ISSN:0975-9646

User Future Request Prediction

Rujuta Panvalkar[#1], Namrata Valera[*2], Ami Vashi[#3], Khushali Deulkar[#4]

[#]*Computer Engineering Department,*
Dwarkadas J. Sanghvi College of Engineering, India

Abstract— **The use of Web technology has increased by a great extent in the recent times. Millions of users spend time surfing net to obtain information or for recreational activities. Along with satiating their own purpose, the users leave behind a detailed trail of all the web pages accessed and the frequency with which they are accessed. This information is of paramount use to several commercial domains like e-commerce websites, social networking sites, entrepreneur franchises, etc. Web usage mining satisfies the basic purpose of amalgamating information from a user's web history and finding patterns in the usage characteristics. Web usage mining is the extension to the traditional data mining and also forms the base of our paper. Our paper emphasizes on predicting a user's future request. The attributes of web usage mining help to determine a pattern depending on a user's navigational footprints and actions. This pattern is then analyzed, giving us tools to predict requests that the user is likely to make in the future.**

Keywords— Future Request, Prediction, Surfing, Web History, Web Usage Mining **Introduction.**

1. Introduction

Today Internet has become more of a necessity than a luxury to people. Every day, inestimable websites are created, visited, and merged. People use internet for almost everything and hence, it is but natural that Web domain is a domain that hasn't yet been explored to its full potential. Whenever anything is accessed on web, that particular page is saved in the device's web history. This as we call is the temporary trail of the web accesses. The more permanent trail is that which is stored in the web logs. Web logs act as the chart of user's behavior on a particular machine. Most of the web log data is generally automatically generated by the web servers. A colossal amount of data is uploaded, downloaded or just viewed on the World Wide Web. This also means a number of users are very much interested in carrying out communications and transactions using this medium. This in turn has a tantamount effect on the commercial businesses having a huge interest in using the web services. As users form the base of this commercial build-up, the comfort of the end users is given ultimate importance. Now, as the advances in graphic content can only satiate the user to some extent, new advances are being made in field of web usage mining which is more appealing to the user as it affects in reducing the efforts on the end-user side. Future user prediction is one such field which benefits the user as well as the Web sites.

The basis of the paper is to correctly predict what the end user wants to see on the page he opens. This is accomplished by first accumulating the users' data from the web logs generated by the servers. This data is then cleaned so that only the relevant information is made available for the further process. Once this step is done, the remaining data is grouped into clusters according to the various users using the same machine and then classified according to the domains to which the accessed information belongs. For example, all the sports data of user A will be classified together, all the data relating to politics will be classified together. In another cluster, will exist the sports data of user B, the educational data of user B and so on. Thus according to the variance and access frequency of the users, the total accessed data will be clustered and classified and will be made available for future processing. Now, when any of the existing users logs in the machine, using the algorithm discussed below, new suggestions will be made for the user's future requests in accordance to the pages accessed by the user on all the previous instances.

We mainly focus on Web Usage Mining. It is a type of web mining which extracts interesting and useful patterns about the user's navigational behavior. This activity helps developers to understand individual user's psyche and helps them to customize the services provided to that particular user. This customization is now an integral and attractive quality about software or website as it makes the user's navigation easy and smooth.

The rest of the paper is divided as follows- Section [2] explains the Problem Formulation. Section [3] states the review of literature. Section [4] states the Proposed Solution. Section [5] gives the FP Tree algorithm. Section [6] states the overall architecture of the proposed system. Finally, section [7] concludes the paper and states the thought of future scope.

2. Problem Formulation

In today's world information on Internet is increasing day by day and web administrator's continuously trying to make their website more users friendly and efficient. Pattern extracted from web server log helps them in a big way to make decision about restructuring of websites and implementation of new applications which will increase their traffic and eventually business. In this report the problem defined is the extraction of patterns from web server log file. It is an excellent way to define the usage mining using pattern recognition techniques.

