User Navigation Pattern Prediction using Statistical Classifier and Modern Techniques

Yogesh Rajaram Bhalerao¹, Prof. P. P. Rokade²

¹Student, Department of Computer Engineering, SND COE AND RC, Yeola, Nasik (Maharashtra)-India.
²Head, Assistant Professor, Department of Information Technology, SND COE AND RC, Nasik (Maharashtra)-India.

Abstract—Web Usage Mining (WUM) are intentionally formed to carry out the user navigation pattern in web log files. Organizations collect huge volumes of data in their daily operations through online Interface, generated automatically by web servers and collected in server logs and this idea is used for analyzing the data representing usage in domain and to predict the future actions by users. The improved pairwise nearest neighbor algorithm is used to group the potential clusters and the maximum likelihood classification algorithm with decision tree is to predict users future requests. The Experimental result shows improvement in the quality of clustering for user navigation pattern in web usage mining and for the prediction of users next request. This document gives formatting guidelines for authors preparing papers for publication in the International Journal of Emerging Technology and Advanced Engineering. The authors must follow the instructions given in the document for the papers to be published.

Keywords—User navigation pattern; Web log data; statistical classifier; Data Mining; improved navigation

I. INTRODUCTION

World Wide Web is huge repository of web pages and linking between them. They provide us wide information regarding Internet users. This area has tremendous growth and development in online users behavior recognition web usage mining is application of mining techniques in server logs. Log files are growing more in recent view, its size is becoming large preprocessing over them has vital role in efficient mining process traditionally for noise and appropriate analysis sessions and paths of missing pages are reconstructed in processing. All the transactions illustrate the behavior of users are in processing by calculating the reference length of user access by means of measurement unit. Using clustering techniques multiple types of objects clustered, different groups with common characteristics are clustered. Online user’s navigation performance is ability to capture behavioral ambiguity. They are useful in predicting future actions. The experimental result showed improved performance in the same through modern techniques.

A. PREPROCESSING

Pre-processing is process of converting the usage, content, and structure information contained in the various available data sources, this data abstractions necessary for pattern discovery. It always necessary to adopt a data cleaning method to eliminate the impact of the irrelevant items to the analysis result. The usage pre-processing probably is the most difficult of web user navigation patterns and proposed a novel approach task to the incompleteness of the available sufficient data; it is very difficult to identify the users.

B. PATTERN DISCOVERY

This intermediate phase consists of different techniques derived from various fields such as User statistics, machine level learning, Information Retrieval, Pattern recognition etc., applied to the Web group and to the available dataset. The task for learning the patterns offer some techniques as statistical analysis, suggestion rules, clustering, classification, sequential pattern, and dependency modeling.

C. PATTERN ANALYSIS

Pattern Analysis is the conclusion stage of WUM (Web Usage Mining), which involves the validation of truth and interpretation of the discovered patterns. Validation: to eliminate the not required rules or patterns and to extract the interesting rules or required patterns from the output of the pattern discovery process. Interpretation: the output of data mining algorithms is mainly in mathematical form and not understood by direct humans. The main objective of the proposed method is to predict the user interest by a classification process that identify the potential user from the web log data and the clustering process to group the similar interest of users and the classification process to predict the future request.
II. LITERATURE SURVEY

A. Motivation

Identifying Web browsing strategies is a crucial step in Website design and evaluation and requires methods that provide information on both the extent of any particular type of user behavior and the motivations for such behavior [9].

B. User Navigation Strategy

Pattern discovery from web dataset is the key component of web mining and it unite algorithms and techniques from several research areas, Baraglia and Palmerini (2002) proposed a Web Usage Mining system called SUGGEST that provide useful information to make easier the web user navigation and to optimize the web server performance.

Liu and Keselj (2007) proposed the automatic classification of web user navigation patterns and proposed a novel approach to classifying user navigation patterns and predicting users future requests and Mo basher (2003) presents a Web Personalized system which provides dynamic recommendations, as a list of hypertext links, to users. Jespersen et. al. (2002) [10] proposed a hybrid approach for analyzing the visitor click sequences. To form extremely dense clusters, it is proposed to do pair-wise nearest neighbor base clustering approach using only k neighbors. A. anitha in 2010 [7], Maximum Likelihood classification for knowledge discovery from users navigation patterns.

