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. 4, April 2017, pg.332 – 340

Use of Fuzzy C-Means Algorithm for Web Proxy Server Performance Improvement

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: ***Now a days the web is loaded with lot of request from users and it creates a lot of traffic on the web. As the requests are increasing the resources in the World Wide Web are also rising to large extent. In addition the services and applications provided by the web are directly proportional to its growth. For this reason, web traffic is huge, and to gain access to these resources incurs user-perceived latency. Web traffic has been the largest portion of Internet traffic. This is determined by the number of visitors and the number of pages they visit. The majority of website traffic is driven by the search engines. Millions of people use search engines everyday to research various topics, buy products, and go about their daily surfing activities. Search engines use keywords to help users find relevant information and each of the major search engines has developed a unique algorithm to determine where websites are placed within the search results. When a user clicks on one of the listings in the search results, they are directed to the corresponding website and data is transferred from the website's server, thus counting the visitors towards the overall flow of traffic to that website. because of the lots of traffic on websites whenever any users make any requests then the response time for users request decreases so to improve the response time the cache server is used . In this paper we will use Fuzzy c-Means algorithm to cluster the user according to their access pattern and usage behavior. And remove the noise from the web loges using firefly algorithm.***

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

I) INTRODUCTION

A) *Cache Server*

A cache server is a dedicated network server or service acting as a server that saves Web pages or other Internet content locally. By placing previously requested information in temporary storage, or cache, a cache server both speeds up access to data and reduces demand on an enterprise's bandwidth. Cache servers also allow users to access content offline, including rich media files or other documents. A cache server is sometimes called a "cache engine." A cache server is almost always also a proxy server, which is a server that "represents" users by intercepting their Internet requests and managing them for users. Typically, this is because enterprise resources are being protected by a firewall server. That server allows outgoing requests to go out but screens all incoming traffic. A proxy server helps match incoming messages with outgoing requests. In doing so, it is in a position to also cache the files that are received for later recall by any user. To the user, the proxy and cache servers are invisible, all Internet requests and returned responses appear to be coming from the addressed place on the Internet. A cache server basically is a dedicated server acting as a storage for web content, usually to have it available in a local area network. This serves to make web browsing and other services that need to go out over the internet, like software updates, faster because all of the usual data that used to be fetched from the outside is made available within the local neighborhood.

B) *Web Server*

Web server is a program that uses HTTP to serve files that create web pages to users in response to their requests, which are forwarded by their computers HTTP connection. Any server that delivers an XML document to another device can be a Web server. A better definition might be that a Web server is an Internet server that responds to HTTP requests to deliver content and services. Always a web server is connected to the internet. Every Web server that connects to the Internet will be provided with a unique address which was arranged with a series of four numbers between 0 and 255 separated by periods. Also, web server enables the hosting providers to manage multiple domains(users) on a single server[1]. A Web server uses HTTP to serve the files that form Web pages to users, in response to their requests, which are forwarded by their computers' HTTP clients[2].

C) Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are called a cluster are more similar in some sense or another to each other than to those in other groups clusters. Clustering is basically the most important unsupervised learning problem so, as with every other problem of this kind, it deals with finding a structure in a collection of unlabeled data[5]. The process of organizing objects into groups whose members are similar in some way is a cluster. A collection of objects which are "related" between them and are "unrelated" to the objects belonging to other clusters[3]. It can be achieved by different algorithms that differ significantly in their idea of what constitutes a cluster and how to efficiently find them[4].

II) TYPES OF CLUSTERING

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

1)Hierarchical clustering

A hierarchical clustering method consists of grouping data objects into a tree of clusters. There are two main type of techniques: a bottom-up and a top-down approach. The first one starts with small clusters composed by a single object and, at each step, merge the current clusters into greater ones, successively, until reach a cluster composed by

all data objects. The second approach use the same logic, but to the opposite direction, starting with the greatest cluster, composed by all objects, and split it successively into smaller clusters until reach the singleton groups. The bottom-up approach is called agglomerative and e top-down approach is called divisive [6]. An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion means the requested number k of clusters is achieved. In hierarchical clustering the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place[8].

