IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, S.Iss 6, 2022

ANALYSIS OF WEB USAGE MINING BASED ON WEB LOG PARTITION

[1]Abhay Shukla, [2]Shail Dubey, [3]Rituraj Kushwaha, [4]Ashish Shukla, [5]Pooja Dwivedi

Department of Computer Applications, Axis Institute of Higher Education, Kanpur, Uttar Pra desh, India

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 involves applying data mining techniques to extract valuable and interesting information from the World Wide Web. This process addresses issues such as identifying relevant information, developing web-based knowledge, understanding customer or user behavior, and personalizing information. It utilizes various data mining techniques to automatically locate and gather data from web documents. Web mining is divided into four main tasks: resource discovery, information selection and preprocessing, pattern generalization and discovery, and pattern analysis. It incorporates not only data mining techniques but also artificial intelligence, information retrieval, natural language processing, information analysis, and machine learning to extract useful knowledge from the internet.

Web mining, compared to traditional data mining, often focuses on information collection and processing. Information retrieval involves indexing and searching text to find relevant documents, though it may also retrieve some irrelevant records. Data mining integrates data collected through traditional methods with knowledge obtained from the Web to understand customer behavior, improve website quality, and enhance business competitiveness.

Web mining can be categorized into three forms: web usage mining, web content mining, and web structure mining. Web usage mining involves collecting and analyzing web log data to discover patterns. Web content mining focuses on extracting text and images, while web structure mining examines website structure based on internal and external links. This paper aims to provide an overview of web usage mining and the analysis of log files to study user behavior and improve website efficiency. Data for this analysis is collected from web servers, which maintain access logs for all hosted websites. This data is cleaned and formatted appropriately for various types of information extraction.

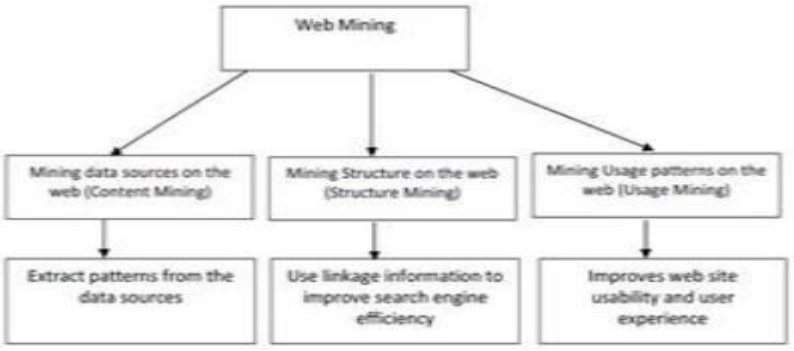


Figure 1: Web mining and its types

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

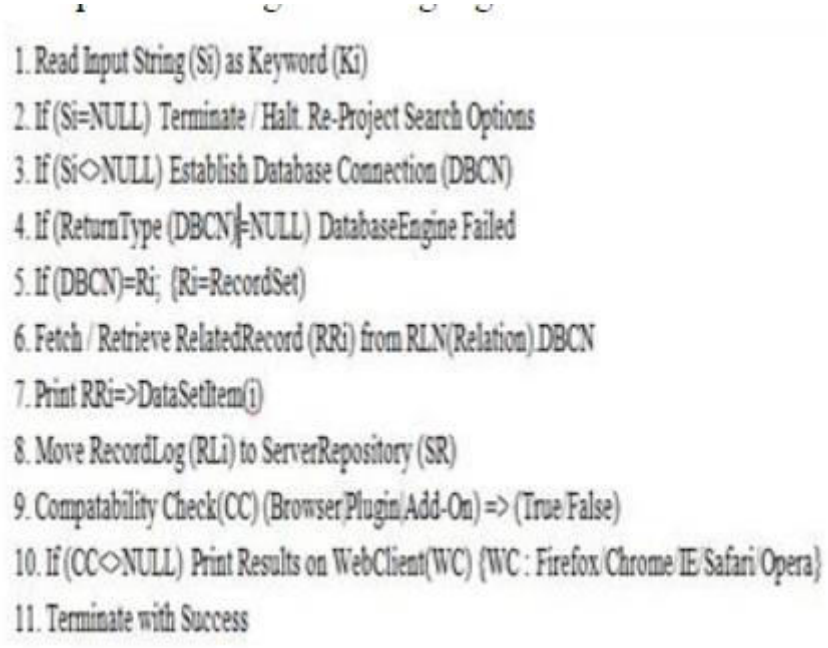
used as the individual demands during the specific customer session. Research[17] finds that — Weblog is the way assessments of IT managers to ensure adequate data transfer as well as availability of web servers on client websites. In the past 5 years, log record evaluation has progressively progressed 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. Human attitude prediction can analyze investigative errors in the contact system and identify potential cognitive needs. In the context of rough set theory, one author proposes an indiscernibility method to extract information from extended web logs, determining visitor origins and the keywords used to access a website. This approach aims to enhance website design and optimize search engines.

