ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Improved Web Prediction Algorithm Using Web Log Data

Megha P. Jarkad, Prof. Mansi, Bhonsle

Assistant Professor, Department of Computer, GHRCEM, Wagholi, Pune, India

Professor, G.H.Raisoni College of Engineering and Management, Wagholi, Pune, India

ABSTRACT: Modeling user web navigation pattern is challenging task that is continuing to gain importance as the size of the web as well as its user-base increase. Web Usage Mining (WUM) is the automatic discovery of user access pattern from web servers. This paper focuses on predicting user future request in smaller time using clustering, classification and backtracking algorithm. In the first step web log file is preprocessed to remove unwanted entries. In the second step potential users are identified from non potential users. In third step clustering is performed using graph partitioned algorithm. In fourth step backtracking algorithm is applied on smaller unit of data and in the last step prediction is given using longest common subsequence algorithm.

KEYWORDS: Web usage mining; Navigation pattern; classification; weblog; clustering; Graph partitioning; Backtracking

I.INTRODUCTION

Web Mining can be broadly classified into three different categories, according to the kinds of data to be mined. They are web structure mining, web content mining and web usage mining Many web analysis tools exist but they are limited as well as the efficiency of these tools is still to reach the state of perfection. When user browses different websites then browsing behavior of the user automatically save into web log file. Web usage mining deals with such log files and extract information about user browsing behavior on internet. This extracted information is used in Personalization, Improving the website design, Business intelligence and predicting the user future requests. Improving the prediction process can decrease the user's access times while browsing, as well as it can ease network traffic by avoiding visiting unnecessary pages. User future request prediction is a web usage mining technique for predicting user next request. For this purpose, first web log files are analyzed and user's future requests are predicted according to the earlier related activities.

The main objective of the proposed system is predict user future request in less time using clustering, classification and backtracking technique. The classification categorizes users into potential and non potential user using decision rule. The clustering groups users with similar interest together. Backtracking algorithm reduces the prediction time. Using the result of classification, clustering and backtracking user future request is predicted.

II.RELATED WORK

V. Sujatha, Punithavalli [1] proposed the Prediction of User navigation patterns using Clustering and Classification (PUCC) from web log data. For clustering they used graph partitioned algorithm and for classification they used longest common subsequence. This PUCC system basically has four steps. The first step is cleaning or preprocessing which removes unwanted data/entries from web log file and reduces the size of web log file. In second step,users are classified into potential users and non potential users and only potential users are only considered for further processing. In third step clustering is done using graph partitioned algorithm and in last step user future request is predicted using LCS algorithm.[1].

Dilpreet Kaur, A.P. Sukhpreet Kaur[2] predict the browsing behavior of user using fuzzy Clustering methods FCM and KFCM.First step is read web log file. In the second step preprocessing of web log file is performed where all unwanted entries are removed from web log file. In third step data is divided into clusters using Fuzzy C-Means and Kernelized Fuzzy C-Means algorithms. In the fourth step User future request is predicted using Fuzzy CMeans and Kernelized Fuzzy C-Means algorithms [2]. Maryam Jafari, Shahram Jamali, Farzad Soleymani Sabzchi[3] proposed a novel algorithm called PD-FARM is proposed for FP-tree mining process. This algorithm uses fuzzy FP-tree to find fuzzy association rules and obtain desired access patterns from a database that contains user's sessions[3]. Samir S. Shaikh Pravin B. Landage D. B. Kshirsagar[4] proposed user future request using longest common subsequence algorithm.[4] S. Vigneshwari, M. Aramudhan[5] developed a model for web information gathering using ontology mining method.[5] L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai[6] gives a detailed information about the web log file, its types, its contents, its location and it also gives a detailed information of how the file is being processed in the case of web usage mining.[6] Xiaohui Tao, Yuefeng Li,Ning Zhong [7]proposed an ontology model for representing user background knowledge for personalized web information gathering.[7] Ramakrishnan Srikant, Yinghui Yang[8] proposed a novel algorithm to automatically discover pagesin a website whose location is different from where visitors expect to find them for this they used the concept of backtracking which says that user will backtrack if they do not get information where they expect it.[8]Ketul B. Patel,Dr. A. R. Patel[9] described data collection,preprocessing of data,pattern discovery as well as pattern analysis task of web usage mining[9]. Amit Kumar Mishra, Mahendra Kumar