2)Partitioning clustering

Given D, a data set of n objects, and k, the number of clusters to form, a partitioning algorithm organizes the objects into k partitions(k<=n),where each partition represents a cluster. The cluster are formed to optimize an objective partitioning criterion, such as a dissimilarity function based on distance, so that the objects within a cluster are "similar", whereas the objects of different cluster are "dissimilar". K-means is an example of this type of clustering method In k- means algorithm[10], the number of clusters k is the user-specified parameter. With k number of initial mean vectors, k-means algorithm iteratively assign the objects into one of cluster whose centre is the closest with itThe process continues until there is no other re-assignment or it depends on the user-specified threshold. k means is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

3)Density based clustering

Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the concept of density reachability and density connectivity. DBSCAN, DENCLUE and OPTICS [9] are examples for this algorithm. Density-based spatial clustering of applications with noise (DBSCAN) is a density-based clustering algorithm in which there is a given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN is one of the most common clustering algorithms.

4)Grid-based clustering

Grid based algorithm quantize the object space into a finite number of cells that forms a grid structure[7].Operations are done on these grids. The advantage of this method is its lower processing time. In this method Clustering complexity is based on the number of populated grid cells and does not depend on the number of objects in the dataset. The major features of this algorithm is that there is no distance computations in this method. One example of this method is STING algorithm.

5)Fuzzy clustering

Fuzzy clustering is a class of algorithm for cluster analysis in which the allocation of data points to clusters. It is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. Depending on the nature of the data and the purpose for which clustering is being used, different measures of similarity may be used to place items into classes, where the similarity measure controls how the clusters are formed. Some examples of measures that can be used as in clustering include distance, connectivity, and intensity. Fuzzy c- clustering [11] is an unsupervised technique that has been successfully applied to feature analysis, clustering, and classifier designs in fields such as astronomy, geology, medical imaging, target recognition, and image segmentation. Many fuzzy clustering algorithms had been are polpular[12].Those algorithms include fuzzy ISODATA, fuzzy C-means, fuzzy K-nearest neighborhood algorithm, potential-based clustering, and others [13].

Working Principle of FCM Algorithm:

1. Initialize U=u_{ij} matrix, U^0

2. At k-step: calculate the centroid C^k=C_j with U^k

$$C_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

3.update U^k, U^{k+1}

$$U_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}^{\frac{2}{m-1}} \right)}$$

4.if $\|U^{k+1} - U^k\|$<θ then STOP; otherwise return to step 2.Here, 'm' is the fuzziness parameter.

III) FIREFLY ALGORITHM

The Firefly Algorithm is a metaheuristic, nature-inspired, optimization algorithm which is based on the social flashing behavior of fireflies, or lighting bugs, in the summer sky in the tropical temperature regions . It was developed by Dr. Xin-She Yang at Cambridge University in 2007, and it is based on the swarm behavior such as fish, insects, or bird schooling in nature. In particular, although the firefly algorithm has many similarities with other algorithms which are based on the so-called swarm intelligence, such as the famous Particle Swarm Optimization , Artificial Bee Colony optimization , and Bacterial Foraging algorithms, it is indeed much simpler both in concept and implementation. The firefly algorithm has three particular idealized rules which are based on some of the major flashing characteristics of real fireflies. There are three rules for firefly algorithm

1) All fireflies are unisex, and they will move towards more attractive and brighter ones regardless their sex.
2) The degree of attractiveness of a firefly is proportional to its brightness which decreases as the distance from the other firefly increases due to the fact that the air absorbs light. If there is not a brighter or more attractive firefly than a particular one, [14]it will then move randomly.
3) The brightness or light intensity of a firefly is determined by the value of the objective function of a given problem. For maximization problems, the light intensity is proportional to the value of the objective function.

I,1)initialize objective function f(xi)

I(r) =$\frac{I_s}{r^2}$,where I(r) is the intensity at the source and r is the observers distance from source.

