iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 3 **Issue:** VI **Month of publication:** June 2015

**DOI:**

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

# An Efficient Approach for Extracting Usage Pattern from Web Log Using Web Usage Mining

Nirali H. Panchal[1], Ms. Ompriya kale[2]
[1]*M.E in Computer Science and Engineering, L.J.Institute of Engineering and Technology*
[2]*Assistant professor in M.E Computer Science and Technology, L.J.Institute of Engineering and Technology*

***Abstract— In today's world most of the tasks are done by Internet. User spends number of hours over the Internet. So by knowing the behaviour of user to access the Internet we can get lots of information about interest of user. Web usage mining is the process to identify the user's behaviour in terms of access the Internet. It is the sub part of web mining. It focuses on the techniques that discover the usage pattern from web log. It uses the data mining techniques like association rule analysis, clustering, classification and machine learning. Before applying these techniques to web log some data pre-processing steps are needed for prepare the data. After that pattern discovery techniques are applied to find some interesting pattern. Clustering is one of the data mining techniques to make a group of user or user session having similar characteristics. Data pre-processing are perform using java and sql. After that proposed a new approach for web usage mining is hybridize a Black-Hole with K-means algorithm for clustering of web log data. Black-Hole approach is inspired by nature and relay on concept of black-hole phenomena. From analysis it is proved that Black-hole with KMeans algorithm is an efficient approach for clustering.***
***Keywords—Black-hole, Clustering, Data pre-processing, KMeans algorithm, Web usage mining***

## I. INTRODUCTION

Data mining is the process to discover some unknown patterns or knowledge from data. Web mining is part of data mining and Web usage mining is part of web mining. Web usage mining includes usage characteristics of the user of web on the internet. Data mining techniques are applied to web data to discover web usage pattern. Analyzing such data can help organizations or web site admin to determine the interest level of customer, making some marketing strategies, personalization of web site and find more user friendly logical structure for their web sites. As we know internet is big source of information. The web users while performing their actions also leave back their records of their work. Lots of juicy knowledge can be obtained from this huge amount of global data. Various advance data mining procedures are essential for the knowledge to be obtained, understood and benefit from it. Web usage mining processes are particularly framed to carry out the task of analyzing the data that represents the usage data.

## II. BACKGROUND THEORY

Web mining refers to the use of data mining techniques to automatically retrieve, extract and analyze information for knowledge discovery from web documents and services [1]. Web mining contains three sub parts: i) Web structure mining. ii) Web content mining. iii) Web usage mining. Web structure mining is the process of extract knowledge from the World Wide Web organization and links between references and referents in the web. Normal web graph contains web pages as nodes and hyperlinks as edges connecting associated pages. Web structure mining is the process of analyze the node and connection structure of a web site using graph theory [11]. Web content mining is the one that used to find out useful knowledge from the page content. Generally, the numbers of types of data are included in web content, for example, image, hyperlinks, textual, video, audio and metadata [11]. Web usage mining is the process of the discovery and analysis of patterns in click stream associated and related data that are collected and generated as a result of user relations from one or more web sites [11].

*A. Types of Data Source*

Server-side log is the major data source in web usage mining. There are some additional data sources like client-side log and proxy-side log are also use for some user and some application.

*B. Data Pre-processing*

Data preprocessing is the first phase of web usage mining in as shown in figure 2. It is the most important task of web usage mining. Data that are collected from data source are generally semi structured or unstructured. These logs contain redundancy and some unwanted data. So data preprocessing are needed for clean the data. It is main used for reduce the size of data by

removing redundancy and irrelevant data. Data preprocessing contains some technique like: User identification, Data cleaning, Session identification, Path Completion and Pageview Identification.

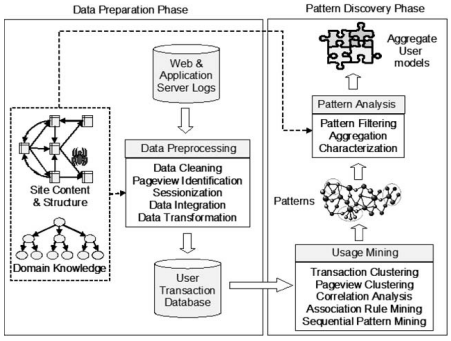



Fig 1. Web Usage Mining Architecture[21]

