An Improved User Browsing Behavior Prediction using Regression Analysis on Web Logs

Vedpriya Dongre
Institute of Engineering & Technology
Devi Ahilya University, Indore, India

Jagdish Raikwal
Institute of Engineering & Technology
Devi Ahilya University, Indore, India

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. In this paper, we have proposed a new model for predicting the next web page. K-means clustering and Regression Analysis algorithms are used to predict the future request. These two algorithms in combination produce efficient and accurate results.

General Terms
Web Usage Mining, Personalization, Recommendation, Clustering, Prediction.

Keywords
Web Access Logs. Navigation Pattern,

1. 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 action that makes the web experience of a user personalized according to the user's behavior.

Web mining is an application to discover useful information or knowledge from the Web hyperlink structure, page content, and usage data. Also Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web [1]. According to objective and purpose, web mining can be classified into three types, which are Web content mining, Web structure mining and Web usage mining. Web Content Mining or Web Text Mining is the fetching and mining of text, images and graphs of web page to determine the relevance of contents to the search query. When Web Content Mining is used in relation to content data base associated with specific topic, it becomes very effective [2]. Web structure mining is a tool used to recognize the connection between web pages linked by information or direct link connection. Web Usage Mining is that part of Web Mining, which deals with the extraction of knowledge from server log files. Server Logs contains textual data which are available in standard formats [2].

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]. IN Web access log files, data is represented by different formats which contain various attributes like host IP address, requested URL, date & time of access and other information.

Two main applications of Web Usage Mining in this research we have used are Personalization and Recommendation. Personalization can be viewed as a system which is used for filtering the information [6]. It is aimed to provide ease to user as they are using internet with their own tastes on the systems. Users and Web objects are main elements in web personalization, thus it includes categorization, matching of/between these two [5]. We have user profile, based on user's navigation activities, to determine the Personalization actions.

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. In the process of Web Usage Mining, prediction is done in two phases- Training and Testing. In training phase, knowledge base is prepared by using web log files captured from web server, proxy server or client's cookies. And in testing phase, prediction is done by using knowledge base and using current navigation pattern so as to recommend the next web page to users [7].

2. 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] emphasize 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.

3. PROPOSED WORK

3.1 Problem Domain

There are rich variants of browsing behavior analysis techniques are available but most of them are suffers from the different issues. Thus the proposed web access log analysis technique finds the solutions for the following issues in this work.

- Web server access log based technique only contains the partial user behavior therefore need to improve the log management scheme
- 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
- Huge data needs large time and space complexity
- Inaccurate predictive methodology due to less number of feature availability on the user navigation pattern.

3.2 Proposed Solution

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. In order to achieve the proposed aim the following work involved in the proposed study.

- *Study of different web browsing pattern analysis technique:* in this phase different browsing pattern analysis technique is evaluated and most optimum technique is distinguished for prediction and browsing pattern analysis.
- *Design and implementation of the proposed data model:* In this phase a new data model for improving the performance of web page prediction is developed and implemented.
- *Performance study of proposed data model:* in this phase the proposed data model is evaluated using different performance parameters such as accuracy, error rate and time and space complexity.

Previous section described the problem definition and the proposed solution to the problem. Next section demonstrates the proposed system architecture and their components.

