Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# Clustering Analysis based on Greedy Heuristic Algorithm

**M Ashok Kumar**,
Assistant Professor,
V R Siddhartha Engineering College,
Kanuru, Vijayawada, India

**Y.Sandeep,**
Student Dept of Information Technology
R Siddhartha Engineering College,
India

**K.Sujitha**
Student Dept of CSE,
Prasad V. Potluri Siddhartha Institute of Technology, India

***Abstract: Due to increase of the mining of compressed data from various types of applications. Clustering is the process for resolving that type of applications. Traditionally we are using different clustering applications for grouping elements with equal priority. And we represent single clustering process with multi dimensional data grouping using clustering but the only problem is performance of grouping individual elements with time queries. In this paper we are introducing greedy heuristic algorithm for analyzing multi dimensional data representation. Our approach can be worked with efficient data sharing in the commercial data compression process. By using aggregation functions in greedy heuristic algorithm for every clustering techniques to increase the performance of divisible data representation.***

***Index Terms: Clustering, Multi dimensional, Greedy Heuristic algorithm, Aggregation, information retrieval, k-means algorithm,***

## I. INTRODUCTION

Document data clustering is the increasing clustering technique for indicate our both supervised and unsupervised document organization, automatic information extraction for filtering that data. Clustering techniques can be used for automatically retrieved grouping documents. Clustering has been used for fiding latend concepts in unstructured text documents, summarize with all the collections with label data representation. Clustering usually inherently useful in organizing and searching large text, collections. for example, in automatically building an ontology like Yahoo! (www.yahoo.com). Furthermore, clustering is useful for compactly summarizing, disambiguating, and navigating the results retrieved by a search engine such as AltaVista (www.altavista.com).

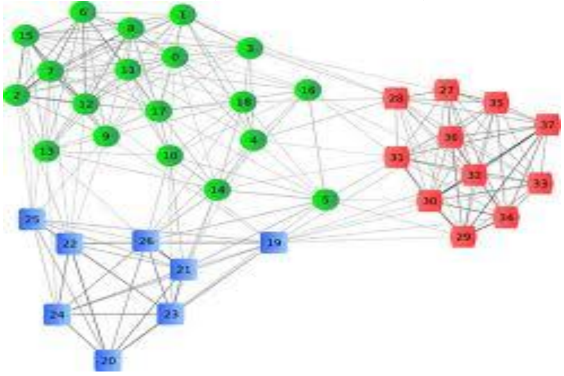


Figure 1: Data mining clustering process.

Conceptual structure generated by clustering is akin to the .Table-of-Contents. in *front* of books, whereas an inverted index such as AltaVista is akin to the .Indices. at the *back* of books; both provide complementary information for navigating a large body of information. Clustering is useful for personalized information delivery by providing a setup for routing new information such as that arriving from news feeds and new scientific applications. According to a recent study in clustering process more than half a century after it was introduced, the simple algorithm i.e., K-means still remains top 10 data mining algorithms now a days. . It is the most frequently used partition clustering algorithm in practice. Another recent scientific discussion states that k-means is the favorite algorithm that practitioners in the related fields choose to use. Needless t mention, k-means has more than a few basic drawbacks, such as sensitiveness to initialization and to cluster size, and its performance can be worse than other state-of-the –art algorithms in many domains. In spite of that, its simplicity, understandability and scalability are the reasons for its tremendous popularity. While offering reasonable results, k-means is fast and easy to combine with other methods in larger systems. An algorithm with adequate performance and usability in most of application scenarios could be preferable to one with better performance in some cases but limited usage due to high complexity. In this paper we introduce the greedy heuristic algorithm. First we get a set of maximal sub queries. We are increasing the data clustering efficiency.

## II. BACK GROUND WORK

In this approach Initially we calculate the weights of the documents and the respective multiview point similarity matrix can be constructed and then cosine similarity can calculated for the keywords in the document with the help of weight calculated for respective documents and then incremental clustering mechanism can be applied for the documents.

***MVS Similarity matrix***

We present analytical study to show that the proposed MVS could be a very effective similarity measure for data clustering. In order to demonstrate its advantages, MVS is compared with cosine similarity (CS) on how well they reflect the true group structure in document collections.

```
For i:=0 to N(total documents)
For j:=0 to N (total documents)
Simvalue :=((doc[i]*doc[j]))/Math.sqrt(doc[i]*doc[j])
Add _simvalue to the list
Build_matrix; Next Next
```

Where N is tpotal number of documents

doc[i] for i=1,2.....n are documents.

***K-Means Variant Algorithm***

Each document is checked if its move to another cluster results in improvement of the objective function. If yes, the document is moved to the cluster that leads to the highest improvement. If no clusters are better than the current cluster, the document is not moved. The clustering process terminates when iteration completes without any documents being moved to new clusters. Due to the following problems like

- fixed weight functions.
- already chosen relations do not effect the weight.

So the better system is required for doing the above functionalities in clustering

## III. PROPOSED WORK

In this section we introduce the greedy heuristic algorithm, as follows.

```
GreedyJoinOrdering-2(R = {R1, . . . , Rn},w : R, R ∗ → R)
Input: a set of relations to be joined and weight function
Output: a join order S = ǫ
while (|R| > 0)
 {
m = arg minRi∈R w (Ri, S) R = R \ {m}
S = S◦ < m >
 } return S.
```

We choose whatever choice seems best at the moment and then solve the sub problems later. The choice made by a greedy algorithm may depend on choices made so far but not on future choices or all the solutions to the sub problem. It probably select one liberal greedy choice after another, reducing each given problem into smaller one. In another words, a greedy algorithm never reconsiders its choices. This is the main difference from dynamic programming, which is exhaustive and is guaranteed to find the solution.

## IV. RELATED WORK

The main definition for clustering is to arrange data objects into separate clusters such that we are maintaining the intra-cluster similarity as well as the inter-cluster dissimilarity checking procedure. Clustering is an unsupervised learning technique, because it groups objects in clusters without any additional information: only the information provided by data is used and no human operation adds bits of information to improve the learning. The support Vectors Clustering is currently subject of active research, as early as possible stage of development. We have accurately analyzed such type of clustering methods and we have also provided some applications and contributions which allow a reduction in no.of iterations and in the computational complexity and a gain in accuracy of data.

## V. PERFORMANCE ANALYSIS

In this section we are giving input as a data set into the relevant data compression process. Using our proposed greedy Heuristic algorithm present 3 different variant steps.

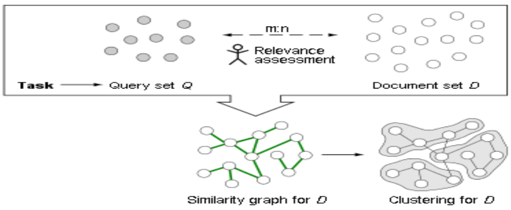
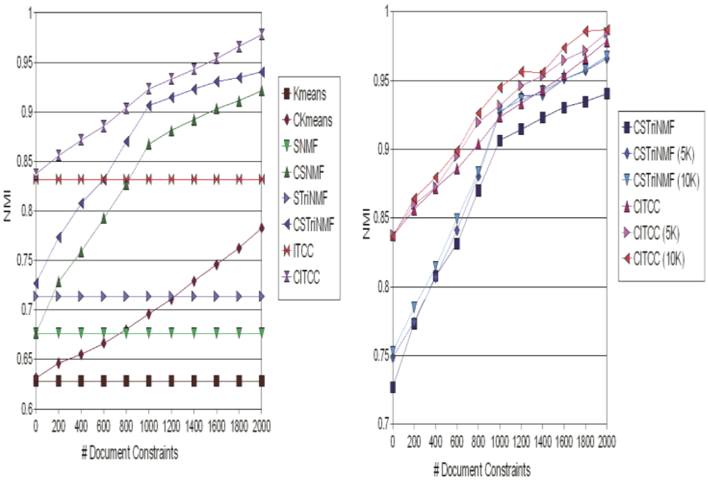


Figure 2: Informational retrieval group clustering process.

As shown in the above figure we are giving relevance dataset with equal assessment. Then it will find similarity functions of each individual graph applications in relevant variance. It will divide group of processes into clustering devices.



(a) Effects of document constraints. (b) Effects of word constraints.

Figure 3: Constrained Text Co-clustering with Supervised and Unsupervised Constraints.

Obtain the results as shown in the above figure with equivalent group clustering process.

## VI. CONCLUSION

In this paper we calculate the weights of the documents and the respective multiview point similarity matrix can be constructed and then cosine similarity can calculated for the keywords. We are introducing greedy heuristic algorithm for analyzing multi dimensional data representation. Our approach can be worked with efficient data sharing in the commercial data compression process. As further improvement we calculating individual cluster results using unsupervised data sets.

## REFERENCES

1) Geetha Sri Yendrapati* , R.N.V.Jagan Mohan,” An Efficient Heuristic Algorithm for Cluster Analysis”, *Geetha Sri Yendrapati et al. / IJAIR ISSN: 2278-7844.*
2) Inderjit S. Dhillon, Dharmendra S. Modha,” Concept Decompositions for Large Sparse Text Data using Clustering”, © 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*
3) El˙zbieta P_ekalska, Artsiom Harol, Robert P.W. Duin,Barbara Spillmann, and Horst Bunke,” Non-Euclidean or Non-metric Measures Can Be Informative”, D.-Y. Yeung et al. (Eds.): SSPR&SPR 2006, LNCS 4109, pp. 871–880, 2006. c_Springer-Verlag Berlin Heidelberg 2006.
4) Shi Zhong, ” Efficient Online Spherical K-means Clustering”, eeexplore.ieee.org › ... › Neural Networks, 2005.
5) Rajeev Guptha,Fellow, “Query Planning for Continuous aggregation queries”, IEEE transactions on knowledge and data engineering volume24,issue:6,2012.
6) http://en.wikipedia.org/wiki/Greedy_algorithm.