# Implementation Of ROCK Clustering Algorithm For The Optimization Of Query Searching Time

Ashwina Tyagi
Department of Computer Science
Amity University, Sector-125
Noida, India
ashi3110@gmail.com

Sheetal Sharma
Department of Computer Science
Amity University, Sector-125
Noida, India
sheetal.meenu@gmail.com

*Abstract*— **Clustering is a data mining technique of grouping similar type of data or queries together which helps in identifying similar subject areas. The major problem is to identify heterogeneous subject areas where frequent queries are asked. There are number of agglomerative clustering algorithms which are used to cluster the data. The problem with these algorithms is that they make use of distance measures to calculate similarity. So the best suited algorithm for clustering the categorical data is Robust Clustering Using Links (ROCK) [1] algorithm because it uses Jaccard coefficient instead of using the distance measures to find the similarity between the data or documents to classify the clusters. The mechanism for classifying the clusters based on the similarity measure shall be used over a given set of data. This method will make clusters of the data corresponding to different subject areas so that a prior knowledge about similarity can be maintained which in turn will help to discover accurate and consistent clusters and will reduce the query response time. The main objective of our work is to implement ROCK [1] and to decrease the query response time by searching the documents in the resulted clusters instead of searching the whole database. This technique actually reduces the searching time of documents from the database.**

***Keywords- Heirarchical clustering;Jaccard coefficient; Algorithm.***

## I. Introduction

The Data Warehouse is a queryable source of data in the enterprise. It comprises of data which is subject oriented, integrated, time-variant and non volatile on which enormous number of queries can be requested by the users. There may be a practicability that the requested queries belong to same subject area. However, it is quite troublesome and time consuming to provide the response even in case of similar queries. Hence the solution to this problem is to make use of the concept of query clustering for the purpose of identifying group of queries belonging to same subject areas. This paper exemplify a clustering method that uses ROCK hierarchical clustering algorithm [3] which follows a more comprehensive approach to clustering that is, two similar points have similar neighborhoods, then only the two points can be merged together in the same cluster. The similarity between two documents is analyzed by Jaccard coefficient [4]. The goal of clustering is to group the similar user data together. This is done to gain the insight of the similar interests of the users and to use this gained knowledge for the future user's queries. The need to process a set of data together often arises in scientific database systems, large bibliography retrieval systems etc. The frequently asked queries (FAQs), index-term selection, query reformulation are the major application areas of clustering. For example, if a number of users have already indicated that a query is related to a particular subject area, Hotels, Banking, for instance then that particular data can be further relevant for the identification of subject area of common interest.

## II. Limitation of traditional clustering with categorical data

Experiments show that the distance measures cannot lead to high-quality clusters when clustering categorical data. Also, most clustering algorithms merge most similarity points in a single cluster at each step and this "localized" approach is prone to errors. The answer to the stated problem is ROCK [1], which takes a more global approach to clustering that is, if two similarity points having similarity neighbourhoods, then only the two points can be merged to the same cluster. Robust Clustering Using Links (ROCK) hierarchical clustering algorithm [7] along with the Jaccard coefficient [4] is being used to determine the group of different subject areas and to obtain the similarity among the various data. The main idea behind this process is to cluster the similarity data together. ROCK algorithm is best suited for clustering categorical data because it does not use distance measures instead of it uses coefficient to find the similarity between the two data.

## III. INTRODUCTION TO HIERARCHICAL CLUSTERING

The methods for hierarchical clustering can be classified as either being agglomerative (bottom-up) or divisive (top-down), based on how the hierarchical decomposition is formed. To compensate for the rigidity of merging or splitting, the quality of hierarchical agglomeration can be improved by analyzing object linkage at each hierarchical partitioning (such as in ROCK and Chameleon), or by first performing micro clustering (that is, grouping objects into micro clusters) and then operating on the micro clusters with other clustering techniques, such as iterative relocation (as in BIRCH). But ROCK plays an important role in clustering as it defines links between neighbors and thus results in good quality clusters. Here is a brief introduction about the ROCK algorithm.

ROCK (RObust Clustering using linKs)[1][6] performs agglomerative hierarchical clustering and explores the concept of links for data with categorical attributes. The various attributes are defined below:-

- **Links** - The number of common neighbours between two objects.