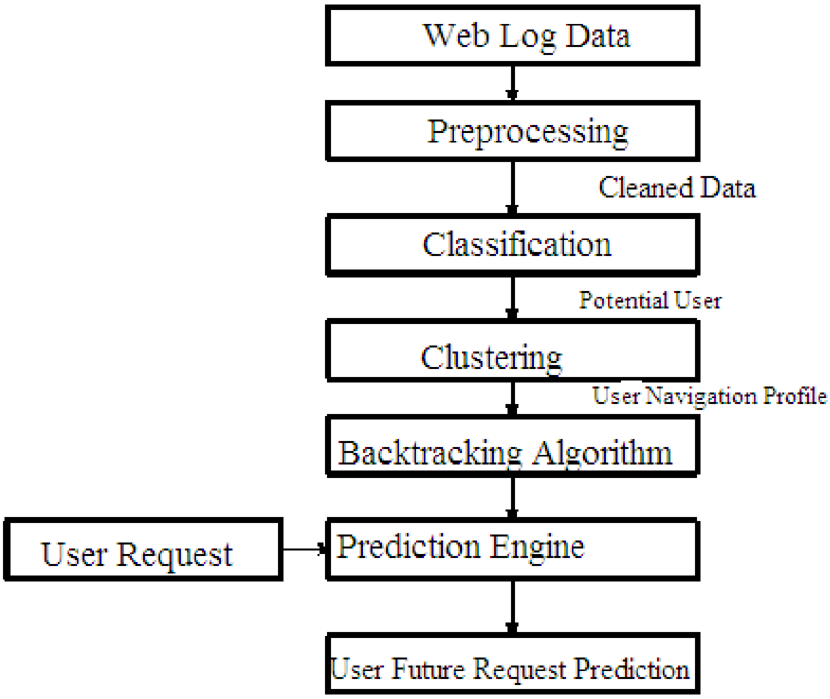
Mishra, Vivek Chaturvedi, Santosh Kumar Gupta, Jaiveer Singh[10]proposed the system that attempts to personalize the website using Self Organized Map and clustering technique. In the preprocessing step transactions are detected and A-priori algorithm is used to detect frequent item set.[10] Bhavna Thakre[11] proposed an algorithm that automatically discover pages in a website whose location is different from where visitors expect to find them. Also proposed an algorithm for finding backtracks that also handles browser caching.[11] K. R. Suneetha, R. Krishnamoorthi[12] primarily focused on group of the frequently accessed patterns of interested users.It helps the web site designers in improvement of the performance of the web by giving preference to the patterns navigated by the regular interested users. In step1 unnecessary as well as junk entries are removed from web log data. In the next step the enhanced version of decision tree C4.5 algorithm is used for identification of interested users from web server log file. The results showed the improvement in both time as well as memory utilization.[12]

Shaily G.Langhnoja1 , Mehul P. Barot , Darshak B. Mehta[13]applied combined effort of both clustering and association rule mining for pattern discovery.In the first step web log file is collected and preprocessing operation is performed on it.In second step combined approach of association rule mining as well as clustering is used for pattern discovery. In the third step association rule mining technique will be used to find user's access patterns.[13].

