Improving Web User Navigation Prediction using Web Usage Mining

Palak P. Patel¹ Rakesh K. Patel²

¹,²Sankalchand Patel College of Engineering, Visnagar-384315, India

Abstract—Web usage mining is the process of automatic discovery of user navigation pattern from the web log files or how the user is accessing page on World Wide Web. The interaction of user with the web gets recorded in the web log file at server side. These log files can be analyzed by various pattern discovery techniques. The discovered patterns can be used for Web Recommendations. By analyzing the web log, the next to be accessed by user can be predicted, and can be pre-send to client to minimize the network latency. There are various techniques like Markov Model, Longest Common Subsequence, Association Rule, Clustering etc. These techniques are used to predict user next request. In this paper, we use longest common subsequence algorithm to predict user navigation. LCS algorithms improve the accuracy and classify current user activities.

Key words: Web Mining, Preprocessing, Longest Common Subsequence

I. INTRODUCTION

A lot of research has been done on web usage mining. Web page usually contain huge amount of information that may not interest the user. Web log contain all user activity which is helpful to find out web navigation behavior of user. Through user behavior, we can find out what kind of information user wants from the websites. The web usage mining is a field of data mining where web usage data is analyzed and based on the analysis, the web browsing trends of the user can be discovered. Web users performs various tasks, such as booking airline tickets, reserving hotel rooms and shopping, using web applications or services. Web miners use an efficient prediction technique to know actual user behavior. There are many techniques like Markov Model, Longest Common Subsequence, Association Rule, Apriori Algorithm etc. All are used to predict user next request of web access.

The rest of this paper is organized as follows: In section II, we describe the background theory of web usage mining. Section III describes the various pattern discovery techniques and section IV summarizes the paper.

II. BACKGROUND THEORY

Data mining is the process of discovering interesting patterns from large amount of data. In data mining, data source can include database, warehouse and other information repository. Data mining can be used in many applications like banking, target marketing, customer relationship management etc. Data Mining is also connected with machine learning, statistics, artificial intelligence, databases, and visualization [1].

A. Web Mining

Web data mining uses data mining technique to automatically discover interesting and potentially useful information from web document and services. Web mining is divided in to three parts [2]:

1) Web Content Mining

2) Web Structure Mining

3) Web Usage Mining

Following section discusses each in detail.

1) Web Content Mining

Web content mining is the process of extracting useful knowledge from the content of web document[3]. Web Content is collection of all web page data like text, image, audio, video, list, or table. Web content mining is an active area of research encompassing many fields like information retrieval, natural language processing, image processing etc.

2) Web Structure Mining

Web structure mining consists of web page node and hyperlink which connects two pages. Web structure mining is the process of discovering the structure knowledge. This structure mining can be divided in to two structure data.

1) Hyperlink

2) Document structure

Hyperlink connects two web pages which are at different location. It might be on same or different page. Document structure is considered as a tree structure. It focuses on the document object model which is automatically extracting structure out of documents.

3) Web Usage Mining

Web usage mining is a process of discovering interesting usage pattern from web data. Web usage mining captures the user browsing behavior of various web sites. Web usage mining process can be divided into three independent tasks: preprocessing, pattern discovery and pattern analysis. There are many techniques like Markov Model, Longest Common Subsequence, Association Rule, Clustering etc. All are used to predict user next request. Web miners use an efficient prediction technique to know actual user behavior.

Fig. 1: Web Usage Mining Process

Web Server Log contains all user activity. Preprocessing steps are applied on web server log and remove all useless data from log file which are not useful for web usage mining. All data mining algorithms are applied on preprocessed data and the knowledge is extracted from that data. Data mining algorithms like association rule mining techniques, clustering, classification etc. are applied on preprocessed data. The third step is pattern analysis, in which tools are provided to facilitate the transformation of information into knowledge. Knowledge query mechanism such as SQL is the most common method of pattern analysis [4].

