# Web Log Mining Based on Soft Clustering FCM and Multi-Objective Genetic Algorithm

**Mr. Harish Bhabad[1], Prof. Pankaj Kawadkar[2]**

[1]Computer Science Engg.Department,
Patel Institute of Engg. Science,
Ratibad,Bhopal,Madhya Pradesh.

[2]Assistant Professor,
Head of the Department Computer Science Engg.,
Patel Institute of Engg. Science,
Ratibad,Bhopal, Madhya Pradesh.

**Abstract:** ***Web log mining is important area of research for the improvement of web efficiency and log cache enhancement. The processing of log in web mining is very important part of view for user profile viewer and improvement of accessing power of website. Also impart a role of security threaten in web mining. Web mining are classify into three domains: content, structure and usage mining.***

***For the processing of web log mining various method of data mining is applied one such method is called clustering. Clustering is unsupervised learning technique of data mining. The part ion clustering technique is used in web log mining. The form of this clustering is k-means and k-median algorithm, but these algorithm are suffered some point of problem now used soft clustering technique such as fuzzy clustering algorithm. Fuzzy clustering algorithm is great advantage over k-means algorithm.***

***Proposed clustering algorithm is based on two specific factors, threshold factor which initial decide the number of cluster and specific factor which merge the clusters according the similarity. The careful selection of threshold value and specific factor which control merging of clusters yields efficient algorithmic results. The purpose of this paper is to provide optimum initial solution for FCM with the help of genetic algorithm For the improvement of FCM clustering technique used multi-objective genetic algorithm for better generation of clustering technique. Our empirical result shows the better result in compression of FCM algorithm for web log mining.***

**Keywords:** *Web Mining, MOGA, clustering technique, FCM*

## 1. INTRODUCTION

The World Wide Web has huge amount information and large datasets are available in databases. Information retrieving on websites is one of possible ways how to extract information from these datasets. Web mining is the extraction of interesting and useful knowledge and implicit information from artifacts or activity related to the WWW. Based on several research studies we can broadly classify Web mining into three domains: content, structure and usage mining. Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified to better suit the demands of the Web. New approaches should be used better fitting to the properties of Web data. Furthermore, not only data mining algorithms, but also artificial intelligence, information retrieval and natural language processing techniques can be used efficiently. The given figure 1 shows the description technique of web mining.

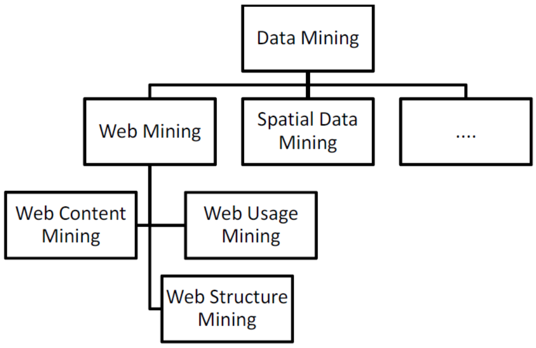



Figure1. Description of Web Miming category

**1.1 Web Mining:** Web mining consists of a set operations defined on data residing on WWW data servers. Such data can be the content presented to users of the web sites such as hyper text markup language (HTML) files, images, text, audio or video. Web mining is mainly categorized into two subsets namely web content mining and web usage mining.

**a) Web Content Mining:**

Web content mining describes the automatic search of information resources available on-line. The focus is on the content of web pages themselves.

**b) Web Structure Mining:**

Web structure mining is the process of discovering the structure of hyperlinks within the Web. Practically, while Web content mining focuses on the inner-document information; Web structure mining discovers the link structures at the inter-document level.

**c) Web Usage Mining:**

Usage mining as the name implies focus on how the users of websites interact with web site, the web pages visited, the order of visit, timestamps of visits and durations of them. The main source of data for the web usage mining is the server logs which log each visit to each web page with possibly IP, referrer, time, browser and accessed page link. Although many areas and applications can be cited where usage mining is useful, it can be said the main idea behind web usage mining is to let users of a web site to use it with ease efficiently, predict and recommend parts of the web site to user based on their and previous user's actions on the website. Figure 2. The General architecture of web uses mining.

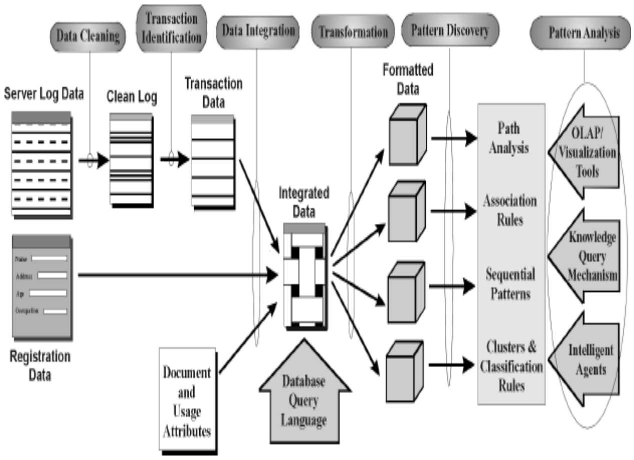



Figure 2: General Architecture for Web Usage Mining

**1.2 Web Usage Mining and Pattern Discovery**: Web usage mining is the application of data mining techniques to discover usage pattern from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. A high level Web usage mining Process is presented in Figure 3.

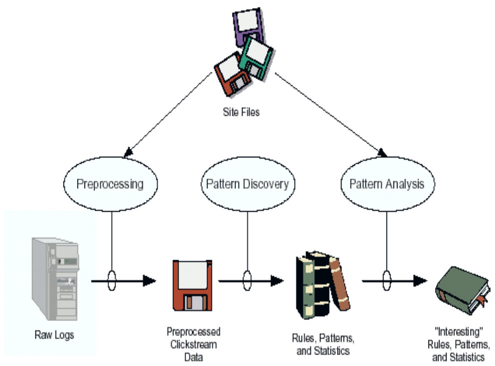



Figure 3. High Level web usage Mining Process

## 2. Related Work

There has been an extensive studied of various research paper on fuzzy c-means algorithm. And the various Initialization methods of FCM clustering algorithm have been more emphasized in this literature.

### 2.1 Initialization method for k-mean algorithm:

Fuzzy clustering plays an important role in solving problems in the areas of pattern recognition and web log mining. A variety of fuzzy clustering methods have been proposed and most of them are based upon distance criteria [6]. One widely used algorithm is the fuzzy c-means (FCM) algorithm. It uses reciprocal distance to compute fuzzy weights. A more efficient algorithm is the new FCFM. It computes the cluster center using Gaussian weights, uses large initial prototypes, and adds processes of eliminating, clustering and merging. In the following sections we discuss and compare the FCM algorithm and FCFM algorithm. The fuzzy c-means (FCM) algorithm was introduced by J. C. Bezdek [2]. The idea of FCM is using the weights that minimize the total weighted mean-square error:

$$J(w_{qk}, \mathbf{z}^{(k)}) = \Sigma_{(k=1,K)} \Sigma_{(k=1,K)} (w_{qk}) \| \mathbf{x}^{(q)} - \mathbf{z}^{(k)} \|^2 \quad (1)$$

$$\Sigma_{(k=1,K)} (w_{qk}) = 1 \text{ for each } q$$

$$w_{qk} = (1/(D_{qk})^2)^{1/(p-1)} / \Sigma_{(k=1,K)} (1/(D_{qk})^2)^{1/(p-1)}, \; p > 1 \quad (2)$$

### 2.2 Random Sampling Methods:

Probably being most widely adopted in the literature, random sampling methods follow a naive way to initialize the seed clusters, either using randomly selected input samples, or random parameters non-heuristically generated from the inputs. Being one of the earliest references in the literature, Forgy in 1965 [10] adopted uniformly random input samples as the seed clusters. The method, named R-SEL in our study, is formalized below.

**R-SEL:**
For i= 1.....K, set $c_i = x_r$ such that
$r$ =uniRand(1;$N$),
Where Nis the total number of input
Samples and uniRand(min; max) is a uniform random generator producing r$\varepsilon$[*min; max*].

### 2.3 Distance Optimization Methods:

Recognizing the characteristics of many clustering methods is to locally minimize the intra-cluster variances without optimizing the inter-cluster separation; it is a natural consideration to optimize the distances among the seed clusters beforehand towards a satisfactory inter-cluster separation in the output. Among some early practices, the Simple Cluster Seeking (SCS) initialization method [12] is adopted in the FASTCLUS
Procedure, which is a K-Means variance implemented in SAS. The SCS method is summarizes below.

**SCS:**
(1) Initialize the first cluster centroid with the first input, i.e. $c_1 = x_1$.
(2) For j= 2…….$N$, if $\|x_j - c_k\| > p$ for all existing seed clusters $c_k$,
Where $p$ is a threshold, and then add $x_j$ as a new seed Cluster.
Stop when $K$ seed clusters are initialized.
(3) If after scanning all input samples, there are less than $K$ seed
Clusters generated, and then decrease $p$ and repeat 1 - 2.

### 2.4 Density Estimation Methods:

This category of initialization methods is based on the assumption that the input data follow a Gaussian mixture distribution. Hence by identifying the dense areas of the input domain, the initialized seed clusters help the clustering method in creating compact clusters.

## 3. PROBLEM STATEMENT

The attractiveness of the FCM lies in its simplicity and flexibility. In spite of other algorithms being available, k-means continues to be an attractive method because of its convergence properties. However, it suffers from major shortcomings that have been a cause for it not being implemented on large datasets. The most important among these are

(I)The Number of Cluster is unknown in K-mean clustering algorithm it specifies by user at run time.

(II)The initial center problem because clusters obtained depend heavily on initial centers [9].

(III) K-means is slow and scales poorly with respect to the time it takes for large number of points.

Because of these shortcomings, proposed efficient FCM clustering algorithm is required to overcome above shortcomings

## 4. PROPOSED METHOD

The process of FCM clustering and multi-objective function takes several process such as population, fitness function, mutation and crossover for propagation of clustering algorithm. Some steps are divided into six phase.

Initializing population

The initial population is done by determining the length of chromosome with size K x d. K is the number of chromosome a lot of d while d is the dimension of the cluster variables Fitness function

K-Means clustering optimization with multi-objective genetic algorithm uses 2 objectives, i.e. minimizing error functions within each cluster (equation 1) and maximizing the centroid value between cluster (equation 2). The calculations used as follows.

$$\sigma_t^2 = \frac{1}{N}\sum_{j=1}^{N}\sigma_{ij}^2 \dots\dots\dots\dots\dots\dots\dots\dots\dots.. \quad (1)$$

Where i = 1, 2... K; K is the number of cluster

$$\sigma^2 \sum_{i=1}^{K}\sigma_i^2 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \quad (2)$$

Whereas:

$\sigma_i^2$ : Error in $i -^{th}$ cluster

$n_i$ : number of data in banyak data pada$i -^{th}$ cluster

$x_{ij}$ : Data in i-cluster, $j -^{th}$variabe

$z_{ij}$ : i-cluster average in j-variable

V : Number of variable

The first function is to minimize error average in cluster which formulated as follow:

$$V(w) = \frac{1}{k}\sum_{i=1}^{k}\sigma^2\, i \qquad , i = 1,2 \dots\dots k \quad (3)$$

Whereas:

V (w) : Error in cluster

k : Number of cluster

The second function is to maximize inter-cluster error which formulated as follow:

$$V(b) = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{v}(z_{ij} - \bar{z}_j) \quad (4)$$

Demean:

V (b) : Error in cluster

$z_{ij}$ : i-cluster average in variable

$\bar{z}$ : Grand mean of $j -^{th}$ variable

**4.1Proposed Architecture for processing of clustering**

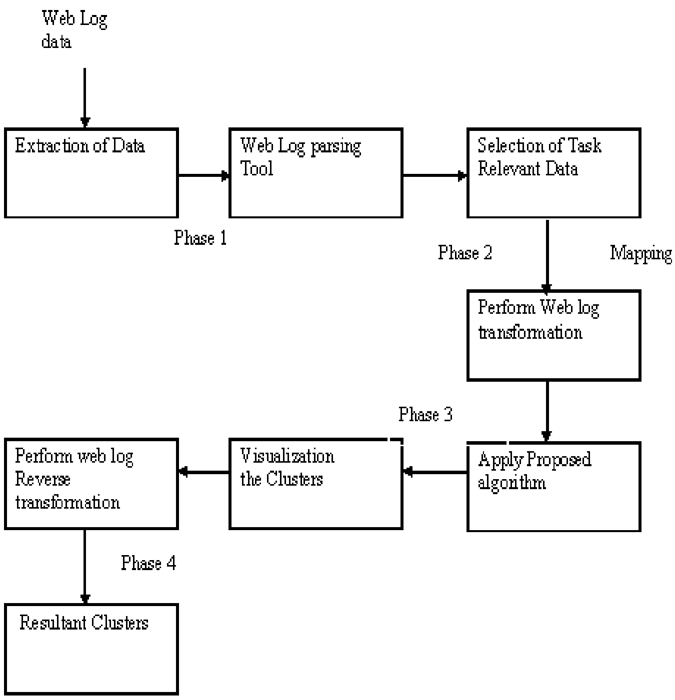



Figure 4. Proposed Architecture for Processing of Clustering

**Phase 1:**

Resource extraction is the process of retrieving the desired web log data files from the web server. These access log files contain information in CERN (Common Log Format).

In this phase of our work:

1. Extract web log data file from web server which contains some common fields:

- User's IP address
- Access date and time
- Request method (GET or POST),
- URL of the page accessed,
- Transfer protocol (HTTP 1.0, HTTP 1.1,),
- Success of return code.
- Number of bytes transmitted.

2. Give this web log data file as an input to web log parsing tool: Web Log Explorer

Web Log Explorer takes your log file, parse it and build reports by grouping or filtering the extracted data. Then we explore page view request in resulting table and export this report in CSV File (Excel).

3. Select task relevant data from excel file. This task relevant data contain 2 fields:

- Web pages
- Frequency of web page

**Phase 2:**
Proposed algorithm work on only numerical data so mapping is performed on web pages. And assign every web page to unique numerical value.

**Phase3:**
After phase 2 proposed clustering algorithm is performed .this algorithm is completed in two stage in first stage according threshold value number of cluster and their centroid is generated. And in second stage according the similarity between the clusters they merge into a one cluster.

**Phase 4:**
After forming the cluster by proposed k-mean clustering algorithm. Remapping the each unique numerical value to web page. So they will form number of cluster according frequency of web pages. And by this algorithm we mining which web pages is highly accessed by client.

## 5. EXPERIMENT AND RESULT

Web log data used FCM and FCM_MOGA clustering algorithm implement in MATLAB 7.10.0 and found the similar pattern of cluster for IP, session, time and data. The cluster found in color group. The formation of cluster gives the information of valid and invalid cluster according to cluster valid index. For the performance evaluation of FCM technique and our FCM-MOGA used MATLAB software package. MATLAB is a software package for high- performance numerical computation and visualization. For the processing of clustering used university web log files in five sections in different size of data. Result table and graphs are as follows.

