Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 1, January 2017, pg.30 – 35

Enhancing the Performance of Web Proxy Server and Cluster the Data using Fuzzy C-Means Algorithm

Sukhvir Kaur[1], Charanjit Singh[2]

[1]M.Tech Student, RIMT-IET, Mandi Gobindgarh

[2]Assistant Professor, RIMT-IET, Mandi Gobindgarh

[1] Sukhvirb0@gmail.com, [2] sehgal_cs@yahoo.com

Abstract: ***A web cache (or HTTP cache) is an information technology for the temporary storage (caching) of web documents, such as HTML pages and images, to reduce bandwidth usage, server load, and perceived lag. The web cache proxy server can serve cached content much more quickly to other computers on the local network. Web caching is used at the proxy server level to reduce the network traffic by caching web pages. But because nowadays World Wide Web has evolved rapidly so caching alone is not sufficient. So to solve this problem prefetching can be used with caching because prefetching transfer data from main memory to temporary storage in readiness for later use and because of this the response time for the user is decreased up to large extent. In this paper we cluster the user according to their access pattern and usage behavior with the help of Fuzzy c-Means algorithm. And remove the noise from the web loges using firefly algorithm.***

Keywords: Clustering, Web server, fuzzy c-means clustering, firefly algorithm

I. INTRODUCTION

A) *Web Server*

A Web server is a program that uses HTTP (Hypertext Transfer Protocol) to serve the files that form Web pages to users, in response to their requests, which are forwarded by their computers' HTTP clients[2]. Dedicated computers and appliances may be referred to as Web servers as well. Web servers often come as part of a larger package of Internet- and intranet-related programs for serving email, downloading requests for File Transfer Protocol (FTP) files, and building and publishing Web pages. Considerations in choosing a Web server include how well it works with the operating system and other servers, its ability to handle server-side programming, security characteristics,

and the particular publishing, search engine and site building tools that come with it. The files stored on web server are read by browsers such as firefox , safari, chrome or internet explorer which convert these files into images and text for the users to view. Basically A web server is simply a computer program that dispenses web pages as they are requested.

B) *Clustering*

Clustering analysis finds clusters of data objects that are similar in some sense to one another[13]. The members of a cluster are more like each other than they are like members of other clusters. Unsupervised classification [5]process [6] is termed as Clustering. The goal of clustering analysis is to find high-quality clusters such that the inter-cluster similarity is low and the intra-cluster similarity is high. Clustering, like classification, is used to segment the data. Unlike classification, clustering models segment data into groups that were not previously defined. Classification models segment data by assigning it to previously-defined classes, which are specified in a target. Clustering models do not use a target. Clustering is useful for exploring data. If there are many cases and no obvious groupings, clustering algorithms can be used to find natural groupings. Clustering can also serve as a useful data-preprocessing step to identify homogeneous groups on which to build supervised models. Clustering can also be used for anomaly detection. Once the data has been segmented into clusters, you might find that some cases do not fit well into any clusters. These cases are anomalies or outliers. . In applications of cluster analysis[9] many types of visualization techniques have been employed to study the structure of datasets .

II. TYPES OF CLUSTERING

There are several different approaches to the computation of clusters. Clustering algorithms may be characterized as:

1) Hierarchical clustering

Hierarchal clustering Groups data objects into a hierarchy of clusters[7]. Hierarchical methods rely on a distance function to measure the similarity between clusters. Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering or HAC. Top-down clustering requires a method for splitting a cluster. It proceeds by splitting clusters recursively until individual documents are reached. Examples of this algorithms are LEGCLUST [3], BRICH [4].

2) Partitioning clustering

This method Partitions data objects into a given number of clusters. Partitional clustering attempts to directly decompose the data set into a set of disjoint clusters. The criterion function that the clustering algorithm tries to minimize may emphasize the local structure of the data, as by assigning clusters to peaks in the probability density function, or the global structure. Typically the global criteria involve minimizing some measure of dissimilarity in the samples within each cluster, while maximizing the dissimilarity of different clusters. A commonly used partitional clustering method is K-means clustering. In K-means clustering [14]the criterion function is the average squared distance of the data items x_k from their nearest cluster centroids.

