EasyChair Preprint

№ 7252

Frequent Item Set Mining from Log Files Using Navigation Pattern and Hadoop Techniques

Nikheel Kasar and Shahrukh Teli

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

December 23, 2021

Frequent Item set mining From Log Files Using Navigation Pattern and Hadoop Techniques

Mr.N.D.Kasar ,Mr.S.S.Teli
Lecturer [Comp Dept]
MSBTE affiliated R.C.Patel.Polytechnic
Nikhilkasar730@gmail.com

Abstract

Because this site log contains a large amount of data, it is preprocessed before modeling. The web log file is preprocessed and transformed into a user web log sequence. Sessions of navigation the web navigation session is the time while you're on the internet. a user's web page navigation sequence over time window. Finally, the user navigation session is modeled. Utilizing a model when the user navigation model is finished, it's time to move on to the next step. Mining is a task that can be carried out in order to discover anything interesting. Pattern. Web log modeling is a critical task in web usage. Mining. A high level of prediction accuracy can be reached by using a to improve the web log by modeling it with an accurate model Caching is used to improve the performance of the servers. Pages that are often visited are cached on proxy servers. Pre-fetching of web pages is a new topic of study. When combined with caching, the performance skyrockets. In A better algorithm for predicting the future is presented in this work.

Keywords: *Semantic Web,Web usage mining, Domain sequential pattern mining ,Recommander System,Markov model,Prediction web log.*

1.Introduction

1.1 Web usage Mining: - Web Usage Mining has recently become a popular method for enabling Web customization. Use of the Internet The goal of mining is to discover user navigational patterns on a website. Collecting facts from web usage to improve the World Wide Web a logs (we will refer to them as web logs). The premise is that a web user can only visit one web page at a time that symbolizes one thing at any one time.

The process of Web Usage Mining goes through the following three phases are .

- Preprocessing phase: The main task here is to clean up the web log by removing noisy and irrelevant data. In this phase also, users are identified and their accessed web pages are organized sequentially into sessions according to their access time, and stored in a sequence database.
- Pattern Discovery phase: The core of the mining process is in this phase. Usually, Sequential Pattern Mining (SPM) is used against the cleaned web log to mine all the frequent sequential patterns

- Recommendation/Prediction phase: Web Usage Mining is a subset of web mining that focuses on identifying interesting usage patterns from log data. The logging data is saved in a file called the web log file. The IP address, date, time, and web page accessed are all included in the web log file.

1.2 Web Log:- The web log is the registry web pages access different method for accessing the user aaat different time . can be maintained at the server-side, client-side or at a proxy server, each having its own benefits and drawbacks on finding the users' relevant patterns and navigational sessions.

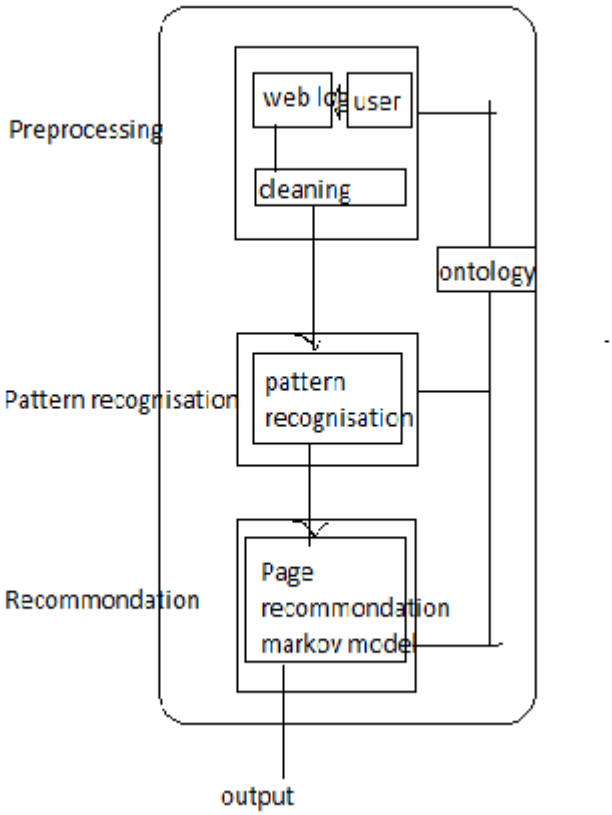

Figure:- Phase web mining

- Server Log: The server saves information about client requests, so data is usually limited to one source. Details about the server logs c an be found n