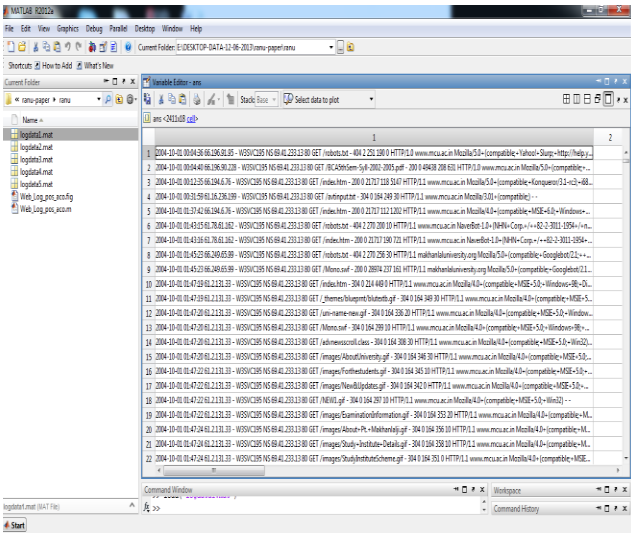


Figure 5. Dataset Web Log Entry

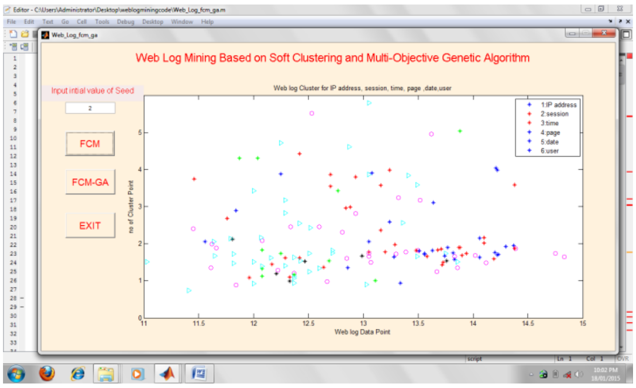



Figure 6. GUI Cluster putting the value of seed by FCM method

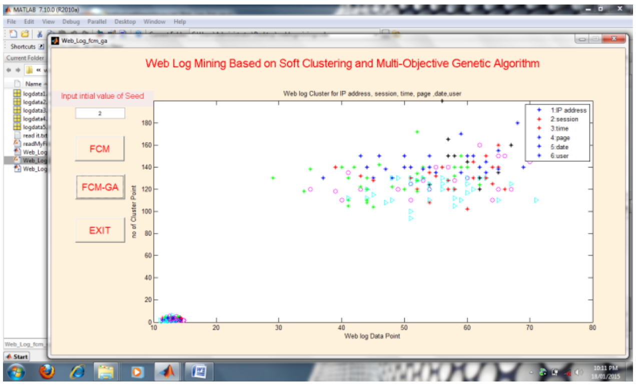



Figure 7. GUI Cluster putting the value of seed by FCM – MOGA method

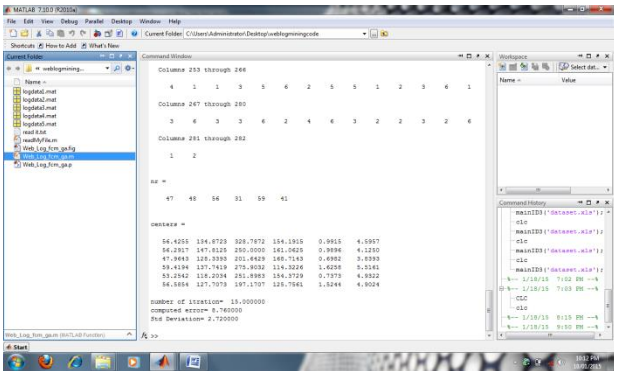


Figure 8. Result window of matlab and computed value of number of iteration, Error and standard deviation

### 5.1 Comparative Result analysis of all data set with two methods.

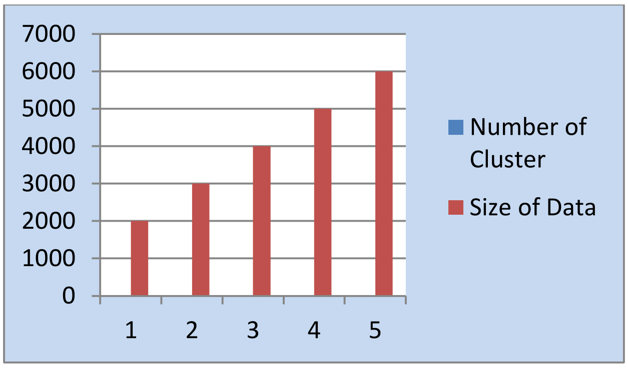
Table1. Performance evaluation of Web Dataset1

| Seed value | Iteration | | Error | |
|---|---|---|---|---|
| | FCM | FCM-GA | Computed Error | Std. deviation |
| 2 | 10.00 | 15.00 | 5.35 | 1.68 |
| 3 | 7.00 | 10.00 | 6.20 | 1.68 |
| 4 | 5.00 | 8.00 | 5.13 | 1.68 |
| 5 | 5.00 | 7.00 | 4.65 | 1.68 |
| 6 | 4.00 | 6.00 | 5.35 | 1.68 |

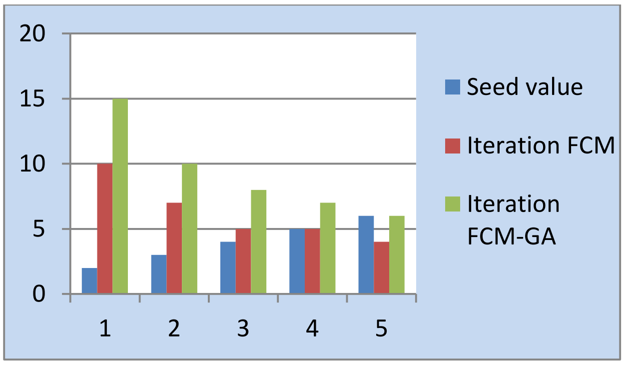
Table2. Performance evaluation of Web Dataset2

| Seed value | Iteration | | Error | |
|---|---|---|---|---|
| | FCM | FCM-GA | Computed Error | Std. deviation |
| 2 | 10.00 | 15.00 | 5.11 | 1.68 |
| 3 | 7.00 | 10.00 | 5.35 | 1.68 |
| 4 | 5.00 | 8.00 | 5.73 | 1.68 |
| 5 | 5.00 | 7.00 | 4.57 | 1.68 |
| 6 | 4.00 | 6.00 | 5.00 | 1.68 |

### 5.2 Result Analysis of all Web log Dataset




Graph 1. Comparative Number of Cluster Generation according to Data size.




