International Journal for Rapid Research in Engineering Technology & Applied Science
Vol VIII Issue III March-2022
ISSN Online :- (2455-4723)

Frequent Item Set Mining from Log Files using Navigation Pattern &Hadoop Techniques

Khushbu Ankil*, Prof. Mohit Jain**
BM College , RGTU Bhopal, Indore, 452001,India*
BM College , RGTU Bhopal, Indore, 452001,India**
ankilkhushbu@gmail.com *, *bmctmohitcs@gmail.com*[**]

ABSTRACT:

This web log contains lot of information so it is preprocessed before modeling. The web log file is preprocessed and converted into the sequence of user web navigation sessions. The web navigation session is the sequence of web page navigated by a user during time window. The user navigation session is finally modeled through a model. Once the user navigation model is ready, the mining task can be performed for finding the interesting pattern. Modeling of web log is the essential task in web usage mining. The prediction accuracy can be achieved through a modeling the web log with an accurate model to improve the performance of the servers, caching is used where the frequently accessed pages are stored in proxy server caches. Pre-fetching of web pages is the new research area which when used with caching greatly increases the performance. In this paper, a better algorithm for predicting the web pages is proposed. Clustering of web users according to their location using clustering is done and then each cluster is mined using FP-Growth algorithm to find the association rules and predict the pages to be pre- fetched for storing in cache.

Keywords: Web Usage Mining, Semantic Web, Domain, Sequential Pattern Mining, Recommender Systems, and Markov Model, Prediction, web log.

1. INTRODUCTION

1.1 Web Usage Mining:

In recent times, Web Usage Mining has emerged as a popular approach in providing Web personalization . Web usage mining is concerned with finding user navigational patterns on the world wide web by extracting knowledge from web usage logs (we will refer to them as web logs). The assumption is that a web user can physically access only one web page at any given point in time, that represents one item.

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: Mined patterns

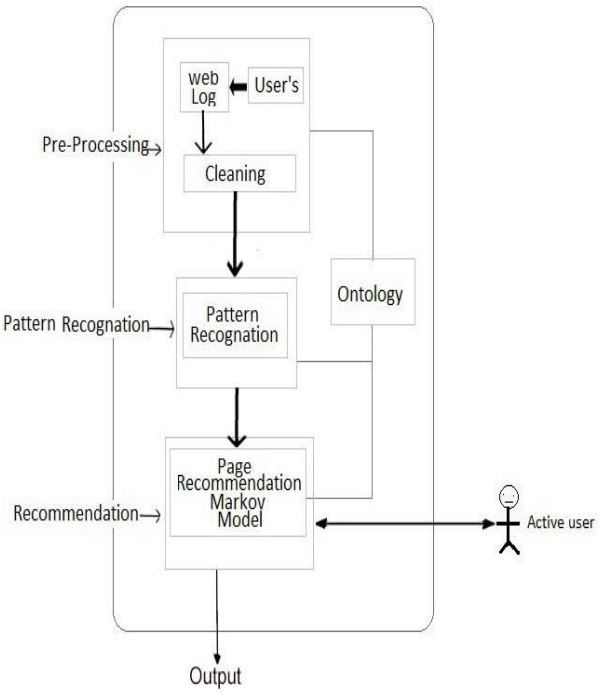
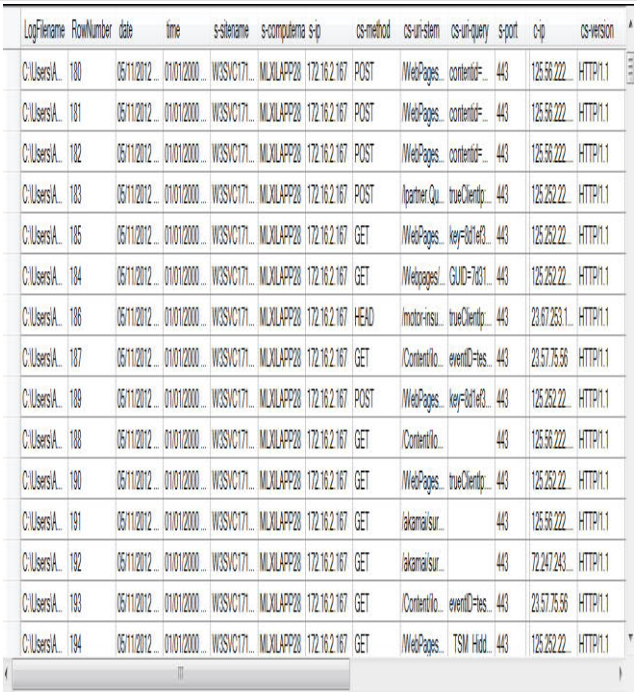


Figure1: Phases of Web Usage Mining



LogFilename	RowNumber	date	time	s-sitename	s-computerna	s-ip	cs-method	cs-uri-stem	cs-uri-query	s-port	c-ip	cs-version
C:\Users\A...	180	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	POST	/WebPages...	contentId=...	443	125.58.222...	HTTP/1.1
C:\Users\A...	181	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	POST	/WebPages...	contentId=...	443	125.58.222...	HTTP/1.1
C:\Users\A...	182	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	POST	/WebPages...	contentId=...	443	125.58.222...	HTTP/1.1
C:\Users\A...	183	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	POST	/partnerQu...	trueClientIp...	443	125.252.22...	HTTP/1.1
C:\Users\A...	185	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	GET	/WebPages...	key=b01ef3...	443	125.252.22...	HTTP/1.1
C:\Users\A...	184	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	GET	/Webpages/...	GUID=7d31...	443	125.252.22...	HTTP/1.1
C:\Users\A...	186	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	HEAD	/motor-insu...	trueClientIp...	443	23.67.253.1...	HTTP/1.1
C:\Users\A...	187	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	GET	/ContentRo...	eventID=tes...	443	23.57.76.56	HTTP/1.1
C:\Users\A...	189	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	POST	/WebPages...	key=b01ef3...	443	125.252.22...	HTTP/1.1
C:\Users\A...	188	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	GET	/ContentRo...		443	125.58.222...	HTTP/1.1
C:\Users\A...	190	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	GET	/WebPages...	trueClientIp...	443	125.252.22...	HTTP/1.1
C:\Users\A...	191	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	GET	/akamaisur...		443	125.58.222...	HTTP/1.1
C:\Users\A...	192	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	GET	/akamaisur...		443	72.247.243...	HTTP/1.1
C:\Users\A...	193	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	GET	/ContentRo...	eventID=tes...	443	23.57.76.56	HTTP/1.1
C:\Users\A...	194	05/11/2012...	01/01/2000...	W3SVC171...	MLXILAPP28	172.16.2.167	GET	/WebPages...	TSM Hdd...	443	125.252.22...	HTTP/1.1

Figure 2 : A Sample of Serer Side Web Log

Web Usage Mining is the field of web mining which dealswith finding the interesting usage pattern from the logging information. The logging information is stored in a file known as web log file. Web log file contains lot of information like IP address, date, time, web pagerequested etc.

1.2 Web Log: The web log is a registry of web pages accessed by different users at different times, which 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.

- Server Log: the server stores data regarding requests performed by the client, thus data regard generally just one source. Server Log details are given in Figure 2.
- Client Log : it is the client itself which sends to a repository information regarding the user's behavior (can be implemented by using a remote agent (such as Java scripts or Java applets) or by modifying the source code of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities.);
- Proxy Log: information is stored at the proxy side, thus Web data regards several Websites, but only users whose Web clients pass through the proxy.

2. LITERATURE REVIEW

Due to rapid growth in the number of internet users, the user perceived latency has become a serious issue for the web service providers. Researches have been done which combines different techniques from multiple domains to overcome this issue.

To reduce perceivable network latency, researchers focused on pre-fetching popular documents. The integration of pre-fetching and caching techniques greatly improves the performance and also reduces the running time of the applications by 50% .

