WEB USAGE MINING USING PATTERNS WITH DIFFERENT ALGORITHMS

SOBIA MEHRBAN

Department of Computer Science, University of Management and Technology, Lahore, Pakistan
Email: f2017108015@umt.edu.pk

ABSTRACT. Web usage mining is a part of data mining. Data usage mining is divided into three parts 1) Data content mining 2) Data structured mining 3) Data usage mining. In this paper I am discussing about log files which are used in data usage mining. Log files are used to store user’s activity in web server using websites. So that websites can be improved by gathering user data. Web usage mining having three sub parts which is reprocessing, data discovery and data analysis. Further, in this paper, details about web log files are discussed. Three algorithms are discussed which are used for patterns of log files. There comparison is showed in this paper with the help of graphs.

Keywords: Log files; apriori; FP-growth; K-means; Data Usage Mining; webserver; Web data; web mining.

1. Introduction Data mining is the procedure to gather data from different locations and analyze it in such a way that it can provide useful information or result, Web mining is the process of extract useful web data from different websites or website documents. Data mining is a part of structured mining while web mining is a part of unstructured and semi structured mining. As we all know, [1] there is vast amount of data available on web pages, so WWW (World Wide Web) is playing a role of fertile area in this technique. In web mining data can be of any kind, like images, text, mp3 audio, history records and mp4 videos. Web mining techniques interacts with different other areas like Artificial Intelligence, Data extraction, Databases and many more. Webb mining provides better opportunities to business organization to understand web data and web users in order to generate more profit from their business because they can improve their product with the help of website user feedback, improve their marketing work strategies and can be up to date with the latest technologies using web browsing techniques. Web mining not only contains web data, it also contains user’s information, number of user’s information, website traffic information and log data. Web mining is divided into three parts. I) Structured Mining II) Content Mining III) Usage Mining.
In this paper, the area at which I am working is Web Usage Mining. In this paper web usage mining techniques, its advantages and best algorithm that is implemented in this will be discussed.

**Web Usage Mining**: Web usage mining is the sub part of web mining techniques. In this type of mining web pages having large group of web access information. Usage defines the webpages patterns which includes date and time and references of web servers and web pages. Usage mining is very much popular in online marketing businesses websites and in e-businesses. In e-businesses these kinds of websites play a very important role to gain users feedback and information. In-depth log is maintained in this case to check productivity flow. The information that is generated by this log, is most important promotion for that product and e-business is very much dependent on this information.

Web usage mining is further divided into three parts. i) Web server data ii) Application server data iii) Application level data.

- **In web server data**, web server collects the record of user’s logs. This type of data contains reference id and IP address of server. **Application server data** is used to keep record of business applications and maintain its log record. In this webserver collects the user’s logs and all information of web traffic. **Application level data** is used to occur new events in applications. These applications generating history using turned on logging of events. In application level is further divided into three parts server, client and proxy side.

- In web usage, most of the work is done at server side. Web usage mining having data of webservers, proxy data, user’s information or logs, number of clicks from mouse, data of registration, having logs of different browsers, many users profile, data of bookmarks and many kinds of data of interactions. This web mining part is used to judge user’s area of interest, so that designer of that website can design their website according to mood of users and attract them by giving their demanding services.

2. **Literature Review**. There is large amount of work done in web mining. Lots of research has been done on content, structural and data usage mining in the field of artificial intelligence, data extraction and in databases. In web usage mining lots of people work on patterns of web usage that help to monitor activities in e-business, e-commerce, e-education and many more [2]. Much research work has been done on how accurate web usage mining help to improve business, motivate clients towards your website, product promotion, guide users to search logical structure in their web pages [3].

Koutris, Avouris and Daskalaki just recently published survey on latest work done in Data Usage Mining [4]. Pierrakos et al research on the tools of usage process [5]. Eirinaki and Vaziriannis published their survey on web usage mining for tools and personalization of web data. Cooley et al applying a frame on web mining using cluster and patterns [6]. Kosala and Bookley research on terms of web mining and lately they represent their work in WebKDD. WebKDD is the workshop which having 10 series. All of the series are related to Web Usage Mining [7]. Zaiane research on the method of OLAP using web usage mining. He also proposed methods for content and structured mining like multimedia data [8]. Spilopoulou proposed his work on applications that are using in web mining [00]. Cooley worked on in-depth log, procedures and tools in web usage mining. Cooley proposed 2 methods Web Miner and Web swift of web mining in practical way [9]. Liu and Lee implement a multi-talented and multi agent java based integrated application platform for e-business and e-commerce for internet shopping users. This application which is intelligent agent can be used in many different applications which are not related to e-business [10].

Kushmerick and Toolan they worked on module of web usage mining. This module compressed the web personalization [11]. The author developed their own pattern system name as SUP. They used SUP for navigation trials of users [12]. Falowski et al proposed algorithm for analyzing evolution of online communities, their languages and divided that communities in sub groups. After divided that, he also gives them levels to analyze their evolution [13].

The author proposed new algorithm using bio-mimetic clustering in web usage mining for visualization of activity of customers and users on website [3]. Khalil Li and Wang proposed prediction
accuracy for web pages using clustering and novel approach. Result shows that’s integration is providing better prediction accuracy than individually [14]. Kumar and Rukami proposed best algorithm between sequence base techniques and fuzzy techniques for pattern of log in frequent sequences [12]. Veeramalai and A.kannan proposed comparison of algorithm between apriori, hash tree and Fuzzy for mining of web log. They proposed end result by comparing all algorithms and show advantages and performance of these algorithms on web log [10]. The author [15] had research on data usage mining technique which is call data preprocessing. In this technique they choose set web log and file log system. Problems that are occurred in their unwanted size and data. In this paper they purposed an algorithm for unwanted size and data name k-means clustering algorithm.

3. Sub Parts of WEB usage Mining. Web usage mining having three parts 1) Pre-Processing 2) Pattern Discovery 3) Pattern Analysis. Further I will explain about these types.

![Figure-2: Process of Web Usage Mining](image)

A. Pre-processing. Pre-processing [5] is a phase in which garbage and noisy data which is present in inconsistent state, processed in a way so that redundant and fuzzy data is removed and now it can be used for getting reliable results. There are few steps in Data Pre-processing.

- **Data Cleaning** is the process in which data is filtered from log file which is irrelevant and full of errors. Data is cleaned from log file by checking its URL to know about which kind of data is going to be cleaned like format of data.

- **User Identification** is the process in which user is identified using his IP address. In this process system identify unique users. [6] With the help of IP address user is identify from which domain they are belonged. Web server can easily identify how many number of users from different domain are using this website. Like Pakistan domain is .pk Australia domain is .au United Kingdom domain is .uk. So with the help of these domains user identify becomes easy.

- **Session Identification** is the process in which time spent by user on each website is recorded. This identification is done by user id and how much users visited that page in that time stamp.

B. Pattern Discovery. In this phase, log data which is organized and converted in cleaned form, different techniques are used for patterns recovery like clustering, association mining rule, sequential patterns, classification and path analysis. Each pattern having different result and methods.

C. Pattern Analysis. In this analysis all the unrelated patterns are removed from this process. Pattern analysis contains query mechanism for data loading, OLAP for visualization of patterns and decision making.

4. Web Log Files. In web usage mining, log files are used. These log files contain record and information of web user’s online activity. Log files contains data which is automatically stored in web servers. Whenever user
open webpage, and he/she request for any action, like they click on any image and link, record of this request or information is attached in current log file. The data in log files are in text form. Log files are saved/located in 3 types of logs. i) Web Server ii) web proxy server iii) client browser.

A. **Log files in Web Server.** When client access the webserver through different websites using their browser, then these webservers record user activities in log file. In this log server all the user activities and users information is recorded. These all information of users is secret. These webserver log files not maintained record of cached pages.

B. **Log files in Web Proxy Server.** Web proxy server is present in-between client and webserver. User information is gathered in different log file. User send HTTP request to webserver, and webserver send user request to web proxy server. In Response web proxy server send response back to webserver, and webserver send response to client.

C. **Log files in Client Server.** Log files in client server are usually generated in browser window automatically. Although these files are generated automatically in user browser window, but this action is performed by some server. The files which are generated by webservice, they contain information for future access of data.

5. **Overview Of Algorithms.** In this paper, I will discuss about different algorithms which are used for patterns in log files for the better result of web application. Results are based on the comparison of different algorithms for time and memory usage patterns. In this paper, comparison of clustering and classification algorithms, K-means, apriori algorithm and FP-growth algorithm will be shown.

I. **Apriori Algorithm.** This [5] algorithm is purposed on patterns of log files. Associations rules are implemented in this algorithm Apriori uses breath first search for the count of item sets and it also use tree structure for item sets. In this algorithm common subset is find for the smallest number ‘X’ of item sets. It generates X from large number of data sets which having X-1 length. Infrequent pattern removed the useless number from list.. Item sets in this algorithm contains frequent number of item sets. In this algorithm there is a rule non-empty subset must also be in frequent form. Frequent items are those which having small sets of items.

**Algorithm:**

\begin{verbatim}
Apriorialgorithm()
{
    X = none;
    Ak = {freq 1-item};
    i= 2;
}
\end{verbatim}
while (Ak-1! = none){
    X = X U Ak ;
    Bk = size I produced for new items
    from Ak-1 ;
    Increment in count for candidates present in Bk that is present in t
    Ak = Min candidates in Bk
    support ;
    i++ ;
} return ( X )

a. Advantages of Apriori. Apriori algorithm is very efficient and it is used for large number of item sets. Implementation of apriori is easy and they are also in parallelized and optimized form. Bottom up approach is used in this algorithm.

b. Limitations of Apriori Algorithm. Apriori algorithm is very easy algorithm but it also having some limitation when it is used in frequent item sets. If there is large number of sets it take large amount of memory and it is very hard to handle large data in this algorithm. This algorithm checks each transaction many times. So, if there is large number of data, than apriori will take huge amount of time and due to which cost will increase. Results of this algorithm in large data is not efficient and mostly not accurate.

To overcome all problems that is occurred in apriori, FP –growth algorithm is now used to compare its result.

II. FP- Growth Algorithm. FP- growth is [3] considered now a day’s fastest algorithm for mining of frequent sets. FP-growth compressed log files into ASCII form using pre-processing module. FP trees is used in this algorithm for storing important, qualitative information regarding frequent patterns. These FP trees contains root, set of sub items roots which is named as children and one header table of frequent items. Sub –trees contains item name, counter and link-node. Item name represents which item is used and in counter number of actions are count and at which number node is reached while link-node checks that item name contains something or either it is null.

Algorithm:

FPGrowth (tree, a)

- Tree having single P path,
- Each combination is indicates as b, datasets will do in path P,
- b U a patterns will generate, with minimum support of datasets in b
- Else variable a will be in header of patterns {
- b=xi U a patterns will generate with support of xi.
- Draw B’s conditional base pattern and FP tree of B’S conditional.
- If B= 65;
- Function of FPGrowth (tree, b, b) will call.
- The datasets of clusters and frequent webpages are mining by FP growth algorithm.

a. Advantages of FP-Growth. FP-growth algorithm is faster than apriori algorithm and tree searching. FP-growth produces a compressed version by reducing the data item sets. FP trees are used in FP-growth which are used for storing efficient information of frequent data item sets. FP trees are complete and compact. [8] Divide and conquer rule is used in FP growth which is used for decomposing of tasks into small number of frequent patterns. FP growth having different mining techniques as compared to apriori and it has different data structure.
b. Limitation of FP growth. Right now there are no limitations are recorded in this algorithm. There is no issue recorded in its result accuracy and complexity. Multiple Local Frequent Pattern Tree (MLFPT) is a new algorithm which is designed for FP-growth. In this algorithm FP trees are going to use in processors.

III. K-means Algorithm. K-means is algorithm of data mining. K-means is a very simple, easy and un-supervised clustering algorithm. K-means consists of two clusters K, one is used for large data items and second one is used for empty data items. K-mean algorithm is much successful in larger number of item sets. K-means are produced randomly in data item sets. [6]K cluster are created by joining them with each nearest mean. Every centroid of k-cluster produce new mean. Convergence is produced when previous steps are continuously repeated. Now I am going to show working of k-means with the help of figures. Log file consist of input from data file. Patterns are used in this algorithm with the name of P1, P2, and P3 and so on. Output contains clusters of defined users. In figure 3, 4 and 5 working of k-means is
Suppose time requires for P1 is 20 minutes, P2 requires 30 minutes P3 requires 25 minutes. Each time is recorded. I am defining the value of K in this algorithm is 2. For k=2 I get result in cluster 1 is p1, p3 and p4 while in cluster 2 I get result p2, p3, p4, p1. With the help of clustering algorithm, useful pattern is obtained. For the partitioning of data mathematical equations are used in k-means.

a. Equation

\[
J(Y, V) = \sum_{N=1}^{M} N_i (y_k, v_n) = \sum_{N=1}^{M} \sum_{i=1}^{u} o_{ij} \cdot d^2(y_i, v_i)
\]

\[
l^2(d_i, v_j) = \left\| \sum_{k=1}^{n} d^2_k - v^2_k \right\|^2 \text{ (eq 2)}
\]

\[
d^2(y_i, z_j) = \left\| \sum_{l=1}^{n} y_i - z_j \right\|^2 \text{ (eq 3)}
\]

\[
x_j = 1 / |m_j| \sum_{i, x \in m} x_i
\]

The k-means randomly initializes the center of clusters, which is used to find minimum distance between points of clusters and data sets. This equation will give better result for average center point of cluster.

b. Advantages of K-means. K-means is very good in large number of data. K-mean is faster clustering algorithm if we keep the value of K is small. If the cluster are scalable and global than k-means is tightly clustered. K-means provides result in group.

c. Limitation of K-means. There are few limitations in k-means like we must know about the value of in starting. K-mean can’t handle the garbage and noisy data which is present in datasets. K-mean did not give good result in non-global cluster. If the data and size of data is different, in this form k-mean did not perform well.

6. Comparison of all Algorithms. In this section, I am going to compare results of all three algorithms with the help of 400 log files that are used in specific server. Comparison of apriori, FP-growth and k-means is take place. Results are in the forms of graphs. In this comparison, results of recall, results of time taking by algorithm and results of accuracy and prediction is shown.
**Conclusion.** Data mining is a process of gathering data from different areas and analyze it in a form, so that it can be used for get useful result. Log files are patterns are used in web usage mining. Different kind of algorithms are implemented on patterns of log files to check which algorithm gives good result. Apriori, FP growth and k-means algorithm is used. Apriori and k-mean cannot be used for larger number of data sets. But FP-growth is more advanced algorithm and it provides faster results on large data. In the future algorithms for web content mining and structured mining will also be purposed. And in future, data and information can be contained from video and images.
REFERENCES