Graph 2. Comparative valid cluster Generation of Cluster size of Seed value of FCM and FCM-GA

## 6. CONCLUSION

The new algorithm for FCM-MOGA clustering is proposed which efficiently overcome the major drawbacks viz. right number of cluster and initial seed (center point) problem. Proposed clustering algorithm is based on two specific factors, threshold factor which initial decide the number of cluster and specific factor which merge the clusters according the similarity. The careful selection of threshold value and specific factor which control merging of clusters yields efficient algorithmic results. Our empirical result shows the better result in compression of FCM algorithm for web log mining.

## References

[1]Qingtian Han, XiaoyanGao ,Wenguo Wu, "Study On Web Mining Algorithm Based On Usage Mining", Computer- Aided Industrial Design And Conceptual Design, 2008.CAID/CD 2008. 9th International Conference On 22-25 Nov. 2008.

[2]Xuan SU ,Xiaoye WANG ,Zhuo WANG ,Yingyuan XIAZO, "An New Fuzzy Clustering Algorithm Based On Entropy Weighting", Journal Of Computational Information Systems,3319-3326,2010.

[3]Mohanad Alata, Mohammad Molhim, And Abdullah Ramini, "Optimizing Of Fuzzy C-Means Clustering Algorithm Using GA" , World Academy Of Science, Engineering And Technology 39, 2008.

[4]K.Suresh, R.MadanaMohana, A.Ramamohan Reddy, "Improved FCM Algorithm For Clustering On Web Usage Mining",IJCSI International Journal Of Computer Science Issues, Vol. 8, Issue 1, January 2011.

[5] TasawarHussain, Dr. SohailAsghar "A Hierarchical Cluster Based Preprocessing Methodology For Web Usage Mining" In IEEE 2010.

[6]TasawarHussain, SohailAsgharAnd NayyerMasood" Hierarchical Sessionization At Preprocessing Level Of WUM Based On Swarm Intelligence" In 2010 6th International Conference On Emerging Technologies.

[7] Alam, S., G. Dobbie, Et Al. (2008). Particle Swarm Optimization Based Clustering Of Web Usage Data. 2008 IEEE/WIC/ACM International Conference On Web Intelligence And Intelligent Agent Technology 978-0-7695-3496-1/08 2008 IEEE DOI 10.1109/WIIAT.2008.292 2008 IEEE/WIC/ACM International Conference on Web.

[8]Kumar, R. "Mining Web Logs: Applications And Challenges" KDD'09, June 28–July 1, 2009, Paris, France. ACM 978-1-60558-495 2009.

[9] Lu, H. And T. T. S. Nguyen "Experimental Investigation Of PSO Based Web User Session Clustering. 2009 International Conference of Soft Computing And Pattern Recognition "978-0- 7695-3879-2/09 IEEE 2009

[10] Makanju, A., A. N. Zincir-Heywood" Clustering Event Logs Using Iterative Partitioning". KDD'09, June 28–July 1, 2009, Paris, France. ACM 978-1-60558-495,2009.

[11] Mr.V.K.Panchal, M. H. Kundra, "Comparative Study Of Particle Swarm Optimization Based Unsupervised Clustering Techniques." IJCSNS International Journal Of Computer Science And Network Security, VOL.9 No.10, October 2009.

[12] B. Pavel, "A Survey Of Clustering Data Mining Techniques," In Grouping Multidimensional Data. Springer Berlin Heidelberg, 2006, Pp. 25–71.

[13] K. Suresh, R. MadanaMohana, A. Rama Mohan Reddy And A. Subramanyam, "Improved FCM Algorithm For Clustering On Web Usage Mining," International Conference On Computer And Management (CAMAN), Pp. 1 – 4, 2011.

[14] Yanfeng Zhang, XiaofeiXuAndYunming Ye, "An Agglomerative Fuzzy K-Means Clustering Method With Automatic Selection Of Cluster Number," 2nd International Conference On Advanced Computer Control (ICACC), Vol. 2, Pp. 32-38, 2010.