All rights reserved by www.ijsrd.com | 3463
III. RELATED WORK

Hemant Kumar Singh et al. (2012): In this paper author advanced the online recommender system by using a Longest Common Subsequence (LCS) classification algorithm to classify users’ navigation pattern. Classification using the proposed method can improve the accuracy of recommendation and also proposed an algorithm that uses LCS method to know the user behavior for improvement of design of a website. This paper describes longest common subsequence algorithm used to classify the user navigation pattern for predicting the recommendation set for the online users that can make the website of any organization more efficient and user friendly. The quality of the recommendations is measured by the two parameters that are accuracy, coverage and the length of subsequence increases according to requirement.

P. Nithya et al. (2012): In this paper authors continues the line of research on Web access log analysis is to analyze the patterns of web site usage and the features of user’s behavior. It is the fact that the normal Log data is very noisy and unclear and it is vital to preprocess the log data for efficient web usage mining process. Preprocessing is the process comprises of Three phases which includes data cleaning, user identification, and pattern discovery and pattern analysis. In this paper, a novel pre-processing technique is proposed by removing local and global noise and web robots.

Ravi Bhushan et al. (2013): This paper proposes a web recommendation approach which is based on learning from web logs and recommends user a list of pages which are relevant to him by comparing with user’s feedback and optimize the search result by reranking them. This will reduce the search time of desired web pages. The proposed system proves to be efficient as the pages desired by the user, are on the top in the result list and thus reducing the search time. An optimized recommendation system with two level architecture has been proposed in this paper [8]. A matching query algorithm and Rank Updating algorithm have been proposed for implementing effective web search.

Sneha Y. S et al. (2012): In this paper author present architecture for integrating semantic information about the products with web log data and generate a list of recommended products by using LCS Algorithm. They have created two tier architecture for integrating semantic information with web usage mining. They have used LCS algorithm to generate a list of recommended products to the user. Future enhancement involves measuring the accuracy of the recommended list to characterize the quality of the results obtained. Their comparison is based on 3 different metrics, namely, accuracy, coverage and F1 measure. To overcome the drawbacks of the current recommender system such as intelligence, adaptability, flexibility, limitation of accuracy [9].

Mehrdad Jalali et al. (2008): In this paper online predicting in Web Usage Mining system and propose a novel approach based on LCS algorithm for classifying user navigation patterns for predicting users’ future requests. The Incremental results show that the approach can improve accuracy of classification in the architecture [12]. The prediction list this system must find the cluster based on LCS algorithm. After applying this algorithm, the system finds a cluster with highest degree of LCS in respect to sequence.

Visalia P. Sonavane (2012): In this paper they in this paper author present used LCS algorithm for improving accuracy of recommendation. The author show that the approach can improve accuracy of classification in the architecture [13]. Using LCS algorithm we can predict user’s future request more accurately.

Samir S. Shaikh et al. (2013): They present in this paper author present uses the Longest Common Subsequence algorithm during prediction. The main aim of LCS is to find the longest subsequence common to all sequences in a set of sequences [14]. In this paper three parameter measure using LCS accuracy, coverage and F1 measure.

Mehrdad Jalali et al. (2009): In this paper author develop a model for online predicting through web usage mining system and propose a novel approach for classifying user navigation patterns to predict users’ future intentions. The approach is based on using longest common subsequence algorithm to classify current user activities to predict user next movement [15]. used some evaluation methodologies that can evaluate the quality of the clusters found and quality of the recommendations.

Mehrdad Jalali et al. (2010) in this paper author present longest common subsequence algorithm is used for classifying current user activities to predict user next movement. The proposed system has been tested on CTI and MSNBC datasets. The results show an improvement in the quality of recommendations [16]. Prediction engine is used to classify user navigation patterns and to predict user future requests.

Kaushal Kishor Sharma et al. (2014): in this paper author present three simple method for predict user future request. The Processes are the same but up to Clustering in this paper adapted the process of Longest Common Sub Sequence (LCS) in the prediction Engine part [17]. The main aim of LCS is to find longest subsequence common to all sequence in a set of sequences.

IV. PROPOSED WORK

A. Preprocess Process:

Preprocessing algorithm is used for cleaning the data. In log file there are many data which are not useful for our prediction. We have removed that data using this preprocessing algorithm. First, the main objective of the algorithm was to remove image files of all types, robots, crawlers and error request which we describe below.

```
Begin
1.1 Logfile log_file;
1.2 While not EOF(record = read(log_file))
1.2.1 Rpage = extractRpage (record)
1.2.2 If((Rpage.type=(jpg, gif,ico, exe,ani,css,js, png)) then skip record
1.2.3 Referrer = extractReferrer(record)
1.2.4 If(!Referrer.contains(“abc.ac.in”)) then skip record
1.2.5 RStatus = extractRStatus(record)
1.2.6 If (RStatus > 299) then skip record
1.2.7 If (Referrer.contains(“bot”, “spider”, “crawler”) then skip record
1.2.8 InsertRecord(logrecord)
End
```
1.3 End While
End Procedure

The above algorithm is briefly explained below.

Explicit request for a single page by Web user generates many implicit requests for all objects which are embedded in that particular Web page. Request generated for files like jpg, gif, ico, js, png has been removed from the Web access logs, since they are unnecessary for generation of prediction model of page to page transition. These types of records have been removed from the logs as a preprocessing step.

Also the records, which contain “-” in referrer field, cannot be used for model generation, since this type of records does not contain the page from which the request has arrived. Also the records which are having referring Web site other than abc.ac.in for respective analysis have been removed.

Also Web log records, where request status code indicates failure have been removed from further processing. Request status codes which are of the form 2xx indicate success of the request. Request status codes which are of the form 3xx indicate redirection of the request when multiple choices of the requested resources are available or the requested resource has been moved permanently or temporarily. Request status codes which are of the form 4xx indicate errors in request like bad request, unauthorized access, forbidden error, requested resource not found on the server etc. Request status codes which are of the form 5xx indicate server errors like features necessary to satisfy the client request not implemented, service unavailable etc. Thus the records having the status code greater than 299 have been removed from preprocessing.

Records pertaining to various Web crawlers or spiders have been removed.

B. Proposed Work:
The query is generally, first stored in the form of log files on particular web servers in a particular format according to the servers. This log files are then pre-processed using preprocessing algorithm and the meaningful data are kept while the useless data are wiped out.

On this data, the pattern analysis and prediction algorithms will be applied through which the system predicts the next user input and suggests it that.

In this system, we have created clusters based on network IP address.

Implementing this, the duration for pattern generation of the particular input query can be minimized as the query operation will be done on the particular cluster only. After this, the LCS will be applied in particular clusters records, so that the better prediction can given as per the user input.

Applying LCS on clusters itself before applying on records increases the accuracy of prediction. Also it decreases the time of discovery of pattern from the particular clusters.

V. RESULT OF WORK

A. Original Log File:

Fig. 3: Original Log File

B. Records indicating referrer other than said URL:

Fig. 4: Records Indicating Referrer Other Than Said URL
C. Records Indicating Errors Of Requests:

Fig. 5: Records Indicating Errors Of Requests

D. Records Indicating Request By Robots/Spiders Etc:

Fig. 6: Records Indicating Request By Robots/Spiders Etc

E. Record indicating errors of requests:

Fig. 7: Record Indicating Errors Of Requests

F. Preprocessed Log File:

Fig. 8: Preprocessed Log File:

G. Pie Chart For Pre-Process Data

Fig. 9: Pie Chart For Pre-Process Data

H. Result of Clustering Process:

Fig. 10: Result of Clustering Process

I. Page Access Prediction:

Fig. 11: Page Access Prediction

VI. CONCLUSION

The aim of this paper is to improve the user navigation by decreasing user navigation time so that the user spends less time for searching that particular information. Using the LCS algorithm we will improve user navigation time. Also it decreases the time of discovery of pattern from the particular clusters.

REFERENCE

[1] Hemant Kumar Singh, Brijendra Singh “A Classification Algorithm to Improve the Design of