I.2) generate initial population of fireflies

$$x_{t+1} = x_t + \beta_0 e^{-\gamma r^2} + \alpha\varepsilon$$

I.3)determine the light intensity li at xi via f(xi):Now determine the light intensity of each firefly to find brightness of each firefly

I=$I_o e^{-\gamma r^2}$

I.4)calculate the attractiveness of fireflies

$\beta = \beta_0 e^{-\gamma r^2}$

I.5)movement of less brighter fireflies towards more brighter flies:this movement is determined by:

$x_i = x_i + \beta_0 e^{-\gamma r_{i,j}} + (\text{xi-xj}) + \alpha\varepsilon$

IV) PROBLEM FORMULATION

In this paper we will develop an client server architecture where user from client machine search anything the our main server. We will first check in cache server whether the data has been existed regarding the search by user. If exist then all stuff will return to the client from cache server otherwise main server. Then it will return the result to client from Web and store it's one copy to their cache server for future usages.

V) EXPERIMENTAL WORK PROCEDURE

1) First of all user will request for update from the server or download the file from server.
2) Then in this next step predict based web caching algorithm will be used to check in cache server that whether the file exist in the cache server or not if the file which was requested by user exist in the cache server then the requested file will be fetched from cache server and response will be returned back to user.
3) Otherwise if the requested file is not present in cache server then the file will be fetched from main server and one copy of the requested file will be stored on cache server and send one copy to respected user.
4) On Cache server we are running Fuzzy c means algorithm to cluster the document based on multiple request raised by user for same document.

VI) RESULT ANALYSIS

In this section, result analysis has discussed on the basis of experimental work in which we have tested out the hit and byte ratio for predefined scheme i.e using k-means clustering. Simulations have been carried out by implementing the proposed framework in Java language. In the experiment we have implemented our prefetching approach and tested out the hit and byte hit ratio. In the first experiment we have compared our approach using c-means algorithm for page search.

1) Hit Rate

The Hit rate is the ratio of the number of cache hits to the number of client requests. Greater the hit rate, greater is the efficiency of the system[15,16].

HR = Number of objects found in the cache/Total number of requests

.

	Hit ratio value(%)	Cache size(%)
Using fuzzy c means clustering	0.2	5.5
using k-means clustering	0.2	4.0

Figure no.1

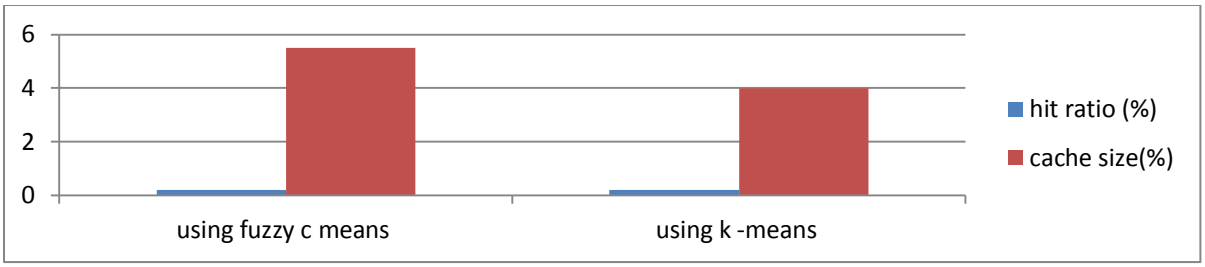

Figure no.2

Here the two diagrams 1 ,2 describes the comparison between the hit –ratio values using the different approaches in the fuzzy c-means algorithm hit ratio value is more in comparison with the predefined approach that is using k-means clustering .

2)byte hit-ratio

The ratio of bytes served by the cache over the total number of bytes requested by the clients. BHR can be significantly different from HR in a case where only few, but large files are being served by the cache.

	Byte hit ratio(%)	Cache size(%)
Using fuzzy c-means algorithm	0.2	6.0
Using k-means algorithm	0.2	4.5

Figure no.3

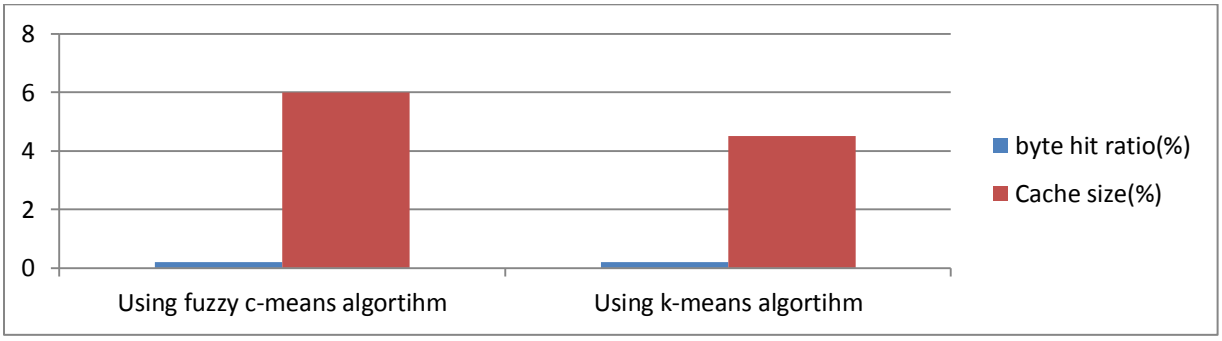

Figure no.4

As shown in the figures no. 3 and 4 the comparison is shown between the value of byte hit ratio of using two different approaches. As in new approach in which the fuzzy c –means algorithm is used to cluster the user requests the byte hit ratio value for cache size is more so this method is more effective than the previous method in which the k-mean algorithm was used.

3)Precision

Precision is the positive predictive value that gives the ratio of the good prefetch hits to the total number of prefetches or predictions made. This measure gives the exactness or the quality. It is the fraction of retrieved instances that are relevant.
Pc = Number of Good Predictions/Total Predictions

4)Recall
precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

5)F-measure

The weighted average of precision and recall is calculated called as *F*-measure.

F-measure = $2*\frac{precision*recall}{precision+recall}$

The *F*-measure obtained for different user requests is observed to be higher when precision is higher. Higher the measure, greater is the efficiency of the model.

	Number of user	recall	f-measure	Precision
Using fuzzy c means clustering	10	0.99	0.92	0.99
Using adaptive clustering	10	0.96	0.85	0.96

Figure no.5

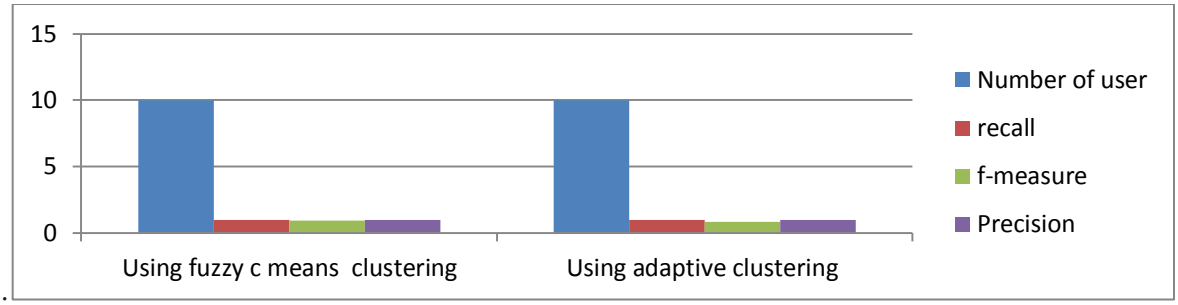

Figure no.6

The precision obtained through fuzzy c-means technique is higher compared to all other technique in which adaptive clustering is used. This is because of the accuracy of fuzzy c-means clustering. Moreover, the *F*-measure obtained for different user requests is observed to be higher for fuzzy c-means method . Further, the *F*-measure obtained through fuzzy c-means method is higher than other clustering techniques. From this, it is justified that the *F*-measure is higher when the precision value is high.

6) Response time

Response time is the total amount of time it takes to respond to a request for service from the user .that means whenever any user makes any request then how much time the system takes to response to the user request.

	Response time in seconds
Using fuzzy c means	18
prediction by partial matching	20

Figure no.7

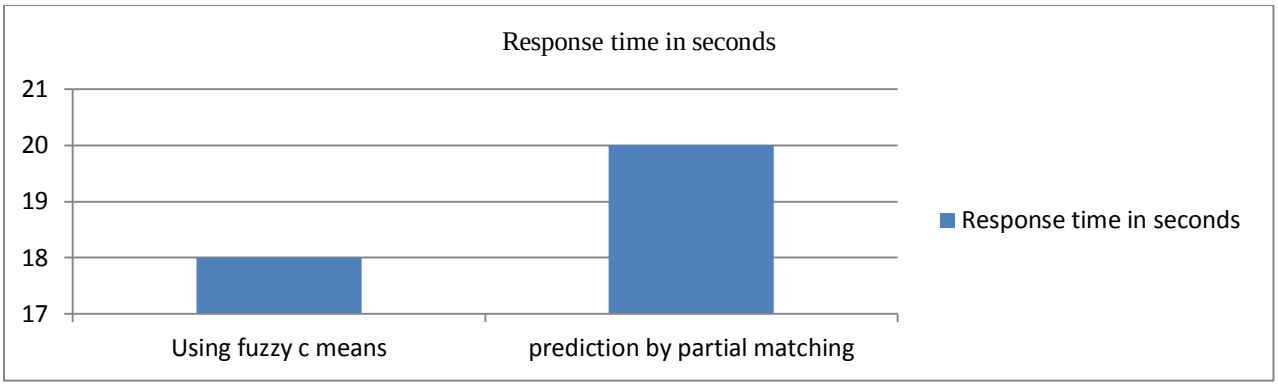

Figure no.8

In the fuzzy means clustering the response time to the user request is less than the other method so it can be determine that the in the system in which fuzzy c –means clustering is used is more effective as compared to the other method in which partial matching method was used.

VII) CONCLUSION AND FUTURE WORK

In this paper we made a cache server which will help in reducing the problems regarding the requests of user to main server. This will not only improve the response time but also will improve the accuracy regarding the requests . The results which are calculated in this paper are compared with k-means ,adaptive clustering, partial matching prediction method . After evaluation the results it can be said that the research that we have done in this paper is much more effective than the other compared techniques. For the future purpose it will effective if the push based caching will be used To keep cached data close to those clients requesting that information, in the push-based data delivery, the server tracks all proxies that have requested objects. If a web page has been modified, it notifies each proxy and when the client requests for the file, it is served from the proxy's cache instead of the request going directly to the server. This improves network utilization.

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] k.geetha, M.santhiya ,"a comparative study on data mining approachs" ,International Journal of Advanced Research in Datamining and Cloud Computing Volume 2, Issue 8, ISSN 2321-8754, August 2014.

[4] R.Suganya, R.Shanthi , "Fuzzy C- Means Algorithm- A Review", International Journal of Scientific and Research Publications, Volume 2, Issue 11, ISSN 2250-3153, November 2012 .

[5] Abdolreza Abhari, Sivarama P. Dandamudi, Shikharesh Majumdar, "Web object-based storage management in proxy caches", Future Generation Computer Systems 22,pp.16–31,2006.

[6] Survey of Clustering Data Mining Techniques Pavel Berkhin Accrue Software, Inc.

[7] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques" Elsevier Publication.

[8] 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.

[9] M.Vijayalakshmi, M.Renuka Devi, "A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3, March 2012.

[10] Bill Andreopoulos, Aijun An, Xiaogang Wang, and Michael Schroeder, "A roadmap of clustering algorithms: finding a match for a biomedical application.Brief Bioinform", February 2009.

[11] Keh-Shih Chuang , Hong-Long Tzeng, Sharon Chen , Jay Wu, Tzong-Jer Chen, "Fuzzy c-means clustering with spatial information for image segmentation", Computerized Medical Imaging and Graphics 30 ,pp. 9–15,2006.

[12] Subhagata Chattopadhyay, Dilip Kumar Pratihar, Sanjib Chandra De Sarkar, "a comparative study of fuzzy c-means algorithm and entropy-based fuzzy clustering algorithms", Computing and Informatics, Volume. 30, pp. 701–720, 2011.

[13] Pratihar, D. K, Soft Computing, Narosa Publishing House, New-Delhi, India, 2008.

[14] Xin-She Yang, "Firefly Algorithms for Multimodal Optimization".

[15] Gracia,C.D, Sudha, S. "ensemble prefetching through classification using support vector machine", Intelligent Systems Technologies and Applications, pp. 261–273. Springer, London ,2016.

[16] Berkhin, P.: "A survey of clustering data mining techniques" ,Grouping Multidimensional Data, pp. 25–71. Springer, Berlin, 2006.