- **Neighbors** - If similarity between two points exceeds certain similarity threshold (θ), they are neighbours i.e., if similarity(A,B)≥θ then only two points A, B are neighbours, where similarity is a similarity function and θ is a user-specified threshold.

- **Criterion Function** - The objective is to maximize the criterion function to get the good quality clusters. By maximizing we mean maximizing the sum of links of intra cluster point pairs while minimizing the sum of links of inter cluster point pairs.

$$E_l = \sum_{i=1}^{k} n_i \times \sum_{p_q, p_r \in C_i} \frac{link(p_q, p_r)}{n_i^{1+2f(\theta)}}$$

Where Ci denotes cluster i, ni is the number of points in Ci, k is the number of cluster, θ is the similarity threshold.

- **Goodness Measure** - While performing clustering the motive of using goodness measure is – to maximize the criterion function and to identify the best pair of clusters to b merged at each step of ROCK.

$$g(C_i, C_j) = \frac{link[C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}}$$

## IV. IMPORTANCE OF JACCARD'S COFFICIENT

Jaccard's coefficient is a good similarity measure because it can find the similarity between the categorical data. For sets A and B of keywords used in the documents, the Jaccard coefficient [4] may be defined as follows:

Similarity (A, B) = (|A ∩ B|) / (A U B|)

After determining the similarity for each pair of documents in the sample data set of N documents, the calculated similarity will be represented in the form of N x N Similarity Matrix, which will be transformed into an Adjacency Matrix with the help of similarity threshold θ≥0.4 (i.e., if similarity (A, B)≥θ then only two documents A, B are neighbours). The Adjacency Matrix is then multiplied by itself (i.e. A x A) to generate Link Matrix. Finally, the approach is to apply Criterion Function and Goodness Measure in an iterative fashion until we get clusters of similar documents representing different subject areas as well as noise, if any. Jaccard coefficient also known as Tanimoto coefficient is the best suited similarity coefficient for finding the similarity between the categorical data. It finds out the similarity by finding the intersection among the two documents divided by the union of the two documents [6]. It works on the mechanism of finding the similar strings among the two documents. If the value of the strings matched between the two documents are more, then they are similar to each other but if the value is less then they both are dissimilar. The Jaccard's value lies between 0 and 1. And if the value is 0 then both the documents are different and if the value is 1 then both the documents are just same. A threshold value has to be defined to get the desired results of similarity.

## V. ROCK ALGORITHM

ROCK algorithm is the best suited algorithm for clustering categorical data because it may use Jaccard or Cosine similarity coefficients to find out the similarity between the two data points and moreover it uses the idea of links to determine the neighbors. It is difficult to manage and handle the large chunks of data; therefore clustering can help grouping them in order. What we have observed in general is that the task of finding or searching some document out of the large amount of data is cumbersome. Also, the response time of searching the document is very high due to high scale searching among the data. So our approach is to cluster the data in order to divide the data into some groups with similar features and hence to decrease the query response time by searching the clusters obtained instead of whole database or data warehouse.
This project works in two steps:-
1. Clustering of data by ROCK algorithm and to store the clusters.
2. Reducing the query response or query search time by providing the results from the obtained clusters instead of the database.

### A. *Input*

A sample set of documents. Number of k clusters to be found. The similarity threshold for this task: θ≥0.4.

### B. *Method*

```
Do for All Data
{
Take k and θ≥0.4
Begin
```

1. Initially, place each document into a separate cluster.

2. Construction of Similarity Matrix: Constructing the similarity matrix by computing similarity for each pair of queries (A,B) using measure for instance i.e.
   Similarity (A, B) = (|A ∩ B|)/(A U B|)

3. Computation of Adjacency Matrix : Compute Adjacency Matrix (A) using similarity threshold θ≥0.4 i.e.
   if similarity(A, B)≥θ then 1;
   else 0

4. Computation of Links: Compute Link Matrix by multiplying Adjacency Matrix to itself i.e. A x A to find the number of links.

5. Calculation of Goodness Measure: The goodness measure for each pair of documents is calculated by using the following function:

$$g(C_i, C_j) = \frac{link[C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}}$$

Where f (θ) = (1-θ)/(1+θ).

6. Merge the two documents with the highest similarity (goodness measure).

7. When no more entry exists in the goodness measure table then stop algorithm by resulting in k number of clusters and noise (if any) otherwise go to step 4.

### *C.* Output

A group of documents i.e. clusters.

