
Ashwini Jadhav¹, Prof. Vani Hiremani², Prof. P T Borse³

¹,²,³ Asst. Professor

Abstract- The web mining is very interesting research topic which supports two of the activated research areas namely Data Mining and World Wide Web. The process of extracting knowledge and information from World Wide Web is defined as web mining. This paper is a survey based on recent work in the field of Web Usage Mining, an important type of web mining. It extracts required information from web log files. The essence of user future request prediction using web log record emphasizes on current evaluation and update in web usage mining. This article views comparison and differentiation of various methods focusing on user future request prediction. Analyzing these request predictions will help the organizations to realize user’s navigational behavior. We therefore present a survey of most recent work in the field of users future request predictions focusing on the overview of the development in the research.

Keywords-- Web logs, User session, Prediction Engine, WebPUM

I. INTRODUCTION

Web has become the largest information source which excitedly impacts on human society which effects on the ways of doing business, educational environment and most importantly the media of information retrieval, conveyance and exchange. Web service providers are in need to trace the ways to predict user’s behaviors and information personalization [9].

It is very important to analyze web site usage by users and web traffic due to growing rate of World Wide Web which leads to different client side and server side tools which extract knowledge from available information. The process of extracting knowledge and information from World Wide Web is defined as Data Mining. It has three types of forms namely, Web Content Mining, Web Structure Mining and Web Usage Mining. Web Content Mining defined as retrieval of information from manuals, documents and descriptions from Web. Web Structure Mining emphasizes on information discovery through hyperlinks. Extracting required information from the Web Log Files is known as Web Usage Mining. Web Usage mining supports four phases namely Data collection, Data Pre-Processing, Pattern Discovery and Pattern Analysis [7].

While browsing the Web Pages by the user, he/she leaves some valuable information in web log files. Web Log Files are used to trace user’s web navigation behavior, through which we can easily analyze which type of information user frequently navigates from the web sites [2].

The proposed model is a literature survey on "Users Future Request Prediction" based on Web Usage Mining. The sections that we have organized in this survey paper are as follows: Section 2 in the paper represents related work, goal and motivational aspects. Section 3 introduces literature review; Section 4 describes prediction system architecture and the literature survey on users request predictions. Section 5 discusses Review Paper's Summery and Section 6 concludes literature survey.

II. MOTIVATION AND RELATED WORK

Discovering the patterns while browsing and navigation data of users are traversed by web usage mining research that has been focused on the web server side which improves the server's performance and web site's service. Web server logs are used as data sources for web usage mining. Initially server side issues are important to investigate but equally fruitful aspect of web usage mining is client side mining of usage data. This term is named as Personal Web Usage Mining [1]. Personal Web usage mining is used to enhance individual user's web use.
User web activities are depicted clearly and accurately through client side data. As all aspects are implemented at client's side, users have full control to access data from anywhere, anytime with privacy of users protected [9][1]

In all over the world, millions of users access Web Sites. While accessing these Web sites, a large amount of data is generated in the Log Files. These Log files are important because user repeatedly access same type of web pages multiple times and this record is maintained in log files. Such web access pattern is helpful to search out user’s information navigation behavior, through which we can predict the accurate next request of user which reduces browsing time of Web Pages. Near about all the web browsers cache the currently fetched web pages so that it results into enhanced performance and avoids network traffic. There is no provision to make entry for pages that has been accessed by using back button. Server log does not maintain this entry information and it gets lost which results into unreliability of data [7]. The exact theme happens with proxy web servers in which identifier are used for all user request that represent same name. Therefore it is very critical to distinguish between dynamic user sessions [3][7].

Anand Sharma from Rochester Institute of Technology suggested using Java agent instead of that act as markers which keeps track of users visiting web pages which will guide to differentiate users from user sessions. But these suggested approaches causes privacy issues which are harmful for data secrecy and storage.

The motivational theme behind this proposed literature survey is to understand future request prediction through Web Usage Mining. The key source of inspiration is an opportunity to know which researches has been done on web usage mining in future request predictions.

III. LITERATURE REVIEW

Alexandra’s Nanopolous, Dmitri’s Katsaros and Yannis Manolopolous researched on 'Web Prefetching' to reduce user perceived latency which is present in web based application. We are going to study available techniques which are used to predict the user’s future movements.

There is heavy traffic in the Internet due to Web Popularity, which results in response delay. Numbers of reasons are there which causes delay in response such as Network congestion, Low bandwidth, Bandwidth underutilization, propagation delay and Web servers under heavy load.

Increase in the available bandwidth is one of the solutions for above mentioned problems but due to its economical cost, proposed solution of increase in bandwidth is not proper solution. Therefore the proposed technique is used to reduce delay of client future requests for web objects and keeping storage of that web objects into buffer or cache before an explicit request is made for them.

There is a proposed architecture of a prediction enabled Web Server with a pre-fetch engine which gives hints for specific client request of document to server.

Association Rules, Frequent Sequences and Frequent Generalizes Sequences are web mining approaches for web log most often referenced web pages are kept together in one single server session with the use of association rule generation. It discovers unordered correlation between items found in data navigation. Association Rules work as a trigger to prefetch data to reduce user perceived latency. Frequent sequences intend to find time ordered sequences of URLs that have been followed by past users [6]. Generalized sequence allows wildcards to reflect navigational pattern according to user frequent access. [10]

IV. PREDICTION SYSTEM ARCHITECTURE

Nien-yijian and Nancy P. Lin [9] represented trend based application system to traverse behaviors of users and to analyze for predicting future travelling path based upon trends similarity. With the consideration of single browsing cluster of older user, it is not appropriate solution to predict user’s navigational behavior. We have been proposed trend based prediction model to predict future navigation path through ordered sequence of browsing patterns.

Out system works in two phases as pictured in figure namely constructing phase and predicting phase. An expert user relies in constructing phase to discover information through browsing patterns.
Expert users generalize browsing sequences using these browsing patterns. In second phase, novice users browsing behavior can be obtained by comparing similarity with prediction model.

Hence candidate’s pages could be pre-fetching to improve the browsing performance. Cookies and proxy servers survey published by researchers [3][7][9] have been replaced by proposed model which is useful to pre-fetch candidates pages and to predict users navigation behavior.

In proposed model we are going to take reference of a new classification model for online predicting users future movements published by Mehrdad Jalali, Norwati Mustapha, Ali Mamat, Md. Nassir B. Suleiman [10]. Two components are generally structured according to web usage mining such as online and offline. The extraction of knowledge from the historical log files is termed as offline structure. This extracted information is used by online component.

We are going to implement Longest Common Sequences (LCS) algorithm for classification of mentioned architecture. Our architectures both the parts that is offline and online components are going to be work together. Offline component supports two modules namely data pretreatment and navigational pattern mining. Weblog is processed and reformatted again to identify all web access sessions through data pretreatment. The group sessions are clustered according to common properties in navigation pattern mining model. URL requested and session identified in online phase according to which user belongs to that particular session, so that the historical information storage base is updated and a list of suggestion, modifications are applied to the requested page and prediction list.

V. FUTURE SCOPE

Mehrdad Jalali, Norwati Mustapha, Md. Nassir Suleiman, Ali mamat [8] advanced [10] work and renamed there architecture as WebPUM. In WebPUM they proposed a novel formula for assigning weights of edges of undirected graph to classify the current user activity. Markov model based sequential pattern mining with clustering was proposed by A. Anitha [1]. This integrated model approaches near about 12% of prediction accuracy. The condition of such clustering type was that every object must be a candidate of only cluster. It has lack of accuracy because of poor historical storage and therefore it suffers from space complexity problems.

Whereas our model is referenced from [8] and [10] which is based on Longest Common Sequence Architecture to predict user future request predictions. We are going to conclude our survey through two experiments such as for navigation pattern mining and user next request prediction. This has been found through quality clustering for user navigation pattern.

VI. CONCLUSION

In this paper, we survey the researches in the area of Web Usage Mining, focusing on its application of, ‘future prediction of web user accesses’. Around the key topic of this paper we stated detailed description of the mining of client side Web usage data, which is termed personal Web usage mining analysis. A framework for navigation patterns discovery and users future request predictions analysis are discussed by studying different research papers.
REFERENCES


