Digital Assignment 1

# Name :- Loukik Bhangale

# Registeration Number :- 17BCE0961

# Slot :- F2-TF2

# Faculty :- Prof. Anuradha J

Pattern Mining in Web Logs

Pattern Mining is one of the widely researched areas for data mining. And it is particularly famous for pattern mining in a web logs because that gives us a user behaviour analysis and helps you choose the right tools and strategy for targeting a company’s advertisement to be effective. Ever since the boom of the Internet we have large amount of data available and accessible. Several data mining methods are used to discover the hidden information in the web which can not only be used to analyse but also help with other fields like artificial intelligence, NLP, etc. Web Mining’s biggest application so to extract the hidden information in data stored on the Web. But it has another application where is need to access data more efficiently.

# Frequent Pattern Mining

There are three patterns to be searched are frequent itemsets, sequences ad tree patterns.

Frequency itemset is found using ItemsetCode which is efficient and fast in discovering small frequent itemsets hence faster in long ones too. SM-Tree algorithm is used for discovering tree like pattern called PD-Tree algorithm.

# Web Mining Approaches

Web mining can be categorized into 3 parts: -

1. Web Content Mining
2. Web Structure Mining
3. Web usage mining

Web Content Mining is itself quite self-explanatory, it discovers important information from web based content. It is basically used so that the user efficiently gets the necessary information he/she was looking for. It’s function also includes task of organizing and clustering the document and providing search engines for accessing the different places in the document for quicker access.

Web structure Mining is basically the process of finding the path between links and pages and making a map out of it. This helps in identifying the authoritative pages and original pages which usually have several pages pointing towards it. This process can help web content mining as well as it can provide a structure and direction to the algorithm on where to find the most related information hence improving the performance.

Web usage mining is the task of monitoring the actions of a user while they are navigation through web. Aim is the understand the user behaviour better and use this in marketing and help the E-commerce website lay their websites in a way that they can do their adjustments and increase their sales.

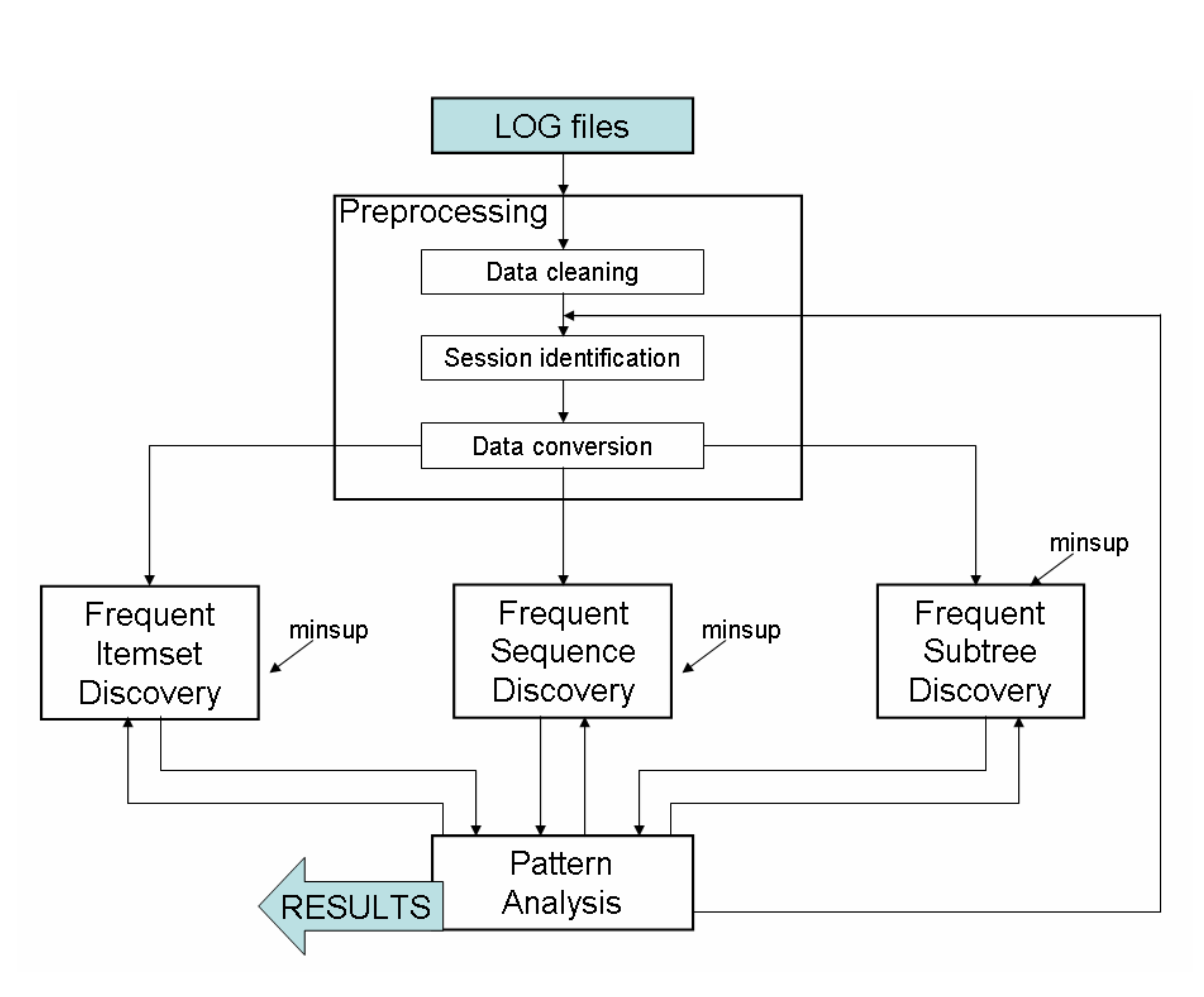
There are basically three types of log files that can be used for web mining. Log files stored on server side, client side and on the proxy servers. Most of the web mining is done on the server side log files as they are more reliable and have the information on which we can work on as each log files have different information stored on them.