4. PROPOSED 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.

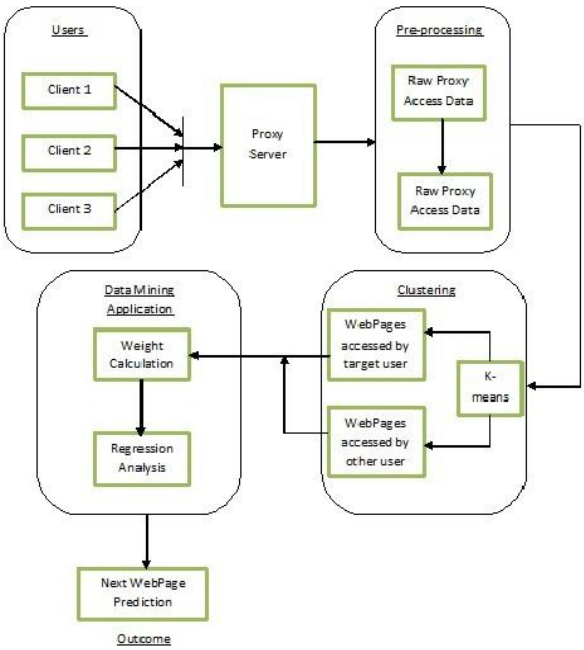

Figure 1: System Architecture

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.

5. ALGORITHM STUDY

In this section the algorithms that are used in our proposed system, for web page prediction are provided.

5.1 K-means Clustering Algorithm:

The K-means calculation is a basic iterative strategy to segment a given dataset into a client determined number of groups, k. The calculation works on an arrangement of d-dimensional vectors,

$$D = \{xi \mid i = 1 , ... , N\}$$

where xi ∈ _d means the ith information point. The calculation is instated by picking k focuses in _d as the beginning k group delegates or "centroids".

Systems for selecting these introductory seeds incorporate examining indiscriminately from the dataset, setting them as the arrangement of bunching a little subset of the information or irritating the worldwide mean of the information k times. At that point the calculation repeats between two stages till meeting:

Step 1: Data Assignment. Every information point is appointed to its nearest centroid, with ties broken discretionarily. These outcomes in a dividing of the information.

Step 2: Relocation of "means". Every bunch agent is moved to the inside (mean) of all information focuses allocated to it. On the off chance that the information focuses accompany a likelihood measure (weights), then the movement is to the desires (weighted mean) of the information segments. The calculation focalizes when the assignments (and consequently the cj values) no more change. Note that every emphasis needs N × k examinations, which decides the time many-sided quality of one cycle. The quantity of emphases needed for union shifts and may rely on upon N, yet as a first cut, this calculation can be viewed as straight in the dataset size. One issue to determine is the manner by which to evaluate "nearest" in the task step. The default measure of closeness is the Euclidean separation, in which case one can promptly demonstrate that the non-negative expense capacity, will diminish at whatever point there is an adjustment in the task or the movement steps, and subsequently merging is ensured in a limited number of cycles. The eager plunge nature of k-means on a non-arched cost likewise suggests that the union is just to a neighborhood ideal, and undoubtedly the calculation is commonly truly delicate to the beginning centroid areas.

5.2 Regression Analysis:

Relapse is the endeavor to clarify the variety in a subordinate variable utilizing the variety as a part of autonomous variables. Relapse is consequently a clarification of causation. On the off chance that the autonomous variable(s) adequately clarify the variety in the reliant variable, the model can be utilized for expectation. The capacity will make a forecast for each watched information point. The perception is meant by y and the expectation is indicated by y.

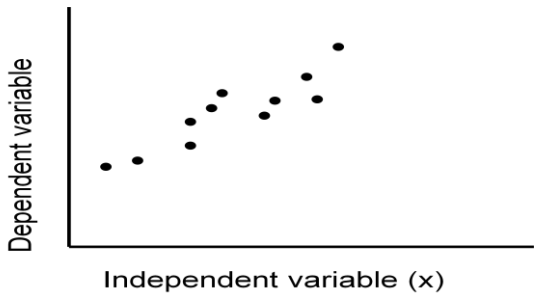

Figure 2: Dataset

The yield of a relapse is a capacity that predicts the subordinate variable heaps of the free variables. Straightforward relapse fits a straight line to the information.

