ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

An Analysis of the Query Point in Neighborhood Service using K-means Clustering Algorithm

A.Roslin Deepa, Dr. Ramalingam Sugumar

Ph.D Scholar, Dept. of Computer Science, Christhu Raj College, Bharathidasan University, Trichy, Tamilnadu, India.

Professor, Dept. of Computer Science, Christhu Raj College, Trichy, Tamilnadu, India

ABSTRACT: In the latest improvement of web technologies, it is a very challenging issue to find web services. Recently UDDI (Universal Description, Discovery and Integration) provide keyword based searches for web services. But this kind of search functionality is fails to account for relationship between web services. Because, the users are busy by the huge number of irrelevant returned services. So the improvement of finding appropriate web services is the main thing on web services. Clustering is a principal data discovery technique in data mining that segregates a dataset into subsets or clusters so that data values in the same cluster have some common characteristics or attributes. In this paper, a group of clustering semantic algorithms like K-means clustering algorithm analysis is discussed to eliminate the irrelevant services with respect to a query known as query point service provider. The analysis presents an incremental density-based clustering technique which is based on clustering algorithm to enhance its computational complexity. This paper also contains the concept level analysis of two main problems which are introduced by keyword based search approach in query point service provider.

KEYWORDS: Clustering, Query processing, Web services, Key word based search, Query point service provider.

I. INTRODUCTION

Clustering means the act of partitioning an unlabeled dataset into groups of similar objects. The goal of clustering is to group sets of objects into classes such that similar objects are placed in the same cluster while dissimilar objects are in separate clusters (5). Basically clustering is used as a data processing technique in many different areas, including artificial intelligence, bioinformatics, biology, computer vision, data mining, data compression, image analysis, image segmentation, information retrieval, object recognition, pattern recognition, spatial database analysis, statistics and web mining (5). Any cluster analysis it should exhibit two main properties; low inter-class similarity and high intra-class similarity.

Cluster analysis is a primary method for database mining. It is used to identify natural groupings of cases based on a set of attributes (1). Cases within the same group have more or less similar attribute values. Most clustering algorithms build the model through a number of iterations and stop when the model converges, that is, when the boundaries of these segments are stabilized. Cluster analysis could be divided into hierarchical clustering and non-hierarchical clustering techniques. Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters. Hierarchical clustering can be further categorized into Agglomerative (bottom-up) and Divisive (top-down). The widely used hierarchical clustering algorithm is CURE (Clustering Using Representatives) and BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies). Non-hierarchical techniques include Partitioned clustering algorithms which obtain a single partition of the data instead of a clustering structure (2). The K-means algorithm is one of the popular data clustering algorithms in data mining. To use it requires the number of clusters in the data to be pre-specified.

The K-means algorithm requires the number of clusters to be specified by the user (6). To find a satisfactory clustering result, usually, a number of iterations are needed where the user executes the algorithm with different values of K. The

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

validity of the clustering result is assessed only visually without applying any formal performance measures. With this approach, it is difficult for users to evaluate the clustering result for multi-dimensional data sets. The performance of a clustering algorithm may be affected by the chosen value of K. Therefore, instead of using a single predefined K, a set of values might be adopted. It is important for the number of values considered to be reasonably large, to reflect the specific characteristics of the data sets (6). At the same time, the selected values have to be significantly smaller than the number of objects in the data sets, which is the main motivation for performing data clustering methodology.
The rest of the paper is structure as follows, section 1 provide the basic information and importance of data mining and K-means clustering methods. Section 2 embraces the various existing papers which are based on clustering and query analysis processing solutions. It is followed by section 3 includes the K-means algorithm details with its procedures. The next section 4 contains the query point analysis using K-means algorithm partition techniques and its evaluation of clustering solution selection schemes. Finally section 5 brings to a close with conclusion of the analysis of query point in neighborhood service using K-means algorithm methodologies.

II. RELATED WORK

In 2015, Kedar B. Sawant (7)reviews some existing methods for selecting the number of clusters as well as selecting initial centroid points. The author says, this overview of the existing methods of choosing the value of K i.e. the number of clusters along with new method to select the initial centroid points for the K-means algorithm has been proposed in their paper along with the modified K-Means algorithm to overcome the deficiency of the classical K-means clustering algorithm. The author followed by a proposed method for selecting the initial centroid points and the modified K-mean algorithm which will reduce the number of iterations and improves the elapsed time. The new method in this paper is closely related to the approach of K-means clustering because it takes into account information reflecting the performance of the algorithm. The improved version of the algorithm uses a systematic way to find initial centroid points which reduces the number of dataset scans and will produce better accuracy in less number of iteration with the traditional algorithm.
In 2015, Ahmed M. Fahim (1) presents a new method namely enhanced DBSCAN which clusters spatial databases that contain clusters of varying densities effectively. The idea is to allow varied values for the Eps parameter according to the local density of the starting point in each cluster. The clustering process starts from the highest local density point towards the lowest local density one. And the value of Eps varies according to the local density of the initial point in current cluster. For each value of Eps, DBSCAN is adopted to make sure that all density reachable points with respect to current Eps are clustered. In this paper, the author has introduced a simple idea to improve the results of DBSCAN algorithm by detecting clusters with variance in density without requiring the separation between clusters. The experimental results in this paper showed the efficiency of this method.
In 2011, Alaa H. Ahmed and WesamAshour (2) proposed an initialization method to select initial cluster centers for K-means clustering. This algorithm is based on reverse nearest neighbor (RNN) search and coupling degree. Reverse nearest neighbor search retrieves all points in a given data set whose nearest neighbor is a given query point, where coupling degree between neighborhoods of nodes is defined based on the neighborhood-based rough set model as the amount of similarity between objects. The initial cluster centers of this method is computed using this methodology are found to be very close to the desired cluster centers for iterative clustering algorithms. The author says, the application of this algorithm to K-means clustering algorithm is also demonstrated in this paper. The experiment result of this paper is carried out on several popular datasets and the results show the efficiency of this method.
In 2013, DepaPratima and NiveditaNimmakanti (4) have used pattern recognition algorithms for representing the graphs and solve the cluster identification problem using K-Means-Mode, single linkage clustering and K-Nearest Neighbor Algorithm. In this article, it shows the analysis of unsupervised learning, pattern recognition, and parallel algorithms to identify the dense clusters in the noisy data. To identify the dense cluster, the author needs to construct one of the clustering algorithms in their analysis using the software CLUMP has proved to be a highly useful tool for clustering large quantities of dataset. According to analysis of the article for a typical data clustering problem with 1,000,000 data points in 30 dimensions, CLUMP calculates in 40 minutes on 105 Intel x 86 processors. CLUMP is an open source software developing tool by adding new distance functions with software features.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

III. K-MEANS CLUSTERING ALGORITHM

K-means clustering algorithm works on the assumption that the initial centers are provided. The search for the final clusters or centers starts from these initial centers (3). Without a proper initialization the algorithm may generate a set of poor final centers and this problem can become serious if the data are clustered using an on-line k-means clustering algorithm. After having chosen the distance or similarity measure, we need to decide which clustering algorithm to apply. There are several agglomerative procedures and they can be distinguished by the way they define the distance from a newly formed cluster to a certain object, or to other clusters in the solution. The most popular agglomerative clustering (3) procedures include the following:

a. Single linkage (nearest neighbor): The distance between two clusters corresponds to the shortest distance between any two members in the two clusters.

b. Complete linkage (furthest neighbor): The oppositional approach to single linkage assumes that the distance between two clusters is based on the longest distance between any two members in the two clusters.

c. Average linkage: The distance between two clusters is defined as the average distance between all pairs of the two clusters' members.

d. Centroid: In this approach, the geometric center (centroid) of each cluster is computed first. The distance between the two clusters equals the distance between the two centroids.

In K-means clustering algorithm, the clusters are fully dependent on the selection of the initial clusters centroids (8). The basic data elements of K is selected as initial centers, then the distance of all data elements are calculated by Euclidian distance formula. In data elements it having less distance to centroids is moved to the appropriate cluster. The process id continued until no more changes are occurred in cluster. In general, there are three basic problems that normally arise during clustering namely dead centers, local minima and centre redundancy. Dead centers are centers that have no members or associated data. These centers are normally located between two active centers or outside the data range (8).

The problem may arise due to bad initial centers, possibly because the centers have been initialized too far away from the data. Therefore, it is a good idea to select the initial centers randomly from the training data or to set them to some random values within the data range. However, this does not guarantee that all the centers are equally active (9). Some centers may have too many members and be frequently updated during the clustering process whereas some other centers may have only a few members and are hardly ever updated.

The basic K-means algorithm steps are defined as follows,

Step 1: Choose a number of desired clusters, *k*.

Step 2: Choose *k* starting points to be used as initial estimates of the cluster centroids. These are the initial starting values.

Step 3: Examine each point in the given dataset and assign it to the cluster whose centroid is nearest to it.

Step 4: When each point is assigned to a cluster, recalculate the new *k* centroids.

Step 5: Repeat steps 3 and 4 until no point changes its cluster assignment, or until a maximum number of passes through the data set is performed.

IV. QUERY POINT ANALYSIS USING K-MEANS CLUSTERING ALGORITHM

In this paper, we have discussed a query point analysis service provider using K-means algorithm. This paper has consists two types of query point analysis namely, Centroid determination method and the evaluation of clustering algorithm method.

Centroid Determination Method

Given the value of K which is the number of clusters to be formed by the user as an input to the K-Means algorithm, next important work is to select K numbers of cluster centroid points from the given dataset which the normal K-Means algorithm does randomly (9). But the random selection leads to more number of iterations also choosing different centroid points every time gives different clusters thus leading to almost wrong output. To overcome these problems centroid determination method has been used which uses neighborhood distance between the points to determine the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

best possible cluster centroid points which will at least reduce the number of times K-Means algorithm need to re-iterate. Considering a dataset with n points, take the very first point of the given dataset and find out its distance to all the other (n-1) points in the dataset. Next step is to sort all the points and arrange it based on this sorted distance. Assuming that the user has entered the value of K i.e. the number of clusters to be formed, we will divide the entire dataset into k numbers of equal proportion and select the very first point of each proportion as a cluster centroid. Once this selection is done, next step is to apply the normal k-Means algorithm.

A. Partitioning of K-means clustering algorithm

Another important group of clustering procedures are partitioning methods. As with hierarchical clustering, there is a wide array of different algorithms; of these, the k-means procedure is the most important one for market research (10).The k-means algorithm follows an entirely different concept than the hierarchical methods discussed before. This algorithm is not based on distance measures such as Euclidean distance or city-block distance, but uses the within-cluster variation as ameasure to form homogenous clusters. Specifically, the procedure aims at segmenting the data in such a way that the within-cluster variation is minimized. Consequently, we do not need to decide on a distance measure in the first step of the analysis. The clustering process starts by randomly assigning objects to a number of clusters.The objects are then successively reassigned to other clusters to minimize the within-cluster variation, which is basically the (squared) distance from each observation to the center of the associated cluster. If the reallocation of an object to another cluster decreases the within-cluster variation, this object is reassigned to that cluster (10).

In the partition of K-means clustering algorithm the following steps are processed to find the query point analysis.
Step 1: Accept the number of clusters K to group data into and the dataset to cluster as input values.
Step 2: calculate the distance of first point to all other points in the dataset.
Step 2.1: Arrange all the points in the dataset according to the above sorted distance
Step 3: Divide the entire dataset into K equal proportion.
Step 3.1: Choose the first point of every proportion as the K different initial cluster centroid points.
Step 4: Examine each point in the given dataset and assign it to the cluster whose centroid is nearest to it based on Euclidean distance.
Step 5: Calculate the arithmetic means of each cluster formed in the dataset and recalculate the new K centroids.
Step 6: Repeat steps 4 and 5 until no point changes its cluster assignment, or until a maximum number of passes through the data set is performed.

B.Evaluation of clustering solution selection

The following criteria are very useful to make an evaluation choice for selecting a clustering solution (10).
Substantial: The segments are large and profitable enough to serve.
Accessible: The segments can be effectively reached and served, which requires them to be characterized by means of observable variables.
Differentiable: The segments can be distinguished conceptually and respond differently to different marketing-mix elements and programs.
Actionable: Effective programs can be formulated to attract and serve the segments.
Stable: Only segments that are stable over time can provide the necessary grounds for a successful marketing strategy.
Familiar: To ensure management acceptance, the segments composition should be comprehensible.
Relevant: Segments should be relevant in respect of the company's competencies and objectives.
Compactness: Segments exhibit a high degree of within-segment homogeneity and between-segment heterogeneity.
Compatibility: Segmentation results meet other managerial functions' requirements.
The final step of any cluster analysis is the interpretation of the clusters. Interpreting clusters always involves examining the cluster centroids, which are the clustering variables' average values of all objects in a certain cluster. This step is of the utmost importance, as the analysis sheds light on whether the segments are conceptually distinguishable. Only if certain clusters exhibit significantly different means in these variables are they distinguishable – from a data perspective, at least.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

V. CONCLUSION

In data base processing system, query point analysis in neighborhood service is the main thing to search the appropriate database. K-means algorithm is an important algorithm to cluster the large databases. Existing methods of selecting the number of clusters and the initial centroid points for K-means clustering algorithm have a number of drawbacks. In this paper we have discussed a query point analysis system by using K-means clustering algorithm. An overview of the existing methods of choosing the value of K i.e. the number of clusters along with new method to select the initial centroid points for the K-means algorithm has been proposed in the paper along with the modified K-Means algorithm to overcome the deficiency of the classical K-means clustering algorithm. In this paper it also presented a survey on analysis of personalized recommendation techniques based on clustering. The evaluation method of the cluster solution selection criteria is used to provide a simple and proficient way of method to offer the better analysis of query point in neighborhood service using K-means clustering algorithm.

REFERENCES

1. Ahmed M. Fahim, “A Clustering Algorithm for Discovering Varied Density Clusters”, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 02 Issue: 08, Nov-2015. Pp.566-573.
2. Alaa H. Ahmed and WesamAshour, “An Initialization Method for the K-means Algorithm using RNN and Coupling Degree”, International Journal of Computer Applications (0975 – 8887), Volume 25– No.1, July 2011, pp. 1-6.
3. M.P.S Bhatia, DeepikaKhurana, “Analysis of Initial Centers for k-Means Clustering Algorithm”, International Journal of Computer Applications”, Volume 71– No.5, May 2013.
4. DepaPratima and NiveditaNimmakanti, “Pattern Recognition Algorithms for Cluster Identification Problem”, Special Issue of International Journal of Computer Science & Informatics (IJCSI), ISSN (PRINT): 2231–5292, Vol. - II, Issue-1, pp. 25-32.
5. Fahim A. M., Salem A. M., Torkey F. A. and Ramadan M. A., “An efficient enhanced k-means clustering algorithm, Journal of Zhejiang University Science, 2006, pp. 1626–1633.
6. Jieming Zhou, J.G. and X. Chen, "An Enhancement of K-means Clustering Algorithm," in Business Intelligence and Financial Engineering, BIFE '09. International Conference on, Beijing, 2009.
7. Kedar B. Sawant, “Efficient Determination of Clusters in K-Mean Algorithm Using Neighborhood Distance”, International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015, PP 22-27.
8. N. Mittal, R. Nayak, M. C. Govil, and K. C. Jain, “Recommender system framework using clustering and collaborative filtering,” in Proc. 3rd Int. Conf. Emerging Trends Eng. Technol., Nov. 2010, pp. 555-558.
9. Zhang Yao, Feng Yu-qiang, “Hybrid Recommendation method IN Sparse Datasets: Combining content analysis and collaborative filtering”, International Journal of Digital Content Technology and its Applications (JDCTA) Volume6, Number10, June 2012.
10. Z. Zhou, M. Sellami, W. Gaaloul, M. Barhamgi, and B. Defude, “Data providing services clustering and management for facilitating service discovery and replacement,” IEEE Trans. Autom. Sci. Eng., vol. 10, no. 4, pp. 116, Oct. 2013.