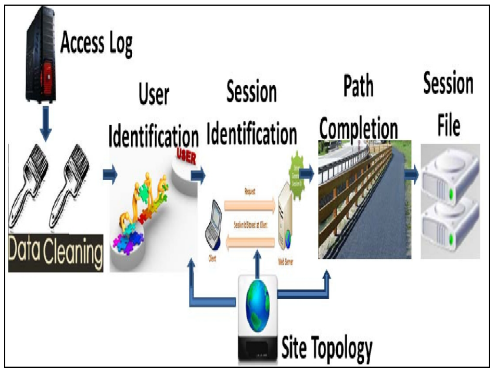



Fig 2. Various Phase of Data preprocessing [2]

*C. Pattern Discovery*

Pattern discovery is the next step after completing data preprocessing step. In this phase using data mining techniques like association rule analysis, clustering, classification and prediction discover some usage pattern.

*D. Pattern Analysis*

In the overall web usage mining process pattern analysis comes at last. The inspiration behind pattern analysis is to filter the unnecessary patterns and rules from the set of pattern generated at the pattern discovery phase. Knowledge query mechanism like SQL is the most general type of pattern analysis [11]. Another way is to enter usage data into a cube of data in order to perform various OLAP operations like roll-up, drill-down etc. Some visualization techniques, for example, graphing or chart of patterns or assignment of colors to various values for highlighting general patterns or trends. Content and structure knowledge can also be used to filter the patterns contains page of certain content type and usage, or pages that equivalent to some hyperlink structure.

## III.RELATED WORK

Xidong Wang, Yiming Ouyang ,Xuegang Hu and Yan Zhang[3] has proposed a new algorithm to discover frequent access pattern from web browsing behavior of user. Algorithm name was “FAP-Mining” algorithm. It has two steps. One was Construction of FAP-Tree and second was FAP-Growth. Lin Feng, Baohua Guan[4] has proposed a new navigation approach known as “Web Usage Mining based on Variable Precision Rough Set Model” for web user browsing a website. First, Log data sets are reduced with attribute reduction module by rough set. After that, a reduced Log data set is trained to create a rough classifier. K.Suresh, R.MadanaMohana, A.RamaMohanReddy, A.Subrmanyam[5] has proposed a algorithm to find user group with similar characteristics using “Improved FCM algorithm”. Anna Alphy, S. Prabakaran [6] has proposed cluster optimization technique to ART1 Neural network. Nayana Mariya Varghese, Jomina John [7] has provided the cluster optimization technique using the algorithm “Fuzzy Cluster-chase algorithm for cluster optimization”. V. Diviya Prabha, R. Rathipriya[8] has introduced an algorithm “Gravitational Search Algorithm” to extract highly correlated Bicluster. The correlation based fitness is used to identify the correlated bicluster with large volume. jinHuaXu, HongLiu[9] has presented vector analysis and KMeans based algorithms for mining user clusters. Abdolreza Hatamlou [10] has proposed a new heuristic algorithm that is motivated by the black hole phenomenon. In this algorithm black hole algorithm (BH) a randomly generated population of candidate solutions

"the stars "are placed in the search space of some problem or function. After initialization, the fitness values of the population are evaluated and the best candidate in the population, which has the best fitness value, is selected to be the black hole and the rest form the normal stars. The black hole has the ability to absorb the stars that surround it.

## IV. PROPOSED WORK

Proposed approach is as shown in figure 3.This is the basic flow of web usage mining using data pre-processing technique and pattern discovery using hybridize Black-hole clustering technique. Here Black-hole clustering with KMeans is the proposed algorithm to identify clusters.

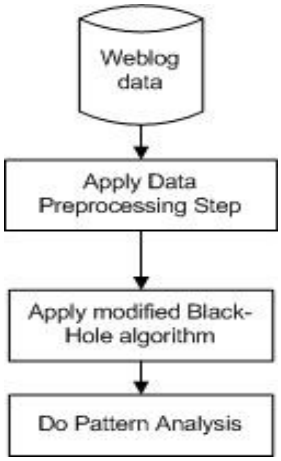



Fig 3. Proposed Approach

Clustering is one of the data mining techniques for pattern discovery in web usage mining. After data preprocessing step clustering will applied to preprocessed data. Here Black-hole clustering algorithm use with k-means algorithm for clustering. Black-hole clustering algorithm is a nature-Inspired algorithm. It is used for real world data set like iris, vowel, CMC etc. It has been proved that black-hole clustering algorithm gives better solution than other techniques like k-means, GSA, PSO etc. So, here by using this black-hole algorithm tries to discover clusters from web log.

### *A. Basic Concept Of Black-Hole*

In the 18th century John Michell and Pierre Laplace were the founder to discover the black-hole concept. They formulated the theory of a star becoming invisible to the eye integrating Newton's law, though, during that period black hole was not known and it was only in 1967 that John Wheeler the American physicist first named the information of mass collapsing as a black hole [10]. A black hole is created in space when a star of huge size falls. The gravity is too much strong so that the gravitational power of the black hole is in massive amount that even the light cannot get away from it. Anything that comes near to the black hole, that things will be swallowed by it and disappear because of its enormous power. The event horizon is identified as he sphere-shaped limit of a black hole in space. Schwarzschild radius is known as the radius of the event horizon. At this radius, the escape speed is the same as the speed of light, and once light passes through, even it cannot get away. Nothing can get away from within the event horizon because nothing is as faster as light. If anything come nearer to the event horizon and crosses the Schwarzschild radius it will be swallowed by the black hole and permanently disappear [10]. This Concept of Black-hole is used to find the cluster of real world data set by A. Hatamlou [10]. Here this concept is use for find the cluster from web log data rather than simple data. So by changing algorithm's input use it for web usage mining.

### *B. Basic Concept Of K-Means Algorithm*

K-means algorithm is well known algorithm and it is very simple. It follows a simple and easy way to categorise a given data set through a certain number of clusters. The core concept is to define k centroids, individual for each cluster. The next step is to assign each point be appropriate to a given data set and associate it to the nearest centroid. When no point is remaining, the first step is completed. At this step calculate k new centroids again. After that a new group is created between the same data set points and the nearest new centroid. The k centroids change their location step by step until no more changes are done.

### *C. Black-Hole With Kmeans Clustering Algorithm*

Step 1: Select Web log data for clustering

Step 2: Calculate fitness function for each data.

Fitness Function F= $\sum_{i=1}^{n}(x_i - c_j)^2$ (1)

Where Xi is the data, Cj is the initial cluster and N is the Number of cluster

Step 3: Select the data as black hole having best fitness value

Step 4: Change the location of data using equation 1

$$X_i(t+1) = X_i(t) + rand\,(X_{BH} - X_i(t)) \quad (2)$$

Where $X_i(t+1)$ is location of $X_i$ data at t+1 iteration

$X_i(t)$ Is location of $X_i$ at t iteration

$X_{BH}$ is location of Black-Hole

Rand is rand number between [0,1]

Step 5: If data reaches to location with lower cost than black-hole, exchange their location

Step 6: Calculate event Horizon of each data by equation :

$$R = fBH / \sum fi \quad (3)$$

Step 7: If distance between data and Black-Hole is less than R then new candidate is generated

Step 8: Initialize the k-means center with data

Step 9: Allocate each data to a cluster by equation:

$$dis(x_i, c_j) = \sum_{i=1}^{n}(x_i - c_j)^2 \quad (4)$$

Step 10: Refine the cluster centroid by equation:

$$Mean = \frac{1}{n}\left(\sum_{i=0}^{n} x_i\right) \quad (5)$$

Step 11: Repeat step 9 and 10 until all data is assigned.

Step 12: If termination Criteria met, exit.