Another study focuses on preprocessing data for web-use mining, introducing an algorithm called USIA (User and Session Recognition Algorithm). This algorithm identifies users and sessions using IP addresses and User IDs, assuming requests from the same IP address originate from the same user. Sessions are determined based on login and logout times. The research highlights user awareness within specific sessions and sequences of web pages accessed by users. Clustering techniques are used in another study to group client transactions, allowing for the quick identification of customer behavior. This helps website analysts understand customer needs and make websites more user-friendly and personalized. Pattern-based clustering is employed to group related transactions.

Additionally, two types of clustering groups are discussed: Web Clustering Groups, which cluster similar pages from web server log files, and User Clustering Groups, which categorize users based on similar web pages they visit. The Divisive Hierarchical Clustering Algorithm is used to group web log files and associated users. To refine the relationship between these groups, association rule mining with support and confidence measures is applied.

Proposed Architecture: This research work introduces a cluster formulation approach for site mining. It compares cluster-based web log search results with full web log-based scanning results, specifically focusing on web log document filtering.Complete web log searching algorithm



Requirement for Using the Clustering Approach

When users search for information using specific keywords, they often receive numerous results, many of which are irrelevant. Typically, users focus on the top-ranking results and disregard the rest. This behavior also applies to searching a complete web log, where a large number of results can increase search time. The clustering approach organizes the web log results based on their ranking or popularity (determined by the number of user visits to a webpage). Initially, the search is conducted within the top-ranked cluster, and if sufficient results are not found, the search proceeds to the next cluster. This method significantly reduces search time by providing users with the most relevant results first.

Web Log Generation

The server generates a web log based on user activities or access patterns for web usage mining.

The generated web log is then used to extract data based on the site's popularity.

While a full web log is typically searched to retrieve data or links, the proposed method searches fragments or partitions of the web log.

Steps Used in the Proposed Approach

1. When a user visits a web page, this activity is logged in the site log and recorded in a rank relevancy report. This report documents all referenced web pages along with their ratings.
2. The rank of a web page is determined by the number of users visiting that page; each visit increases the page's visit count.
3. The web page rank is then used to select specific web pages within a particular cluster.

Algorithm for the Proposed Approach

Consider there are nnn web pages.

Phase 1: Relevancy Rank Report Generation

1. Initialize visit counts: for i=1 to n, set visit[i]=0\text{for } i = 1 \text{ to } n, \text{ set } \text{visit}[i] = 0for i=1 to n, set visit[i]=0

2. If the kkk-th web page is visited by a user, increment the visit count: visit[k]=visit[k]+1\text{visit}[k] = \text{visit}[k] + 1visit[k]=visit[k]+1
3. Sort the web pages by visit counts: for i=1 to n, sort visit[i]\text{for } i = 1 \text{ to } n, \text{ sort } \text{visit}[i]for i=1 to n, sort visit[i]
4. Assign ranks based on sorted visit counts: for i=1 to n, rank[i]=i\text{for } i = 1 \text{ to } n, \text{ rank}[i] = ifor i=1 to n, rank[i]=i

Phase 2: Cluster Formation Assume there are ccc clusters and each cluster contains ppp pages. 5. Calculate the number of pages per cluster: p=ncp = \frac{n}{c}p=cn 6. Distribute pages into clusters based on rank:
for i=1 to c, for j=(p×(i−1))+1 to p×i, assign (Cluster[i], rank[j])\text{for } i = 1 \text{ to } c, \text{ for } j = (p \times (i - 1)) + 1 \text{ to } p \times i, \text{ assign } (\text{Cluster}[i], \text{ rank}[j])for i=1 to c, for j=(p×(i−1))+1 to p×i, assign (Cluster[i], rank[j]) 7. If a user visits the kkk-th page, proceed to step II.

Flowchart for the Proposed Approach

1. User visits a web page.
2. The visit is logged and recorded in the rank relevancy report.
3. Web page ranks are updated based on visit counts.
4. Web pages are sorted and clustered according to their ranks.
5. Pages within clusters are accessed based on user visits.

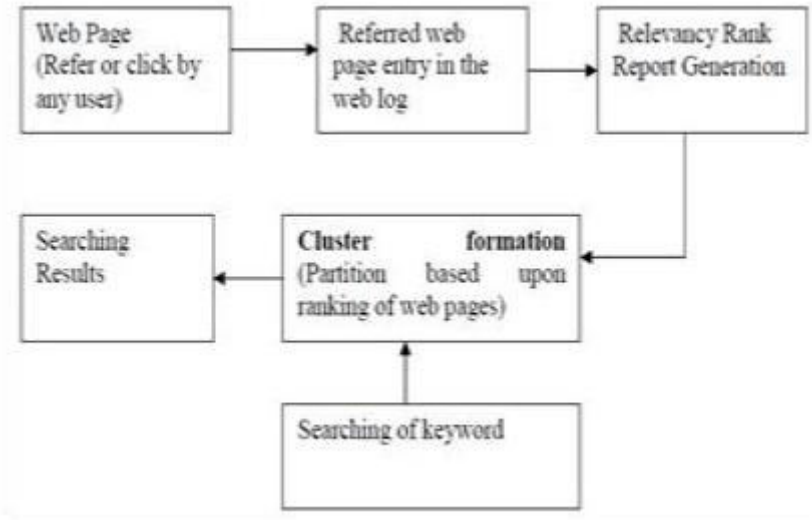


Figure 2: Flowchart for proposed approach

Searching Time Calculation

The time taken to display results was measured based on the user's query or keyword entry and the retrieval of results from the web log database.

RESULTS AND DISCUSSIONS

A 15-page website was developed using PHP to demonstrate the effectiveness of the proposed approach. A search engine was created to collect user input or keywords (Figure 3). A relevancy rank report was generated by comparing these web pages, ranking them based on the frequency of user visits (Figure 4). Figure 5 illustrates the analysis comparing current methods with the proposed method.

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, S.Iss 6, 2022

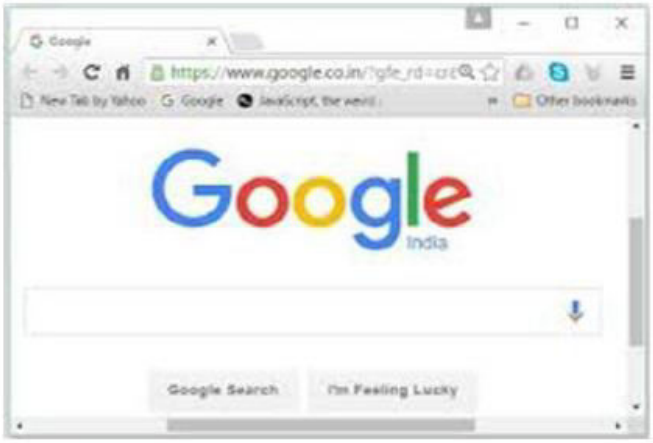


Figure. 3 : Search engine for user input or keyword

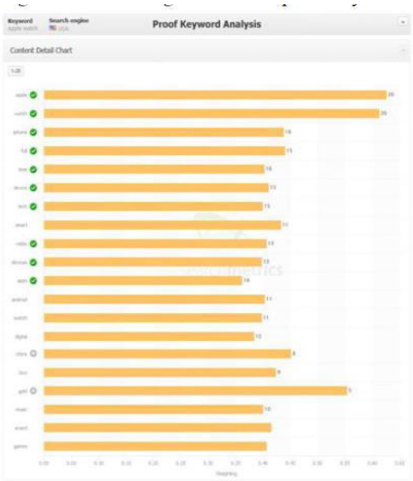


Figure 4: Relevancy rank report

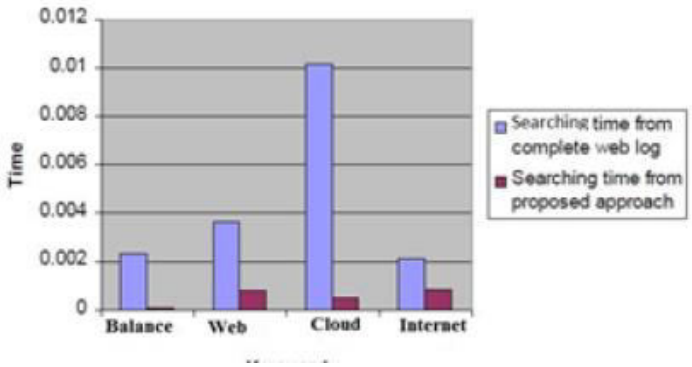


Figure 5: Comparison of complete web log searching and cluster searching time(ms)

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