International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 4

A Survey on: Clustering Algorithms in Web Usage Mining

Anandha Jothi[1], Gayathri. P[2]
[1]PG Student, [2]Assistant Professor
Department of Computer Science & Engineering, IFET College of Engineering,
Villupuram, Tamil Nadu, India

ABSTRACT

Clustering an information is gathering an data together as indicated by their likeness. There are numerous clustering algorithms are availble here to cluster web logs. In this paper, it mainly focusing on clustering algorithm that is used for clustering web logs. Also this survey is about an algorithm which is used to cluster a web log or web usage data.

Keywords: *Clustering algorithms , web usage data, web logs*

1. INTRODUCTION

In Today life there is large volume of data are arrived in web as per user request. These data require gigantic measure of limit. Gathering those data is a giant method. Not all gathering estimations that are related to cluster a huge measure of data. A couple of figurings are unable that are used to aggregate significant measure of data. This paper is generally focusing on bundling figuring that is used for to amass web logs or web usage data. The wellspring of web utilize data are, for instance, web logs and application level logs.

2. CLUSTERING ALGORITHMS USED FOR WEB LOGS

2.1 HPSO CLUSTERING ALGORITHM

Hierarchical Particle Swarm Optimization count in which its in light of PSO(Particle Swarm Optimization) is used gathering data in perspective of its Euclidean Measure. It is familiar with partioning data into pack in 'n' chain of significance mastermind.

Steps Followed by Hpso Algorithm:

1) Swarm represents as a center of one cluster
2) Clustering swarm is represented by entire swarm.
3) Initial swarm size is kept larger.
4) Particle initialized either randomly or in uniform criteria
5) Good initial position can decrease the convergence time. Initialization done in large number of cluster are generated.

Follow criteria to initialize the particles

loc(X(i)) = i × (N/K − 1)
N - Total number of data vectors
K - Number of a patticle
i – Represents index of the particle that ranges from 1 to N.

According to this equation after initialization the particles moves to new position.

They search better position using 3 learning components such that are cognitive component , social and self organizing component

(i) Cognitive component is called as personal best position and is determined by computing pbest equation.

(ii) The social component gBest which is the knowledge of the group that comes from the experience of all the particles of the swarm is calculated by equation.

$$pBest_i(t+1) = \begin{cases} pBest_i(t) \text{ if } f(X_i(t+1)) \\ \text{is not better than } f(pBest_i(t)) \\ X_i(t+1) \text{ if } f(X_i(t+1)) \\ \text{is better than } f(pBest_i(t)) \end{cases}$$

$$gBest = argBest_{i=1}^{n}\{f(pBest_i(t))\}$$

(iii) The self-organizing is very vital as it keeps the particle inside the cluster according to each new data vector the cluster gets. The fresh location of the particle is intended by tallying the rapidity to the current position to the particle using an Equation[1].

$$V_i(t+1) = \omega \times V_i(t) + q_1 r_1(pBest_i(t) - X_i(t)) + q_2 r_2(gBest - X_i(t)) + q_3 r_3(Y_i(t) - X_i(t))$$

2.2 K-HARMONIC ALGORITHM

This algorithm which is similar to k-means but its objective is different. KHM objective function uses an harmonic mean of the distance from each navigational path to all navigational paths which enters on group of clusters.

$$KHM(X,C) = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|x_i - c_j\|^p}}$$

where p is an input parameter and must be P>=2. The harmonic mean provides a worthy score intended for one and all sequence navigational path when that sequential navigational path is adjacent to a few one focus of the group. Harmonic mean is not similar to KM minimum function. The k-Harmonic mean minimizes the square quantization errors within cluster variance.[2]

2.3 K-Means Algorithm

Clustering is an unsupervised classification or dividing of patterns into groups or subsets (i.e. clusters). Here the objects are grouped keen on modules of parallel things based on their location and connectivity within an n dimensional space. K-Means clustering algorithm is widely used in the domain of data mining for clustering the data. Now it's also used to cluster web usage data. [3]

Steps in K-Means Algorithm:

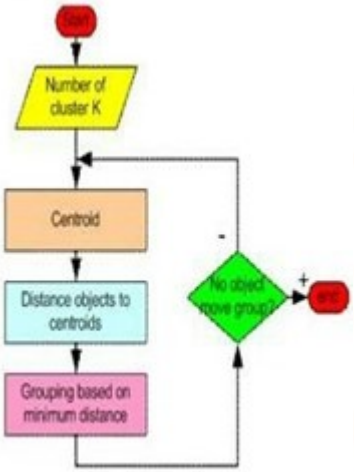
1. Using Jenks optimization algorithm it moves k centroid.

2. Criteria for doling out things to the group is having centroid that are nearest to them.

3. The new group centroid should be refreshed to the things has been incorporated or prohibited from bunches and the enrollment of things to the bunch updated.

4. Proceed with stage 3 until there are no more things that to change their group enlistment.

K-means gives more firmly gathering if bundles are globular and its count is speeder when the K is pretty much nothing.



2.4 SOM ALGORITHM

Self Organization Map which is resembling to K-means algorithm. Self Organizing Map (SOM) are those which have processes that allow automatically the internal organization to grow without being guided or controlled by any external source. SOM usually show emergent properties which allow, starting from simple rules, to acquire composite arrangements. Self organizing concept is a basis in the depiction of genetic schemes, after sub-cellular level to ecosystem level. [4]

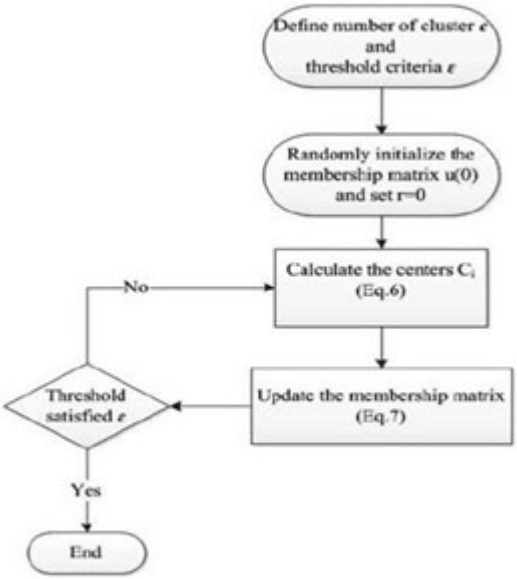
Steps followed in SOM:

1) Initialize the centroid.
2) Replicate
3) Choose the side by side object.
4) Find out the nearest centroid of the object.
5) Modify this centroid and the centroids that are near, i.e. in a specific region till the centroids do not vary much or a threshold is exceeded.
6) Allot all entity to its nearest centroid and Come back to the centroids and clusters.

It is used as a trend analysis to identify customers patterns in the process of Web Usage Mining.

2.5 FUZZY C-MEANS

Fuzzy C-means (FCM) is a scheme of overlapped grouping which permits individual data item to fit to dualistic or additional groups.



The accompanying advances clarify the working of FCM:

Input : The component vector X I that speak to the navigational examples of every client and the quantity of groups

Output: The bunches having clients with comparative access designs.

Step 1: Start
Step 2: Initialize or refresh the fuzzy parcel grid U with condition

$$U_{ij} = \frac{1}{\sum_{k=1}^{C} \left[\frac{\| X_i - C_j \|}{\| X_i - C_k \|} \right]^{2/m-1}}$$

Step 3: Calculate the inside vectors utilizing condition

$$C_j = \frac{\sum_{k=1}^{C} U_{ij}^{m} \cdot X_{ki}}{\sum_{k=1}^{C} U_{ij}^{m}}$$

Step 4: Repeat step (2) and (3) till the closure norm is fulfilled.
Step 5: Stop

The fuzzy c-implies system proceeds until the point that the end foundation is satisfied.[5]

2.6 BIC-GSA

BIC-GSA is which is based on Gravitational Search algorithm. GSA enactment is dignified by their quantity. It is hand-me-down to catch a best solution for bi-clustering problem.

Steps in BIC-GSA

1. Randomized initialization of bi-cluster as agents
2. Evaluate the fitness of the agents (bi-cluster) using equation.

$$F(I, J) = \begin{cases} |I| * |J|, \text{ if } ACV(I, J) \geq \delta \\ 0, \text{ otherwise} \end{cases}$$

3. Update best (t), worst (t) and M i (t) for i = 1, 2,...,N respectively.
4. Figuring of the aggregate power in various ways.
5. Estimation of quickening and speed.
6. Refreshing bi-cluster utilizing condition

$$x_i = \begin{cases} 1, & \text{if } r_3 < S(v_i(t+1)) \\ 0 & \text{otherwise} \end{cases}$$

7. Rehash stages 3 to 6 until the point when the stop model is come to. (i.e., the predefined number of emphasis is come to or the wellness work is fulfilled)
8. Yield the best operator, which has the best an incentive for the wellness work as the last arrangement; generally come back to the stage 2..

Bi-clustering of web utilization information utilizing gravitational hunt calculation (GSA) to extricate the exceptionally related bicluster.[6]

References

1) Shafiq Alam,Gillian Dobbie, Yun Sing Koh, Patricia Riddle (2013) ,"Clustering Heterogeneous Web Usage Data Using Hierarchical Particle Swarm Optimization".
2) R. Gobinath , M. Hemalatha (2013),"An Optimized k-Harmonic Mean Based Clustering User Navigation Patterns".
3) Hiral Y. Modi, Meera Narvekar (2015),"Enhancement of online web recommendation system using a hybrid clustering and pattern matching approach"
4) Sheetal Sahu, Praneet Saurabh, Sandeep Rai (2014), "An Enhancement in Clustering for Sequential Pattern Mining Through Neural Algorithm Using Web Logs".
5) Nayana Mariya Varghese , Jomina John (2012), "Cluster Optimization for Enhanced Web Usage Mining using Fuzzy Logic".
6) V. Diviya Prabha, R. Rathipriya ,"Biclustering of Web Usage Data Using Gravitational Search Algorithm"