## VI. APPLYING ROCK ALGORITHM TO THE DATASET

To implement the ROCK algorithm we have taken a categorical dataset i.e. some documents containing the data. We have taken different type of data with different attributes as it will check the functionality and applicability of this algorithm. This dataset contains documents related to various conferences and journals (national and international). As this data is heterogeneous in nature so we took this challenge to make some clusters out of it which will decrease the complexity of this dataset. We wanted to get the clusters of similar data. The aim was to decrease the intra-cluster similarity and to increase the inter-cluster similarity. Similarity between the two documents is calculated by using Jaccard coefficient. The similarity values lies in between 0 and 1. So we set the threshold value to 0.4 so that we can get the value of adjacency matrix. The values above 0.4 will be converted to 1 and the value below 0.4 will be converted to 0. This means the higher the similarity value, the more is the relation between the documents and lesser value of similarity denotes some or no relation between the two documents. Jaccard coefficient finds out the similarity between each and every documents present in the dataset. The next step is to calculate the adjacency matrix that is by converting the lesser value to 0 and higher value to 1. After getting the adjacency matrix, our aim is to find the link matrix. Link matrix can be obtained by multiplying the adjacency matrix with itself. Clustering points based only on the closeness or similarity between them is not strong enough to distinguish two not so well-separated clusters because it is possible for points in different clusters to be neighbors. The link-based approach adopts a global approach to the clustering problem. It captures the global knowledge of neighboring data points into the relationship between individual pairs of points. Since the ROCK clustering algorithm utilizes the information about links between points when making decisions on the points to be merged into a single cluster, it is very robust. Goodness measure; while performing clustering the motive of using goodness measure is – to maximize the criterion function and to identify the best pair of clusters to be merged at each step of ROCK. This is an iterative step because clusters are merged according to the goodness measure values. At every iteration the value of goodness measure increases and the more clusters are merged. ROCK works on agglomerative bottom up approach so there are major chances that only a single cluster is obtained in the end, so to avoid this consequence we have applied a threshold value, after which the merging process of clusters stops and the desired number of clusters can be obtained.

## VII. ADJACENCY MATRIX

The adjacency matrix is calculated from the similarity matrix. The similarity matrix is calculated using Jaccard coefficient. After selecting the threshold value theta= 0.4, we converted the value greater than 0.4 to 1 and the value lesser than 0.4 to 0. This is how we got the Adjacency matrix.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

Figure 1. Adjacency Matrix.

## VIII. Computation of Link matrix

Clustering points based on only the closeness or similarity between them is not strong enough to distinguish two not so well-separated clusters because it is possible for points in different clusters to be neighbors. The link-based approach adopts a global approach to the clustering problem. It captures the global knowledge of neighboring data points into the relationship between individual pairs of points. Thus, since the ROCK clustering algorithm utilizes the information about links between points when making decisions on the points to be merged into a single cluster, it is very robust. Link Matrix is calculated by multiplying Adjacency Matrix to itself i.e. A x A to find the number of links between the two documents. The Link Matrix of few documents is shown below:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 |
| 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 |
| 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

Figure 2. Link Matrix

## IX. Goodness measure

The best clustering points were those that resulted in the highest values for the criterion function. Since our goal is to get a clustering that maximizes the criterion function, we use a measure similar to the criterion function in order to determine the best pair of clusters to merge at each step of ROCK's hierarchical clustering algorithm. The pair of clusters for which the above goodness measure is maximum is the best pair of clusters to be merged at any given step. It seems intuitive that pairs of clusters with a large number of cross links are, in general, good candidates for merging. However, using only the number of cross links between pairs of clusters as an indicator of the goodness of merging them may not be appropriate. This naive approach may work well for well-separated clusters, but in case of outliers or clusters with points that are neighbors, a large cluster may swallow other clusters and thus, points from different clusters may be merged into a single cluster. This is because a large cluster typically would have a larger number of cross links with other clusters [1].

| Pair | Goodness Measure |
|---|---|
| C0, C1 | 0.62 |
| C2,C0 | 0.62 |
| C3,C4 | 0.62 |
| C5,C6 | 0.62 |

Figure 3. Goodness Measure

## X. DISCUSSION ABOUT THE RESULTS

This algorithm results in the specified number of clusters. The number of clusters depends on the user specified criteria. ROCK algorithm gives provision to specify the number of clusters. As this algorithm works on bottom up approach and the merging of clusters is done iteratively so it is mandatory to initialize the value of k otherwise it will result in a single cluster. In our experiment we initialized the value of k with 4. So we got four clusters with high intra cluster similarity and low inter cluster similarity. ROCK also identified some noise which was totally dissimilar with every other document. The clusters obtained were of good quality as in the related documents were merged in the same cluster and the unrelated documents were merged in the different clusters. We incorporated the query searching module in this project which matches the keywords of the query with the documents. Query optimization is done by firing the query on the resulted clusters instead of the initial database. Because of the clustering mechanism, the query response time has been eventually decreased. It just took one second to compute the clusters and provided exact results according to the query being fired. So we found Robust Clustering using links algorithm is the best suited algorithm for clustering the categorical data as it provides the exact results.

## XI. CONCLUSION AND FUTURE WORK

Various clustering algorithms [7] had been used to cluster categorical data but the main problem with these algorithms is that they make use of distance measures which in turn does not produce high quality clusters. Thus, the general structure of the ROCK [1][6] hierarchical clustering algorithm has been used which performs agglomerative hierarchical clustering and it merges two points to the same cluster only when having similar neighborhoods. Clearly, if clusters are to be meaningful, the similarity measure should be selected precisely. ROCK assumes that similarity function is normalized to return a value between 0 and 1. The target is to maximize the criterion function so that the intra cluster similarity can be maximized and inter cluster similarity can be minimized. Therefore, the suitable choice for basket data would be: $f(\theta) = (1-\theta)/(1+\theta)$. And during clustering, goodness measure function helps in merging clusters which have highest goodness measure at each step of ROCK algorithm with the intent of maximizing criterion function. In our future project we will try to implement more clustering algorithms.

## REFERENCES

[1] Sudipto Guha, Rajeev Rastogi and Kyuseok Shim, “ROCK: A robust clustering algorithm for categorical attributes”.In: IEEE Internat. Conf. Data Engineering, Sydney, March 1999.

[2] I Dutta M,Kaskoti Mahanta A,Pujari Arun K, “QROCK:A quick version of the ROCK algorithm for clustering of categorical data,” Pattern Recognition Letters,Vol.26,Nov.2005, pp. 2364-2373, doi: 10.1016/ j.patrec. 2005. 04. 008

[3] Z. Huang. “A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining”, CSIRO Mathematical and Information Sciences, Australia.

[4] Anna Huang, “Similarity Measures for Text Document Clustering”, Volume: 2008, Issue: April, Pages: 49–56, Mendeley.

[5] Shyam Boriah Varun Chandola Vipin Kumar, “Similarity Measures for Categorical Data: A Comparative Evaluation”, 2008 Volume: 30, Issue: 2, Publisher: Citeseer, Pages: 3.

[6] Rizwan Ahmad,Dr. Aasia Khanum,Document, “Topic Generation in Text Mining by Using Cluster analysis with EROCK”, 2010, International Journal of Computer Science & Security (IJCSS), Volume (4) : Issue (2).

[7] Rui Xu, Donald Wunsch “Survey of clustering algorithms”, Volume: 16, Issue: 3, Publisher: Institute of Electrical and Electronics Engineers, Inc, 445 Hoes Ln, Piscataway, NJ, 08854-1331, USA,, Pages: 645-678.

[8] Florian Beil,Martin Ester,Xiaowei Xu, “Frequent Term-Based Text Clustering” ,in 2002 Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.

[9] Athena Vakali , Jaroslav Pokorny , Theodore Dalamagas, “An Overview of Web Data Clustering Practices”,2004, Proceedings of the 9th International Conference on Extending Database Technology - EDBT’04, Springer-LNCS 3268.

[10] Qiongbing Zhang, Lixin Ding, Shanshan Zhang, “A Genetic Evolutionary ROCK Algorithm” 2010 International Conference on Computer Application and System Modeling (ICCASM 2010).

## AUTHORS PROFILE

Ashwina Tyagi has received her B.Tech degree in Information Technology from Vishveshwarya Institute Of Engineering & Technology (UPTU) in the year 2009. Currently she is pursuing M. Tech degree in Computer Science and Engineering from the Department of Computer Science and Engineering in Amity University, Uttar Pradesh, India (2010-2012).

Sheetal Sharma is an Assistant Professor in the Department of Computer Science and Engineering in Amity University, Uttar Pradesh, India.