III. PROPOSED SYSTEM ARCHITECTURE

The general architecture of the proposed system is given in figure 1. The heart of the proposed system is web log data, which contain all the successful hit made in the Internet while browsing. A hit is nothing but a request to view a HTML document or image or any other document. The web log data are automatically generated and can be obtained from either client side server or proxy server or from an organization database (Srivastava *et al*., 2000). Each entry in the web log data include details like the IP address of the computer making the request, user ID, date and time of the request, a status field indicating if the request was successful, size of the file transferred, referring URL (URL of the page which contains the link that generated the request), name and version of the browser being used.

DOI: 10.15680/ijircce.2015.0305176



A. Preprocessing

Proposed work is done on web log file generated by web server. First we have to collect web log file then preprocessing is implemented on web log file. In preprocessing module unformatted web log data is converted into formatted web log data. Formatted data obtained through preprocessing is used for further processing. The Preprocessing includes three steps which are cleaning, user identification and session identification. All entries which will have no use during mining are removed in cleaning. Users are identified on the basis of ip address in user identification. In Session identification process sessions are identified by taking threshold value of time.

- Find and remove all entries which has accessed robots.txt file
- Find and remove all entries with visiting time of access as midnight (commonly used as the
- network activity at that time is light)
- Remove entry when access mode is HEAD instead of POST or GET
- Calculate browsing speed and remove all entries whose speed exceeds a threshold T1 and
- number of visited pages exceeds a threshold T2.

Potential User Identification

Here Potential users are separated from others using decision rule. The decision rule used to find potential user is "If Session Time > 30 minutes and Number of pages accessed > 5 and Method used is POST then the classify user as "Potential" else classify as "Not-Potential". The purpose of introducing classification is to decrease the size of the web log file. This reduction in the size of web log file will help for efficient clustering and prediction.

B. Clustering Process

This paper uses a graph partitioned clustering algorithm to group users having similar navigation pattern. An undirected graph which is based on the connectivity between each pair of web pages is used. Weight is assigned to each edge in the graph. Weight is based on the frequency and connectivity time. Connectivity Time measures the degree of visit ordering for each two pages in a session.

$$TC_{a,b} = \frac{\sum_{i=1}^{N} \frac{T_i}{T_{ab}} X \frac{f_a(k)}{f_b(k)}}{\sum_{i=1}^{N} \frac{T_i}{T_{ab}}} \qquad \text{Equation(1)}$$

Where, Ti is the time duration of ith session that contain both a and b pages, Tab is the difference between requested time of page a and page b in the session, f(k)=k if web page appears in position k. Frequency measures the occurrence of two pages in each sessions (Equation 2).

$$FC_{a,b} = \frac{N_{ab}}{Max\{N_a, N_b\}} \qquad \text{Equation(2)}$$

Where Nab is the number of sessions containing both page a and page b. Na and Nb are the number of session containing only page a and page b. Both the formulas normalize all values for time and frequency are between 0 and 1. Both these are considered as two indicators of the degree of connectivity for each pair of web pages and is calculated using Equation (3).

$$W_{a,b} = \frac{2\ X\ TC_{ab}\ \ X\ FC_{ab}}{TC_{ab} + FC_{ab}} \qquad \text{Equation(3)}$$

The data structure can be used to store the weights is an adjacency matrix M where each entry Mab contains the value Wab computed according to (3) .To limit the number of edge in such graph ,element of Mab whose value is less than a threshold are too little correlated and thus discarded. This threshold is named as MinFreq in this contribution.

C. Backtracking Approach

Whenever any user access the same website after long time which he/she has visited long time ago. In this case we have to search all database linearly which is very time consuming. Instead of searching linearly we are using backtracking algorithm which divides all data subsequently at each stage and repeat the procedure until it find equal. The usual scenario is that we are faced with a number of options, and we must choose one of these. After we make your choice we will get a new set of options; just what set of options we get depends on what choice we made. This procedure is repeated over and over until we reach a final state. If we made a good sequence of choices, your final state is a goal state; if we didn’t make good sequences, it isn’t a goal state.