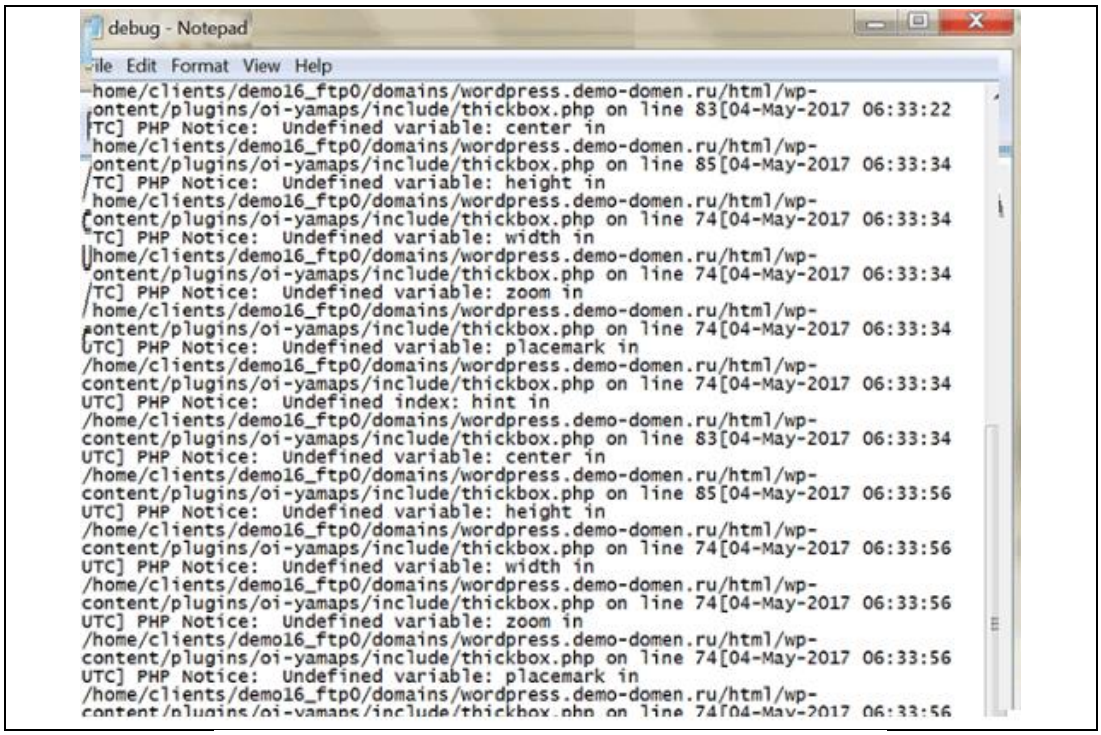

Figure No. 2:-Server Web log files and Records

- Client Log: The client sends information about the user's behavior to a repository (this can be done with a remote agent (such as Java scripts or Java applets) or by changing the source code of an existing browser (such as Mosaic or Mozilla) to improve its data collection capabilities.) ;

- Proxy Log: data is stored on the proxy side, so Web data pertains to several Websites, but only to users whose Web clients travel via the proxy.

II Literature Review:-

Due to the excessive growth of the web user for excessive growth and provide latency For web service providers, this has become a severe problem. To address this problem, studies have been conducted that mix methodologies from many disciplines.

Researchers worked on pre-fetching popular pages to reduce perceivable network delay. The use of prefetching and caching strategies dramatically enhances performance while also reducing the time it takes for the application to run.

By 50%, the number of applications received has increased.

- **Garofalakis:**-Essentially conducted a data mining survey. Methods and algorithms for detecting web structures Hypertext and hyperlink are two terms that are used interchangeably. Clustering based on generalization

A method has been presented that includes attribute information.
Induction that is oriented **Pitkow et al**. predicted the journey of a web user in a pattern. Mechanism of extraction Worked on anticipating future requests. and has created an n-gram model for it.

- **Cooley** divided web mining into categories and then presented his findings conceivable research topics A method for quickly allocating online resources. Websites that use competitive neural networks and data mining approaches. The topic of network is being discussed.

- **Zhang** By establishing hierarchical clustering of web users based on their access habits, suggested an effective data clustering approach for very big databases. In order to respond to user requests for web pages, a clustering technique based on first-order Markov models has been implemented. Short-term pre-fetching employs the Dependency Graph (DG), which is made up of access patterns, as well as Prediction by Partial Matching (PPM).

Short-term pre-fetching has the advantage of lowering user-perceived latency. Aside from that, it also has two flaws.
To begin with, if the pre-fetching policy is not carefully set, it may result in unnecessary network traffic. Second, this pre-fetching approach does not optimize cache space well. Long-term pre-fetching is based on global item access pattern statistics, which identify clusters of important objects. This approach could be utilized in a Content Distribution Network (CDN), mobile computing environments, and other places.
Different advantages of web pre-fetching are discussed, as well as what motivates people to do so.