Eltahir, M.A., Anour, F.A., AllFa D 2013 Extracting knowledge from web server logs using web usage mining have improved approach from Modern Techniques than previous one. In 2012 Weichbroth, P., Owoc, M. Online User Navigation Pattern Discovery from World Wide Web server log files gives various ways to prepossessing; hence it reduces size of log file.

The conclusion based on the literature survey is that various researches had done on User Navigation Pattern Prediction approach. In existing research various algorithms of pattern discovery techniques like graph partition techniques of clustering, LCS and Naive Bayesian Techniques of classification etc. are used for user navigation Pattern Prediction and many types of models are developed. For prediction, World Wide Web has necessitated the users to make use of automated tools to locate desired information resources and to follow and assess their usage pattern. Web page prefetching has been widely used to reduce the user access latency problem of the internet; its success mainly relies on the accuracy of web page prediction. Recommendation model is the most commonly used prediction model because of its high accuracy.

Low order recommendation models have higher accuracy and lower coverage. The higher order models have a number of limitations associated with i) Higher state complexity, ii) Reduced coverage, iii) Sometimes even worse prediction accuracy.

Clustering is one of the best solutions for resolving the problem of worse prediction accuracy of Prediction Engine. It is a powerful method for arranging users’ session into clusters according to their similarity using nearest neighbor. I have discussed some of the techniques to overcome the issues of web page prediction. As the web is going to expand, web usage in web databases will become more and more. The above findings will become good guide in web page prediction effectively. In this project, I have presented a comprehensive survey of up-to-date researchers of web page prediction. Besides, a brief introduction about web usage mining, clustering, Classification and web page prediction have also been presented. However, research of the web page prediction is just at its beginning and much deeper understanding needs to be gained.

III. PROPOSED WORK

A. WEB LOG FILES

Web log file is log file automatically created and maintained by a web server. Every hit to the Web site, including each view of a HTML document, image or other object, is logged. These files maintain the browsing performance of site visitors.
Figure 2: Sample Web Log File

Dataset can be collected from multiple groups of users on a single site. Log files are stored in various formats such as Common log [8] (fig 2) or combined log formats. This line consist the following attributes.

i. Client IP address
ii. User id (-if anonymous)
iii. Access time
iv. HTTP Request Methodology
v. Path of the resource utility on the Web server
vi. Protocol stack used for the transmission
vii. Status code returned by the respective server
viii. Number of bytes transmitted in unit

B. DATA CLEANING

Irrelevant information which is not useful for mining purposes can be removed from the HTTP server log files e.g. entree performed by spiders, crawlers, and robots [7]. These are spontaneous agents that surf the web to collect and store the information e.g. search engine spiders and files with extension name jpg, gif, css for various system.

C. USER IDENTIFICATION

Identification of individual users who access a web site is an important step in web usage mining. Various approaches are to be followed for identification of potential users. The easiest method is to assign different user ID to different IP. If the IP of a handler is same as previous entry and user agent is different then the user is assumed as a new user. If both IP and user agent are same then referrer URL and site topology is checked. If the requested page is not directly reachable from any of the pages visited by the user, then the user is identified as a fresh user in the same address [8].

D. SESSION IDENTIFICATION

The browsing speed is calculated as the number of viewed pages / session time. After handling the network robot entries, a series of judgment rules are applied to group the users as potential and not-potential users. Given a dataset of training data having valid log attributes.

C4:5 classification algorithm is used to group the potential users and The attributes selected are

Input:- User Session Time(t),Number Pages Accessed(n),Browsing Method.

Process:-
1. Time > 30 seconds, Number of pages referred in a session
2. (Session time= 30minutes) and the access method used.
3. The decision rule for identifying potential users is,
4. If (session Time > 30 minutes) and Number of pages accessed > 5) and Method used is POST
5. Then the classify user as Potential else classify as Not-Potential.
6. The determination of introducing classification is to decrease the size of the log file.
7. This decrease in size will help for effective clustering and prediction.

