VISUALIZING TRENDS USING CLUSTERING ALGORITHM

Abhishek Kotian[1], Vinit Korgaonkar[2], Ankit Mhapankar[3]

Sujay Pednekar[4] , Sejal D'mello[5]

Information Technology Department, Mumbai University
Atharva College of Engineering, Malad, Mumbai,India

fabregas_0430@yahoo.com, ,*vinitkorgaonkar@yahoo.com* ,ankit.mha1807@gnail.com,pednekarsujay@yahoo.in
dmello.sejal@gmail.com

Abstract- **Organizations and firms are capturing increasingly more information about their customers, suppliers, competitors and business environment. Most of this data is multidimensional and temporal in nature. Data mining and business intelligence technique are often used to discover in such data. We propose a new data analysis and visualization technique for representing trends and temporal data using K-means clustering based approach. In our paper, we present a temporal clustered based technique with its implementation and performance [1].**

Keywords-Clustering, Data knowledge and visualization, Data mining, Temporal data mining, Trend analysis.

I.INTRODUCTION

Business intelligence applications represent an important opportunity for data mining techniques to help firms gather and analyze information about their performance, customers, competitors, and corporate environment. Data visualization and Knowledge representation tools constitute one form of business intelligence techniques that present information to users in a manner that supports business decision-making processes. In our paper, we develop a new technique for analysing and visualizing data that presents complex multi attributes temporal data in a cohesive graphical manner by building on well-established data mining methods [2].

The basic idea of our project is combating visualization difficulties of large volume of data via clustering algorithms.Section 2 gives an explanation of hierarchical clustering to manage the large amount of data. Since the project mainly focuses on K-means algorithm so with the help of section 2.1 we present a full detailed concept along with the algorithm. In section 3 we show the results and analysis of the algorithm. Since every algorithm has its advantages and disadvantages, so we have mentioned them in section 4. Finally we conclude this paper in our section 5 with the future scope of our project explained in section 6.

II.HIERARCHIAL CLUSTERING

Hierarchical clustering creates a hierarchy of Clusters which may be represented in a tree structure called a dendrogram.

A. K – Means algorithm

This algorithm takes the input parameter k and partitions are a set of n objects into k clusters. It selects k at random which is initially specified as center or cluster mean of the objects. For all the remaining objects, an object is assigned to the cluster to which it is the more similar depending on the distance between the object and cluster mean.

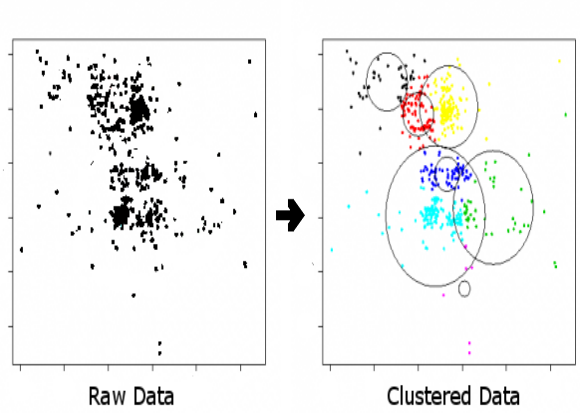

Figure 1: Clustering [3]

k-means clustering is a technique of quantization of vectors, derived from signal processing, that is renowned for cluster analysis in data mining. *k*-means clusters mainly aims to partition *n* observations into *k* clusters in which each findings is possessed by the cluster with the nearest mean,

serving as a prototype of the cluster. Which results into splitting up of the data space into voronoi cells as shown in figure 1.

B. Steps of k-means algorithm

1: Select the initial centroid points as k
2: repeat
3: Assign all points to nearest centroid from k clusters
4: Calculate the centroid of each cluster again
5: Until the centroids don't change.
The details of K-means clustering algorithm is
a) Initial centroids are often chosen randomly
b) The centroid is mean of the points in the cluster
c) Distance is measured by Euclidean distance
d) K-mean centroids move at each iteration

Finally, this algorithm aims at minimizing an objective function knows as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

Where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers.[4]

Flow Chart of k-Means Algorithm:

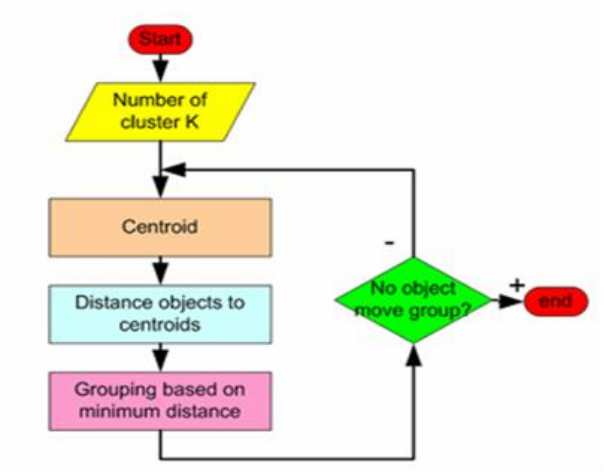

Figure 2: Flow chart of k-Mean [4]

III.RESULTS & ANALYSIS

The temporal cluster graph method is applied on share values of different companies. The share values of all different companies, basically which are in un-clustered format are grouped using k-means clustering algorithms so that the data of all different companies share values are compared at once and the comparison is shown by using some visualization techniques so that it is very easy to study the trends in share values. So the ability to identify trends in general temporal data can provide great values, such as competitive advantages to an organization performing forecasts or making decisions on future investments and strategies for the same.
The implementation process involves.

- Creation of dataset
- Clustering of relevant data
- Extracting values according to user input N - (N - no of elements)
- Performing Visualization on clustered data for single company and multiple companies.

For creation of data set the data is collected from internet and stored in the excel sheet as per our preferred order. Data is stored into dataset with the help of import and export data wizard present in the SQL-Server and irrelevant data which is present in the dataset is separated and grouped using the clustering algorithm and Values are extracted from the clustered dataset details according to the value given by the user and display it in same page through which the comparison of current trend analysis of each company with

the other companies takes place. Extracted values from the above process are given as the input to this

Module and the results are shown in the form of graph. The figures shows sample temporal cluster graph process for different company share values.

The proposed method is developed using JAVA programming language and a SQL-Server database. The Dataset is handled with the support of SQL-Server. For our work in the dataset, data is taken for five years with respect to every month for some companies [5].

An Example of Share Values of Different Companies:

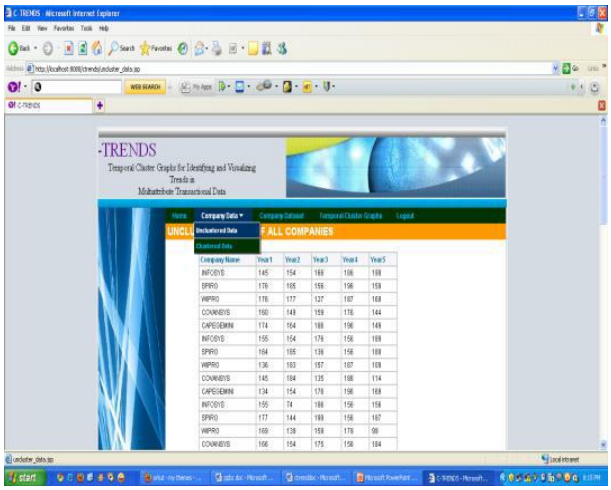

Figure 3: Un-clustered dataset details [2]

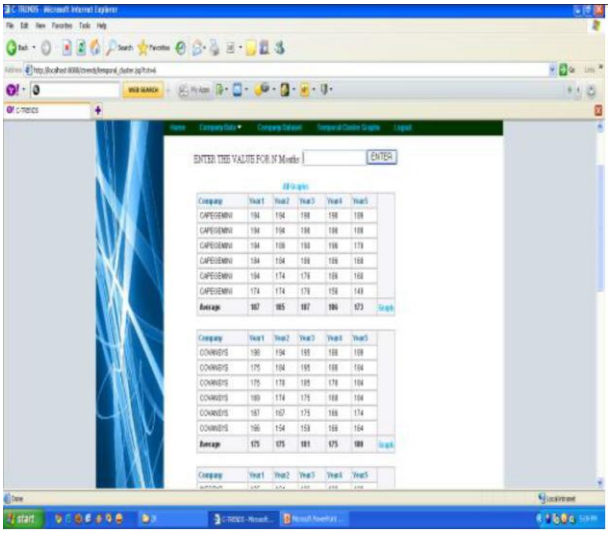

Figure 4: clustered dataset by company details [2]

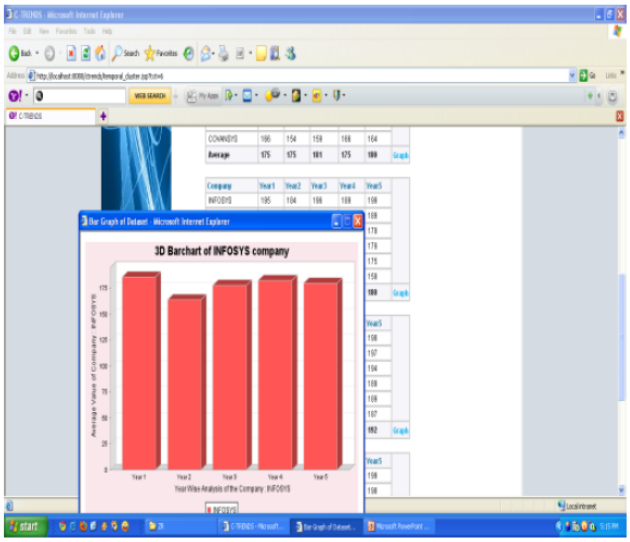

Figure 5: clustered data depends on Value N. [2]

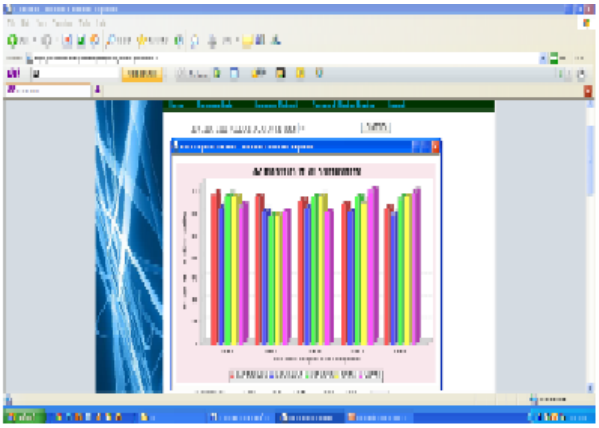

Figure 6: Comparison graph for Multi attribute representation [1]

IV.PROS AND CONS OF K-MEANS ALGORITHM

A. PROS

- Easy to understand
- Easy to implement
- Fast and robust
- K-mean clustering is extremely simple and flexible to use.
- If variables are large, then K-Means most of the times is computationally quicker than hierarchical clustering, if we keep k smalls.
- If the clusters are globular then, K-Means produce tense clusters than hierarchical clustering.
- Gives best output when data set are dissimilar or well separated from one another.

B. CONS

- The use of Exclusive Assignment - If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters.
- Algorithm is a waste for data set which are non-linear
- Algorithm is not useful for handling noisy data and outliers
- It does not work well with clusters (in the original data) of Different size and Different density
- K-means didn't work well for global clusters
- Not applicable only when mean is undefined , which means it fails for categorical data.
- Difficult to predict K-Value.

V.CONCLUSION

In this project we have learned the way of presenting the unclustered data in a proper graphical manner and how it can be benficial to the entrepreneur in setting up business of his wish in respective areas and with that we have come across a concept of how a large chunk of unclustered data can be properly organised with the help of the k-means algorithm.

VI.FUTURE SCOPE

This paper explains how to tackle the problems off visualising large amount off data.With the addition off extra features as explianed below would prove beneficial for entrepreneur.

- Shortlisting of best companies –Through the result of clustering the entrepreneur can view best companies in the respective area along with the statistics .
- Visualising companies in a specific area- Not only we can search the companies on yearly basis but with this feature the entrepreneur can view the companies in a respective area.

REFERENCE

[1] J. Abello, J. Korn, "MGV: A System of Visualizing Massive Multi-Digraphs," in IEEE Transactional Visualization and Computer Graphics, vol. 8, no. 1, pp. 21-38, Jan.-Mar. 2001.

[2] R. Agrawal, K.I. Lin, H.S. Sawhney, and K. Shim, "Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases," in Proceedings 21st Int'l Conf. Very Large Data Bases, June 2003.

[3] M.S. Aldenderfer and R.K. Blashfield, "Cluster Analysis", Sage Publications, vol. 65, February 2003.

[4] *Pritesh Vora, Bhavesh Oza, "A Survey on K-mean Clustering and Particle" , ISSN 2319-6386,vol 1, February 2013.*

[5] C. Apte, B. Liu, E. Pednault, and P. Smyth, "Business Applications of Data Mining," in Communication ACM, vol. 45, no. 8, pp. 49-53,August 2002.

[6] G.C. Battista, P. Eades, R. Tamassia, and I.G. Tollis "Temporal Cluster Graphs For visualizing trends" in Graph Drawing. Prentice Hall, vol 14,no 6,pp.1277-1284,June 1999.

[7] Jiawei Han and MichelineKamber," Data Mining Concepts and Techniques", Wiley publications.

[8] Ralph Kimball, "The Data Warehouse lifecycle toolkit" ,Wiley student edition.