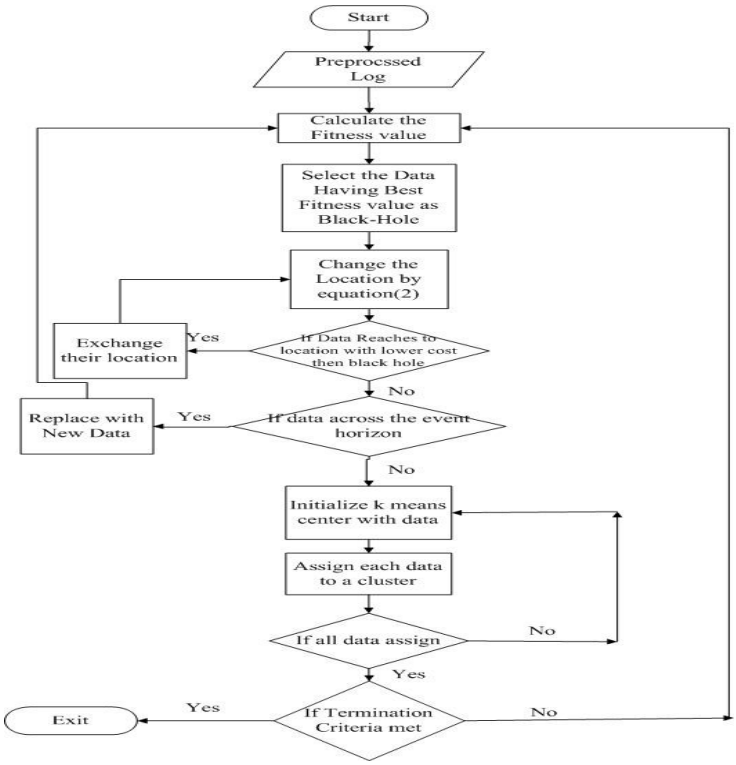



Fig 4.Flowchart of Balckhole with Kmeans Clsuteirng Algorithm

## V. EXPERIMENT ANALYSIS

Using this approach we obtaining three clusters of session time_interval between access urls:

**Cluster 1:** [6000, 1000, 2000, 3000, 7000, 9000, 6000, 14000, 22000, 12000, 3000, 6000, 8000, 15000, 11000, 16000, 22000,

20000, 6000, 8000, 7000]
**Cluster 2:** [26000, 28000, 25000, 30000, 26000, 23000]
**Cluster 3:** [33000, 47000, 221000, 72000, 44000, 213000, 210000, 81000, 91000, 184000, 60000, 33000, 179000, 128000, 40000, 146000, 141000, 228000, 35000, 118000, 182000, 74000, 182000]

*A. Centroid of Cluster*

| | Black-hole | Black-hole with k-means | KMeans |
|---|---|---|---|
| cluster 1 | 9714.2857 | 19817.07 | 18000 |
| cluster 2 | 26333.333 | 121424.91 | 89142.857 |
| cluster 3 | 119217.39 | 228406.43 | 188600 |

Table 1: Centroid of cluster

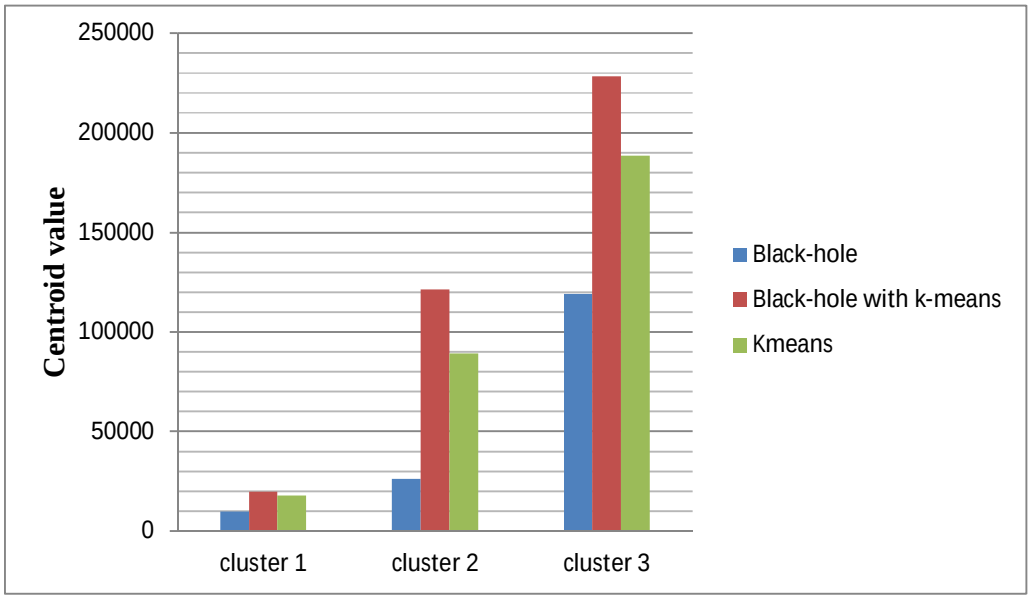



Fig. 5: Centroid of cluster

Here Black-hole with KMeans algorithm gives the best centroid value among Black hole and KMeans algorithm as shown in figure 5. Centroid is defined as mean value of each cluster.