Garofalakis basically provided a survey on data mining techniques and algorithms for discovering structures of web, hypertext and hyperlink .Ina generalization based clustering approach has been presented, which also incorporates attribute oriented induction.

Pitkowetal.predicted the web surfer's path in pattern extraction mechanism.Worked on prediction of future requests and has built n-gram model for the same.

Cooleycategorized the web mining and then presented possible research areas. A scheme for fast allocation of web pages using data mining techniques and competitive neural network is being discussed in.

Zhang proposed an efficient data clustering approach for very large databases, by generating hierarchical clustering of web users based on their access patterns. In order to user's web page requests, clustering technique using first-order Markov models has been provided in Short-term pre-fetching uses Dependency Graph (DG), where graph consist of access patterns and Prediction by Partial Matching (PPM) is used. The merit of short-term pre-fetching is that it reduces the user-perceived latency. Other than this, it also has two demerits. Firstly, it may cause excessive network traffic, if pre-fetching policy is not designed cautiously. Secondly, optimization of cache space is not good in this pre-fetching scheme. The long-term pre-fetching uses global object access pattern statistics, where clusters of valuable objects are identified. This scheme may be used in places like as Content Distribution Network (CDN), mobile computing environments etc.

Different benefits of web pre-fetching are provided in, whereas motivates in research in web caching.

Vakalidescribed 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 .

NanhaySingh used the two web mining techniques, K-Means clustering and Apriori algorithm together to predict and pre-fetch the web pages from the proxy server.

3. PROBLEM FORMULATION

3.1 Problem Definition:

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.

3.2 Description

In the existing works pre-fetching the likely pages and caching them in the web caches improve the performance of the servers. Prediction of the pages can be performed using different algorithms such as Markov model, Apriori algorithm etc. The recent works also includes the integration of more than one of algorithms to overcome the limitations of each other.
In the Existing work two different data mining approaches to predict the web pages, which are likely to be accessed in near future, are used. The existing works try to cluster the data based on the user interests or the time taken by the server to respond back to the requests. In this work improvement of the performance is achieved by clustering the users in different group based on the location from which the request is sent.
Clustering the users based on the location improves the hit ratio. Suppose the user staying in Delhi searches for Cineplex in Delhi. It is most likely that he may also search for restaurants near him. By adding the location in the algorithm we can provide him better search results, which he is looking for.

3.3 Development Methods/ Methodology

In this Existing work an improvement in the performance is also achieved by using the FP growth algorithm for finding the frequent item set instead of the Apriori algorithm.