Conceptually, we start at the root of a tree; the tree probably has some good leaves and some bad leaves, though it may be that the leaves are either all good or all bad. We want to get to a good leaf. At each node, beginning with the root, we choose one of its children to move to, and we keep this up until we get to a leaf.

Algorithm.

Here is the algorithm (in pseudo code) for doing
backtracking from a given node n:

```
boolean solve(Node n) {
    if n is a leaf node {
        if the leaf is a goal node, return true
        else return false
    } else {
        for each child c of n {
            if solve(c) succeeds, return true
        }
        return false
    }
}
```

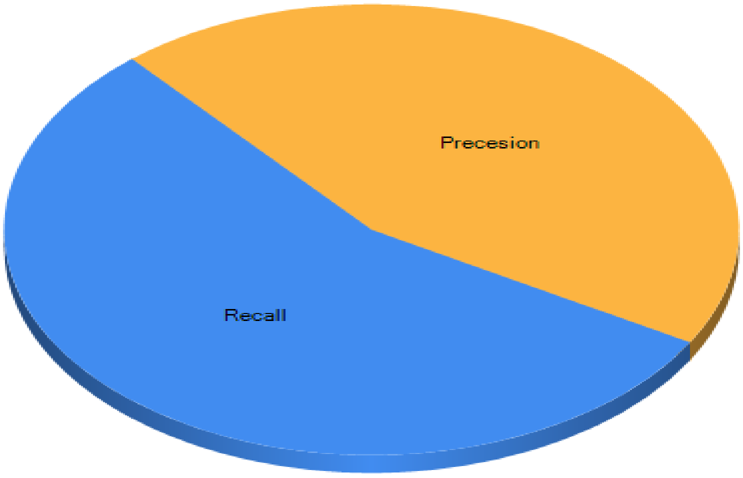
E Prediction Engine

The main objective of prediction engine is to classify user navigation patterns and predicts user's future requests. In this paper Longest Common Subsequence algorithm is used during prediction. The main objective of LCS is to find the longest subsequence common to all sequences in a set of sequences. This method is discussed in this section. The algorithm works with two distinct features.

- The first property says that if two sequences X and Y both end with the same element, then their LCS will be found by removing the last element and then finding LCS of the shortened sequence.
- The second property is used when the two sequences X and Y does not end with the same symbol. Then, the LCS of X and Y is the longest sequence of LCS (Xn, Ym-1) and LCS (Xn-1,Ym).

IV. EXPERIMENTAL RESULT

We have created our own website deployed it on server and collect web log file from server and do processing on that web log file. Following two graphs shows the result for prediction.



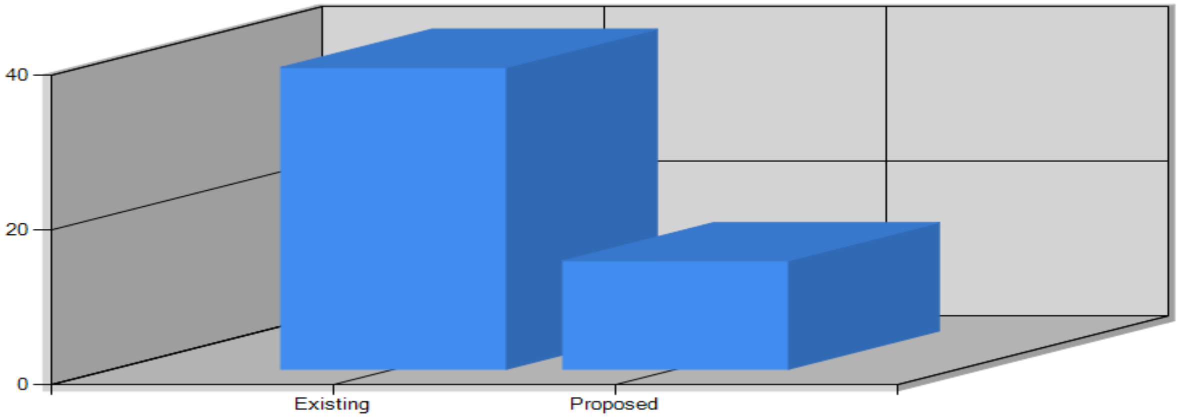


Figure 3:Time Complexity measure

DOI: 10.15680/ijircce.2015.0305176

The time complexity measure shows that time complexity of the system decreases.

V. CONCLUSION

In this paper, a user navigation pattern prediction system was presented which predict user future request. The Proposed system consists of mainly Five steps. In first step data is collected from Web Log file and that data is preprocessed to reduce size of file. In Second step users are classified into potential user and non potential user. In the third step clustering is done using graph partitioning algorithm and users having similar behaviors are grouped into one cluster. In fourth step backtracking algorithm is used. In fifth step User's future request is predicted. Proposed System also uses backtracking algorithm which improves the performance and decreases the time complexity.

REFERENCES

[1] Amit Kumar Mishra, Mahendra Kumar Mishra, Vivek Chaturvedi, Santosh Kumar Gupta, Jaiveer Singh," Web Usage Mining Using Self Organized Map", *International Journal of Advanced Research in Computer Science and Software Engineering,* Volume 3, Issue 6, June 2013

[2] Bhavna Thakre," Mining Web Logs to Improve Website Organization" *International Journal Of Core Engineering & Management(IJCEM),*Volume 1, Issue 1, April 2014

[3] Dilpreet Kaur, A.P. Sukhpreet Kaur," User Future Request Prediction Using KFCM inWeb Usage Mining", *International Journal of Advanced Research in Computer and Communication EngineeringVol. 2, Issue 8, August 2013*

[4] Ketul B. Patel,Dr. A. R. Patel,"Process of Web Usage Mining to Find Interesting Patterns from Web Log Data",*International Journal Of Computers and Technology,*Volume 3,No. 1,2012

[5]L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai," Analysis of Web Logs And Web User In Web Mining*", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011*

[6]Maryam Jafari, Shahram Jamali, Farzad Soleymani Sabzchi," Discovering User's Access Patterns for Web Usage Mining from Web Log Files", *Journal of Advances in Computer Research, Vol. 4, No. 3, August 2013*

[7] Mehrdad Jalali, Norwati Mustapha, Ali Mamat, Md. Nasir B Sulaiman," Web User Navigation Pattern Mining Approach Based On Graph Partitioning Algorithm", *Journal of Theoretical and Applied Information Technology*

[8] Ramakrishnan Srikant, Yinghui Yang," Mining Web Logs to Improve Website Organization"

[9] S. Vigneshwari, M. Aramudhan," A Technique to Ontology Mining for Semantic Web Information Extraction", *European Journal of Scientific Research ISSN 1450-216X Vol. 94 No 1 January, 2013*

[10]Samir S. Shaikh Pravin B. Landage D. B. Kshirsagar," User Navigation Pattern Prediction Using Longest Common Subsequence", *International Journal of Computer Applications (0975 – 8887)*

[11]Suneetha, K.R. and Krishnamoorthi, R., "Classification of web log data to identify interested users using decision trees*", International Conference on Computing, Communications and Information Technology Applications,* (CCITA 2010), Coimbatore, India.

[12] Shaily G.Langhnoja , Mehul P. Barot , Darshak B. Mehta," Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery, *International Journal of Data Mining Techniques and Applications,* Vol 02, Issue 01, June 2013

[13] V. Sujatha,Punithavalli," Improved User Navigation Pattern Prediction Technique From Web Log Data", *International Conference on Communication Technology and System Design 2011*

[14] Xiaohui Tao, Yuefeng Li,Ning Zhong," A Personalized Ontology Model for Web Information Gathering", *IEEE Transactions On Knowledge And Data Engineering, Vol. 23, No. 4, April 2011*