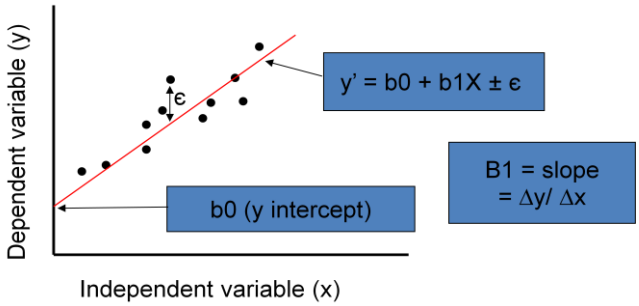

Figure 3: Regression line for Prediction

The capacity will make a forecast for each watched information point. The perception is meant by y and the expectation is indicated by y. For every perception, the variety can be depicted as:

$$y = y^{\wedge} + \varepsilon$$

where, y is perception, $y^{\wedge}$ is forecast and ε is expectation blunder. A minimum squares relapse chooses the line with the least aggregate whole of squared expectation blunders. This worth is known as the Sum of Squares of Error, or SSE.

The Sum of Squares Regression (SSR) is the total of the squared contrasts between the expectation for every perception and the populace mean. The Total Sum of Squares (SST) is equivalent to SSR + SSE.

Numerically,

SSR = $\sum (y^{\wedge} - y)$ (measure of clarified variety)

SSE = $\sum (y - y^{\wedge})$ (measure of unexplained variety)

SST = SSR + SSE = $\sum (y - y)$ (measure of aggregate function).

6. RESULT ANALYSIS

This chapter provides the understanding of the proposed systems performance evaluation therefore a number of different experiments are performed and their performance is evaluated in terms of space and time complexity and compared with the traditionally available data model.

6.1 Accuracy

The amount of data which is correctly identified by the classifier is termed as the accuracy of the system. That can be evaluated using the following formula.

$$accuracy = \frac{correctly\ classified\ samples}{total\ samples\ input} \times 100$$

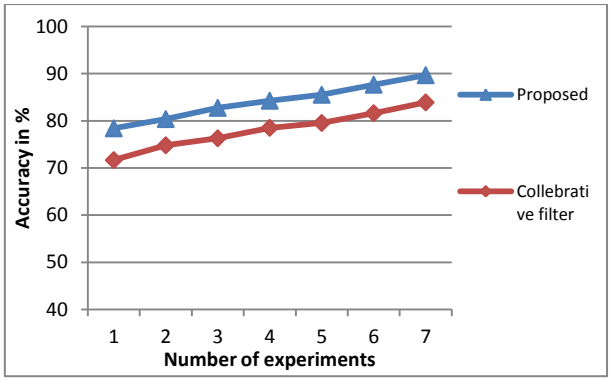

Figure 4: Accuracy

The given figure 5.1 shows the comparative performance among the proposed recommendation engine and the traditional collaborative filter. In order to demonstrate the performance the X axis contains the number of experiments performed and the Y axis contains the percentage accurately classified data. according to the obtained results the performance of the proposed classification technique is provide efficient outcomes as compared to the traditional technique.

6.2 Error rate

The amount of data which is not accurately classified is known as the error rate of the system. That can be evaluated using the following formula.

$$error\ \text{rate} = \frac{\text{total incorrectly classified data}}{\text{total input samples}} \times 100$$

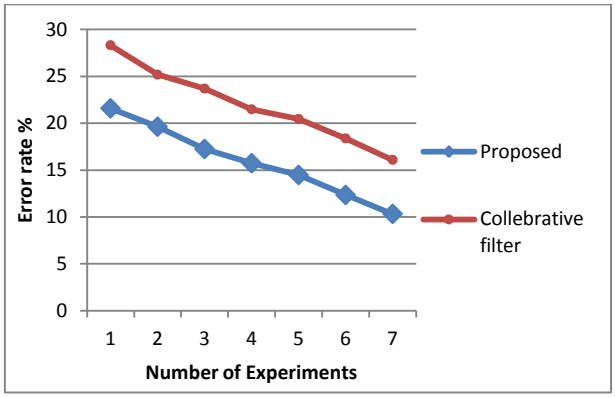

Figure 5: Error rate

Figure 5.2 shows the error rate of the proposed and traditional collaborative filter for recommendation. In this diagram the blue line shows the performance of the proposed technique and the red line shows the performance of the traditional filter. According to the obtained results the error rate of both the filters are reducing as the amount of data for training is increases for learning.

