Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Web Usage Data Clustering Using Improved K-Means Algorithm

[1]K. Sangeetha, [2]D. Mohanapriya, [3]D. Gowathami, [4]M. Mujeeb
[1]M.Sc., M.Phil., SSNC Arts and Science –, Kanavaipudur, Salem, Tamil Nadu, India
[2]M.Sc., M.Phil., Sri Sakthi Kailash Women's College– Salem, Tamil Nadu, India
[3]M.Sc., M.Phil., ERK Arts and Science College– Harur, Dharmapuri Dt, Tamil Nadu, India
[4]M.B.A.(IT)., M.Li.Sc., M.Phil – AVS- Arts & Science College – Salem, Tamil Nadu, India

Abstract: Web Clustering is a data mining technique, which is a demanding field of research in which its latent applications create their own special requirements. The K-means is a widely used partitioned clustering method. The benchmark K-means clustering algorithm is sensitive to the selection of the initial centroids and may converge to a local minimum of the criterion function value. K-means clustering utilizes an iterative procedure that converges to local minimum. This local minimum is highly sensitive to the selected initial partition for the K-means clustering. The classical K-means clustering algorithm is sensitive to the initial clustering centroid selection. World Wide Web (WWW) clustering it can be classified into three different types: Web Content Mining, Web Structure Mining and Web Usage Mining.The proposed improved K-means algorithm helps to identify good seeds. It avoids outlier to be seed. It also avoids seeds that are very close to each other. Thus, the proposed algorithm converges fast due to the selection of good seeds instead of random seeds.

Keywords:

I. INTRODUCTION

A cluster is a collection of objects which are 'similar' between them and are 'dissimilar' to the objects belonging to other clusters; and a clustering algorithm aims to find a natural structure or relationship in an unlabeled data set. There are several categories of clustering algorithms. The K-means clustering algorithm identifies K number of distinct clusters.

The K-means clustering algorithm is widely used for many practical applications. But the original K-means algorithm is computationally expensive and the quality of the resulting clusters heavily depends on the selection of initial centroids.
In this thesis, an improved K-means algorithm has been applied to find initial seeds (centroids).

The good initial starting points allow k-means to converge to a better local minimum; also the numbers of iteration over the full dataset are being decreased. Experimental results show that initial starting points lead to good solution reducing the number of iterations to form a cluster.Web mining is currently being used to extract the knowledge about user's requirements while visiting the internet. Web data mining is an important area of data mining which deals with the extraction of interaction knowledge from the

II. K MEANS ALGORITHM

K-means (KM) clustering is a heuristic algorithm that can minimize sum of squares of the distance from all samples emerging in clustering domain to clustering centers to seek for the minimum *k* clustering on the basis of objective function . First and foremost, the *k* as input is accepted, and then data objects which are belonging to clustering domain (including *n* data objects, *n*>*k*) are divided into *k* types.

As a result, the similarity between same cluster samples of is higher, but lower between hetero-cluster samples. *K* data objects, as original clustering centers, are randomly selected from clustering domain by KM algorithm. K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The K-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

I. K-Means Clustering Algorithm Properties

i. There are always K clusters.
ii. There is always at least one item in each cluster.
iii. The clusters are non-hierarchical and they do not overlap.
iv. Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

II .K- Means Clustering Algorithm Process

i. The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.
ii. For each data point calculate the distance from the data point to each cluster.

iii. If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.

iv. Repeat the above step until a complete pass through all the data points' results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.

v. The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intra cluster distances and cohesion

Analysis of the Performance of K-means Algorithm

I. Advantages:

i. K-means value algorithm is a classic algorithm to resolve cluster problems; this algorithm is relatively simple and fast.

ii. It provides relatively good result for convex cluster.

iii. Because the limitation of the Euclidean distance. It can only process the numerical value, with good geometrical and statistic meaning.

II. Disadvantages:

i. The K value is most important for K-means clustering algorithm. There is no applicable evidence for the decision of the value of K (number of cluster to generate), and sensitive to initial value, for different initial value, there may be different clusters generated.

ii. K-means clustering algorithm has a higher dependence of the initial cluster centers. If the initial cluster center is completely away from the cluster center of the data itself, the number of iterations tends to infinity, but also makes it easier for the final clustering results into local optimization, resulting in correct clustering results.

Toy Example

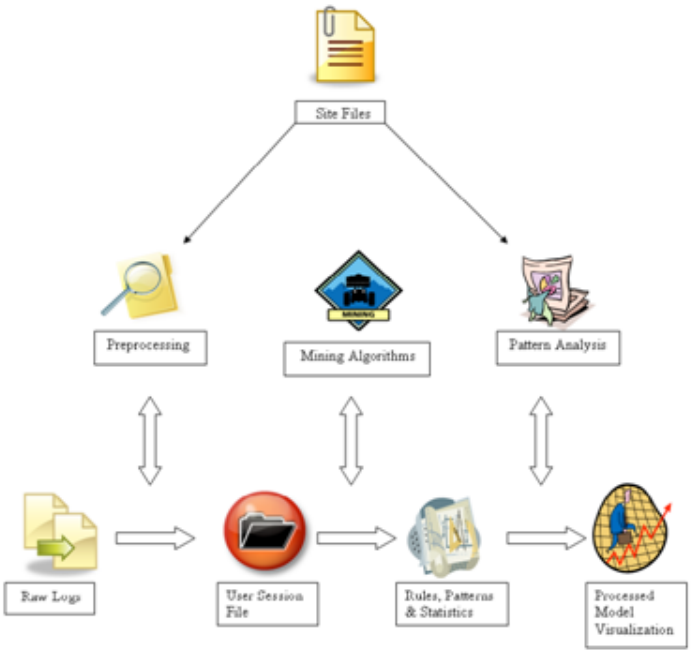
In this dataset dimensions are 21 objects with 2 dimensions. Objects are represented (o1, o2, o3……o21).

No. of clusters K = 3
No. of objects = 21
No. of Attributes = 2

Sample Dataset

O	a1	a2	o	a1	a2	o	a1	a2
o_1	2	3	o_8	20	20	o_{15}	1	1
o_2	11	14	o_9	1	2	o_{16}	50	50
o_3	3	3	o_{10}	8	8	o_{17}	100	50
o_4	2	2	o_{11}	30	25	o_{18}	40	30
o_5	15	15	o_{12}	10	30	o_{19}	50	50
o_6	13	15	o_{13}	15	20	o_{20}	12	12
o_7	25	24	o_{14}	10	10	o_{21}	60	30

Web Usage Mining Process



III. EXPERIMENTAL ANALYSIS

Data Set Information

The data comes from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for the entire day of September, 28, 1999 (Pacific Standard Time). Each sequence in the dataset corresponds to page views of a user during that twenty-four hour period. Each event in the sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail that is, at the level of URL, but rather, they are recorded at the level of page category (as determined by a site administrator).

Category	Code	Category	Code	Category	Code
front page	1	Misc	7	Summary	13
News	2	Weather	8	Bbs	14
Tech	3	Health	9	Travel	15
Local	4	Living	10	msn-news	16
Opinion	5	Business	11	msn sports	17
on-air	6	Sports	12		

In the case of the msnbc data only the rows have to be converted into item sets. A row is converted into an item set by omitting the duplicates of the pages, and sorting them regarding their codes. In this way the ItemsetCode Algorithm can be executed easily on the dataset.

Traditional K-means Algorithm

The K-means cluster algorithm was proposed by J.B. Mac Queen in 1967,which is used to deal with the problem of data clustering, the algorithm is relatively simple, so generate a widely influence in the scientific field research and industrial applications.

It is based on decomposition, using K as a parameter, divide n object into K relatively low similarity between clusters. And minimize the total distance between the values in each cluster to the cluster center. The cluster center of each cluster is the mean value of the cluster.

The calculation of similarity is done by mean value of the cluster objects. The measurement of the similarity for the algorithm selection is by the reciprocal of the Euclidean distance.

Procedure of K-means Algorithm

i. Distribute all objects to K number of different cluster at random;
ii. Calculate the mean value of each cluster, and use this mean value to represent the cluster;
iii. Re-distribute the objects to the closest cluster according to its distance to the cluster center;
iv. Update the mean value of the cluster. That is to say, calculate the mean value of the objects in each cluster;
v. Calculate the criterion function E, until the criterion function converges.
vi. Usually, the K-means algorithm criterion function adopts square criterion, be defined as;

$$E=\sum_{j=1}^{k}\sum_{i=1}^{n} \| x_i - m_j \|^2$$

In which, E is total square error of all the objects in the data cluster, x_i bellows to data object set, m_i is mean value of cluster C_i (x and m are both multi-dimensional). The function of this criterion is to make the generated cluster be as compacted and independent as possible.

Improved K-means Clustering Algorithm

The research point of K-means clustering algorithm is mainly from the following two aspects:

First, about the determination of K value. Through the above analysis, the K value of the initial cluster centers to determine the far-reaching impact throughout the clustering process and the final clustering results, while the K value in practical applications is very difficult to direct or one time determination. Especially, if the amount of data tends to infinity which is pending, the K value of the K-means algorithm to determine will be very difficult.

At present, there are two clustering algorithms to determine the K value is relatively effective which is the cost function based on distance and propagation clustering algorithm based on nearest neighbors.

In this paper, the improvement of K-means algorithm is mainly reflected in the following two aspects:

i. Optimize the initial cluster centers, to find a set of data to reflect the characteristics of data distribution as the initial cluster to support the division of the data to the greatest extent.
ii. Optimize the calculation of cluster centers and data points to the cluster

IV. SIMULATION EXPERIMENT AND RESULTS ANALYSIS

To further validate the effectiveness of the improved clustering algorithm, this thesis uses a part of the data set MSNBC for experiment , the data set name, the number of attributes, as well as the time duration for data set contained in Table

Time Duration

Algorithm	Dataset	Time
Traditional K-means	MSNBC	10.879593 sec
Improved K-means	MSNBC	0.527325sec

All of its data objects have been assigned to the corresponding class in the data set in improved clustering algorithm clustering results can be intuitive and easy to gets accuracy, which is the validity of improved clustering algorithm. Before running the algorithm, set the value of K the number of categories of the standard data set, then turns to run the traditional K-means clustering algorithm and improved K-means clustering algorithm, the experimental results as shown in Table

Accuracy of the traditional and improved K-means clustering algorithm

Algorithm	Dataset	Inter cluster distance	Intra cluster distance
Traditional K-means	MSNBC	15.33	30.23
Improved K-means	MSNBC	20.78	27.30

As seen from Table, compared to the traditional K-means clustering algorithm, the improved K-means clustering algorithm on the data set MSNBC clustering results accurate rate has improved significantly. Consider the K-means clustering algorithm, the basic idea is to make the data in the same cluster have high similarity, a relatively low similarity data will be divided into different clusters. This indicates that the improved K-means clustering algorithm clustering result, each cluster is better cluster.

V. CONCLUSION

The traditional K-means algorithm is a widely used clustering algorithm, with a wide range of applications. The proposed method, improves traditional K-means clustering algorithm by selecting good initial seed points, which is one of the challenges of K-means clustering algorithm. The simulation results on MSNBC dataset show the improved clustering algorithm is not only the clustering process clustering algorithm to reduce or even avoid the impact of the noise data in the data set object to ensure that the final clustering result is more accurate and effective.

REFERENCES

[1] Aastha Joshi and Rajneet Kaur., "*Comparative Study of Various Clustering Techniques in Data Mining*", vol(3), pp:55-57, 2013.

[2] Aarti singh and Ravi dutt mishra., " *Exploring Web Usage Mining With Scope of Agent Technology*", International Journal of Engineering Science and Technology(IJEST), vol(4), pp:4283-4289, 2012.

[3] Abdul Nazeer.K.A and Sebastian.M.P., " *Improving the Accuracy and Efficiency of the K-means Clustering Algorithm*", Vol(1), pp:1-3, 2009.

[4] Abraham.A., "*Neuro-Fuzzy systems: State-of-the –art modeling techniques,connectionist models of neurons, learning processes, and artificial intelligence*", vol(2084), pp:269-276, 2001.

[5] Arun K Pujari., "*Data Mining Techniques, University press india(Private Limited)",* vol 4(2), pp:28-34, 2005.

[6] Ahamed Shafeeq B.M and Hareesha K S., " *Dynamic Clustering of Data with Modified K-means Algorithm*", IPCSIT, Vol(27), pp:221-225, 2012.