Output:- Set of Potential and Non Potential Users(P[],NP[])．

E. CLUSTERING PROCESS

Clustering is the process of grouping the objects in such a way that, intra cluster similarity is high and inter cluster similarity is low. The pair wise nearest neighbor approach is a bottom up hierarchical clustering technique, by which every object belongs to separate clusters initially, pair wise merging of objects is done at every stage based on their likeness. Lastly, resultant in a single cluster. The distance calculations are replaced by likeness measure. Similarity between two transact ions are given by the ratio of, total number of unique pages referenced by them to the number of common references.

For every operation, its first k-nearest access sequences are identified. Among the complete set of k - neighbors, the pair of sequences having high likeness is identified and merged. For this new cluster, the new k neighbors are identified from 2k neighbors and its back neighbors are also updated [4]. The merging is continued until no more integration is possible. In this process, only the pair of access sequence having likeness values greater than a redefined. Threshold is selected for inclusion process.
By this method, the unfriendly objects that are irrelevant to mining process are eliminated resulting in homogeneous Access patterns.

**Input:** - Collect the access log information from various Type of web servers

**Preprocessing:** - Retrieve only IP and URL details from access log by removing noise and filtering irrelevant details.

**Optimizing:**-Form clicks stream transactions and place each transaction in individual cluster Perform pattern Discovery by improved pair-wise nearest neighbor method

On k neighborhood as follows:

Find out similarity values between transactions

i. Identify first k neighbors having similarity greater than Threshold, for every transaction and remove other neighbors
ii. Cluster the pair with highest similarity
iii. Update similarity for objects in the neighborhood of Merged pair
iv. Find out new set of k neighbors from 2k neighbors of merged pair (vi) Update the neighbors in the back list of merged pair
v. Repeat steps (iii) to (iv) until no more merging is possible.

**Output:** - Similar Clusters

### F. PREDICTION PROCESS

Maximum likelihood classification (MLC) algorithm is a statistical decision rule that examines the probability function of a dataset for each of the modules and assigns the pixel to the class with the highest probability. This technique is most often used in image classification and is used to classify user requests from web log data in the present. The maximum relative possibility equation used is called Mahalanobis minimum distance (MD) and is equation 1.

\[ MD = (x - m)^T C^{-1}(x - m) \]

Where C and I forms the covariance matrix for the specific imagined movement considered, left or right and T stands for the transposition Operator. The Mahalanobis distance is used in a minimum distance classifier as follows:

Let \( mR, mL \) be the means for the right and left imagined association classes, and let \( CR, CL \) be the corresponding covariance matrices. A feature vector \( x \) is classified by measuring the Mahalanobis distance \( d \) from \( x \) to each of the means assign \( x \) to the class for which the Mahalanobis Distance is minimum. The full covariance matrix is used to compute the MD. The MLC algorithm is used to categorize user requests into potential and Non-potential, based on the amount of time spent by them in a website.

If the amount of time expended is more than at least 30 seconds, then they are considered as genuine users. According to [9], interested users exhibit certain access patterns, they access certain web pages for a rather long time because they need time to spend on its contents. The user who does not have attention simply accesses many web pages rapidly to browse content. This algorithm results in the prediction of users next request.

### G. Mathematical Module:

The proposed system S is defined as follows:

\[ S = \{I, O, F, U\} \]

Where,

\[ I: Input \ O: Output \ F: Functions \ U: User Recommendation \]

\[ I = \{UL, UN, FE, UP\} \]

Where,

\[ UL= User Logs, \]

\[ UN = User Navigation, \]

\[ FE = Different Classification and Clustering Functions, \]

\[ UP = User Profile for assigning Decision Rule to Online Users. \]

\[ O = \{SC,NN, FE, UH, FA, UN\} \]

Where below are the output generated from system processing:

\[ SC = Statistical Classifier, \]

\[ NN = Nearest Neighbor, \]

\[ FE = Different Classification and Clustering Functions \]

returning result,

\[ UH = User History, \]

\[ FA = Future Action Prediction, \]

\[ UN = User Next URL or Action Recommendation result set with highest probability. \]

\[ U = \{SU, WU\} \]

Where

\[ SU = Survey URL, \]

\[ WU = Web User. \]

\[ F= \{F1, F2, F3, F4, F5\} \]

Where,

Function F1 : Web log File recognition from www Server as shown in figure Sample log file.