6.3 Memory consumption

The amount of main memory is consumed during the algorithm execution is termed as the memory consumption. The memory consumption of both the algorithms are given using figure 5.3 shows the comparative outcomes of the algorithms in terms of memory consumption.

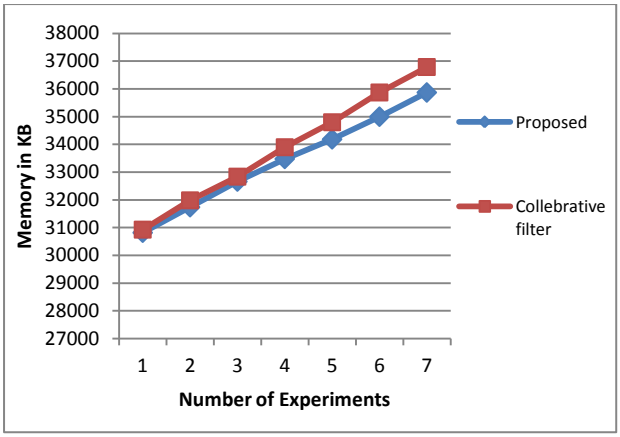

Figure 6: Memory consumption

According to the obtained results the performance of the proposed technique is much efficient than the traditional algorithm because the proposed algorithm consumes less amount of memory (as given in blue line) then the traditional approach (as shown in red line). In order to show the performance the X axis contains the number of experiments performed with the system and Y axis contains the memory consumption in terms of KB.

6.4 Time complexity

The amount of time consumed for data evaluation and prediction of the URL is known as the time complexity of the selected algorithm. Figure 5.4 shows the comparative performance of the algorithm in terms of seconds. Therefore Y axis contains the time in seconds and the X axis contains the number of experiments performed with the system. according to the obtained results the performance of the proposed system is much adoptable than the traditional approach of URL recommendation.

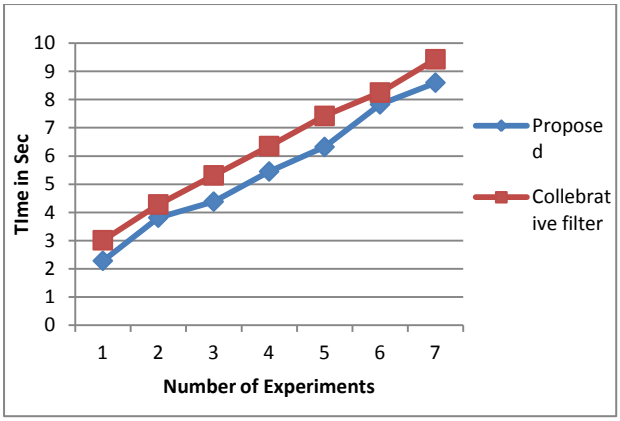

Figure 7: Time complexity

7. CONCLUSIONS

In this paper, the proposed system predicts the requested web page for the users. To evaluate and work with the web mining application, we design basic proxy server based architecture. In this work, first required to make connection with the service provider and then make request by client end. Any request made through the client browser makes an entry to the proxy log. Now the time to introduce the data mining application to over collected data from different users, according to the classification of data system able to predict the user who is login for network access. Two basic algorithms are used K-means clustering algorithm and Regression analysis algorithm. By using these algorithms we got more accurate and efficient results. The system is adoptable due to high accurate predictive results and the less resource consumption in terms of time and space complexity. The given model can be enhanced for improving the data evaluation time and also for improving number of data instance for evaluation.

8. 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