International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.6., 2015
(Nov.-Dec.,)

REVIEW ARTICLE

ISSN: 2321-7758

WEB LOG DATA CLUSTERING USING ENHANCED K-MEANS ALGORITHM

AVNEESH TIWARI[1], CHETALI PARVE[2]
[1,2]NATIONAL INSTITUTE OF TECHNICAL TEACHERS TRAINING AND RESEARCH, BHOPAL (M.P)

AVNEESH TIWARI

CHETALI PARVE

ABSTRACT

Log files holds the information such as Name of the user, Access Request, IP Address, number of Bytes Transferred, Time, User Agent, Referred URL and Result Status. All these log files are managed by web servers and it has information of all records of each user request. This information of log files gives a suggestion about the user by analysing the log files. Web log mining is used to determine the patterns which can be useful for solving many real world problems like improving web sites, product recommendation, better understanding the visitor's behavior, etc. The main advantage of log files is that, the data is easily available to be analyzed. In this paper the research is going to apply the enhanced k-means clustering algorithm in web log file. These Web Log files usually contain ambiguous and noisy data. Pre-processing phase involves the removal of unnecessary data from log file by using the method of K-Means Clustering.

Keywords: Web Log, Data mining, clustering algorithms

INTRODUCTION

WEB USAGE MINING: Web usage mining is the process of extracting the useful information from logs and finding out what the users are looking for on the Internet. Different users looking different result as some users might be looking textual data, whereas some others user might be interested in multimedia data. Mining of web log is the application of techniques for finding out the interesting usage patterns from Web data, understanding them and better serving the needs of Web-based applications.

- **Application Level Data:** In application new kinds of event also defined, and logging can be turned on for generating.
- **Web Server Data**: Users logs information collected by Web server. This log data includes IP address, access time, and page reference.
- **Application Server Data**: Commercial application severs have remarkable features to facilitate e-commerce applications with very less effort. Application server logs have ability to track various kinds of events logs.

WEB LOG FILES: Web log files are automatically created and are maintained by a web server. Accessing web site by hitting it create logs including each view of an image, HTML document or other object. In the raw web log file format, for each hit there must be one log line. This log line contains the information about the users who have visited the web site, where they have come from, and exactly what they want.

The log files contain the following information's:

1. Request of the user
2. The IP address of the computer making the request.
3. Visitor's login ID
4. The request method
5. Location and name of the requested file
6. Time and date of hit
7. HTTP status code.
8. The web page which referred the hit.
9. Size of the requested file

LOG FILE TYPES: Access Log, Agent Log, Error Log and Referrer Log.

CLUSTERING WITH K MEANS

K-means algorithm is one of the unsupervised learning algorithms that determines the well known clustering problem.[1][2] The Procedure of K means algorithm is very simple and easy way to distribute a given data set to fixed priori number of clusters. The main objective is to find centroids of the cluster. Then arrange all the centroids in cunning way as different places cause different result. So it is better to place them far away from each other as much as possible. Then finalize the points which belong to dataset set and compare it to the centroid that is the most nearest. When all points covered and no points pending, the first step is completed and an early group age is done. The next step is recalculating the centers of the cluster as k new centroids come from the earlier step. With the k new centroids, make sure, is there any need to a new biding to be done with the same given data set points and the new centroid that is nearest to it. Now generate a loop at this point. The result of this loop, all centroids changes their position until there are no more changes. These centroids do not shift any more. Finally, the objective function use to minimize the aim of this algorithm, for this purpose squared error function is used. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $\| X_i^{(j)} - C_j \|^2$ explain the distance of selected data point $X_i^{(j)}$ and the cluster centreC_j, it is a display of the distance of *n* data points from their respective cluster centers. Euclidean distance between two multi-dimensional data points X and Y is described as follows:

$$D(X, Y) = \sqrt{(x1 - y1)2} + \sqrt{(x2 - y2)2} + \cdots + \sqrt{(xm - ym)2}$$

Where X = (x_1, x_2, x_3 … x_m) and Y = (y_1, y_2, y_3 … y_m) are two multi-dimensional data points.

The proposed method comes from the statement in which cluster centre moves as new points are added to or removed from it. All the process related to data points and centroid of clusters. Below figure explains the idea behind it.

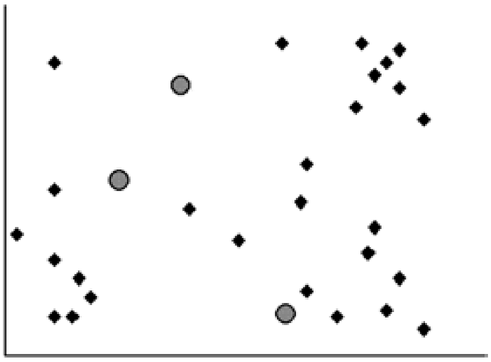

Figure1 (a) Initial centres of a given dataset

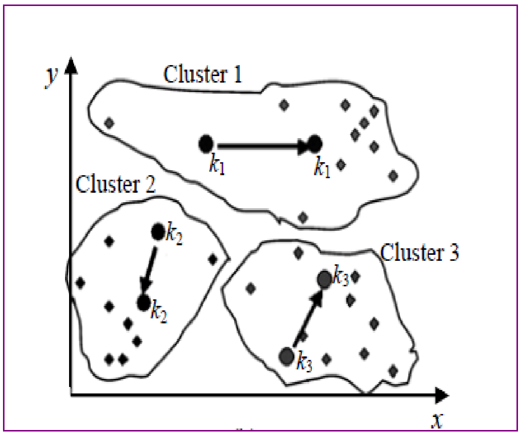

Figure1 (b) Again calculate the position of the centroids

Fig1 (a) Shows the three centroid of given dataset. Fig1 (b) explains point of dataset is distributed to the next iteration over the starting three centroids, and the new centroids. Fig1(c) shows the final clusters and their centroids.

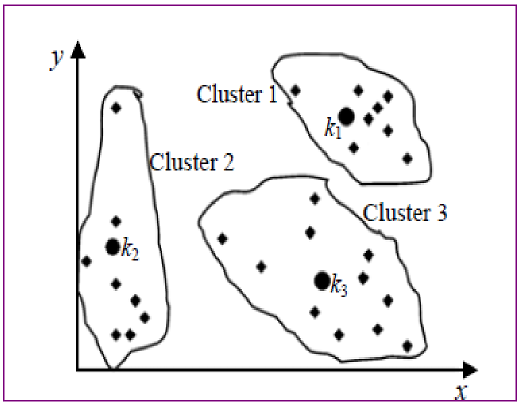

Figure1 (c) Final positions of the centroids

Advantages of K-mean clustering:

1. Processing is very simple and flexible of K-mean clustering algorithm.
2. Method to implement K-mean clustering algorithm is easy and understandable.

Disadvantages of K MEANS Algorithm [1]:

1. Fix the number of cluster in advance and then decide midpoint on the first cluster.
2. Data clustering is not working with different forms and density.
3. This method having problem to work with data collection and then it is not describable calculate average.
4. Selection number of optimal cluster for problem is difficult.

RELATED WORK:

Paper Title & Approach

1. Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity

In [3] K-mean clustering algorithm has a very common approach in which initial centroids select randomly. In This paper author proposes a new method of clustering using K-mean algorithm in which initial centroids selected instead of random selection, this method cause reduction of the number of iterations and improved the elapsed time. Author gives two phases in phase-I, fix the size of cluster in priori and get the initial cluster as output of the first phase. Here, all elements of array scanned and divide into sub array; this is forming the initial clusters. In second phase change the size of the cluster and then get the final cluster. In all clusters divide the centroids on the basis of distance from the other elements data calculated. Having same and less distance between data element remains in the same cluster or moved to suitable clusters. Repeat this process until there is no change in the clusters is detected.

2. Data Clustering with Modified K-means Algorithm

In [8] the technique of data clustering by K-means algorithm is proposed in which initial mean of cluster according to this algorithm whole data is divided into segments with frequency of data points in each segments is calculated.

3. Enhancing K-means Clustering Algorithm with Improved Initial Centre

In [5] proposed a method for finding better initial centroids and also gives an idea to present a well-organized way to allocate data points to the appropriate clusters with compact time density. This method explains more precision with less time complexity compare to original k-means clustering algorithm. In this paper authors had proposed an enhanced method that has improved the k-means clustering algorithm efficiency. But the initial centroids are selected randomly in this method. Therefore this method for the initial starting points is very sensitive and it may or may not produce the unique results of clustering. Authors proposed algorithm have proved that it is more exact and efficient compared to the original k-means algorithm. It ensures about method of the clustering in O(nlogn) time without any loss in the correctness of the clusters.

4. An Efficient K-mean Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points

In [9] authors has introduced the technique called K-means enhanced Clustering algorithm for improving the running time complexity .They developed the cluster in two phases. In the first phase they find initial cluster and then in the second phase they finalize the clusters.

5. K-means Clustering using Max-min Distance Measure

In [10] introduced the Max-min measure that is an alternative of distance measure. In this method entire data is adjusted in limits using Max-min normalization. Then normalization clustering is done.

ENHANCED K MEANS ALGORITHM:

Require: Dp = {a1, a2, a3,..., ai,..., an } // priory setup the n data points

k // priory fixed the number of desired clusters.

Ensure: k number of clusters.

Steps

1. Identify the total cluster size Si (1= i = k) by using Floor (n/k)

   Where n denotes for number of all data points Dp (a1, a2, a3, …… an) and k stands for total number of clusters..

2. Make Arrays Ak of k number
3. From the Input Array shift all data points Dp to Ak until Si =Floor (n/k).

Repeat this step until all Dp transferred from input array

Exit with having k initial clusters.

4. Calculate the distance between each data object dp (1 <= i<=n) and all k clusters k j(1 <= j<=k) and
5. Allocate data object di to the nearest cluster.
6. For each cluster j (1 <= j<=k), recalculate the cluster center.
7. until no change in the center of clusters.

Figure 2 shows the flow chart of proposed method.

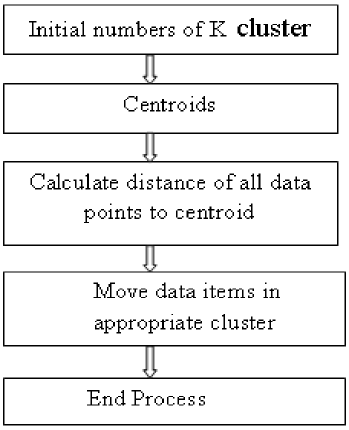
Initial numbers of K cluster
Centroids
Calculate distance of all data points to centroid
Move data items in appropriate cluster
End Process

Figure 2 Flow chart of proposed Algorithm process

Conclusion and Future Work: Web log files records all the information of each user request. Our proposed work is focusing on web log file format, its type and location. The main advantage of log files is that, the data to be analyzed is easily available. Web Log files usually contain ambiguous and noisy data. Data pre-processing is an important step to organize and filter appropriate information before using the web mining algorithm. Pre-processing involves the removal of unnecessary data from log file by K-Means Clustering method. Log files are used for the purpose of debugging. This paper presents an efficient k means algorithm for field extraction and data cleaning with process [fig.1]. Pre-processing web log files are used as input in intrusion detection system to detect intrusion and also used in data mining techniques. Efficient retrieval of knowledge from web logs, this is main motive of our approach. Future work will be focusing on the research of more efficient clustering and noise removal techniques. In Further the proposed algorithm would be more efficient by doing some research, and then it will be compared with other clustering algorithms on mining the logs.

The proposed algorithm will be implemented to provide an optimized solution for pre-processing of the log files.

References:

[1]. Amin Rostami and Maryam Lashkari "Extended PSO Algorithm for Improvment problems K-MEANS clustering Algorithm" International Journal of Managing Information Technology (IJMIT) Vol.6, No.3, August 2014

[2]. Supinder Singh, Sukhpreet Kaur "Web Log File Data Clustering Using K-Means and Decision Tree" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 8, August 2013

[3]. Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity" IDOSI Publications, 2012 DOI: 10.5829/idosi.mejsr.2012.12.7

[4]. T.Chandrasekher, K.Thangavel and E.Elayaraja " Performance Analysis of Enhanced Clustering Algorithm for Gene

Expression Data" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011 ISSN (Online): 1694-

[5]. Madhu Yedla, Srinivasa Rao Pathakota, T M Srinivasa "Enhancing K-means Clustering Algorithm with Improved Initial Center" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (2) , 2010, 121-125

[6]. M.V.B.T,Santhi, V.R.N.S.S.V.Sai Leela, P.U.Anitha, D.Nagamalleswari "Enhancing K-Means Clustering Algorithm" IJCST Vol. 2, Iss ue 4, Oct . - Dec. 2011

[7]. Fahima.M , Salema.M , Torkey F.A , Ramadan M.A "An efficient enhanced *k*-means clustering algorithm" Journal of Zhejiang University SCIENCE A ISSN 1009-3095 (Print); ISSN 1862-1775

[8]. Singh,R.V and M.P Bhatia,2011."Data Clustering with Modified K-means Algorithm" in International Technology (ICRTIT),Chennai,Tmil Nadu.

[9]. Napolean.D and P.G Lakshmi, "An Efficient K-mean Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points"in Trendz in Information Science and Computing(TISC)

[10]. Visalakshi.N,K and J. Suguna "K-means Clustering using Max-min Deistance Measure" in Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS)