Approach for Processing of Web Usage Data

Mr.Ankit Rathi[1], Prof. Abhijeet Raipurkar[2]

[1]M.tech, Computer science and Engineering, RCOEM Nagpur, India.

[2]Assistant Professor, Computer Science and Engineering, RCOEM Nagpur,India.

Email: ankitr5691@gmail.com, raipurkarar@rknec.edu

---***---

Abstract - *The Web has recently become a powerful stand for retrieval of Information and discovering knowledge from web data. The Web mining is one of the applications of data mining techniques to depict the knowledge out from web log data. Web mining is generally defined as Preprocessing, discovery and analysis of useful information from the Web. Web Usage Mining consists as the process of Preprocessing, Pattern Discovery and pattern Analysis. The memory and time usage is compared by means of the pattern discovery algorithms such as Apriori and Frequent Pattern Growth algorithm. The aim of this paper is to understand the web usage mining process such as preprocessing of web usage data and also the finding of frequent Patterns and their analysis. And also the comparison of both algorithms on the same dataset is done. Due to more use of internet, the log files are increasing at higher rate in according to size. The Preprocessing plays an important role in efficient mining process because data in Log files is normally noisy and not distinct*

Key Words: Web Usage Mining, Preprocessing, Pattern discovery, Pattern analysis.

1. INTRODUCTION

The Data mining is defined as the extraction of unidentified, useful and understandable patterns from user transactions. The design of website matters a lot. The interests of the users also help in designing enhanced Websites. The Web service providers desire to find the technique to guess the behavior of the users and create the Website suited for the different users. The analysts in business area want to have tools to learn needs of the consumers. All of them are expecting techniques to help them suit their needs and solve the problems occurs on the Web. Therefore, Web mining becomes a popular area and is taken as the research area for this analysis. Web mining is used to retrieve, extract and assess data for discovery of useful data from documents on Website. Web mining process consists of types as Web content mining, Web structure mining and Web usage mining. Web Content Mining done with the information discovered from the web documents. Web Structure Mining mines the hyperlink structure within the web data itself. Web Usage Mining mines data stored in web log file at web server.

2. WEB USAGE MINING

Web usage mining is the process of mining the data in web log files which are generated during the usage on web by the different users. This mined data from the web is used to find the useful information about users and their behavior also for the improvement of the website design. The stages of web usage mining process are:

1. Data Collection:

This step includes different data sources. The log file at web server is one of the data source. There are also other sources of data like log file on client or on the Proxy servers. The client side log is also used for the tracking purpose in some cases.

2. Data Preprocessing:

The Contents on the web is not in structured format always, hence preprocessing phase is required. The use of this process is that it converts the raw data into useful data that is data is used for drawing the knowledge about the behavior of the user. Data preprocessing consists of the following –

A. *Data Cleaning*: Data Cleaning helps in removing the irrelevant items or files such as images and audio files. The quality of data is very important for analysis purpose, so cleaning is an important task. If user tries to access some web page then the images or audio files are also downloaded with that page. Hence only html files are useful for us and these images or audio files get deleted. Also in this, the Checking of the Status codes in log entries is done that is successful or not, if it is not successful, then remove that entry.

B. *User Identification*: The important step is to identify individual users who access a website. The user is identified from his IP address. And also the unique users are identified. If the IP address of the user is same as previous entry in log file then it is considered as only one user. And if it is different, then it is assumed that there is a new user.

C. *Session Identification:* The pages surfed by the user at a particular visit or time then it is considered as the session of that user. There are many sessions are possible of the same user. There are different methods available for identifying the sessions of the user. There is one method which depends on time and another which depends on navigation in web used for identifying sessions. The method which depends on time which is calculated by difference between two time stamps of the same user. These methods are not reliable because users may engage in some other work after opening that web page. While in method depends on navigation, the web page connectivity is find out. If a web page is not connected with page which is opened previously in a session of the same user, then it is considered as a new session of that user. Both the methods are widely used according to the applications.

3. *Pattern Discovery:*

The pattern discovery task is applied to the preprocessed data in web usage mining process. There are different methods or techniques performed in this phase like Classification, Clustering, Association rule mining, etc. Clustering is a technique of grouping items having same property. The patterns discovered in this phase by using these methods are helpful in Ecommerce area or for the improvement of website. By using such methods, web analyzers can predict about behavior of user which can help in placing advertisements intended for certain user groups.

4. *Pattern Analysis:*

The Pattern Analysis is last stage in web usage mining process. The patterns which are found in pattern discovery phase are analyzed in this phase. So it is important to take only these patterns which are useful for our analysis purpose according to applications. There are also many tools available for the knowledge transformation. The accurate analysis is governed by the application for which web mining is done.

3. IMPLEMENTED WORK

Our approach that can be given with the help of block diagram –

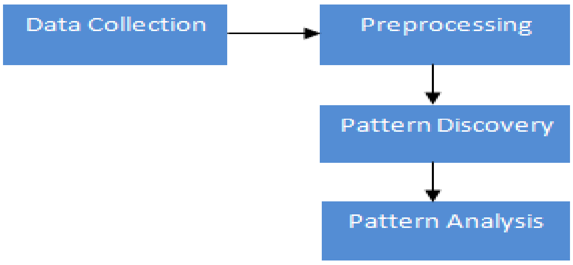


Fig 1: Proposed Approach

A) Data Collection –

The web usage log file is collected from the web Server. The entry in the log file is given in the format as-

```
155.33.62.85 - - [02/sep/1995:16:01:12 -0400]

"get /pub/rsjdfg/pictures.html http/1.0" 200 679
```

Fig 2: Dataset

The format of Web log data which is stored in web log file is shown above. The row in the log file contains information as the IP address, the date and time, method, http request, http status code and total bytes sent, as shown in fig 2. The row in fig 2 informs that the user having IP address 155.33.62.85 have access to pictures.html on September 2, 1995 at 16:01:12 with the http status code 200 that means successful and size of file transferred is 679 bytes and so on.

B) Preprocessing –

The first stage is Data Cleaning in this phase. The irrelevant and unwanted data is sort out in this stage. This process is done based on criteria as –

The first criteria based on the extension of a file. The file with extensions html, php, jsp are taken. And file having extensions such as jpeg, gif, jpg, txt are not taken, because they does not give any solid information about user.
The second criteria based on response code from web server. The http status code 200 shows that the users request is accepted and displayed web page to user by web server. Therefore, all data items having status code other than 200 are not taken.
The third criteria based on access method that is with the only GET method access is allowed, because it shows behavior of the user. With other methods such as HEAD and POST, the data item will be deleted.

The Second stage in the Preprocessing is the User Identification. After data cleaning the next step is User Identification. In this stage, we find out the number of unique users and also considered users based on frequency of accessing a web page. The users having access more than five times were used in our system, because it gives the exact description about the user behavior.

The third stage is session identification. The session identification is the process of dividing the sessions of the user. If user having two accesses which are separated by an interval exceeds than a set threshold they are considered as different sessions. We considered here a threshold of 30 minutes. If access of same user carrying a difference of more than 30 minutes then it is considering as a new session of that user.

C) Pattern Discovery –

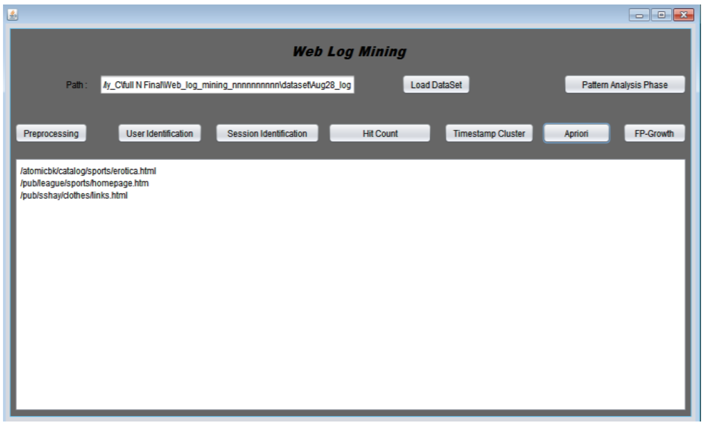
After first phase that is preprocessing phase, the method of pattern discovery should be used, because the aim of this phase is to know the frequent patterns from web log files. This phase consist the process of association rules, statistical analysis, clustering and so on applied to the web data. We perform processes to discover patterns from preprocessed data.

First we did the Clustering. The Clustering is a process to grouping elements having similar property. We did clustering according to the timestamp. It means that which web pages are accessed on which time that is at morning, afternoon or at night. According to given timestamp in data, the web page gets clustered.

Also we find out page hit count that is how many times the particular web page gets accessed by the user. According to this we get idea that which pages are most popular pages in dataset accessed by the users.

Then we applied two algorithms for finding the frequent patterns as Apriori and FP-Growth Algorithms on same dataset. Both are well-known algorithms for frequent patterns mining. Apriori Algorithm uses the "bottom-up" technique. In Apriori algorithm, there is a minimum support which is set and also the candidate generation takes place. The different passes which performs candidate generation. And at last prune all candidate item sets which are lower than minimum support. And scan the transaction database to find out the support and then save the frequent item sets.
Whereas in FP-Growth algorithm, the specific data structure that is frequent-pattern tree (FP-tree) is used. The FP-tree attributes are: It includes one root marked as "root", a set of prefix sub-trees as the child of the root, and a frequent items header chart. And each node in the prefix sub-tree consists of three areas: Item name, Count, Node Link. And there is no candidate generation in this algorithm. In simple words, this working of this algorithm is as follows:
First calculate minimum support then find frequency of occurrence. Prioritize the item and order the item according to that priority. Then draw the FP-tree and finally generate frequent patterns.

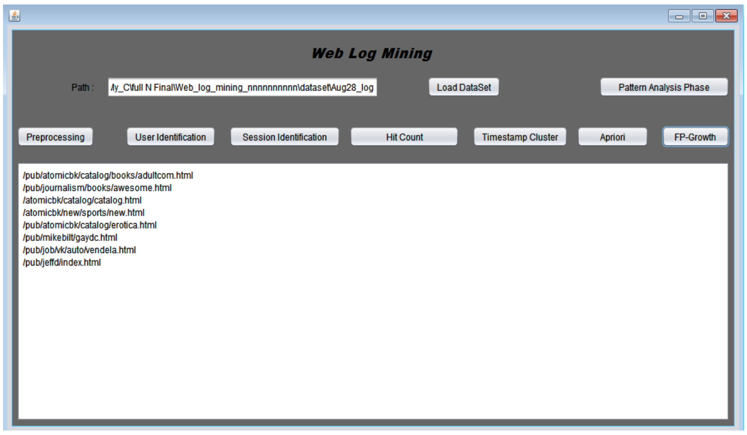
D) Pattern Analysis –
The last phase of this Process is Pattern Analysis phase. The patterns which are mined are not suitable for interpretations. So we removed patterns or rules which are not interesting from the set found in the pattern discovery phase. And most frequent pattern we get using these algorithms as –



Path : /y_C\full N Final\Web_log_mining_nnnnnnnnnn\dataset\Aug28_log

```
/atomicbk/catalog/sports/erotica.html
/pub/league/sports/homepage.htm
/pub/sshay/clothes/links.html
```

Fig 3: Output of an Apriori Algorithm



Path : /y_C\full N Final\Web_log_mining_nnnnnnnnnn\dataset\Aug28_log

```
/pub/atomicbk/catalog/books/adultcom.html
/pub/journalism/books/awesome.html
/atomicbk/catalog/catalog.html
/atomicbk/new/sports/new.html
/pub/atomicbk/catalog/erotica.html
/pub/mikebilt/gaydc.html
/pub/job/vk/auto/vendela.html
/pub/jeffd/index.html
```

Fig 4: Output of an FP-Growth Algorithm

We calculated the hit count for each web page and depending on that we can generate report on which pages are most popular. In our experimental web log file ‘/pub/journalism/books/awesome.html’ this page is most popular and we can also see the number of visitors count of the page.

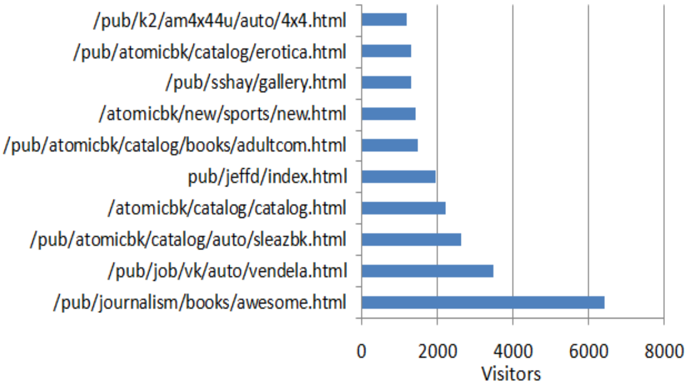


Fig 5: Popular Pages

Also by the highest click we also estimated the daily activities. We calculated the total number of visitors by date that is on which date how many users accessed the web pages. We describe this activity by a graph as –

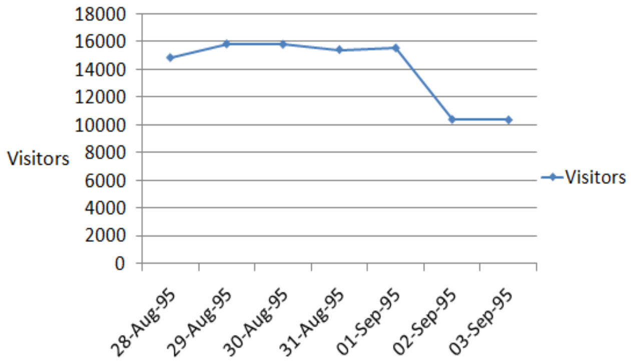


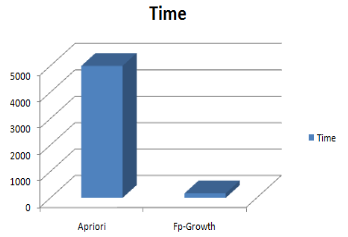
Fig 6: Daily Visitors

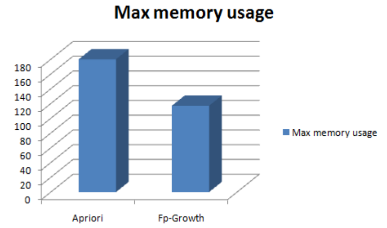
4. RESULTS

Web usage mining is an essential tool for realizing personalized user-friendly and business-optimal Web services. Web usage mining is used by e-business websites to organize their sites and to increase the profits as well. We observed some results after performing the web mining on web data as –

Total entries in Log File	1654934
Entries after cleaning	266701
No. of total Users	90513
Users after cleaning	79762
Unique Users	11797
Total no. of Sessions	28096

Apriori - the typical mining algorithm is a way to find out certain capable, regular knowledge from the massive ones. But there are two more defects in the mining process of this algorithm. The first that it needs every time the scanning of the database and the second that it produces a irrelevant candidate sets which seriously occupy the system resources. An improved method is used on the basic of the defects above. The improved that is FP-Growth algorithm reduces the database scan, and we prune the candidate item sets in accordance with the provided minimum supporting degree and we easily get the frequent item sets. The Comparison of both algorithms is as follows –





After analysis, the FP-Growth algorithm reduces the resources occupied and improves the efficiency and gives better quality. Under large minimum supports, the improved FP-tree runs faster than Apriori algorithm.

5. Conclusion

The web is an important channel of conducting business and E-commerce. Therefore the designing of these pages on web is very important for the web designers. This feature has large impact on the number of visitors. So the web analyzer has to analyze with the data of log file for finding the useful patterns. In this paper we tried to give a clear understanding the preparation of data and pattern discovery process. One of the algorithms which is very simple and easy to implement is the Apriori algorithm. The output of the system was in terms of memory usage and speed. The main disadvantage of Apriori algorithm is that the generation of candidates which is very costly. In future the algorithm can be extended up to web content mining, web structure mining, etc. More research needs to be done in future in area as Computer Security, Bioinformatics, Web Intelligence, Database Systems, Finance, Marketing, Healthcare and Telecommunication, etc.

REFERENCES

[1] Uma Maheswari, Dr. P.Sumathi, *A New Clustering and Preprocessing for Web Log Mining*, World Congress on Computingand Communication Technologies, IEEE, 2014

[2] Theint Theint Aye, *Web Log Cleaning for Mining of Web Usage Patterns,* International conference on Computer Research and development IEEE, 2011.

[3] Navin Kumar Tyagi, A.K. Solanki, Sanjay Tyagi, *An Algorithmic Approach to Data Preprocessing in Web Usage Mining*, International Journal of Information Technology and Knowledge Management, Volume 2, July-December 2010.

[4] Suresh R.M., Padmajavalli R, *An Overview of Data Preprocessing in Data and Web usage Mining,* IEEE, 2006.

[5] Chitraa, Dr. Antony Selvdoss Davamani, *A Survey on Preprocessing Methods for Web Usage Data*, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010.

[6] Shahnaz Parvin Nina, Md. Mahamudur Rahaman, Md. Khairul IslamBhuiyan, Khandakar and Entenam Unayes Ahmed, *Pattern Discovery of Web Usage Mining*, International Conference on Computer Technology and Development, Volume 2, IEEE,2009.

[7] Tasawar Hussain, Dr.Sohail Asghar, Dr. Nayyer Masood, *Web Usage Mining: A Survey on Preprocessing of Web Log File*, IEEE, 2010.