Over a period of time, millions of web accesses are made. Many users have many interesting aspects of web searched and studied. This information can be of prime importance to the commercial enterprises and in general to the websites rooting to provide for a better end-user experience. The main aim of the paper is to provide a better system for the analysis of the user's future wants even before the user has a chance to search for them. This enables better customer service and efficiency on the part of the website owners and effortless work on the part of the end users. The basic steps that are to be carried out include- gathering data, filtering data into information, sketching patterns and finally studying the patterns and making predictions.

3. Review Of Literature

Some of the important references used in this project are :
The explosion of data all over the world has led us to strive to find ways to manage and use the available data in more than one way. The aim is to convert heaps of 'data' to 'useful knowledge'. Hence it was important for researchers all over to create client and server side technologies which can carry out this conversion.

Thus came the concept of web mining. What we are going to deal with is web usage mining, a mining approach for user browsing and access patterns. Analysing this browsing data can help organisations to better understand their customers, design strategies, evaluate user response and conduct surveys. [1]

The major tasks that are to be handled are :

a. Preprocessing
b. Pattern discovery

Preprocessing is mainly data cleaning. It basically retains data that can be useful and discards or eliminates data that is irrelevant to the purpose. The second major preprocessing task is transaction identification. Before any mining is done on Web usage data, sequences of page references must be grouped into logical units representing Web transactions or user sessions. A user session is all of the page references made by a user during a single visit to a site. Identifying user sessions is similar to the problem of identifying individual users, **as** discussed above. A transaction differs from a user session in that the size of a transaction can range from a single page reference to all of the page references in a user session, depending on the criteria used to identify transactions.

Content Preprocessing handles content like Images, text, scripts and other files such as multimedia files are converted into useful data for Web Mining Processes. The process involves classification and clustering. Result of a classification is such that what type of pages has been visited or what class of products has been searched.

Pattern discovery has wide range of applications like on statistical data, data mining and machine learning .The author has limited the coverage of pattern Discovery in the field of Web Domain. The Pattern Discovery in the Web Usage Mining to analyze and Discover the Pattern that has been generated by Server sessions which is the sequence of pages requested by the user. Statistical analysis, association rule, clustering, classification, sequential pattern, etc are some of the techniques used in pattern discovery.

RenataIvancsy and IstvanVajk [4] presented discovering frequent patterns in Web log data is to obtain information about the navigational behavior of the users. The different patterns in Web log mining are page sets, page sequences and page graphs. Devinder Kaur and Ravneet Kaur [3] talk about today the World Wide Web is popular and interactive medium to distribute information. The web is huge, diverse, dynamic and unstructured nature of web data, web data research encountered lot of challenges for web mining. Information user could encounter following challenges when interacting with web.

a. Finding Relevant Information- People either browse or use the search service when they want to find specific information on the web. Today's search tools have problems like low precision which is due to irrelevance of many of the search results. This results in a difficulty in finding the relevant information. Another problem is low recall which is due to inability to index all the information available on the web.

b. Creating new knowledge out of the information available on the web- This problem is basically sub problem of the above problem. Above problem is query triggered process (retrieval oriented) but this problem is data triggered process that presumes that already has collection of web data and extract potentially useful knowledge out of it.

c. Personalization of information- When people interact with the web they differ in the contents and presentations they prefer.

d. Learning about Consumers or individual users- This problem is about what the customer do and want. Inside this problem there are sub problem such as customizing the information to the intended consumers or even to personalize it to individual user, problem

e. related to web site design and management and marketing

TABLE 1-_Comparison of various algorithms for clustering and classification of data