# Web Usage Mining

There are three steps that web usage mining consists: -

1. Pre-processing
2. Pattern Discovery
3. Pattern Analysis

The data from the log files first need to be first pre-processed and only then can we apply the algorithms and then it would output the pattern discovery we need. Pattern Analysis is the process of going through the output of the algorithms and making our inference.



For frequent patterns discovery needs only the web pages that are visited, the sequences don’t matter. So the pages are set in a pre-defined manner while omitting the pages that are visited more than once. Where as in case of sequence mining, the ordering of the pages is important as well as how many times and their time stamps also matters. This is why pre-processing matters. For subtree mining both the sequence and the structure of the web pages visited matters.

In the above figure, depending on the output of the analysis the input parameters can be adjusted to obtain the best results.

# Mining Access Patterns Efficiently from Web Logs

A data structure called, Web access pattern tree (WAP-tree), I used for efficient mining of access patterns from pieces of logs. This method is in general an order of magnitude faster than other conventional methods.

Applications to knowledge discovered from web log mining includes improving designs of web sites, analysing system performance as well as network communication, understanding user reaction and motivation, and building adaptive web sites.

Srikant and Agrawal generalized their de ion of sequential patterns in to include time constraints, sliding time window, and user-defined taxonomy and developed a generalized sequential pattern mining algorithm, GSP, which out performs their AprioriAll algorithm. Sequential pattern mining.

In general, web log can be regarded as sequences of pairs of user identifier and event. Each piece of Web log is a sequence of events from one user or session in timestamp ascending order, i.e event happened early goes first.

# Contents of a Log File

* User name: Identification of the user in terms of IP address assigned by Internet Service Provider(ISP)
* Visiting Path: The path taken by the user while visiting the web page, either by typing in the URL, or clicking on links or search engine.
* Path Traversed : Path taken by the user
* Time Stamp : The time spent by the user in each page
* Page last visited
* Success rate: rate can be determined by the number of downloads made and the number of copying activity under gone by the user.
* User Agent: String describing the type and version of the browser software being used
* URL: the resourced accessed by the user
* Request type: The methods like GET, POST.

# Location of Log File

* Web Servers
* Web Proxy Servers
* Client Browsers

# Types of Web Server Logs

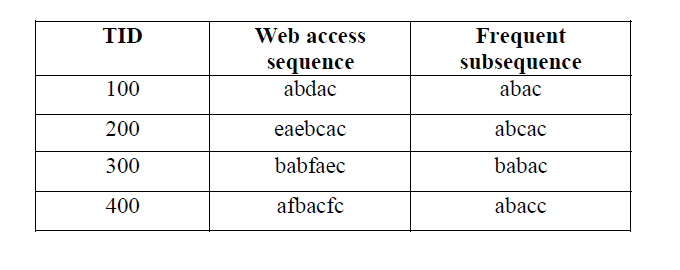
* Transfer Log
* Agent Log:- This records all request that are processed by the server. A log format is specified.
* Error Log: - When an error is occurred while the page is being requested by the client to the web server the entry is made in the error log. Eg. [Wed Oct 11 14:32:52 2000] [error] [client 127.0.0.1] client denied by server configuration: /export/home/live/ap/htdocs/test
* Referrer Log

# Extended log file

The log file contents would be even more efficient if it provides the details of the clicks made by the user while visiting the web site. If the user opens a particular web site and does some other work outside the system then it may also be considered as the usage of the web site. The details regarding the clicks made by the user and the time he or she scrolled or did any other operation can also be noted for effective mining of the web usage data. We shall consider one more situation were the user clicks to open a web site and also works with some other web site in a different browser window. In this situation we can analyze that the user can only read details of only one web site at a time. Then it is understood that the other web site is said to be ideal. But even in this situation the details in the log file would note the input as the web page is being used. By taking these small differences in the time or the session of the web page being used, still an efficient web mining can be done.

# Mining Process of PLWAP Tree

Every one of this WAP tree's hubs has a parallel position code relegated for legitimately mining the consecutive examples without re-building the middle of the road WAP trees. The method introduces a greatly improved exhibition than that accomplished by the WAP-tree strategy. The web log get to arrangement database (WASD) in Table 1 is utilized to indicate how to build the PLWAP-tree and do PLWAP-tree mining.



## Construction

Position code is a twofold code used to show the position of the hubs in the WAP-tree. In information structure, when executing a general tree structure, a tree is more often than not changed into its proportional parallel tree, which has a fixed number of tyke hubs.

* Rule 1: Given a WAP-tree with some nodes, the position code of each node can simply be assigned following the rule that the root has null position code, and the leftmost child of the root has a code of 1, but the code of any other node is derived by appending 1 to the position code of its parent, if this node is the leftmost child, or appending 10 to the position code of the parent if this node is the second leftmost child, the third leftmost child has 100 appended, etc. In general, for the nth leftmost child, the position code is obtained by appending the binary number for 2n−1 to the parent’s code.
* Rule 2: Suppose any ancestor node in WAP tree has no more than n (n<100) children node Given a WAP-tree with some nodes, the layer code of each node can simply be assigned following the rule that the root has 0 layer code, and the leftmost child of the root has a code of 001, the second leftmost node has a code of 002, and so on. So the layer code of a node in WAP tree is its parent node layer node plus its position (from left to right) 01-99.

The BFWAP calculation adjusts the WAP-tree structure for putting away successive consecutive examples to be mined. Be that as it may, to enhance mining productivity, the paper proposes to discover layer designs rather than prefix designs as done by PLWAP-tree mining. In addition, so as to discover progenitor relative relationship of hub in WAP tree effectively, Breadth-First incessant header hub linkages and layer codes are proposed. While the Breadth-First linkage gives an approach to navigate the occasion line without going in reverse, layer codes are utilized to recognize the degree of hubs in the BFWAP tree. With these two techniques, the next successive occasion in each addition tree is found without

navigating the entire WAP-tree and more straightforward and effective than PLWAP calculation. Hence, it maintains a strategic distance from re-building WAP-tree recursively. The investigations demonstrate that mining web log utilizing BFWAP calculation is much more effective than with WAP-tree and GSP calculations, particularly when the normal continuous grouping moves toward becoming longer and the first database increases.

# Data Mining and Pattern

For the mining procedure, other than the information, the base help limit worth is required. It is one of the key issues, to which worth the help limit ought to be set. The correct answer can be given uniquely with the client connections and numerous emphasess until the fitting qualities have been found. Hence, in particular, that the association of the clients is required in this period of the mining procedure, it is prudent executing the regular example disclosure calculation iteratively on a generally little piece of the entire dataset as it were. Picking the correct size of the example information, the reaction time of the application stays little, while the example information speaks to the entire information precisely. Setting the base help edge parameter is certainly not a unimportant errand, and it requires a great deal of training and consideration with respect to the client.

The frequent itemset discovery and the association rule mining was accomplished using the ItemsetCode algorithm with different minimum support and minimum confidence threshold values.

Analyzing the results, one can make the advertising process more successful and the structure of the portal can be changed such that the pages contained by the rules are accessible from each other.

# User’s Area of interest

Learning the clients desire is an extremely monotonous procedure. A solitary word may have various perspectives by various client. In the event that the clients territory of intrigue is recognized, at that point we can have a proficient mining process. How is this done. On the off chance that questions are presented to the client it would be a tiring procedure for a client to address the inquiry each time he makes a hunt. In this way the clients intrigue can be investigated by the main endeavor made to open a page. At that point the following stage done by the excavator is to mine the web by and by and give the rundown of result implied uniquely for the clients territory of intrigue. This may inturn limit the rundown of alternatives and make the looking through procedure significantly increasingly viable. This should be possible alongside examination of the log documents to have utility as one of the factor.

# References

1. Mining Web Log Sequential Patterns with Layer Coded Breath-First Linked WAP-Tree
2. Mining Access Patterns Efficiently from Web
3. Fast Algorithm for mining assication rules
4. Mining sequential patterns
5. Mining temporal relationships with multiple granularities in time sequences
6. Data Mining and Knowledge discovery
7. Frequent Pattern Mining in Web Log Data
8. Frequent Sequence mining in web log data
9. Mining constraint-based Multidimensional Frequent Sequential Pattern in Web logs
10. Web Access Sequential Pattern Tree Mining
11. Web Mining: Information and Pattern Discovery on the World Wide Web.