3) Density based clustering

Partitioning and hierarchical methods are designed to find spherical-shaped clusters. They have difficulty finding clusters of arbitrary shape .The problem of detecting clusters of points in data is challenging when the clusters are of different size, density and shape. Many of these issues become even more significant when the data is of very high dimensionality and when it includes noise and outliers .The algorithm which is popular in Density based clustering is DBSCAN algorithm[12] .The DBSCAN algorithm relies on a density-based notion of clusters. It is the one-scan algorithms[16].Clusters are identified by looking at the density of points. Regions with a high density of points depict the existence of clusters whereas regions with a low density of points indicate clusters of noise or clusters of

outliers. This algorithm is particularly suited to deal with large datasets, with noise, and is able to identify clusters with different sizes and shapes.

4) Grid-based clustering

These methods partition[15] the space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. Grid-based [11] Algorithm define a set of grid-cells,it assign objects to the appropriate grid cell and compute the density of each cell and eliminate cells, whose density is below a defined threshold t The main advantage of the approach is its fast processing time.

5) Fuzzy clustering

Traditional clustering approaches generate partitions; in a partition, each instance belongs to one and only one cluster. Hence, the clusters in a hard clustering are disjointed. It is an [18]unsupervised type of clustering[17]. Fuzzy clustering for instance extends this notion and suggests a soft clustering schema. In this case, each pattern is associated with every cluster using some sort of membership function, namely, each cluster is a fuzzy set of all the patterns. Larger membership values indicate higher confidence in the assignment of the pattern to the cluster. A hard clustering can be obtained from a fuzzy partition by using a threshold of the membership value. The most popular fuzzy clustering algorithm is the fuzzy c-means algorithm. Even though it is better than the hard K-means algorithm at avoiding local minima, FCM can still converge to local minima of the squared error criterion. The design of membership functions is the most important problem in fuzzy clustering; different choices include those based on similarity decomposition and centroids of clusters. A generalization of the FCM algorithm has been proposed through a family of objective functions. A fuzzy c-shell algorithm and an adaptive variant for detecting circular and elliptical boundaries have been presented. e reflect the reality. Now a days, the fuzzy c-means clustering algorithm is the most widely used[1]. FCM produces almost close results to K-Means clustering but this method requires more calculation time than K-Means because of the fuzzy measures calculations involvement in the algorithm[19].

5.1) Working of fuzzy clustering

Fuzzy *C*-means (FCM----Frequently C Methods) is a method of clustering which allows one point to belong to one or more clusters.. The FCM algorithm attempts to partition a finite collection of points into a collection of *C* fuzzy clusters with respect to some given criteria. Thus, points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center of cluster. The FCM algorithm attempts to partition a finite collection of *n* elements X=$\{x_1, \ldots . x_n\}$ into a collection of c fuzzy clusters with respect to some given criterion.

Given a finite set of data, the algorithm returns a list of *c* cluster centres C=$c_1, \ldots c_c$ and a partition matrix W=$w_{ij} \in$[0,1],i=1,….,n,j=1,…..,c where each element w_{ij} tells the degree to which element x_i , belongs to cluster c_j.

The FCM aims to minimize an objective function :

$$\arg\min_{c} \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^{m} \left\| x_i - c_j \right\|^2$$

The FCM is also known as fuzzy *c*-means nebulous because it uses fuzzy logic so that each instance is not associated with only one cluster, but has a certain degree of membership for each of the existing centroids. For this, the algorithm creates a matrix **U** associativity, where each term μ_{ij} represents the degree of membership of sample *i* to cluster *j*. In this FCM algorithm have a variable fuzziness *m* such that 1.0 < *m* < where *m* and being areal number. The closer *m* is to infinity (∞), the greater the fuzziness of the solution and the closer to 1, the solution becomes increasingly similar to the clustering of binary *k*-means [11]. A good choice is to set *m* = 2. fuzzy theory[8] can handle uncertainties in the data more efficiently than traditional data mining techniques.Fuzzy relational equations play important roles in many areas of applications, such as intelligence technology[10]. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership

towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. After each iteration membership and cluster centers are updated according to the formula:

$$\mu_{ij} = 1 \Big/ \sum_{k=1}^{c} \left(d_{ij}/d_{ik}\right)^{(2/m-1)}$$

$$v_j = \left(\sum_{i=1}^{n} (\mu_{ij})^m / x_i\right) / \left(\sum_{i=1}^{n} (\mu_{ij})^m\right), \forall_j = 1,2,3,\ldots\ldots,c$$

Where

'n' is the number of data points.
'vj' represents the j^{th} cluster center.
'm' is the fuzziness index m € [1, ∞].
'c' represents the number of cluster center.
'μij' represents the membership of i^{th} data to j^{th} cluster center.
'dij' represents the Euclidean distance between i^{th} data and j^{th} cluster center.

Main objective of fuzzy c-means algorithm is to minimize:

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} \left(\mu_{ij}\right)^m \left\| x_i - v_j \right\|^2$$

where $'||x_i - v_j||'$ is the Euclidean distance between i^{th} data and j^{th} cluster center.

5.2) Algorithmic steps for Fuzzy c-means clustering

Let $X = \{x_1, x_2, x_3 ..., x_n\}$ be the set of data points and $V = \{v_1, v_2, v_3 ..., v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

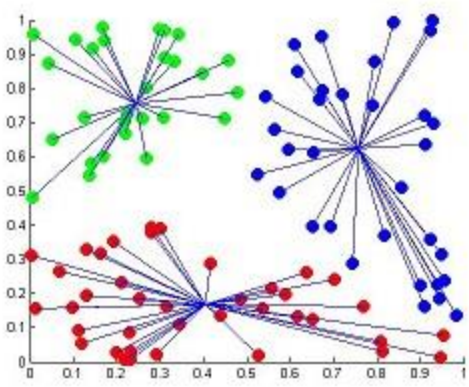
2) Calculate the fuzzy membership $'\mu_{ij}'$ using:

$$\mu_{ij} = 1 \Big/ \sum_{k=1}^{c} \left(d_{ij}/d_{ik}\right)^{(2/m-1)}$$

3) Compute the fuzzy centers $'v_j'$ using:

$$v_j = \left(\sum_{i=1}^{n} (\mu_{ij})^m / x_i\right) / \left(\sum_{i=1}^{n} (\mu_{ij})^m\right), \forall_j = 1,2,3,\ldots\ldots,c$$

4) Repeat step 2) and 3) until the minimum *'J'* value is achieved or $||U^{(k+1)} - U^{(k)}|| < \beta$.
where, *'k'* is the iteration step.
'β' is the termination criterion between [0, 1].
'$U = (\mu_{ij})_{n*c}$' is the fuzzy membership matrix.
'J' is the objective function.



Result of Fuzzy C-means clustering

III. FIREFLY ALGORITHM

The firefly algorithm is an optimization[20] algorithm which is used to find out the best solution from all feasible solutions. It is a met heuristic algorithm, inspired by the flashing behaviour of fireflies. The primary purpose for a firefly's flash is to act as a signal system to attract other fireflies. The primary purpose for a firefly's flash is to act as a signal system to attract other fireflies. In the firefly algorithm all fireflies are unisexual, so that any individual firefly will be attracted to all other fireflies. In this method attractiveness is proportional to their brightness, and for any two fireflies, the less bright one will be attracted by (and thus move towards) the brighter one; however, the intensity (apparent brightness) decrease as their mutual distance increases.in the case If there are no fireflies brighter than a given firefly, it will move randomly. The brightness should be associated with the objective function.

The main update formula for any pair of two fireflies $x_i$and y_j is

$$x_i^{t+1}=x_i^t+\beta\left[-\gamma r_{ij}^2\right]\left(x_j^t-x_i^t\right)+\alpha_t \in_t$$

It can be shown that the limiting case $\gamma \to 0$ corresponds to the standard Particle Swarm Optimization (PSO). In fact, if the inner loop (for j) is removed and the brightness I_j is replaced by the current global bestg^* then FA essentially becomes the standard PSO.

IV. WEB PERFORMANCE PARAMETERS

1) Hit ratio:-int his paper we will try to improve the hit ratio of the user's request.

2) Response-time:-in this paper we are trying to improve the response of any request which is made by the user to the server.

REFERENCES

[1] Deguang Wang, Baochang Han, Ming Huang, "Application of Fuzzy C-Means Clustering Algorithm Based on Particle Swarm Optimization in Computer Forensics", International Conference on Applied Physics and Industrial Engineering,pp.1186 – 1191,2012.

[2] Nanhay Singh, Arvind Panwar, and Ram Shringar Raw ,"Enhancing the Performance of Web Proxy Server through Cluster Based Prefetching Techniques", IEEE,pp.1158-1165, 2013.

[3] Jorge M. Santos, Joaquim Marques de Sa , and Luıs A. Alexandre, "LEGClust- A Clustering Algorithm based on Layered Entropic subgraph", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 30, No. 1, pp.62-75, january 2008.

[4] M. Livny, R.Ramakrishnan, T. Zhang,"BIRCH: An Efficient Clustering Method for VeryLarge Databases", ACMSIGMOD Workshop on Research Issues on Data Mining andKnowledge Discovery,pp.103-114, 1996.
[5] K.GEETHA, M.SANTHIYA,"a comparative study on data mining approachs" ,International Journal of Advanced Research in Datamining and Cloud Computing Volume 2, Pp.8-18, Issue 8, August 2014.
[6] Z. Huang, D. W. Cheung and M. K. Ng, "An Empirical Study on the Visual Cluster Validation Method with Fastmap", Proceedings of DASFAA01, Hong Kong, pp.84-91 ,April 2001.
[7] Periklis Andritsos, "Data Clustering Techniques",Qualifying Oral Examination Paper.
[8] Boldeanu Silviu , "fuzzy clustering".
[9] Swagati Julka,"Cluster Analysis for Multidimensional Data Using Visualization Technique", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8 ,pp.720-724, August 2013.
[10] Jianxiong Yang and Junzo Watada , "fuzzy clustering analysis of data mining-application to an accident mining system", ICIC International ,Volume 8, Number 8, pp.5715-5724, 2012.
[11] James c. Bezdek ,Robert Ehrlich,William Full, "FCM: The Fuzzy c-means clustering algorithm" Computers & Geosciences, Volume 10, pp. 191-203, 1984.
[12] Abdelghani Guerbas, Omar Addam, Omar Zaarour, Mohamad Nagi, Ahmad Elhajj, Mick Ridley,Reda Alhajj,"Effective web log mining and online navigational pattern prediction", Elsevier, pp.50–62,2013.
[13] S. Anitha Elavarasi, "A survey on partition clustering algorithms",International Journal of Enterprise Computing and Business Systems,Volume 1 ,Issue 1 ,January 2011.
[14] K. Naveen Kumar, G. Naveen Kumar,Ch. Veera Reddy, "Partition Algorithms– A Study and Emergence of Mining Projected Clusters in High-Dimensional Dataset", International Journal of Computer Science and Telecommunications ,Volume 2, Issue 4, pp.34-37, July 2011.
[15] K.Kameshwaran,K.Malarvizhi, "Survey on Clustering Techniques in Data Mining",International Journal of Computer Science and Information Technologies, Volume 5 , pp.2272-2276, 2014.
[16] Amandeep Kaur Mann ,Navneet Kaur , "Review Paper on Clustering Techniques",Global Journal of Computer Science and TechnologySoftware & Data Engineering ,Volume 13 ,Issue 5 ,2013.
[17] Soumi Ghosh, Sanjay Kumar Dubey ,"Comparative Analysis of K-Means and Fuzzy CMeans Algorithms", International Journal of Advanced Computer Science and Applications, Volume 4, No.4, pp-35-39,2013.
[18] R.Suganya, R.Shanthi, "Fuzzy C- Means Algorithm- A Review", International Journal of Scientific and Research Publications, Volume 2, Issue 11, November 2012.
[19] Tejwant Singh,Mr.Manish Mahajan, "Performance Comparison of Fuzzy C Means with Respect to Other Clustering Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, pp. 89-93, May 2014 .
[20] Sankalap Arora, Satvir Singh, "Performance Research on Firefly Optimization Algorithm with Mutation", International Conference on Communication, Computing & Systems,pp.168-172, 2014.