- **Vakali** described an extensive range of web data clustering schemes, in most of the cases clusters belongs to intra-site web pages . In grouping inter-site web pages, web clustering performance reduces due to increase in complexity of web. If there is some change in web user's pattern, then it must to be updated in the resulted clusters.

- **Schloegel** used graph theory for working with web log files. The paper represented the web log files using web navigational graph and then using web partition techniques .

- **Nanhay Singh** combined the KMeans clustering and Apriori algorithm, two web mining approaches, to forecast and pre-fetch web pages from the proxy server.

III Problem Formulation

3.1Problem Defination

In today's world, the amount of information available on the Internet is growing by the day, and web administrators are constantly working to make their websites more user-friendly and efficient. Extracting a pattern. The information gleaned from the web server log greatly aids them in making decisions. Concerning website reorganization and the introduction of new technologies applications that will boost their traffic and, eventually, their profits business. The problem mentioned in this study is the extraction of patterns extracted from the log file of a web server it's a fantastic method to Using pattern recognition tools, define the usage mining. Hundreds of millions of web accesses are performed over time. Many individuals are searching for a variety of interesting parts of the internet. And research This knowledge could be extremely valuable to you. Commercial enterprises and, more broadly, websites to deliver a better end-user experience, rooting is used. The most important purpose of this paper is to develop a better analytical system even before the user has a chance to look for them. of the user's future wants even before the user has a chance to seek for them. This allows website owners to provide greater customer care and efficiency, while end users may work more efficiently. Gathering data, refining data into information, sketching patterns, and lastly studying and making predictions are the main stages that must be followed.

3.2 Explanation

Pre-fetching the most likely sites and caching them in web caches improves server speed in existing operations. Different techniques, such as the Markov model and the Apriori algorithm, can be used to predict the pages. Recent work has also included the integration of multiple algorithms to overcome each other's constraints.

Two separate data mining algorithms are employed in the existing work to anticipate the web sites that are likely to be accessed in the near future. Existing efforts have attempted to cluster data based on user interests or the time it takes the server to react to queries. The performance of the users is improved in this work by grouping them into distinct groups based on the location from where the request is issued. The hit ratio is improved by grouping users depending on their location.

3.3 Methodologies/Development Methodologies

There is a performance improvement in this existing work.The FP growth a lgorithm was also used to obtain the solutioninstead of using the Apriori tec hnique, use the frequent item set

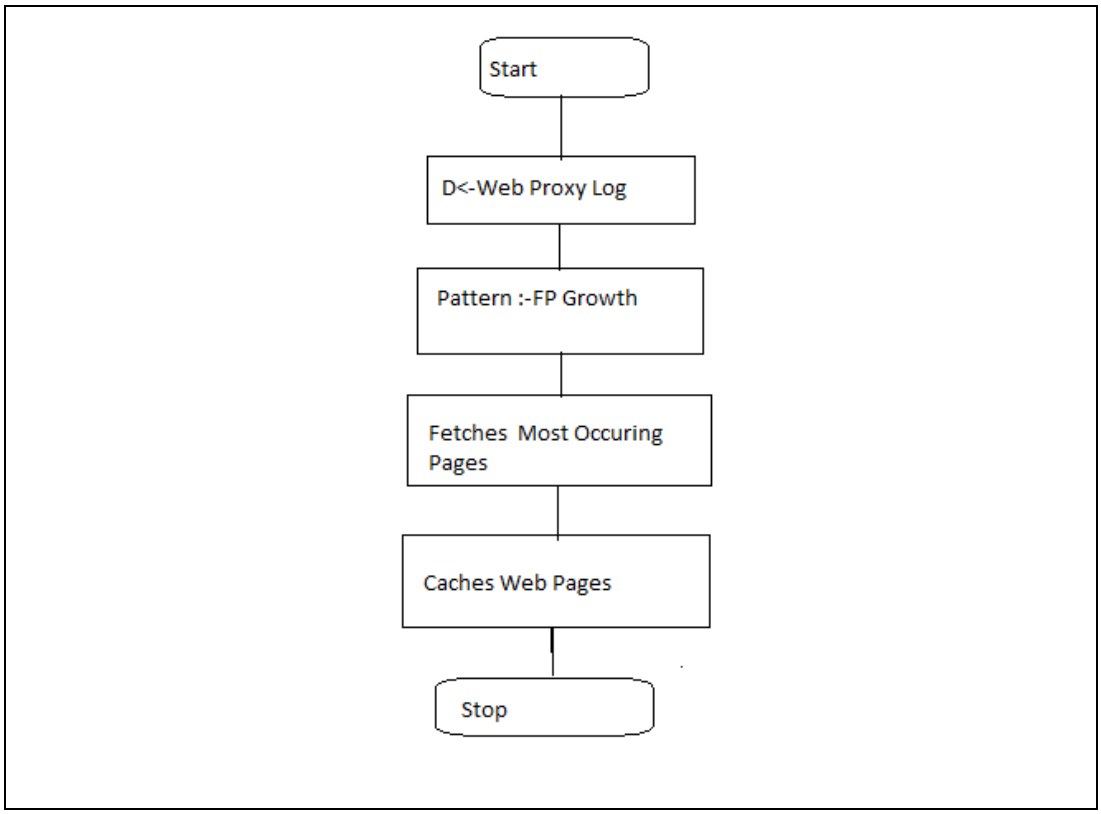

Figure 2:- Existing System for Fetching the Most Frequently Occurred webpages

IV. WORK PROPOSED

To apply the solution to the matter formulation, we propose developing an e nhanced F.P tree rule.
The phases that we usually think of to implement a planned method are as follows

Stage 1: Knowledge is gathered from the online log file in the first step, and then Preprocessing is applied. The information is loaded and processed during preprocessing. Being reborn into an information set with fields Client-IP, Session ID, Country, Date and Time of Access, Method Bytes transmitted, URL, URL ID, Protocol, Status The When calculating a session, half-hour intervals of your time are used. After a half-hour, the system can recognize a similar user as the next.

Stage 2: During this step, the Frequent Pattern (FP) does Pattern Discovery, which involves an FP Tree that grows in a sequential manner. In data processing, the FP tree technique is used. It consists of two runs through the data set. It analyses knowledge in the first pass and notices the bare minimum of support for each thing. The item set with a support value less than the minimum is discarded. The electronic computer or URL that the User is visiting is the Data item that is contained. The following phases in the FP tree's initial pass are to get a decreasing order based on the idea of frequency of prevalence of the Item Set, which is the URL visited by the

User. The FP Tree group action is being browsed in the Second Pass. The group action in this work is that the variety of

Algorithm: Modified FP_Tree (WebLog [][])

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 finds the patterns that appears analysed also used for distinguishing different categories.

The web usage mining model is similar to server log mining. Web use mining is a technique for expanding the usability of a website design while also improving system performance and customer relationships. The suggested system's main purpose is to extract usage patterns from web log files. For this, the FP Growth Algorithm is employed. Apriori is a well-known method for mining association rules. The fundamental disadvantage of the Apriori technique is that it is expensive to generate candidate sets, especially when there are a high number of patterns and/or extensive patterns. One of the fastest ways for frequent item set mining is the FP-growth algorithm. The FP-growth method employs a divide-and-conquer strategy by utilising the FP-tree data structure to achieve a condensed representation of the database transaction. 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.

V] References

[1] MayankKalbhor [1] “Fuzzy Based Hybrid Approach for User Request Prediction Using Markov Model” [IEEE International Conference on Computer, Communication and Control (IC4-2015).]

2] PriyankaBhart [2] “Prediction Model Using Web Usage Mining Techniques “[International Journal of Computer Applications Technology and Research -2014]

[3] Garofalakis M. N., Rastogi R., Sheshadri S., and Shim K.,“Data mining and the Web: past, present and future.” InProceedings of the second international workshopon Web information and data management, ACM, 1999.

[4] Fu Y., Sandhu K., and Shih M., “Clustering of Web Users Based on Access Patterns.” International Workshop on Web Usage Analysis and User Profiling (WEBKDD’99), San Diego, CA, 1999.

[5] Pitkow J. and Pirolli P. Mining longest repeating subsequences to predict www surfing. In Proceedings of the 1999 USENIX Annual Technical Conference, 1999. International Journal of Computer Applications (0975 – 8887) Volume 115 – No. 16, April 2015

[6] Z. Su, Q. Yang, Y. Lu, and H. Zhang. Whatnext: A prediction system for web requests using n- gram sequence models. In Proceedings of the First International Conference on Web Information System and Engineering Conference, pages 200-207, Hong Kong, June 2000.

[7] Phoha V. V., Iyengar S.S., and Kannan R., “Faster Web Page Allocation with Neural Networks,” IEEE Internet Computing, Vol. 6, No. 6, pp. 18-26, December 2002.

[8] Zhang T., Ramakrishnan R., and Livny M., “Birch: AnEfficient Data Clustering Method for Very Large Databases.” In Proceedings of the ACM SIGMODConference on Management of Data, pages 103-114, Montreal, Canada, June 1996.

[9] Cadez I., Heckerman D., Meek C., Smyth P., and Whire S., “Visualization of Navigation Patterns on a Website Using Model Based Clustering.” Technical Report MSR- TR-00-18,