Function F2 : Preprocessing and Data Cleaning which removes unwanted entries from F1 and separate required one.

Function F3 : Session Identification based time and Statistical Classifier Analysis Parameter i.e. C4.5 Algorithm.

Function F4 : URL and Action scoring is done by using profile matching i.e User history analysis and user action recommendation, User set is a class of objects. Let X be a user ID of point or objects.
User Set = x, f(x) where x is recommendation index and fA(x) is membership function returning probability.

Function F5 : URL or Action with highest Match is selected for final Recommendation by using Statistical Classifier.

IV. EXPERIMENTAL RESULTS ANALYSIS

A. CLUSTERING RESULT

This data set has pre-processed web logs of the site www.microsoft.com. It records 38,000 randomly selected anonymous users of the site of which 31,501 are used for pre-processing out of which 10,481 weblog data are pre-processed and are used to find the potential user with 4000 are used for testing the ensemble model for clustering and classification.

The number of visits made by the browsers in 24 hours to these 15 pages. It is presented in Figure 3 number of clusters found is another parameter that was used to analyze the performance of the clustering algorithm in Fig: 4.

B. PREDICTION RESULT

The first set is used for generating prediction and the second set is used to evaluate the predictions. Let ASnp denote the navigation pattern obtained for the active sessions and let T be a threshold value. The prediction set is denoted as P (ASnp; T) and the evaluation set is denoted as evalnp. The three parameters can then be calculated using Equations 1 and 2 and the results are projected in Fig: 5 for accuracy and in Fig: 6 for coverage.

The first set is used for generating prediction and the second set is used to evaluate the predictions. Let ASnp denote the navigation pattern obtained for the active sessions and let T be a threshold value. The prediction set is denoted as P (ASnp; T) and the evaluation set is denoted as evalnp. The three parameters can then be calculated using Equations 2 and 3 and the results are projected in Figures 5 for accuracy and in Fig: 6 for coverage.

Accuracy = |P (ASnp - T)evalnp|P(ASnp - T) ............(2)

Coverage = |P(ASnp - T)evalnp|evalnp| ............... (3)
V. CONCLUSION

A user navigation pattern prediction system was presented which consists of three stages. The first stage is the leaning stage, where unwanted log entries were removed. The second stage removes cookies while identified. The result was segmented to recognize potential users, then improved pairwise nearest neighbor clustering algorithm was used to discover the navigation pattern. A Maximum Likelihood classification algorithm was then used to predict future requests. The experimental results prove that the proposed combination of methods is efficient both in terms of clustering and classification. In future work, the proposed will be compared with existing systems to analyze its performance effectively. Plans in the way of using suggestion rules for prediction engine are also under consideration.

VI. ACKNOWLEDGEMENT

I am very much thankful to my respected project guide Prof. P. P. Rokade, H.O.D., Department of Information Technology, for his ideas and help, proved to be valuable and helpful during the creation of this paper and set me in the right path. I am also very much thankful to Head of Computer Department Prof. S. R. Durugkar and P. G. Coordinator Prof. K. P. Gaikwad, Computer Engineering Department, for helping while selecting and preparing for dissertation work. I would also like to thank all the Faculties and Friends who have cleared all the major concepts that were involved in the understanding of techniques behind my seminar.

References

[1] Liu Xinyi ; Sun Hailong ; Wu Hanxiong ; Zhang Richong ; Liu Xudong, Publication Year: 2014,Using Sequential Pattern Mining and Interactive Recommendation to Assist Pipe-like Mashup Development Service Oriented System Engineering (SOSE),IEEE 8th International Symposium on DOI: 10.1109/ SOSE.2014.24


[19] Roger S. Pressman’s, Software Engineering A Practitioner’s Approach [Reference is the SEPA, 4/e, see risk checklists contained within this Web site]

[20] Tom Pender, UML 2 Bible Version 1.4 and 2.0 UML specifications PART I-VII.