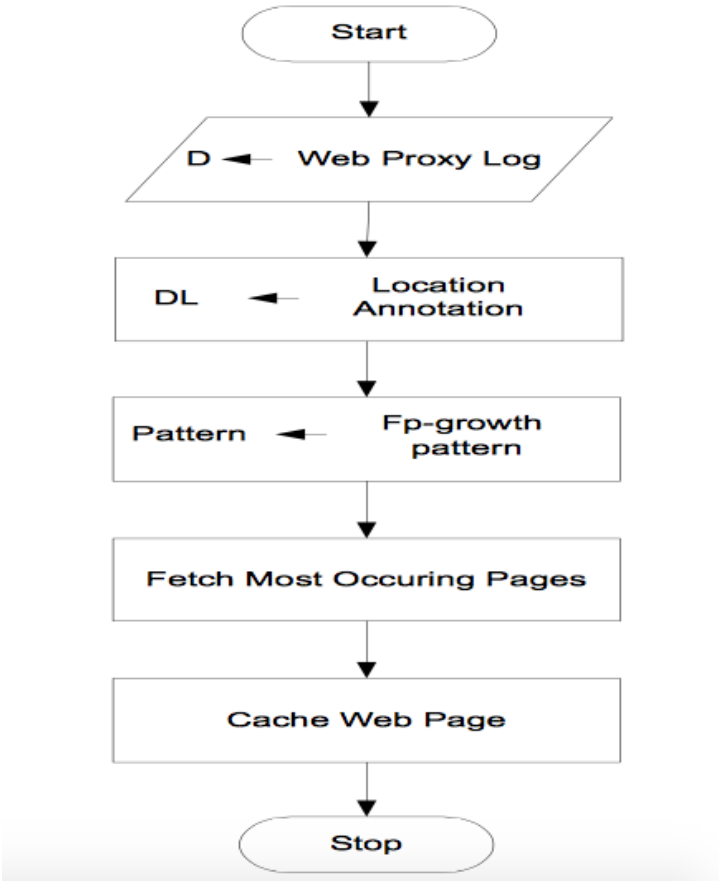


Figure 2 : Existing System with Fetch Most Occurring Pages

4. PROPOSED WORK

We propose to formulate AN improved F.P tree rule to implement the answer to the matter formulation.

The F.P Tree rule works as follows-

The planned approach that we tend to conceive to implement follows the subsequent steps:

Step1: within the initial steps knowledge is being collected from the online log file then Preprocessing is applied. within the Preprocessing the information is being loaded and it's being born-again in to the information set having fields Client-IP, Session ID, Country, Access Date Time, Method, URL, URL_ID, Protocol, Status, Bytes transferred. The session is calculated in half-hour interval of your time, when half-hour the system can acknowledge a similar user as next user.

Step 2: during this step there's Pattern Discovery that is performed by the Frequent Pattern (FP) that involves FP Tree that successively FP growth .FP tree technique is employed in data processing .It consists of 2 passes over the information Set .In the initial Pass it scans knowledge and notice the minimum support for the every item. The item set whose support is a smaller amount than minimum is discarded .The Data item that's enclosed is that the electronic computer or the URL that's being visited by the User. Next steps within the initial Pass within the FP tree ar to get a decreasing order on the idea of frequency of prevalence of the Item Set that is that the URL visited by the User. within the Second Pass of the FP Tree group action is being browse .In this work the group action is that the variety of user visited the actual electronic computer. The browse group action is iterated till all the group action is being completed. when Reading all the group action discards all the group action that has lees support or support than the minimum threshold worth.

Step3: during this step Pattern analysis is completed and during this Candidate rule is generated and on the idea of candidate rule confidence is generated. On the idea of pattern analysis Prediction is completed of the User's Future request.

The integration of linguistics info directly within the transition chance matrix of lower order Markov models, was bestowed as an answer to the current exchange drawback. This integration additionally solves the matter of contradicting prediction. , we tend to propose to use linguistics info as a criteria for pruning states in higher order (where k > 2) Selective Markov models, and compare the accuracy and model size of this idea with semantic-rich models and with ancient Markov models.

Markov Model as a planned resolution to proven semantically meaty and correct predictions while not exploitation difficult all ^{K}th order. The linguistics [distance matrix Weight Matrix and Transition Matrix is directly utilized in Markov model

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

5. CONCLUSION

Web usage mining model is kind of mining to server logs. Web usage mining used for the improvement of improving the requirement of the system performance, the customers relation and realizing enhancing the usability of the website design. 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

[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, Microsoft Research, March 2002.

[10] Podlipnig S, Boszormenyi L. A survey of Web cache replacement strategies. ACM Comput Surveys 2003;35(4):374–98.

[11]Rabinovich M, Spatsheck O. Web caching and replication. Addison Wesley; 2002.

[12] Teng WG, Chang CY, Chen MS. Integrating Web caching and Web prefetching in client-side proxies. IEEE Trans Parallel Distributed Syst 2005;16(5):444–55.

[13] Schloegel K, Karypis G, Kumar V. Parallel multilevel algorithms for multi-constraint graph partitioning. In: Proceedings of 6th international Euro-Par conference. September 2000. p. 296– 310.

[14] Vakali A, Pokorny J, Dalamagas T. An overview of Web data clustering practices. In: Proceedings of the EDBT Workshops 2004. Heraklion, Crete; 2004. p. 597–606.

[15] Nanhay Singh, ArvindPanwar and Ram Shringar Raw.Enhancing the performance of Web Proxy Server using Cluster Based Pre-fetching technique.IEEE 2013.