Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
REVIEW ARTICLE
Available Online at www.ijarcs.info

Analysis of Web Usage Mining techniques to predict the user behavior from Web Server Log Files

Anmol Kaur
M.Tech Student, Department of Computer Engineering
Punjabi University, Patiala Punjab, India

Dr. Raman Maini
Professor, Department of Computer Engineering
Punjabi University, Patiala, Punjab, India

Avneet Singh Ahuja
B.Tech Student, Department of Computer Engineering
Punjabi University, Patiala, Punjab, India

***Abstract*:** Web usage mining can be defined as the approach to find the fascinating patterns from the web information in order to comprehend the web based implementation. The fundamental target of present research is to decide a few new web use mining approaches which are valuable for extracting the learning which is covered up in the web utilization logs. To be more particular the exploration work centers to mine the web server logs. The present study has been taken up with a view to study the different algorithms. It has been observed that Apriori algorithm is easy to implement and k-means clustering produces tighter clusters whereas FP-Growth pattern is faster than the Apriori algorithm. These algorithms are studied to find the interest of the students according to the sites that are frequently visited by the students.

.

Keywords: Web log file, frequency, clustering, pre processing.

1. INTRODUCTION

The World Wide Web is continuously rising day by day with the great speed both in the terms of traffic volume and the web site issues. It is mainly identified that the separation of the information from the web data is imperative to find the interest of the clients. Web mining helps in the process of the inferring the vital from the documents of the web, hyperlinks and from the web server logs with the use of data mining algorithms. Now a day, Web data mining is used by individual or by the organizations for the advancement of their profession. The trend is developing among organizations, associations and people alike to accumulate data through web information mining to use that data to their greatest advantage [14] [1].

The web mining additionally enables organizations in such a way that companies, banks etc can keep check on the fake installments. Data mining can be used to investigate this kind of information and data. This data can be helped to grow better progressed and defensive techniques which could be used in the prevention of all the fraudulent activities. There are introduction of new colossal advancements, techniques and more enhanced innovations are introducing in web data mining which helps to gather the relevant and resultant information. Web data mining innovation is not only concerned about simply assembling the information but rather it is likewise providing a great deal of concerns identified with the security of the crucial information. On the internet abundant of personal data is available and to secure it, web mining played a major role [14].

The three basic categories of web mining are:

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

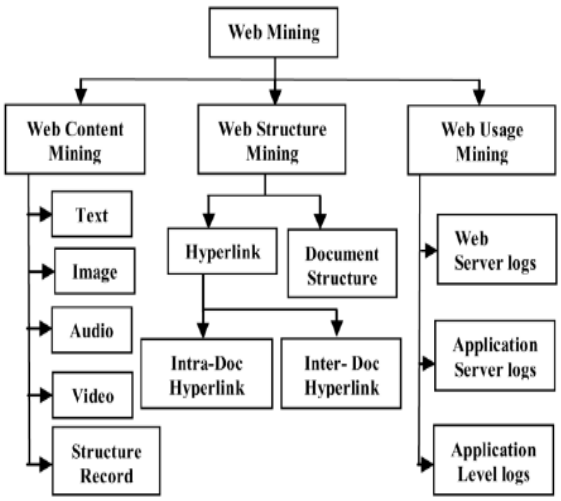

Fig 1: Web Data Mining Structure [10]

Web Content mining is utilized to investigate the information gathered via web crawlers. It is mainly concerned with the finding of important information or data from the unstructured; semi structured or structured content from the web documents. Unstructured documents contain images, text, audio and video. HTML documents are comprised by the semi structured documents and the structured documents include tables and lists [4].

Web Structure mining is utilized to taking in the information, associated with example of specific site and utilization mining are utilized to investigate information identified with a specific client programs are and also the

information to assemble the types of the clients. This data can be assembled by means of Web mining. A utilizing customary information mining defined techniques are the association, classification and clustering, and then examining of the successive patterns [4] [5].

Web Usage Mining is an information mining strategy that mines the data by breaking down the log records that contains the client get to designs. Web Usage Mining mines the optional information which is exhibit in log records and got from the associations of the clients with the web. Web use Mining systems are connected on the information display in web server logs, program logs, treats, client profiles, bookmarks, mouse clicks and so forth. This data is regularly assembled consequently get to web log through the Web server [3].

Predominantly there are four types of information sources present in which usage information is recorded at various levels they are:

a) ***Customer Level Collection:*** At this level information is assembled together by methods for java scripts or java applets. This information demonstrates the conduct of a solitary client on single site. Customer side information accumulation requires client interest for empowering java scripts or java applets. The upside of information gathering at customer side is that it can catch all snaps including squeezing of back or reload catch [2] [6].

b) ***Browser Level Collection****:* Second technique for information accumulation is by changing the program. It demonstrates the conduct of single client over numerous locales. The information accumulation abilities are upgraded by adjusting the source code of existing program [4]. They give a great deal more adaptable information as they consider the conduct of single client on numerous sites [2].

c) ***Proxy Level Collection:*** Proxy servers are utilized by web specialist organization to give World Wide Web access to clients. These server stores the conduct of various client at various site. These server capacities like store server and they can deliver reserved online visits. By analyzing the usage pattern of the guest, Web Usage Mining better the nature of web based business administration customizes the web [1] or, then again upgrades the execution of web structure and web server [6].

d) ***Server information*** that are gathered from web servers; it incorporates log documents, cookies and unequivocal client input. Servers contain distinctive sorts of logs, which are considered to be the primary date asset for web utilization mining. The most well known logs are:

- **Regular Log Format (CLF)**: made to monitor demands that happen on a site in sequential request. It contains the IP address of the customer, hostname, username, time stamp, record name and document estimate. CLF has the accompanying components [4].

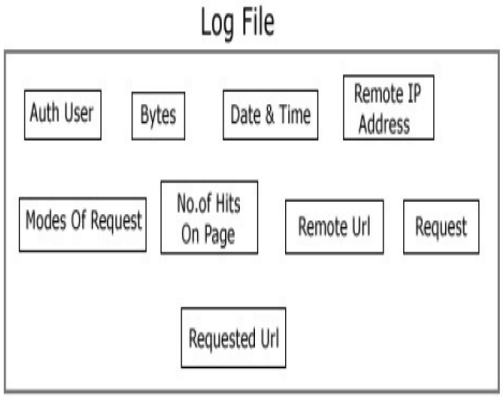

Fig:2 Practical outlines of web log documents [10]

- **IP address**: It is recognized as 32 bit number. IP address is interestingly recognized in the system. IP locations are twofold numbers; however they are generally put away in content records furthermore, showed in comprehensible documentations.
- **Remote log name:** This will return "- " unless client confirmation set on the web server.[4]
- **Confirmed client name:** Only accessible when getting to substance which is secret word secured by web server confirm framework.
- **Timestamp**: Entering and leaving date and time and time zone of the web server.
- **Access asks for**: Different strategies for demand like GET, POST, HEAD, PUT, DELETE and TRACE are utilized [4] [6].
- **Referrer URL**: This is the page address connects by which a guest is clicked to result in these present circumstances page. It is additionally conceivable that guest sort this connection address in the address bar. Some client specialists not supply this data dependably.
- **User Agent:** User specialist gives the data about client's program, working framework and program adaptation.
- **Protocol**: the protocol used [6].

II. WHY WEB USAGE MINING?

The essential objective of web utilization mining is to help individuals to use sound judgment to enhance organization execution and to keep up upper hand in the commercial center, i.e., it helps organizations to settle on the best choices rapidly and effortlessly. Web use mining is the appropriate method for removing data and building a valuable and proficient database about client practices. Likewise, it is vital in deciding powerful promoting techniques, i.e., those that expansion deals and place the organization's items on a more elevated amount [4].In this way, it can be easily discovered that the web usage mining has important utilizations for the analyzing the web client practices [7] [9]. The main focus of the paper is to learn the web usage mining procedures and the calculations utilized for the utilization mining and then using a few information mining strategies to predict the behavior of the student [8].

III. ALGORITHMS TO FIND THE INTERSET OF THE STUDENTS

a) K means Clustering Algorithm [15]

K-means Clustering algorithm comes under the category of unsupervised machine learning. It is mainly used when there is information of unlabeled type means which is not properly categorized. The objective of this calculation is to discover K clusters in the information. It is better to place the centers at some distant places from each other, because it gives different results at the different places. Now take the every point which belonged to the given set and try to relate it with that center which is nearer to it. If no point remains the left then calculate the new centroid. As a result number of iterations will be performed and stop when there is no choice to choose the new center. The defined function, commonly called as mean squared error function can be purposed by;[15][11]

$$J(v) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\| x_i - v_j \| \right)^2 \qquad 3.1$$

Where

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers.

Steps of algorithm:

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

1. Choose arbitrary centers of cluster.
2. Compute the gap between the every data point and center of the cluster.
3. Select the center of the cluster in such a way which is least from all the centers of cluster and allocate the data point to it [17].
4. Again calculate the center of newly formed cluster by:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i \qquad 3.2$$

5. Again calculate the newly formed centers of the clusters and distance of every data point.
6. Stop when there is no need of further reassignment of data points, if not then step 3 will be repeated [15].

b) Apriori Algorithm [4]

Apriori algorithm is defined as a step by step method. The n count of number of thing sets used to found n+1 thing sets. An arrangement of common sets, examining of database for the accumulation of the count of everything, and gathering those thing sets is fulfilling the base support. The resultant set is considered as the ordinary item set. Then, usual time set is utilized to discover coming fascinating thing sets, this procedure can be preceded until better item set is obtained. A last emphasis, you will wind up with numerous n-thing sets, this is fundamentally called association rules [4]. To find the repeated item sets it is consider that Apriori algorithm is useful.

Steps of Algorithm:

1. Define *Ak* as the set of the replicated item sets with size of k with the minimum support.
2. Also define *Ck* as the set of candidate item with the k size which is commonly defined as the item set of replicated items.
3. The items to be repeated are defined with the variable L1.
4. Perform the iteration with the for loop: for (*k* = 1; *Ak*!=; *k*++).
5. Choose *Ck*+1 = competitors created from the *Ak.*
6. The exchanges t which occurs in the database, then increment those *Ck*+1 candidate which are included in t.
7. Assign the *Ak*+1 to the applicants of the *Ck*+1 which having the least support.
8. Returnk *Ak* [4] [6].

c) FP-Growth algorithm [16]

FP –Growth algorithm is well planned and flexible technique to mine the pattern which occurs frequently .It can be performed by constructing the prefix tree to store the critical and compressed information about all patterns which occur frequently together called Frequent Pattern Tree. In terms of performance FP Growth calculation is better when contrasted with Apriori algorithm [13].

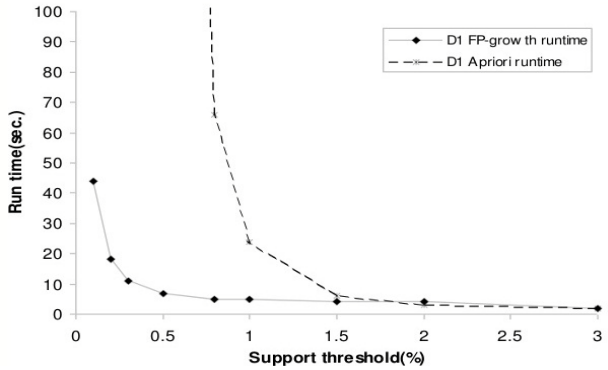

Fig: 3 Correlation of Apriori algorithm and FP – Growth algorithm [18]

Steps of Algorithm: FP-Growth permits the finding of repeated item sets without the generation of candidate item sets. It is a proposal of two steps.

Step 1- Generation of the dense structure which is called as the FP Tree.

Step 2- Pruning of the item sets which occur frequently from the FP Tree directly [16].

Construction of FP Tree requires the two passes to build over the data set which are:

Pass 1: Includes the following steps:

- Inspect the whole data and then notice the support for every item.
- Reject the items which are not occurring frequently.
- On the basis of their support, arrange all the frequent items in the decreasing manner.

Pass2: Includes the following steps:

Items have nodes and counters

- At a time, FP-Growth reads a single transaction and it is mapped to the path.
- The order is kept fixed to perform the overlapping of the paths when the same items are shared by the transactions. Incrimination in the counters is performed.
- Maintenance of pointers is done between the nodes which contain the similar items and single linked list is created.
- Finally pruning the frequent items from the FP tree [16].

IV.COMPARISON OF ALGORITHMS

Table I: Comparison of the algorithms [11]

Parameters	Apriori algorithm	Kmean Clustering algorithm	FP Growth algorithm
Methodology	Apriori is basically an algorithm for learning the rules of the Associatio n. It is intended to work on databases containing exchanges	Fundament al thought from the K-Means algorithm is to give the classificatio n of data based on its own data	FP Growth is designed on the divide and conquer scheme for the production of frequent item sets without the generation of candidate.
Time complexity	O(MN+ (1-R^M)/(1-R))	O(kn)	O(n*n).
Application used for	Used for market basket analysis	Used in search engines, Academics, Wireless sensor activities	Medical treatments, Market basket analysis, web page classification
Advantages	Can be easily implement ed	Faster than hierarchical clustering,	Compress data set, no generation of candidate
Disadvantage s	Need to scan the database many times	Prediction of k-value is difficult	Expensive to build, initial takes time to build.

It has been observed that breadth first search strategy and large item set property is used in Apriori algorithm. It is widely used up in Market basket analysis. It examines the whole database every time for the generated candidate item set. It is stored in array form. Its execution time is high as it consumes lot of time to scan the database for every item set. Whereas K-means clustering forms the tighter clusters and produce the different results on choosing the different initial partitions. FP-Growth algorithm is widely used in medical field. Divide and conquer ideology is used in it. There is no need of candidate generation. It gives the preferred outcomes over the Apriori algorithm calculation as far as execution. FP-Growth storage structure is tree. It examine the database two times whereas Apriori algorithm scan the database each time .That's why Apriori algorithm is somewhat time consuming. Another benefit of FP-Growth algorithm is that it does not requires calculating the pairs. This makes FP-Growth O(n) which makes it faster than the Apriori algorithm [11][12].

V. CONCLUSION

In this work, various web usage mining algorithms have been studied. It has been concluded that Apriori algorithm is easy to implement and consumes large memory space because of large candidate sets. K-means clustering is faster than the hierarchical clustering on keeping the cluster k size small. But it is difficult to predict the value of k. FP- Growth scanned the database twice. It is faster than the Apriori algorithm. Further new techniques can be used to predict the student behavior by analyzing the log files.

VI. REFERENCES

[1] V.Anitha and P.Isakki Devi, "A Survey on Predicting User Behavior Based on Web Server Log Files in a Web Usage Mining", 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), ppno:1-4,2016.

[2] G. Neelima and Sireesha Rodda, "Predicting user behavior through Sessions using the Web log mining", 2016 International Conference on Advances in Human Machine Interaction (HMI) ,ppno:1-5, 2016.

[3] Ashwini Ladekar, Dhanashree, Raikar, Pooja Pawar,"Web Log Based Analysis of User's Browsing Behavior", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 3 Issue 11**,**ppno: 3895-3899 2014.

[4] Anurag kumar, Vaishali Ahirwar, Ravi Kumar Singh, "A Study on Prediction of User Behavior Based on Web Server Log Files in Web Usage Mining", International Journal Of Engineering And Computer Science, Volume 6 Issue 2**,**ppno: 20233-2026,2017.

[5] ZAKARIA SULIZAKARIA ZUBI and SULIMAN MUSSAB SALEH EL RIANI, "Applying Web Mining Application for User Behavior Understanding", 1st WSEAS International Conference on Acoustics, Speech and Audio Processing (ASAP) ppno:217-224,2013.

[6] Amit Pratap Singh,Dr. R. C. Jain, "A Survey on Different Phases of Web Usage Mining for Anomaly User Behavior Investigation", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 3, May –June 2014.

[7] Virendra R. Rathod and GOVIND V.PATEL, "Prediction of User Behavior using Web log in Web Usage Mining", International Journal of Computer Applications (0975 – 8887), Volume 139 Issue No.8, ppno: 4-7 April 2016.

[8] Deepti Kapila, Prof. Charanjit Singh, "Survey on Page Ranking Algorithms for Digital Libraries", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6,ppno: 1263-1268June 2014.

[9] Ananthi.J," A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites", International Journal of Computer Science and Information Technologies, Volume 5, Issue 3, ppno: 4091-4094, 2014.

[10] Ashwini Ladekar, Dhanashree Raikar and Pooja Pawar, "Web Log Based Analysis of User's Browsing Behavior", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)Volume 3 Issue 11,ppno: 3895-3899, 2014..

[11] Dr.C.Kumar Charliepau and G.Immanual Gnanadurai, , "Comparison of K-mean algorithm & Apriori algorithm-An Analysis", International Journal On Engineering Technology and Sciences – IJETS volume 1 Issue 3,ppno: 2349-3968,2014.

[12] Shamila Nasreen atl, "Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A

Survey", The 5th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN),ppno : 109 – 116,2014.

[13] A. Vishwakarma, and K.N. Singh, "A survey on web log mining pattern discovery", IJCSIT – International Journal of Computer Science and Information Technologies Volume 5 Issue 6, pp: 7022-7031, 2014.

[14] http://www.web-datamining.net/

[15] https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

[16] https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=7&ved=0ahUKEwicmbzo9NXTAhWBuY8KHR7UDvcQFghOMAY

[17] http://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_425

[18] https://www.google.co.in/search?q=fp+growth+algorithm&source=lnms&tbm=isch&sa=X&ved=0ahUKEwicmbzo9NXTAhWBuY8KHR7UDvcQ_AUICCgD&biw=1366&bih=657#tbm=isch&q=comparison+and+fp+growth+algorithm