	Apriori Algorithm	F.P Tree algorithm	K-Means Algorithm
1. Type of algorithm	An association rule learning algorithm.	An Association Rule that finds frequent patterns. Its a Scalable Mining method.	Clustering algorithm, Unsupervised algorithm
2. Technique	It scans the database multiple times to obtain the support of each itemset and discards the items which are below the threshold value.	The infrequent data items are discarded and frequent item data is converted into a different data structure called FP-tree on the basis of the support of each item	The data is clustered together depending on its proximity to the center of the cluster. Each cluster denotes either a separate user or a separate field of interest to the user
3. Number of scans	Multiple	2	Multiple
4. Time Complexity	O(Number of items)	O(|Number of items in a transaction|)	O(Number of items)
5. Size of data set	Large datasets are involved in this type of algorithm	The *large* set of evolving and distributed data *can* be handled efficiently by FP-Tree algorithm.	K-Means algorithm works well on a large data set. It does not work well on a small dataset.
6. Advantages	1. It is simple to implement	1. Database is needed to be scanned only twice. 2. It finds the items that are used more frequently by the users from Web logs and increases the usability of a website or a system.	1. Large data to work on 2. It works well for clustering of users as well as the content of the users

4. Proposed Solution

We propose to formulate an improved F.P tree algorithm to implement the solution to the problem formulation. There are multiple steps involved in this process. We aim to make a website by using the basics of HTML, CSS and JAVASCRIPT. The website will be the host to the implementation of the web usage mining solution. This particular website is just a host to the otherwise universally implementable algorithm. The main aim is to make the website dynamically customized to each particular user according to the web content searched by the users. The next step is to devise the algorithm. That is done by researching the literature studied in part 2 of the report. Finally, the last step is to use the algorithm in the website to categorize the content.

The following steps are involved

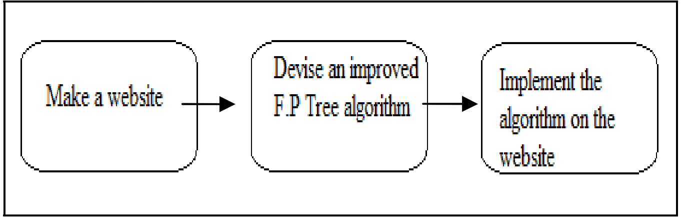

Make a website
Devise an improved F.P Tree algorithm
Implement the algorithm on the website

Fig. 1- Illustration of algorithm

The F.P Tree algorithm works as follows-

The proposed approach that we plan to implement follows the following steps:

Step1: In the first steps data is being collected from the Web log file and then Preprocessing is applied. In the Preprocessing the Data is being loaded and it is being converted in to the Data set having fields Client-IP, Session_ID, Country, Access Date Time, Method, URL, URL_ID, Protocol, Status, Bytes transferred. The session is calculated in 30 minutes interval of time, after 30 minutes the system will recognize the same user as next user.

Step 2: In this step there is Pattern Discovery which is performed by the Frequent Pattern (FP) which involves FP Tree which in turn FP growth .FP tree method is used in Data Mining .It consists of two passes over the Data Set .In the first Pass it scans data and find the minimum support for the each item. The item set whose support is less than minimum is discarded .The Data item that is included is the Web Site or the URL that is being visited by the User. Next steps in the First Pass in the FP tree are to generate a decreasing order on the basis of frequency of occurrence of the Item Set Which is the URL visited by the User. In the Second Pass of the FP Tree Transaction is being read .In this work the Transaction is the number of user visited the particular Web Site. The Read Transaction is iterated until all the Transaction is being completed. After Reading all the Transaction discards all the transaction which has lees support or support than the minimum threshold value.

Step3: In this step Pattern analysis is done and in this Candidate rule is generated and on the basis of candidate rule confidence is generated. On the basis of pattern analysis Prediction is done of the User's Future request.

DFD Diagram-

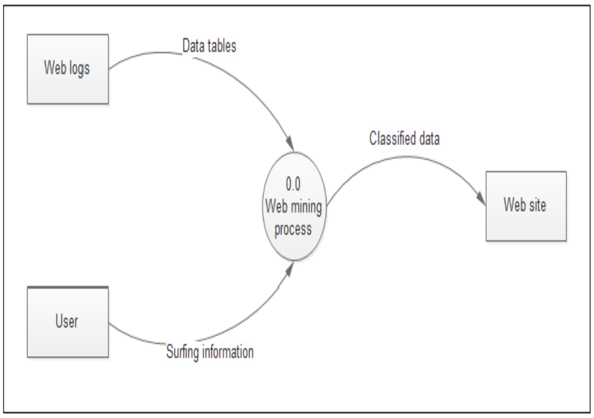


Fig. 2 – Level 0 DFD

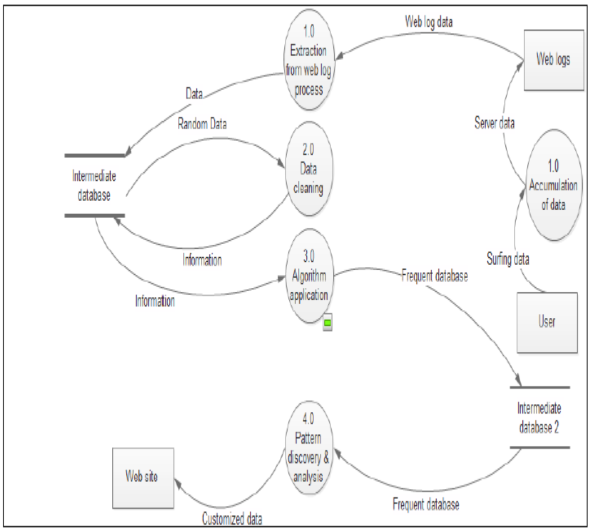


Fig. 3– Level 1 DFD

5. Algorithms / Methods Used

Step 1: Generation of web log data.

The data is generated when the users access/ create any information over the internet. The weblogs are created by the web servers.

Step 2: Extraction of web log data.

The web log data is of prime importance in the entire process. Web log data extraction is done using a software.

Step 3: ETL process.

It does the extraction, transformation and loading of the data extracted from the weblogs. This is also called as cleaning of data. This removes all the abnormalities from the data and makes it ready for use by the algorithm.

Step 4: Application of algorithm.

The algorithm used here is the F.P Tree algorithm. It is applied to the data obtained from the ETL process. It mines the data and finds the frequent patterns in the data. It is a two- step process. It concludes by forming a F.P Tree.

Step 5: Pattern discovery.

The frequent patterns mined by the algorithm are discovered and highlighted.

Step 6: Pattern analysis.

The discovered patterns are analysed and are used for distinguishing different categories of data.

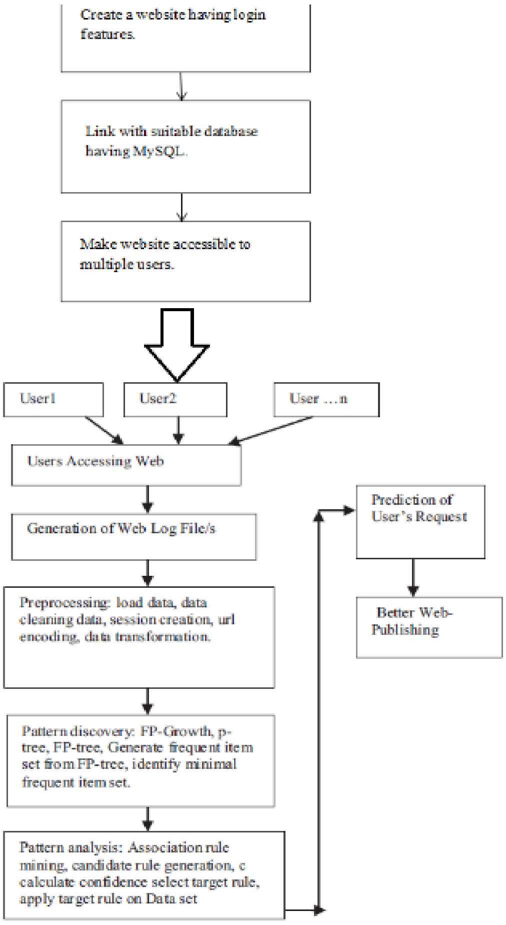
Step 7: Customizations.

The users are provided with dynamic customizations according to their area of internet interest to better their internet experience.

6. Architectural Design

The whole process consists of multiple steps. We plan on making a website which will enable the working of the F.P Tree algorithm. The website includes a login feature and basic functionalities

Fig. 4- Flowchart for the system


Create a website having login features.
Link with suitable database having MySQL.
Make website accessible to multiple users.
User1
User2
User ….n
Users Accessing Web
Generation of Web Log File/s
Preprocessing: load data, data cleaning data, session creation, url encoding, data transformation.
Pattern discovery: FP-Growth, p-tree, FP-tree, Generate frequent item set from FP-tree, identify minimal frequent item set.
Pattern analysis: Association rule mining, candidate rule generation, c calculate confidence select target rule, apply target rule on Data set
Prediction of User's Request
Better Web-Publishing

In the next step, we are making a detailed database using MySQL. We are assuming that multiple users will access a single website and every user will have a charted history to find individual patterns. Each user is provided with a login feature which helps in distinguishing the users at the base step. The users access the web and search multiple things. The entire activity of the user is stored in the web logs generated by the server. The weblogs consists of crude data which has to pre-processed to get information. Data transformation is applied and the data is stored in a separate database. F.P Tree algorithm is then applied to this new database to carry out pattern discovery. After pattern discovery, the last step is pattern analysis. Pattern analysis will give a detailed picture of each user's activities and interests. From this retrieved information, the admin can draw up suggestions and predict the user's future request.

7. CONCLUSION AND FUTURE SCOPE

The simulation result shows that the FP-Growth algorithm is used for finding the most frequently access pattern generated from the web log data, By using the concept of web usage mining we can easily find out the user's interest and we can modify and make our web site more valuable and more easily accessible for the users. The main goal of the proposed system is to identify usage pattern from web log files. FP Growth Algorithm is used for this purpose. Apriori is a classic algorithm for association rule mining. The main drawback of Apriori algorithm is that the candidate set generation is costly, especially if a large number of patterns and/or long patterns exist. The FP-growth algorithm is one of the fastest approaches for frequent item set mining. The FP-growth algorithm uses the FP-tree data structure to achieve a condensed representation of the database transaction and employees a divide-and conquer approach to decompose the mining problem. Our experimental result shows that the FP-growth method is efficient and scalable for mining both long and short frequent patterns. In future the algorithm can be extended to web content mining, web structure mining

REFERENCES/BIBLIOGRAPHY

[1] Cooley, R,.Mobasher,B,&Srivasta,J(1997). "Web Mining: Information and pattern Discovery on the World Wide Web". In Procedings of (IEEE) International Conference on tools with Artificial Intelligence.

[2] V.Sujatha And Punithavalli,"Improved User Navigation Pattern Prediction technique from Web Log Data" doi:10.1016/j.proeng.2012.01.8 35, Elsevier 2011.

[3] Devinder Kaur, Ravneet Kaur, Minimizing the Repeated Database Scan Using an Efficient Frequent Pattern Mining Algorithm in Web Usage Mining, International Journal of Research in Advent Technology, Vol.2, No.6, June 2014.

[4] Magdalini Eirnaki and Michalis Vazirgiannis:"Web Mining for Web Personalization": ACM Transaction on Internet Technology,Vol 3, 1, pp 1-27 February 2003.

[5] Myra Spilopoulou and Luks C Faultstich ,Wum:" A Web utilization Miner .In EDBT Workshop Webdb98, Valencia, Spain, SpringerVerlag,1998.

[6] Jaideep Srivastav,Robert Colley,Mukund Despanade, Pang-NingTan;"Web Usage Mining:Discovery and Application of Usage Pattern from Web Data". SIGKDD Exploration .ACM SIGKDD,Jan2000.

[7] PawelWeichbrth and MieczyslawOwoc, Michal Pleszkum." Web User Navigation Pattern Discover from WWW Server LogFiles".proceding of the Federal Conference on Computer Scienceand Information Systems, pp- 1171-1176.IEEE 2012.

[8] Ujwala Patil and Sachin Pardeshi,"A Servey on User Future Request Prediction: Web Usage Mining", International journal of Emerging Technology and advanced Engineering, ISSN 2250-2459, Volume 2, Issue 3, March 2012.