*B. Intercluster Similarity*

| | **Blackhole** | **Blackhole With Kmeans** | **Kmeans** |
|---|---|---|---|
| **Cluster 1** | **16619.04761904762** | **101607.84368112191** | **71142.85714285714** |
| **Cluster 2** | **92884.0579710145** | **106981.52028093435** | **99457.14285714286** |
| **Cluster 3** | **109503.10559006211** | **208589.36396205626** | **170600.0** |

Table 2. Intercluster Similarity

Volume 3 Issue VI, June 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

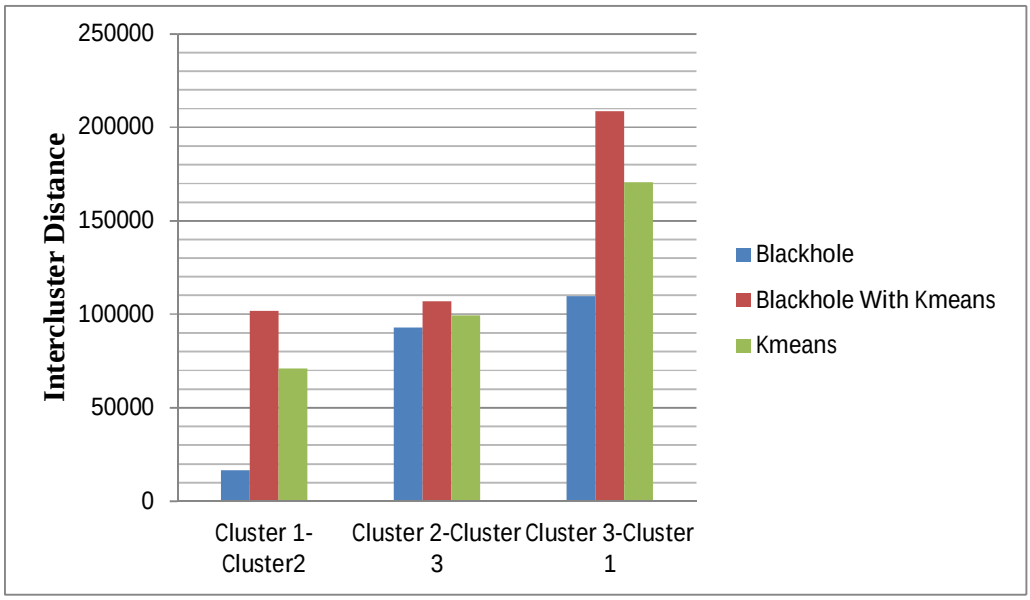



Fig 6. Intercluster Similarity

From figure 6, analysis Black-hole with KMeans has largest inter cluster distance.so it is better than KMeans and black hole algorithm. Here Intercluster distance is the distance between each cluster.

*C. Intra Cluster Similarity*

| | Black-hole | Black-hole With KMeans | KMeans |
|---|---|---|---|
| **Cluster 1** | 5170.068027210885 | 27731.22829861111 | 19877.551020408166 |
| **Cluster 2** | 1777.7777777777774 | 28531.433475378784 | 19877.551020408166 |
| **Cluster 3** | 61096.408317580346 | 12165.709971371907 | 12165.709971371907 |

Table 3: Intracluster similarity

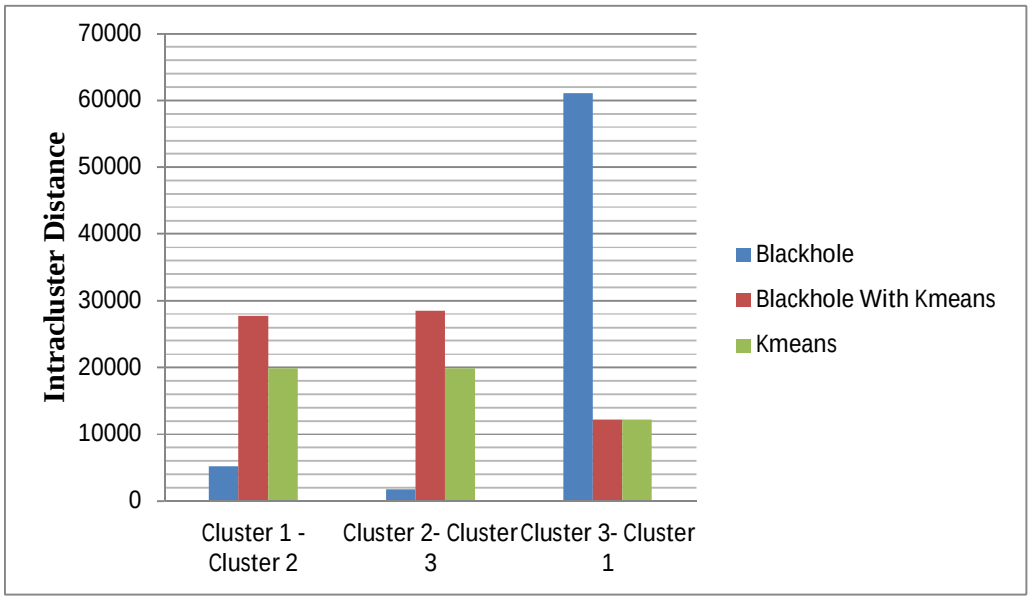



Fig 7. Intracluster Similarity

Intracluster similarity is high in cluster 1 and 2 for black hole with KMeans algorithm but cluster 3 black hole's similarity is

high. Intracluster is defined as average of distance between each cluster element.

## VI. CONCLUSIONS

Web usage mining is a complete process to extract knowledge about browsing behavior of web user from web log. This knowledge is useful in various fields like website customization, personalization and recommendation. There are number of data mining techniques are used to mine knowledge from web log. Each technique has its advantage and limitation. Here in proposed approach Black-Hole algorithm use with K-means algorithm for clustering of web server data. Data preprocessing are applied to web data using MySQL and java. This is very new and efficient approach for web usage mining. In this technique apply modify Black-Hole algorithm using K-means algorithm and analyze cluster using various cluster analysis techniques. From analysis it is proved that black hole with KMeans algorithm is better than KMeans and black hole algorithm. In future we can apply Black-hole algorithm with any other clustering algorithm like k-medoid, PSO etc.

## REFERENCES

[1] Theint Theint Aye, et al, "Web Log Cleaning for Mining of Web Usage Patterns", IEEE, International conference on Computer Research and development, Volume 2, Page-no: 490-494, ISBN no: 978-1-61284-839-6

[2] K. Sudheer Reddy M. Kantha Reddy V. Sitaramulu, et al, "An effective Data Preprocessing method for Web Usage Mining", IEEE, International Conference on Information Communication and Embedded Systems (ICICES), 2013, page-no: 7-10, ISBN no: 978-1-4673-5786-9S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.

[3] Xidong Wang, Yiming Ouyang ,Xuegang Hu, Yan Zhang, et al, "Discovery of User Frequent Access Patterns on Web Usage Mining", IEEE, The 8th International Conference on Computer Supported Cooperative Work in Design Proceedings, 2004, volume 1, page-no:765-769, ISBN no: 0-7803-7941-1

[4] Lin Feng, Baohua Guan, et al, "Web Usage Mining with Variable Precision Rough Set Approach", IEEE, Fourth International Symposium on Knowledge Acquisition and Modeling (KAM), 2011, page no:204-206, ISBN no: 978-1-4577-1788-8

[5] K.Suresh, R.MadanaMohana, A.RamaMohanReddy, A.Subrmanyam, et al, "Improved FCM algorithm for Clustering on Web Usage Mining", International Conference on Computer and Management (CAMAN), 2011, page-no:1-4, IEEE, ISBN no: 978-1-4244-9282-4

[6] Anna Alphy, S. Prabakaran, et al, "Cluster Optimization for Improved Web Usage Mining using Ant Nestmate Approach", International Conference on Recent Trends in Technology(ICRTIT), 2011, page-no:1271-1276, IEEE, ISBN no:978-1-4577-0588-5

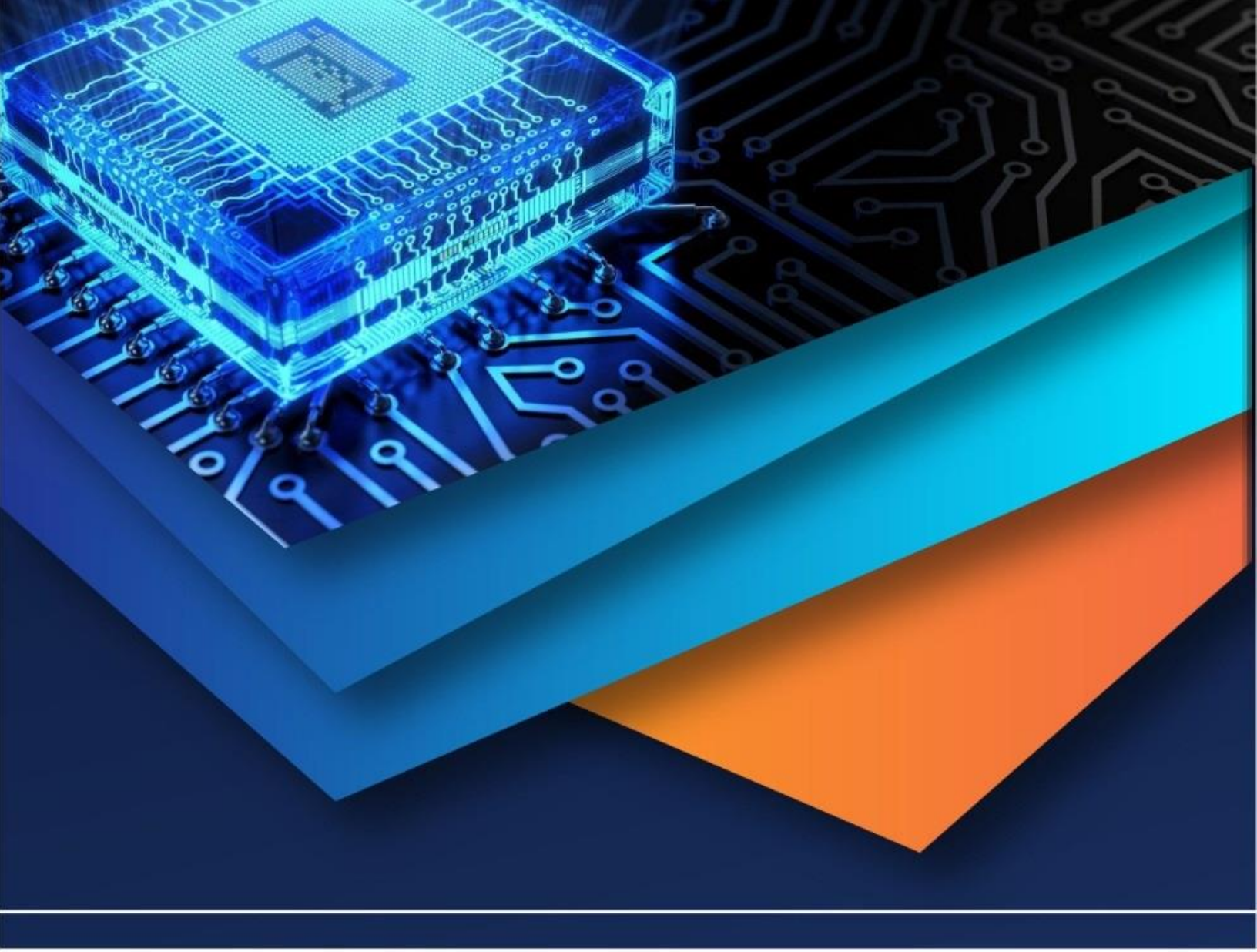
[7] Nayana Mariya Varghese, Jomina John, et al, "Cluster Optimization for Enhanced Web Usage Mining using Fuzzy Logic", World Congress on Information and Communication Technologies (WICT), 2012, page-no:948-952, IEEE, ISBN no: 978- 1-4673-4806-5

[8] V. Diviya Prabha, R. Rathipriya, et al, "Biclustering of Web Usage Data Using Gravitational Search Algorithm", International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013, page-no: 500-505, IEEE, ISBN no: 978-1-4673-5843-9

[9] linHuaXu, HongLiu, et al, "Web User Clustering Analysis based on K Means Algorithm", International Conference on Information Networking and Automation (ICINA), 2010, volume 2, page-no: V2-6 - V2-9, IEEE

[10] Abdolreza Hatamlou, et al, "Black hole: A new heuristic optimization approach for data clustering", Elsevier Science Inc, Information Sciences: an International Journal, Volume 222, 2013, page-no: 175–184

[11] Bing liu, et al, "Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data", 2007, Springer, ISBN no: 13 978-3-540-37881-5



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY