Analysis of Web Usage Mining Based on Web Log Partition

B.Harika[1], Prof .Thatimakula Sudha[2], Rajeswari Sana[3]
[1]*Research Scholar, Dept. of Computer Science, SPMVV, Tirupati*
[2]*HOD & Dept. of Computer Science, SPMVV, Tirupati*
[3] *Department of computer science and engineering, RGUKT, RK Valley, IIIT Idupulapaya, Vempalli, Andhrapradesh*

***Abstract*- The Internet has accrued a great deal of consumer interest. The main applications of web mining are to maintain good ties with customers. Web usage mining is the method of extracting useful knowledge from web logs. This helps to boost the efficiency of the website by evaluating the interest of the customer, and also increases business profitability. Web Use Mining's main aim is to research the navigation habits of the users and their use of web tools. Mining of site use is used to monitor user behaviour. Such documents are also used for data extraction that helps optimize the search engine. We suggest an approach in this research work, in which web logs are used in cluster types. Such clusters are built in Web logs according to records of user behaviour. Therefore, if we search from these clusters instead of the full web log, the search time will be reduced.**

***Index terms*- Web Mining, Web usage, Mining, Clustering, Algorithm**

1.INTRODUCTION

Web Mining is defined as the field of application for data mining which deals with the extraction of useful and interesting information from the World Wide Web. It is used to solve various problems, such as finding relevant information, creating web-based knowledge, learning about customers or individual users, personalizing knowledge etc. Data mining uses various data mining techniques to automatically discover the site and to collect information from the data documents[1]. It is largely decomposed into four subtasks: finding resources, selecting and pre-processing information, generalizing and discovering patterns, and analyzing patterns. Web Mining uses many data mining techniques to collect useful knowledge from the internet. Nevertheless, along with data mining techniques it is also possible to incorporate various other techniques such as artificial intelligence, information recovery, natural language processing, information analysis, machine learning. Comparison with Data Mining, with all such approaches. Web Mining is most often concerned with information collection or information processing. Information retrieval works by indexing text and searching for useful document. It finally retrieved all relevant records, and even some irrelevant[2].

Data mining is the integration of data gathered by traditional methodologies and data mining techniques with knowledge gathered through the World Wide Web. It is used to understand customer behavior, to boost the quality of the website and make it competitive for businesses. Figure 1 demonstrates the possibility of splitting web mining into three forms of web mining, web content mining, and web structure mining. a. Web usage mining: Web log data is collected in this phase, and interesting patterns are discovered and analyzed, b. Web content mining: Text and images are mined in this phase, c. Web structure mining. Website structure is mined on the basis of intra-join and hyperlinks[3]. The layout of the website is mined using intra-links and hyperlinks[3]. The aim of this paper is to concentrate on the overview of web use mining and log file information to study user behavior and improve the efficiency of websites. The data is collected from the web servers, the duty of these web servers is to maintain the online access log for all websites run under the online server[4]. Such data is cleaned and treated in a specific format, and can be used to collect various types of information with different schemes.

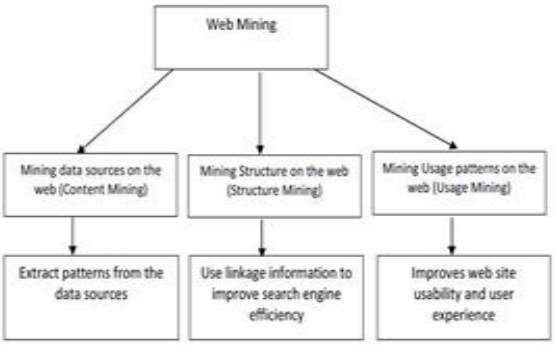

Figure 1: Web mining and its types

2. LITERATURE REVIEW

In [5], estimating the route examples of a individual must scramble for anticipated abuse and describe details from the blog website. The primary fragment of this approach focuses on the separation of people in blog webpage material, just as in the second territory bunch procedure attempts to accumulate clients with similar inclinations and furthermore in the third section the grouping and clutter results are typically conjectured by the people following requests. Work[6] defined sequential availability from blog posts, (Algorithmic Guide Line) GA rule referred to as ALMG. GA supported characteristic technique procedure for design extraction in their work, was used to finding best clarifications for extraordinary disservice over time to find sequential access from understanding blog website. Kim and Zhang use the GA rule to say the significant highlights of HTML labels that are typically re-ranked archives for web structure mining recouped by well-known weight frameworks. The work[7] blesses an inherited technique of searching for a critical one. The work[8] suggested a hymenopterous creepy crawly agglomeration algorithmic rule for identifying Web user designs (data assortments) and a simple inherited programming technique for analyzing the patterns of the explorer[9]. A few frameworks have placed in place web mining for programmed personalization[10],[11] just as [12] usually brings 2 vital methodologies with it: disconnected mining and on-line proposal inside disconnected mining technique, all availability errands of clients in website are taped into log archives by web server. From that point on, scarcely any net mining techniques have applied Web server logs within the online reference strategy to extricate veiled route variants of customers; buyer requests from his current lively meeting have been registered.

The work[13] has created an effort to integrate the use of a website and its content capabilities into the Internet personalization web mining process. To promote flexibility over time, a "postmining" technique was introduced to achieve the same picture for the use of that website as well as the website user accounts. Nonetheless, the strategies envisaged in[14] were restricted to using profiles to independently construct website use and web content profiles. Authors in [15] proposed that for better user navigation, web link forecasting and course analysis also mattered. He suggests a Markov chain method for calculating individual gain access by individual access series to previously collected logs. Work[16] introduces external forward recommendations for victimization fertilization so as to disrupt customer sessions in traversal pattern mining transactions. The actual forward link will be the last page the customer has requested before backtracking occurs, where a web page was previously used as the individual demands during the specific customer session. Research[17] finds that — Weblog is the way assessments of IT managers to ensure adequate data transfer as well as availability of web servers on client websites. In the past 5 years, log record evaluation has progressively progressed with companies; mining files currently contain fine-grained details about site user profiles as well as purchasing behavior. The companies are now looking to use log files to explore the functionality of the website. Data on record logging can provide valuable insight into the usage of websites. It replicates actual use associated with synthetic setups in the all-natural operating environment of a design center. It stands for role of several customers, each for an hour or more over a potentially long period of time being checked out to restricted variety of individuals.

The work[18] has found that assessment of website functionality is a time-consuming physically accomplished task. It offers a platform supporting remote functionality analyzes of websites. It supports information from the client side about individual interactions and JavaScript events. Additionally, the intent of offering custom occasions explores the versatility to include different activities to be observed and even considered for study. By using a

proxy-based template, the tool supports the website and helps the critics execute different individual actions & optimal action sequences. The work[19] suggested that cognitive modeling could test the functionality of intricate vibrant interfaces with the human computer system for its underlying structures. While human attitude prediction can analyze the investigative errors of the contact system, as well as possible cognitive needs. In rough set theory, the author [20] suggests an indiscernibility method for deriving information from extended web logs to establish the origin of visitors and the keywords used to access a website that will lead to improved website design and optimization of search engines. The author [21] has done research on web-use mining pre-processing of the data. They suggested a new Algorithm for User and Session Recognition called – USIA. It may provide user information and session acknowledgment. The same user is marked using an IP address and User ID. Where the request originates from the same IP address, the algorithm assumes the request originates from the same user. The session shall be determined by the time in and time out. This research work has concentrated primarily on user awareness for the particular session and series of web pages accessed by the user. This author [22] used the approach of clustering to concentrate on grouping the transactions of the clients. There are some similarities in the set of transactions within a group, so we can quickly identify the customer's behavior and the website analyst can understand the customer's needs and make the website friendly. To make the website more personalized and user friendly, from a different viewpoint. The researcher used the pattern-based clustering approach to group related sort of transactions.

The author[23] dealt with two types of groups, one being Web Clustering Groups which grouped similar pages from web server log files, the other being User Clustering Groups through which the user related to the same category of web pages. Divisive Hierarchical Clustering Algorithm is employed to group Web Log files and related users. To fine-tune the relationship between them, the association rule of mining with measure of support and trust is then applied to each party.

3. PROPOSED ARCHITECTURE

This research work provides for a cluster formulation approach to site mining use. Cluster-based Web log search results are contrasted with full Web log-based scanning results, i.e. web log filtering of documents [1].

Complete web log searching algorithm

1. Read Input String (Si) as Keyword (Ki)
2. If (Si=NULL) Terminate / Halt. Re-Project Search Options
3. If (Si<>NULL) Establish Database Connection (DBCN)
4. If (ReturnType (DBCN)=NULL) DatabaseEngine Failed
5. If (DBCN)=Ri; {Ri=RecordSet)
6. Fetch / Retrieve RelatedRecord (RRi) from RLN(Relation).DBCN
7. Print RRi=>DataSetItem(i)
8. Move RecordLog (RLi) to ServerRepository (SR)
9. Compatability Check(CC) (Browser/Plugin/Add-On) => (True/False)
10. If (CC<>NULL) Print Results on WebClient(WC) {WC : Firefox/Chrome/IE/Safari/Opera}
11. Terminate with Success

Requirement for using clustering approach

If a user searches for any content or information according to a particular keyword, a lot of results are received by the user, some of these results are useful for the user and most of them are not related. Normal activity of a consumer is always to search out some of the top ranking results and ignore others. The same is true when searching for a complete web log, the results available will be more in numbers and the search time will be longer. Clustering approach group the results available in the web log according to their ranking or popularity (i.e. number of user-specific visits to that web page), then first search inside the top order cluster and if not enough results are not found, then it will go for the next one. In this way the search time for most searches will be much less than the full web log search, and the user will have access to the most common results.

Web log generation

1. The server generates a web log based on the user activity or user access in the case of mining for web use.
2. The created web log is used to extract data as per the popularity of the site.
3. Full web log is usually searched to retrieve data or links but fragments or partitions are searched in the proposed web log method.

Steps used in proposed approach

1. If a user visits a web page, this is reported in the site log as well as in a report on rank relevance. Rank Relevancy Report consists of documenting all the web pages referenced along with their rating.
2. The rank of the web page is determined on the basis of specific users referring the web page (i.e. when any user visits a particular web page, the web page's visit count is increased).
3. Web page rank is used to pick a specific web page within a particular cluster.

Algorithm for the proposed approach

Consider n number of web pages is there

Phase 1: Relevancy rank report generation

1. for i=1,i<=n, i++ visit[i]=0
2. if(kth web page is visited by some user) then visit[k]=visit[k]+1;
3. for i=1;i<=n;i++ Sort (visit[i])
4. or i=1;i<=n;i++ Rank i = i

Phase 2: cluster formation

Consider c number of clusters and p number of pages is present in each cluster

5. p = n / c
6. for i=1; i<=c;i++ for (j=(p*(i-1)))+1;j<=p*i;j++) ,(Cluster[i], rank[j]) (i.e cluster[i] consist of p number of pages according to rank)
7. if (kth page is visited by the user) goto step II

Flowchart for proposed approach

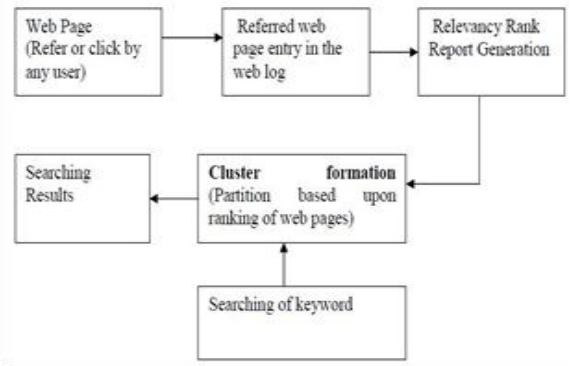

Figure 2: Flowchart for proposed approach

Searching time calculation

The search time was determined for displaying the results on the basis of question or keyword entered by the user and until the results are obtained from the created web log i.e. database.

4. RESULTS AND DISCUSSIONS

A website of 15 web pages is generated in php for displaying the effects of the proposed approach. A Search Engine is built to obtain user feedback or keyword (Figure 3). A relevancy rank report is created based on the comparison of these web pages consisting of a web page rank based on the number of times the user visits the web page (figure 4). Figure 5 shows the analysis of current methods and proposed methods.

Figure. 3 : Search engine for user input or keyword

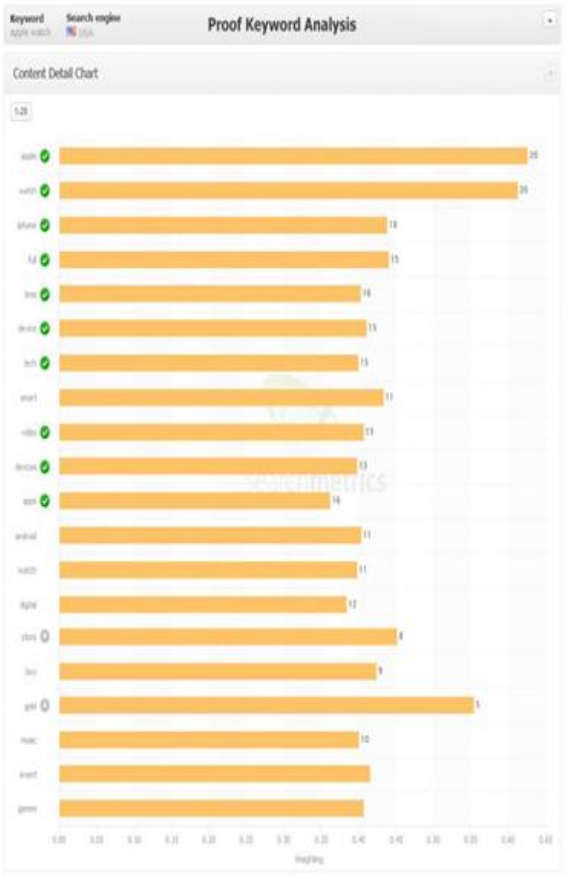

Figure 4: Relevancy rank report

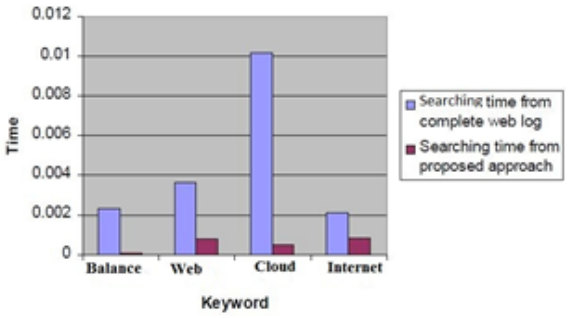

Figure 5: Comparison of complete web log searching and cluster searching time(ms)

Depending on the rank relevance report, more beautiful results will be obtained such as testing the popularity of a web page or web site in a particular time frame that will provide current common data and clusters will be created on that basis.

5 CONCLUSIONS AND FUTURE SCOPE

This research work proposes a network usage mining method focused on partitioned network logs. It takes less time and is generating common results in line with the current approach. Any further results can be obtained if the number of clusters formed is modified i.e. from 4 clusters formed in our approach to 6, 8 or more can be modified. Recall and accuracy, however, can be influenced by increasing the number of clusters, i.e. either improving or decayed.

REFERENCES

[1] Gupta and A. Khandekar, "Development of Weblog Mining Based on Improved Fuzzy C-Means Clustering Algorithm", International Journal of Science, Engineering and Technology Research, Vol.5 (3), pp.688-693, March 2016.

[2] Kapoor and A. Singhal, "A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms", In Proc. of 3rd International Conference on Computational Intelligence & Communication Technology, IEEE, pp. 1-6, 2017.

[3] A. Zahid, A. V. Babuy, W Ahmed and M F Azeemz, "A fuzzy set theoretic approach to discover user sessions from web navigational data", In Proc. of Recent Advances in Intelligent Computational Systems, IEEE, pp. 879-884, 2011.

[4] B. Chandra, M. Gupta, and M.P. Gupta, "A multivariate time series clustering approach for crime trends prediction", In Proc of International Conference on Systems, Man and Cybernetics, IEEE, pp. 892-896, 2008.

[5] B. Maheswari and P. Sumathi, "A New Clustering and Preprocessing for weblog mining" In Proc. of World Congress on Computing and Communication Technologies, IEEE, pp. 25-29, 2014.

[6] B. S. Shedthi, Shetty and M. Siddappa, "Implementation and comparison of K-means and fuzzy C-means algorithms for agricultural data", In Proc. of International Conference on Inventive Communication and Computational Technologies, IEEE, pp. 105-108,2017.

[7] C. Baviskar and S. Patil, "Improvement of data object's membership by using Fuzzy K-Means clustering approach", In Proc. of International Conference on In Computation of Power, Energy Information and Communication, IEEE, pp. 139-147, 2016.

[8] C. T. Baviskar and S. S. Patil, "Improvement of data object's membership by using Fuzzy K-Means clustering approach", In Proc. of International Conference on Computation of Power, Energy Information and Communication (ICCPEIC)IEEE, pp. 139-147, 2016.

[9] C. Yanyun, Q. Jianlin, G. Xiang, C. Jianping, J. Dan and C. Li, "Advances in research of Fuzzy c-means clustering algorithm", In Proc. of International Conference on Network Computing and Information Security, IEEE, vol. 2, pp. 28-31, 2011.

[10] Chen, Y.L. and Huang, C.K., “Discovering fuzzy time-interval sequential patterns in sequence databases”, IEEE Transactions on Systems, Man and Cybernetics, Vol. 35(5), pp. 959-972, 2005.

[11] D. Koutsoukos, G. Alexandridis, G. Siolas, and A. Stafylopatis, "A new approach to session identification by applying fuzzy c-means clustering on weblogs", In Proc. of Symposium Series on Computational Intelligence, IEEE, pp. 1-8, 2016.

[12] G. S. Chandel, K. Patidar and M. S. Mali, "A Result Evolution Approach for Web usage mining using Fuzzy C-Mean Clustering Algorithm", In Proc. of International Journal of

Computer Science and Network Security, Vol.16(1). pp.135-140, 2016.

[13] H. Gulat, and P. K. Singh, "Clustering techniques in data mining: A comparison", In Proc. of 2nd International Conference on Computing for Sustainable Global Development, IEEE, pp.410-4 14. 15, 2015.

[14] H. X. Pei, Z. R. Zheng, C. Wang, C. Li, and Y. H. Shao, "D-FCM: Density based fuzzy c-means clustering algorithm with application in medical image segmentation", Procedia Computer Science, Vol.122(1), pp. 407-414, 2017.

[15] K. Suresh, R. M. Mohana, A. Rama Mohan Reddy, and A. Subramanyam, "Improved FCM algorithm for clustering on web usage mining." In Proc. of International Conference on Computer and Management, pp. 1-4. 2011.

[16] P. Sampath and M. Prabhavathy, "Web Page Access Prediction Using Fuzzy Clustering by Local Approximation Memberships (Flame) Algorithm", Vol.10 (7), pp.3217-3220, 2006.

[17] S. K. Dwivedi and B. Rawat, "A review paper on data preprocessing: A critical phase in web usage mining process", In Proc. of International Conference on Green Computing and Internet of Things, IEEE, pp. 506-510, 2015.

[18] V. Anitha and P. Isakki, "A survey on predicting user behavior based on web server log files in a web usage mining", In Proc. of International Conference on Computing Technologies and Intelligent Data Engineering, IEEE, pp. 1-4, 2016.

[19] Ghada M. Tolan and Omar S. Soliman.: An Experimental Study of Classification Algorithms for Terrorism Prediction. Overall Journal of Knowledge Engineering, Vol. 1, pp. 107-112. 2015.

[20] A. Malathi and Dr. S. SanthoshBaboo.: Evolving Data Mining Algorithms on the Prevailing Crime Trend – An Intelligent Crime Prediction Model. Worldwide Journal of Scientific and Engineering Research, June, Vol. 2, 2011.

[21] A. Sachan and D. Roy.: TGPM: Terrorist cluster estimate model of counter dread based persecution. Overall Journal of Computer Applications, vol. 44, no. 10, 2012.

[22] B. Thuraisingham.: Data mining, national security, assurance and basic opportunities. SIGKDD Explorations, January 2003.

[23] Kumar., V., Zinovyev., R., Verma., Tiwari., P.: Performance Evaluation of Lazy And Decision Tree Classifier: A Data Mining Approach for Global Celebrity's Death Analysis. IEEE Xplore: In International Conference on Research in Intelligent and Computing in Engineering (RICE), pp 1-6, 2018.