information of users' activities is stored at more than one place. But if one wants efficient result, he should take data from all three sources of log file, because web server doesn't keep track of the data, which is kept by proxy server or client

caches [6]. Also Proxy server has some additional information which may not be on Web server. However, Client side caches have all the records of page requests, but it is very difficult to get all the information from client side.

C. Applications of Web Usage Mining

1) Personalization

As Personalization is used for the purpose of customization, which presents a model for pleasing an individual's needs in anticipation and to provide correct results to the users. Another application of Web Usage Mining is Recommender System. Web Usage Mining works on web log files, which have user's navigational details. Using these web log files and user's current navigation pattern, Recommender System recommends next web files to the user in form of recommendation list. Web usage mining is an excellent approach for achieving this objective as described in existing recommendation systems.

2) System Improvement

Performance and other system quality factors are very important for any system to satisfy the user needs. These qualities are expected from the user of web services. Web usage mining provide the way to understand users' behavior by analyzing the network traffic & provide means to improve the performance factors of the web sites.

3) Site Modification

Web site's design is also important factor for website's popularity. User interface attract more users to access that site. Web usage mining analyzes the users' behavior for detailed feedback and also provides the information of the websites' designers. This information can be used to take redesign decisions. The structure of a Website changes periodically on the basis of usage patterns discovered from server logs.

4) Business Intelligence

Customers' usage of the web site is also an important factor for business analysts and marketers of e-businesses. The extraction of web data is done for marketing analysis. Web log data is used in many e-commerce applications with marketing. For knowledge discovery Web usage mining is used in order to identify the customer relationship.

II. BACKGROUND

A. *User Behavior Analysis*

Human behavior is been analyzed within various disciplines, such as economics, marketing, medical science, linguistics etc. To analyze the web users' browsing behavior, many frameworks are available, which are capable of providing high performance. These disciplines are used to extract the knowledge and then to discover the users' browsing behavior. After analyzing the data theoretical models are calibrated. Web log data can be obtained web servers, proxy servers or client cookies, which store the users' action while using the Web. These log files available in many formats and also having millions of records.

Analysis of user behavior includes what pages do users stay longer, what path they use to navigate from one page to another, why they may not reach the target pages, what pages impede search for information, what pages attract the most attention, etc. It is useful to analyze the web server logs at a regular interval, as they contain system information about the server's functions and log many user actions.

Of particular importance understands the behavior that users exhibit in while they using Web, since changes in these behaviors could affect the utility of the web prediction. In this synopsis I am representing the study on the behavior of users using numeric features and to discover the types of behavior, or roles, that uses pattern recognition algorithm. The process of extracting behavioral features, the combined approach of clustering and pattern recognition through which we predict the web pages based on those patterns.

B. *Related Work*

Web Usage Mining is an emerging field in research area. Many algorithms are used for Web Usage Mining in order to get better, accurate & efficient results such as Mehrdad Jalali *et al.* [8], Gang FANG *et al.* [9], Kobra Etminani *et al*. [10], Mamoun A. Awad *et al.* [11], Ashika Gupta *et al.* [12].

In [8], Mehrdad Jalali et al. gave the solution based on LCS algorithm for analyzing and process the user navigation patterns for next web page prediction. Their architecture has improved accuracy of classification & also it provides efficient online prediction . Some evaluation techniques also used for evaluating quality of the prediction found.

In [9], Gang FANG *et al.* proposed a double algorithm of Web usage mining based on sequence number, which is suitable for mining any session patterns in order to improve efficiency of presented algorithms and reduce the time of scanning database. They used the algorithm that turns session pattern of user into binary, and then uses up and down search strategy to double generate candidate frequent item sets. They also computed support by sequence number dimension in order to scan once session pattern of user, which is different from traditional double search mining algorithm. Their experiment indicates that efficiency of the algorithm is faster and more efficient than presented similar algorithms, such as, B_Apriori and B_ARDSM.

Kohonen's SOM (Self Organizing Map) model is applied to pre-processed web logs by Kobra Etminani *et al.* in [10] for

clustering method. They used University's web server logs to extract the frequent patterns.

Markov Model is most widely known algorithm for Web Usage Mining. Mamoun A. Awad and Issa Khalil in [11] presented a new modified Markov model to overcome the issue of scalability in the number of paths. They also presented a new approach for creating classifier EC, which is based on two-tier prediction framework based on the training examples and the generated classifiers. Two-tier framework contributed to preserving accuracy (although one classifier was consulted) and reducing prediction time. The comparative results also show that large number of *N*-grams in the all-*K*th model does not always produce better prediction accuracy. Smaller *N*-gram models perform better than higher *N*-gram models in terms of accuracy.

One of the algorithms, which are very simple to use and easy to implement the Web Usage Mining task, is Apriori algorithm. Ashika Gupta *et al.* in their research work [12] emphasizes on web usage mining and has progress in web utilization with the help of web logs. The bonding of memory and time usage is compared by means of Apriori algorithm and improved Frequent Pattern Tree algorithm. But the main drawback of Apriori algorithm is that the candidate set creation is costly, if the data set is large and a long pattern is recognized. But FP-growth algorithm is not find good enough because it has lack of generating a good candidate method. Future research can combine FP-Tree with Apriori candidate generation method to solve the disadvantages of both apriori and FP-growth.

All these work for Web Usage Mining and recommendation is done to improve the accuracy and efficiency of the system. But still some performance issues are there. We present architecture and propose two prominent algorithms- k-means clustering algorithm and regression analysis. k-means is mostly used algorithm for clustering purpose and hence efficient too. Regression Analysis is an accurate method for prediction that applied on numeric values. Proposed architecture improves the accuracy and efficiency of prediction.

III. PROPOSED WORK

Web access log analysis and user browsing pattern detection is traditional domain of computing analysis. Using these pattern analysis web administrators and web managers are preparing the plans for next generation application development. This domain also helps to improve the computational ability of web servers and recommender system based applications. Therefore in this proposed study the user browsing pattern analysis is improved for finding the more accurate browsing behaviour of end user for providing ease in web data management.

A. Problem Domain

There are rich variants of browsing behaviour analysis techniques are available but most of them are suffers from the following issues:

1. Web server access log based technique only contains the partial user behaviour therefore need to improve the log management scheme
2. More than one pages are navigated in different times, therefore establishing the correlation between each user event and their corresponding web page is complex to learn by an algorithm
3. Huge data needs large time and space complexity
4. Inaccurate predictive methodology due to less number of feature availability on the user navigation pattern.

B. Solution Domain

In order to improve the performance of the traditional web browsing behavior prediction technique a hybrid scheme is presented in this work. The proposed model is able to discover next web using the personalize manner. Therefore k mean clustering, frequent access pattern mining and the similarity matching techniques are incorporated with the proposed solution. The effective solution and their corresponding work are given below.

1. The web log is prepared at the client end which helps to understand the end client browsing and navigational patterns more accurately.
2. Perform analysis of data using frequent mining technique: this may help to recognize those web pages that are frequently navigated by the user
3. Find pattern of navigation using the frequent access pattern and current navigational pattern using support vector machine.

IV. SYSTEM ARCHITECTURE

The proposed system architecture is given in figure 1. In the proposed system, the web access log data is analysed for finding the hidden navigational patterns. Using this navigational pattern the user access behaviour is estimated and a new data model for predicting next user web page is modelled. The main aim of the proposed work is to investigate the technique of web user browsing pattern analysis. Therefore a number of techniques are investigated and a new model is proposed for improving the performance of next user web page access accuracy.

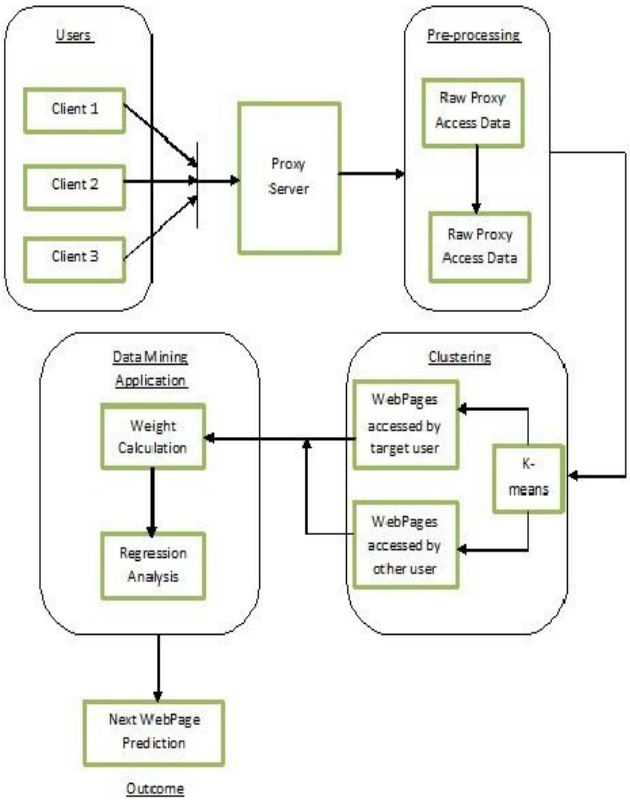

Fig. 1 System Architecture

The concept behind the designed system is to prepare a pre-processed log formation using web proxy log. To achieve low pre-processing time and to get correct classification between two different users from the same IP source, for that below system is designed in two different modules. First client module is designed to gather information from the client machine and submit to the server. There are more than one client is communicate through the server and server generate a log (which is in the form of data base table). At the server end the data mining algorithm prepare a data model using data mining algorithm. Here the actual classification is takes place and user data analysed at the same time system able to predict correct user.

- **End web clients:** These end web clients access the different web pages for finding their data or web pages.
- **Proxy server:** That is an intermediate server, which redirects the client's traffic to the different web servers. Using this server the client's web access data also extracted.
- **Pre-processing:** The raw proxy access data is pre-processed here and only essential attributes are extracted.
- **Access log database:** The accessed data using the proxy server is listed with the help of proxy access log and after pre-processing of data. That is stored in a separate database.
- **K-means Clustering:** That is an unsupervised learning algorithm to solve the clustering problems and to find similar data from huge database.
- **Web page accessed by user:** After process, the data using K-means the similar data according to the target user is separated and their navigational frequency for individual web page is estimated in this phase.
- **Web page accessed by different users:** The K-means also search the targeted user similar web pages accessed by different web users.
- **Weight calculation:** The navigated web pages weights are estimated using the web page accessed by a single user and similar web page accessed by different users.
- **Regression Analysis:** After estimating the weights for a target user, the linear regression analysis is performed using their estimated weights and estimated web pages frequencies.
- **Next web page:** After regression, analysis of data with respect to current navigation pattern provides next accessed web page.

V. CONCLUSION & FUTURE WORK

A. Conclusion

Internet is the fastest growing field with increase in number of data available on it as well as in number of users. Prediction systems provide ease to the user for accessing only needed information. Various techniques have been evolved over years. But most of them are not much accurate and efficient. Therefore a new concept for web page prediction is been proposed in the presented work.

In the proposed model, data will be fetched from the proxy server as web server does not keep track of users' all logs. The extracted data is then preprocessed and saved in an access log database. K-means clustering algorithm is then applied to the refined data to create clusters i.e. web pages accessed by targeted user and web pages accessed by other users. The navigated web pages' weights are calculated. After estimating the weights of web pages, regression analysis algorithm is applied along with their frequencies. Regression analysis is well known method for prediction, it predicts near to accurate data over the numeric values.

B. Future Works

The next work for the proposed model is to implement it correctly. Results will be drawn for number of tests to discuss various parameters like time taken for prediction and memory used to store the data. Based on the result, we will analyze its performance.

Some enhancement can be made in this proposed work. We can also import some data from client caches. Instead of calculating weights of the web pages direct frequencies can be used for prediction. But these enhancements should not affect the performance of the system at any cost.

REFERENCES

[1] Pranit Bari, P.M. Chawan, *Web Usage Mining*, Journal of Engineering, Computers & Applied Sciences (JEC&AS), Volume 2, No.6, 2013.

[2] Amit Pratap Singh, Dr. R. C. Jain, *A Survey on Different Phases of Web Usage Mining for Anomaly User Behavior Investigation*, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 3, May – June 2014

[3] K Sudheer Reddy et al., *An Effective Methodology for Pattern Discovery in Web Usage Mining*, International Journal of Computer Science and Information Technologies, Vol. 3 (2), 2012, 3664-3667.

[4] Resul DAS, Ibrahim TURKOGLU, Mustafa POYRAZ, *Analyzing Of System Errors For Increasing A Web Server Performance By Using Web Usage Mining*, Journal Of Electrical & Electronics Engineering, vol. 7, Number 2, 2007, 379-386.

[5] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava, *Automatic Personalization Based on Web Usage Mining*, Communications of the ACM Volume 43 Issue 8, Aug. 2000, Pages 142-151.

[6] C.P. Sumathi et al., *Automatic Recommendation of Web Pages in Web Usage Mining*, International Journal on Computer Science and Engineering, Vol. 02, No. 09, 2010, 3046-3052

[7] Ms. Dipa Dixit, Mr Jayant Gadge, *Automatic Recommendation for Online Users Using Web Usage Mining*, International Journal of Managing Information Technology (IJMIT) Vol.2, No.3, August 2010.

[8] Mehrdad Jalali1, Norwati Mustapha, Md. Nasir B Sulaiman, Ali Mamat, *A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems*, 12th International Conference Information Visualisation, 2008.

[9] Gang FANG, Jia-Le WANG, Hong YING, Jiang XIONG, *A double algorithm of Web usage mining based on sequence number*, IEEE, 2009.

[10] Kobra Etminani *et al*., *Web Usage Mining: Discovery of the Users' Navigational Patterns using SOM*, IEEE, 2009.

[11] A. Awad and Issa Khalil, *Prediction of User's Web-Browsing Behavior: Application of Markov Model*, IEEE Transaction, 2010, 1083-4419

[12] Ashika Gupta *et al., Web Usage Mining Using Improved Frequent Pattern Tree Algorithms,* IEEE, 2014

