Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Web Document Clustering Approaches Using K-Means Algorithm

Manjot Kaur *
Master of Technology in Computer Science & Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India.

Navjot Kaur
Assistant Professor, Department Of Computer Science & Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India.

Abstract: The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure. In data mining K-means clustering algorithm is one of the efficient unsupervised learning algorithms to solve the well-known clustering problems. The disadvantage in k-means algorithm is that, the accuracy and efficiency is varied with the choice of initial clustering centers on choosing it randomly. So in this paper, less similarity based clustering method is proposed for finding the better initial centroids and to provide an efficient way of assigning the data points to suitable clusters with reduced time complexity. They mainly classified into three main categories: text-based, link-based and hybrid.

Keywords - Data mining, clustering approaches, clustering methods ,k-means clustering algorithm,

I. INTRODUCTION

The term 'data mining' refers to the finding of relevant and useful information from data bases **[01]**. That information can be used to increase revenue and cuts costs. Data mining software is an analytical tool for data analyze. It allows users to verify data from many different dimensions or angles, categorize it, and finally summarize the process of finding correlations or patterns among dozens of fields in large relational databases .One of the techniques that can play an important role towards the achievement of this objective is document clustering. In data mining, Clustering is the process of organizing data objects into a set of disjoint of unsupervised classification. Cluster analysis is classes called clusters. Clustering is an example one of the primary data analysis tool in the data mining. Clustering algorithms are mainly divided into two categories: Hierarchical algorithms and partition algorithms.

A Hierarchical clustering algorithm divides the given data set into smaller subsets in hierarchical fashion. A partition clustering algorithm partition the data set into desired number of sets in a single step.

II. WEB DOCUMENT CLUSTERING APPROACHES

There are many document clustering approaches proposed in the literature. They differ in many parts, such as the types of attributes they use to characterize the documents, the similarity measure used, the representation of the clusters etc. Based on the characteristics or attributes of the documents that are used by the clustering algorithm, the different approaches can be categorized into *i.* textbased, in which the clustering is based on the content of the document, *ii.* linkbased, based on the link structure of the pages in the collection and *iii.* hybrid ones, which take into account both the content and the links of the document.

A. Text-based Clustering

The text-based web document clustering approaches characterize each document according to its content, i.e. the words (or sometimes phrases) contained in it. The basic idea is that if two documents contain many common words then it is likely that the two documents are very similar. The text-based approaches can be further classified according to the clustering method used into the following categories: partitional, hierarchical, graph based.

1.*Partitional Clustering*

The partitional or non-hierarchical document clustering approaches attempt a flat partitioning of a collection of documents into a predefined number of disjoint clusters. Partitional clustering algorithms are divided into iterative or reallocation methods and single pass methods. Most of them are iterative and the single pass methods are usually used in the beginning of a reallocation method, in order to produce the first partitioning of the data. The most common partitional clustering algorithm is k-means, which relies on the idea that the center of the cluster, called *centroid*, can be a good representation of the cluster.

The algorithm starts by selecting k cluster centroids. The advantages of these algorithms consist in their simplicity and their low computational complexity. The disadvantage is that the clustering is rather arbitrary since it depends on many parameters, like the values of the target number of clusters, the selection of the initial cluster centroids and the order of processing the documents. (Steinbach et al., 2000)[8].

2 . Hierarchical Methods

Hierarchical clustering algorithms produce a sequence of nested partitions. Usually the similarity between each pair of documents is stored in a nxn similarity matrix. At each stage, the algorithm either merges two clusters (agglomerative methods) or splits a cluster in two (divisive methods). The result of the clustering can be displayed in a tree-like structure, called a *dendrogram*, with one cluster at the top containing all the documents of the collection and many clusters at the bottom with one document each.

By choosing the appropriate level of the dendrogram we get a partitioning into as many clusters as we wish. The dendrogram is a useful representation when considering retrieval from a clustered set of documents, since it indicates the paths that the retrieval process may follow (Rasmussen, 1992) [2].

3 .Graph based clustering

Graph based algorithms rely on graph partitioning, that is, they identify the clusters by cutting edges from the graph such that the edge-cut, i.e. the sum of the weights of the edges that are cut, is minimized.. Each node represents a document and there exists an edge between two nodes if the document similarity between documents in different clusters. The basic idea is that the weights of the edges in the same cluster will be greater than the weights of the edges across clusters. Hence, the resulting cluster will contain highly related documents. The different graph based algorithms may differ in the way they produce the graph and in the graph partitioning algorithm that they use corresponding to either of the nodes is among the k most similar documents of the document corresponding to the other node. The resulting k-nearest neighbor graph is sparse and captures the neighborhood of each document.. The advantages of these approaches are that can capture the structure of the data and that they work effectively in high dimensional spaces. The disadvantage is that the graph must fit the memory.

B. Link-based clustering

According to Kleinberg (1997)[3], 'the link structure of a hypermedia environment can be a rich source of information about the content of the environment'. The link-based document clustering approaches take into account information extracted by the link structure of the collection. The underlying idea is that when two documents are connected via a link there exists a semantic relationship between them, which can be the basis for the partitioning of the collection into clusters. The use of the link structure for clustering a collection is based on citation analysis from the field of bibliometrics. Citation analysis assumes that if a person creating a document cites two other documents then these documents must be some how related in the mind of that person. In this way, the clustering algorithm tries to incorporate the human judgement when characterizing the documents. There are many uses of the link structure of a web page collection in web IR. Croft's Inference Network Model uses the links that connect two web pages to enhance the word representation of a web page by the words contained in the pages linked to it.

C. Hybrid Approaches

The link-based document clustering approaches described above characterize the document solely by the information extracted from the link structure of the collection, just as the text-based approaches characterize the documents only by the words they contain. Although the links can be seen as a recommendation of the creator of one page to another page, they do not intend to indicate the similarity. Furthermore, these algorithms may suffer from poor or too dense link structures. On the other hand, text-based algorithms have problems when dealing with different languages or with particularities of the language (synonyms, homonyms etc.). Also, web pages contain other forms of information except text, such as images or multimedia.

As a consequence, hybrid document clustering approaches have been proposed in order to combine the advantages and limit the disadvantages of the two approaches. Pirolli et al. (1996)[4] described a method that represents the pages as vectors containing information from the content, the linkage, the usage data and the meta-information attached to each document. The method uses spreading activation techniques to cluster the collection. These techniques start by 'activating' a node in the graph (giving a starting value to it) and 'spreading' the value across the graph through its links. In the end, the nodes with the highest values are considered very related to the starting node. Then, the k-means algorithm is applied to produce the clusters. Finally, Modha & Spangler also provide a scheme for presenting the contents of each cluster to the users by describing various aspects of the cluster.

III. CLUSTERIZATION METHODS

Clustering is a very well-known technique in data mining. One of the most widely used clustering techniques is the k-means algorithms. We categorize the cluster initialization methods into three major families, namely random sampling methods,

distance optimization methods, and density estimation methods.

(i) Random sampling method

RSM follow a naïve way to initialize the seed clusters, either using randomly selected input samples, or random parameters non heuristically generated from the inputs **[9]**.

(*ii) Distance optimization method*

It is to locally minimize the intra-cluster variances without optimizing the inter cluster separation. It is a natural consideration to optimize the distances among the seed clusters before hand towards a satisfactory inters cluster separation in the output **[10]**.

(iii) Density estimation method

This category of initialization method is based on the assumption that the input data follow a Gaussian mixture distribution. Hence by identifying the dens areas of the input domain, the initialized seed clusters help the clustering method in creating compact clusters **[11]**.

IV. K-MEANS CLUSTERING ALGORITHM

This section describes the original kmeans clustering algorithm. One of the most popular clustering methods is k-means clustering algorithm **[5]**. K-means clustering is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. Clustering is defined as grouping similar objects either physical or abstract. Each created group is called a cluster. The objects inside one cluster have most similarity with each other and maximum diversity with other groups.

Clustering is one of the main task in data mining and it can be applied in various fields, including marketing, city planning, biology, earthquake studies and so on. It is a partition clustering algorithm and it is very effective in smaller datasets. First select k initial centers based on desired number of clusters. The user can specify k parameter value. Each data point is assigned to nearest centroid and the set of points assigned to the centroid is called a cluster. Each cluster centroid is updated based on the points assigned to the cluster. The process will be repeated until the centroids remain the same or no point changes clusters. In this algorithm mostly Euclidean distance is used to find distance between data points and centroids. The main drawback of K-means algorithm is the quality of the clustering results highly depends on random selection of the initial centroids. For different runs it gives different clusters for the same input data.

1. The Euclidean distance between two multidimensional
data points X=(x1, x2, x3….xm) and
Y= (y1, y2, y3…ym) is described as follows:

$d(X, Y) = \text{Sqrt}(x1-y1)^2+(x2-y2)^2+\ldots+(Xm-Ym)^2$ (1)

Mean = Sum of elements /
Number of elements…… (2)
=a1+a2+a3+.....+an/n

Algorithm1: The k-means clustering algorithm
Input:
D:{d1,d2….dn}\\set of n items
K //Number of desired clusters
Output:
A set of k-clusters.
Steps:
1. Arbitrarily choose k-data items from
D as initial centroids;
2. Repeat assigns each item di to the
cluster which has the closest centroid,
Calculate new mean for each cluster;
until convergence criteria are met.

The process, which is called "k-means", appears to give partitions which are reasonably efficient in the sense of within class variance, corroborated to some extend by mathematical analysis and practical experience **[6]**. Also, the k-means procedure is easily programmed and is computationally economical, so that it is feasible to process very large samples on a digital computer.

k-means algorithm is one of first which a data analyst will use to investigate a new data set because it is algorithmically simple, relatively robust and gives: "good enough" answers over a wide variety of datasets. The idea is to classify a given set

of data into k-number of disjoint clusters, where the value of k is fixed in advance. The algorithm consists of two separate phases: The first phase is to define k-centroids, one for each cluster. The next phase is to take each point belonging to the given dataset and associate it to the nearest centroid. Euclidean distances generally considered to determine the distance between data points and the centroids. When all the points are included in some clusters, the first step is completed and a nearly grouped. At this point they want to recalculate the new centroids, as the inclusion of new points may lead to a change in the cluster centroids. Once we find k new centroids, a new binding is to be created between the same data points and the nearest new centroid, generating a loop, the k-centroids may change their position in a step-by step manner. Eventually; a situation will be reached where the centroids do not move anymore. This signifies the convergence criterion for clustering. The k-means algorithm, probably the first one of the clustering algorithms proposed, is based on a very simple idea: given a set of initial clusters, assign each point to one of the clustering algorithms proposed, is based on a very simple idea: given a set of initial clusters, assign each point to one of them, and then each cluster center is replaced by the mean point on the respective cluster. These two simple steps are repeated until convergence. A point is assigned to the cluster which is close in Euclidean distance to the point. Although k-means has the great advantage of being easy to implement, it has two big drawbacks **[7]**. First, it can be really slow since in each step the distance between each point to each cluster has to be calculated, which can be really expensive in the presence of a large data set. Second, this method is really sensitive to the provided initial clusters, however, in recent years, this problem has been addressed with some degree of success.

V. CONCLUSION& FUTRE WORK

Data mining has a wide range of applications that are used for various purposes. One of the most popular clustering algorithm is k-means clustering algorithm, but in this method the quality of the final clusters rely heavily on the initial centroids,which are selected randomly moreover,the k-means algorithm is computationally very expensive also. As the same enhanced method also choose the initial centroids based upon the random selection.so this method is very sensitive to the initial starting points and it does not promise to produce the unique clustering results. Finally this proposed method mainly focuses on the less similarity based clustering to find the initial cluster centers efficiently. This method also reduces time complexity. In this less similarity based clustering method the initial cluster centers will not be selected randomly so accuracy will be high. The experimental results show that proposed algorithm provides better results for various datasets.The value of k; desired number of clusters is still required to be given as an input to the proposed algorithm.

REFRENCES

4to 1 BF, 98*"* Refining initial points for k-means clustering.proc.15th internet.conf.on machine learning (ICML'98).

[2] Rasmussen, E. 1992. Clustering Algorithms. Information Retrieval, W.B. Frakes & R. Baeza-Yates, Prentice Hall PTR, New Jersey.

[3] Larson, R.R. 1996. Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace. Proc. 1996 American Society for Information Science Annual Meeting.

[4] Pirolli, P., Pitkow, J., Rao, R. 1996. Silk from a sow's ear: Extracting usable structures from the Web. Proc. ACM SIGCHI Conference on Human Factors in Computing.

[5] [CS, 09],*"*k-means clustering algorithm with improved initial center, "second international workshop on knowledge discovery and data mining, wkdd, pp.790-792, 2009.

[6] F.Keller,*"*clustering", computer university saarlandes, tutorial slides.

[7] [JD, 88],*"*Algorithms for clustering data", prentice hall, New Jersey, 1988.

[8] Steinbach, M., G. Karypis, G., Kumar, V. 2000. A Comparison of Document Clustering Techniques. KDD Workshop on Text Mining.

[11] Jiawei H., and Michelin K,"Datamining: concepts and techniques", morgankaufmaan publishers, 2005.

[12] [JK, 01],"Data mining: concepts and techniques", San Francisco: Morgan kaufmaan, 2001.

[13] Kantardzic, mehmed, Datamining: concepts, models, methods and algorithms.ISMN:0471228524 John wiley&sons, 2003.

