Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 5, May 2013, pg.327 – 331

SURVEY ARTICLE

A Survey on Clustering Principles with K-means Clustering Algorithm Using Different Methods in Detail

Manpreet Kaur[1], Usvir Kaur[2]

[1]Student of masters of technology Computer Science, Department of Computer Science Engineering,
Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India

[2]Assistant Professor, Department of Computer Science and Engineering, Sri Guru Granth Sahib World
University, Fatehgarh Sahib, Punjab, India

[1] preetoor69@gmail.com; [2] usvirkaur@gmail.com

Abstract— Clustering is an essential task in Data Mining process which is used for the purpose to make groups or clusters of the given data set based on the similarity between them. K-Means clustering is a clustering method in which the given data set is divided into K number of clusters. This paper is intended to give the introduction about K-means clustering and its algorithm. The experimental results of K-means clustering and its performance in case of execution time are discussed here. But there are certain limitations in K-means clustering algorithm such as it takes more time for execution. So in order to reduce the execution, time we are using the Ranking Method and Query Redirection. And also shown that how clustering is performed in less execution time as compared to the traditional method. This work makes an attempt at studying the feasibility of K-means clustering algorithm in data mining using different methods.

Key Terms: - Clustering; K-means Clustering; Ranking method; Query Redirection

I. Introduction

In today’s highly competitive business environment, Clustering play an important role. As K- means Clustering is a method for making groups of the data set or the objects that are having similar properties. In this paper the II section includes the introduction part of Clustering and section III contains introduction about K-means Clustering algorithm. This Section also includes how in K-means algorithm the distance between the objects and mean is calculated and the methods of selecting initial points in K-means Clustering algorithm. Section IV contains main steps in K-means clustering algorithm, then Section V includes introduction about methods of algorithm. Then the last Section includes the conclusion and future scope.

II. Clustering

Clustering is a type of unsupervised learning not supervised learning like Classification. In clustering method, objects of the dataset are grouped into clusters, in such a way that groups are very different from each other and the objects in the same group or cluster are very similar to each other. Unlike Classification, in which predefined set of classes are presented, but in Clustering there are no predefined set of classes which means that resulting clusters are not known before the execution of clustering algorithm. In this these clusters are extracted from the dataset by grouping the objects in it [3].

Types of Clustering Algorithms

- Hierarchical Clustering Algorithm
- K-means Clustering Algorithm
- Density Based Clustering Algorithm

- Self-organization maps (SOM)
- EM clustering Algorithm

i. Clustering Principles

Our approach is based on two criteria: one is on the queries themselves, and the other on user clicks. The first criterion is similar to those used in traditional approaches to document clustering methods based on keywords. We formulate it as the following principle:

1. **Principle 1 (using query contents):** If two queries contain the same or similar terms, they denote the same or similar information needs. Obviously, the longer the queries, the more reliable the principle 1 is. However, users often submit short queries to search engines. A typical query on the web usually contains one or two words. In many cases, there is not enough information to deduce users' information needs correctly. Therefore, the second criterion is used as a complement. The second criterion is similar to the intuition underlying document clustering in IR. Classically, it is believed that closely associated documents tend to correspond to the same query. In our case, we use the intuition in the reverse way as follows:
2. **Principle 2 (using document clicks):** If two queries lead to the selection of the same document (which we call a document click), then they are similar. Document clicks are comparable to user relevance feedback in a traditional IR environment, except that document clicks denote implicit and not always valid relevance judgments. The two criteria have their own advantages. In using the first criterion, we can group together queries of similar compositions. In using the second criterion, we benefit from user's judgments. This second criterion has also been used in [1] to cluster user queries. However, in that work, only user clicks were used. In our approach, we combine both user clicks and document and query contents to determine the similarity. Better results should result from this combination. [4]

ii. Clustering Algorithm

Another question involved is the clustering algorithm proper. There are many clustering algorithms available to us. The main characteristics that guide our choice are the following ones:

1) The algorithm should not require manual setting of the resulting form of the clusters, e.g. the number of clusters. It is unreasonable to determine these parameters manually in advance.

2) Since we only want to find FAQs, the algorithm should filter out those queries with low frequencies.

3) Since query logs usually are very large, the algorithm should be capable of handling a large data set within reasonable time and space constraints.

4) Due to the fact that the log data changes daily, the algorithm should be incremental [4].

III. K-MEANS CLUSTERING ALGORITHM

K-means clustering is a well-known partitioning method. In this objects are classified as belonging to one of K groups. The results of partitioning method are a set of K each object of data set belonging to one cluster. In each cluster there may be a centroid or a cluster representative. In case where we consider real-valued data, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases.

Example: A cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset.

K-means is a data mining algorithm which performs clustering of the data samples. As mentioned previously, clustering means the division of a dataset into a number of groups such that similar items falls or belong to same groups. In order to cluster the database, K-means algorithm uses an iterative approach. The input in this case is the number of desired clusters and the initial means and also produces final means as output. These mentioned initial and final means are the means of clusters. If in the algorithm requirement is to produce K clusters then there will be K initial means and final means after termination of this clustering algorithm, each object of dataset becomes a member of one cluster. The cluster is determined by searching throughout the means for the purpose to find the cluster having nearest mean to the object. Cluster with shortest distanced mean is cluster to which examined object belongs. In case of K-means examined object belongs. In case of K-means algorithm, it tries to group the data items in dataset into desired number of clusters. To perform this task well it makes some iteration until some converges criteria meets. After each iteration, recently calculated means are updated such that they become closer to the final means. And at final, the algorithm converges and then stops performing iterations.

IV. STEPS OF K-MEANS CLUSTERING ALGORITHM

K-Means Clustering algorithm is an idea, in which there is need to classify the given data set into K clusters, the value of K (Number of clusters) is defined by the user which is fixed. In this first the centroid of each cluster is selected for clustering and then according to the chosen centriod, the data points having minimum distance from the given cluster, is assigned to that particular cluster. Euclidean Distance is used for calculating the distance of data point from the particular centroid.

This algorithm consists of four steps:

1. Initialization -In this first step data set, number of clusters and the centroid that we defined for each cluster.
2. Classification-The distance is calculated for each data point from the centroid and the data point having minimum distance from the centriod of a cluster is assigned to that particular cluster.
3. Centroid Recalculation-Clusters generated previously, the centriod is again repeatedly calculated means recalculation of the centriod.
4. Convergence Condition

Some convergence conditions are given as below:

4.1 Stopping when reaching a given or defined number of iterations.

4.2 Stopping when there is no exchange of data points between the clusters.

4.3 Stopping when a threshold value is achieved.

5. If all of the above conditions are not satisfied, then go to step 2 and the whole process repeat again, until the given conditions are not satisfied [2].

V. DIFFERENT METHODS FOR K-MEANS ALGORITHM

i. Ranking Method

With regards to Clustering, ranking operations are a natural way to estimate the likelihood of the occurance of data items or the objects. So we propose evaluating ranking overall design of database for student data in order to form the clusters. So Ranking function introduce new opportunities to optimize the results of K-means clustering algorithm.

Need of Ranking Method [2]

Search of relevant records or similar data search is a most popular function of database to obtain knowledge. There are certain similar records that we want to fall in one category or form one cluster. That`s why, we need to rank the more relevance student marks by a ranking method and to improve search effectiveness. In last, related answers will be returned for a given keyword query by the created index and better ranking strategy. So, this method is also having the property to find relevant records. So it is also helpful in creating clusters that are having similar properties between all data points within that cluster.

ii. Query Redirection

To accommodate complex query logic, you can implement a **redirect query**: a named query that delegates query execution control to your application. Redirect queries lets you define the query implementation in code as a static method. When you invoke the query, the call redirects to the specified static method. Redirect queries accept any arbitrary parameters passed into them packaged in a Vector.

Query Redirection in Server explorer to another server:

Query redirection:- Query redirection provides a mechanism for BI Server to determine the set of logical table sources (LTS) applicable to a logical request whenever a request can be satisfied by more than one LTS.

The Oracle BI repository shipped in Oracle Fusion applications contains metadata content for real-time reporting analysis (using Transactional Business Intelligence) and historical reporting (using BI Applications).

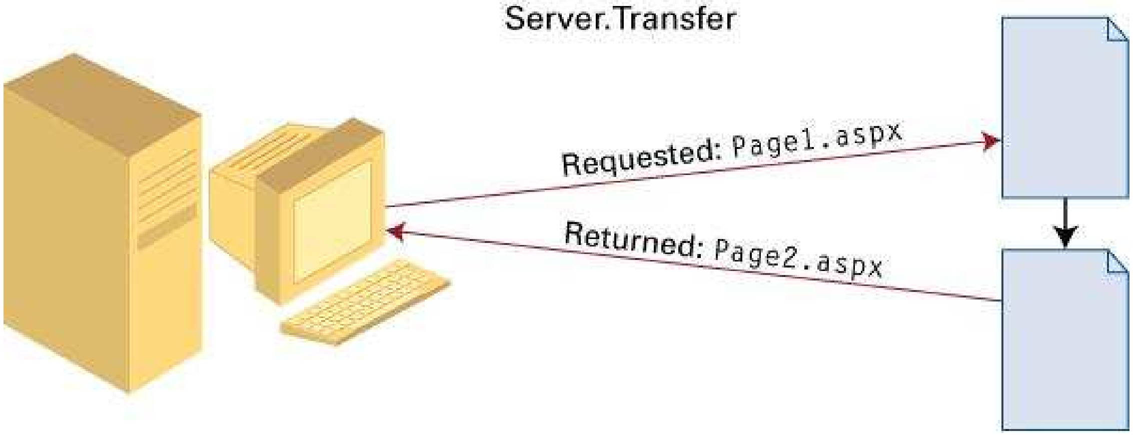
1. Set Priority group in the LTS.

Setting Priority group numbers in the LTS enables you to determine which logical table source should be used for queries in cases where there are multiple logical table sources that can satisfy the requested set of columns in the query.

The values being used for priority group are 0 through 5 for BI Applications and Transactional Business Intelligence respectively. The lower the priority group value, the higher priority it takes for being selected as the underlying source.

2. Set the session variable (REVERSED_LTS_PRIORITY_SA_VEC) and Initialization Block (IB_REVERSED_LTS_PRIORITY_SA_VEC)

A string vector session variable (REVERSED_LTS_PRIORITY_SA_VEC) is defined and initialized with subject area names for which the logical table source priority should be permanently reversed.



Query Redirection between servers [5]

Installation process Of MySQL server Redirdirection:

MySQL Proxy has no GUI, and is setup entirely via the command line. It's a handy little utility that can do a lot more than what we need it for (think: load balancing, query re-writing, and a whole heap more)

To install it, copy the binaries and their dependencies to a folder of your choosing. Then you can go ahead and install the service via the command line. This is the command I ran. It takes requests on the local IP 192.168.13.11 using port 3306 and passes them to port 192.168.13.100 using port 3306. (3306 is the default MySQL port)

sc create "MySQL Proxy" DisplayName= "MySQL Proxy" start= "auto" binPath= "C:\Program Files\MySQL\mysql-proxy\bin\mysql-proxy-svc.exe –proxy-address=192.168.13.11:3306 –proxy-backend-addresses=192.168.13.100:3306"

Before you install this service, you may want to first run this as a simple application to test your parameters work. It is a lot easier to debug this way. You can run "*C:\Program Files\MySQL\mysql-proxy\bin\mysql-proxy.exe" –proxy-address=192.168.13.11:3306 –proxy-backend-addresses=192.168.13.100:3306* from the command line and see if it works as you expected. If not, tweak the parameters and try again.

There are a lot of parameter options. You can find out more about what each of the commands do. Also, make sure you open up the correct ports on your Windows Firewall to allow incoming traffic. If you make a mistake with installing the service, you can edit it in the registry.

Go to *HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\MySQL Proxy*. Just make sure the service is stopped when you do this.

The advantages of using a MethodBasedQueryRedirector are as follows:

- You can specify the static method and its Class dynamically.
- The class that provides the static method does not need to implement Query Redirector.
- Your static method can have any name.
- You can restrict the parameters to your static method to only a Session and a Vector of arguments.

VI. CONCLUSION AND FUTURE SCOPE

This paper conclude that increasing efficiency of k mean algorithm and Users find better results corresponding to queries and Execution time also decreased.

The k-means algorithm is widely used for clustering large sets of data. But the standard algorithm does not always guarantee good results as the accuracy and efficiency is decreased in distributional environment. By using query redirection approach users of web engines got 90% correct queries related to their search in less execution time also in distributed environment as compared to simple k-mean algorithm.

REFERENCES

[1] Ahamed Shafeeq BM and Hareesha K S , "Dynamic Clustering of Data with Modified K-Means Algorithm," proceeding of the 2012 ,International Conference on Information and Computer Networks (ICICN 2012).

[2] Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur ,"EFFICIENT K-MEANS CLUSTERING ALGORITHM USING RANKING METHOD IN DATA MINING," International Journal of Advanced Research in Computer Engineering & Technology ,Volume 1, Issue 3, May2012.

[3] Ricardo Baeza-Yates1, Carlos Hurtado1, and Marcelo Mendoza2, "Query Recommendation using Query Logs in Search Engines" , IEEE,2010.

[4] Ji-Rong Wen Jian-Yun Nie Hong-Jiang Zhang, "Clustering User Queries of a Search Engine" , acm 2009

[5] Bernard J. Jansen, Danielle L. Booth Amanda Spink "Determining the User Intent of Web Search Engine Queries"

[6] K. A. Abdul Nazeer, M. P. Sebastian "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", Proceedings of the World Congress on Engineering 2009 ,Vol. WCE 2009, July 1 - 3, 2009, London, U.K.

[7] D. Napoleon & P. Ganga lakshmi, "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points", IEEE, 2010.

[8] Malay K. Pakhira, "Clustering Large Databases in Distributed Environment ", IEEE 2009 WEE International Advance Computing Conference (IACC 2009) Patialae India, 6-7 March 2009.