Performance based Analysis and Comparison of Multi-Algorithmic Clustering Techniques

Rajesh N. Phursule
Student, M.Tech.(CST)
Shivaji University, Kolhapur.

P. C. Bhaskar
Coordinator, M. Tech. (CST)
Shivaji University, Kolhapur.

ABSTRACT
Clustering the documents based on similarity of words and searching the text is major search procedure and widely used for large set of documents. Documents can be clustered using many clustering algorithms such as Nearest Neighbor, K-Means, Hierarchical, Graph Theoretic etc [4] [5] [7]. The performance measurement in terms of space complexity and execution time and searched output in terms of accuracy and redundancy of these algorithms is a needful study [3]. This paper mainly focuses on performance measurement of Nearest Neighbor, K-Means and Hierarchical agglomerative clustering algorithms on text documents as well as compares them in terms of space complexity, execution time, accuracy and redundancy. In particular, preprocess the input text document and convert it into the document graph represented in the form of matrix. Then convert that document graph into relation matrix which gives relation (similarity score) among all the nodes from 0 to 1 [2]. Implementation and the results of applied clustering algorithms (Nearest Neighbor, K-Means and Hierarchical agglomerative) on documents are discussed and implemented here.

General Terms
Clustering algorithms, Performance Measurement, Space Complexity, Execution time, Accuracy, Redundancy

Keywords

1. INTRODUCTION
Research in performance based analysis and comparison of different clustering algorithm is becoming a very significant in order to provide a good summary result to the different types of users [1] [11]. Different types of data mining techniques are used to mine the appropriate data from the document or the set of documents [4] [7]. This paper mainly focuses on clustering techniques. Some of these clustering techniques are K-Means, Nearest Neighbor, Agglomerative Hierarchical and Graph Theoretic etc [1] [16] [17]. Many efforts have been devoted to search new data mining techniques to get good results, but less attention has been devoted to analyze and compare them to get the effective technique for a particular application. In particular, make analysis of different clustering techniques and compare them in terms of space complexity, execution time, redundancy and accuracy. Final results showed the trade off criteria to choose the clustering techniques for a particular application [3]. Trade of criteria contains accuracy, space complexity, computational loops and redundant data in the result. In short make a framework/model which suggests best clustering technique for a particular application as well as any new clustering technique get compared with the existing.

This paper is organized as follows. Section 2 discusses the previous research work and proposed method to analyze and compare above mentioned three clustering techniques. Section 3 discussed related work. It contains preprocessing of input document and converts it into the relation graph in the form of matrix [2]. It describes the applicability of clustering algorithms on that. It converts the relation graph into the number of clusters [1] [2]. It describes the performance based comparison and trade-off criteria. The results of the experiment evaluating the performance of the three clustering techniques are represented in section 4 and 5. Finally section 6 draws the conclusion and future development work.

2. METHOD
The experimental work architecture is depicted in Figure 1. It shows the basic building blocks for performance based analysis and comparison of Nearest Neighbor, K-Means and agglomerative hierarchical clustering techniques [2] [3]. The basic idea is that preprocessing stage separates the document into a document graph. A document graph contains the node number and data of that node. Then that document graph gets converted into the relation graph. The relation graph shows the similarity score between all the nodes of the document graph between 0 as lowest and 1 as highest [2]. Then three clustering algorithms named as Nearest Neighbor, K-Means and agglomerative hierarchical are applied on the relation graph [1] [16] [17]. This converts relation graph into the number of clusters. To mine the data from the clusters of the document, apply the desired query separately on the cluster set formed by all three clustering techniques mentioned above. It gives the resultant clusters as output. Compare results in terms of space complexity, computational loops, accuracy and redundancy. It gives the trade off criteria for users to choose the clustering algorithm for their application. Trade off criteria shown space complexity and computational loops required for Nearest Neighbor are very high. But accuracy is very high and redundancy is very low. So, if you want exact data from the document, experimental work shown that Nearest Neighbor is best. For K-Means space complexity and computational loops are low. Redundancy and Accuracy depends on master node selection. In Hierarchical Agglomerative, space complexity and computational loops are moderate, but redundancy is very high. Proposed system is described below.
The system accepts input text file. The file is read and stored into a string. The string is then split by the newline keyword. The split file is assigned to the string array as the split function returns the string array. The array contains paragraphs which are further treated as nodes [2].

### 2.2 Adding Weighted Edge to Document Graph

A weighted edge is added to the document graph between two nodes if they either correspond to adjacent node or if they are semantically related, and the weight of an edge denotes the degree of the relationship. Here two nodes are considered to be related if they share common words (not stop words) and the degree of relationship is calculated by “Semantic parsing”. Also notice that the edge weights are query-independent, so they can be pre-computed.

The following input parameters are required at the pre computation stage to create the graph.

1. **Threshold for edge weights.** Only edges with weight not below threshold will be created in the document graph. (A threshold is user configurable value that controls the formation of edges). Adding weighted edge is the next step after generating document graph. Here for each pair of nodes \( u, v \) we compute the association degree between them. It is the score (weight) \( E_{Score}(e) \) of the edge \( e(u,v) \). If Score \( (e) \geq \text{threshold} \), then \( e \) is added to \( E \). The score of edge \( e(u,v) \) where nodes \( u, v \) have text fragments \( t(u), t(v) \) respectively is [2]:
computational cycles than Nearest Neighbor. In case of K-
Means both parameters depends on selection of master nodes
and applied query.

2.6 Trade-off Criteria
Space and time complexity are two important parameters.
Different clustering techniques performance is compared with
each other on the basis of these parameters. That already been
discussed in above section. Some other parameters are also
equally important with space and time complexity to compare
the performance of different clustering techniques. These
parameters are accuracy of the resultant cluster and redundant
data (redundancy) present in the result. But trade-off criteria is
required if you want to use these four parameters to use the
clustering techniques for a particular application.
Experimental results proved that while you are choosing a
particular clustering technique, these four parameters are very
important. There is trade-off relation among them. Nearest
Neighbor clustering technique has lowest redundancy almost
nil and highest space utilization and computational cycles.
Accuracy of Nearest Neighbor is also high. Agglomerative
Hierarchical clustering technique has lowest space utilization
and computational cycles but highest redundancy. Accuracy is
moderate. In case of K-Means space utilization and
computational cycles totally depends on master node selection
and query applied. Accuracy and redundancy are also depends
on master node selection and input query. User can choose
clustering techniques according to the application and trade-
off criteria.

3. RELATED WORK
The process of grouping a set of physical or abstract objects
into classes of similar objects is called clustering [7]. A
cluster is a collection of data objects that are similar to one
another within the same cluster and are dissimilar to the
objects in other clusters. Data clustering is under vigorous
development. Contributing areas of research include data
mining, statistics, machine learning, spatial database
technology, biology, and marketing. Owing to the huge
amounts of data collected in databases, cluster analysis has
recently become a highly active topic in data mining research.
In this paper performance based analysis and comparison of
three clustering algorithms are discussed which guides user
while selection of clustering algorithm for their application.
These three clustering algorithms are a) Nearest Neighbor b)
Simple K Means c) Agglomerative Hierarchical
Performance of these algorithms are analyzed and compared
with four parameters named space utilization, computational
cycles, accuracy and redundancy.

3.1 Preprocessing of Input Text File
System accepts the input text file. File is split into the number
of nodes or paragraphs by the newline character. These nodes
are represented in the form of table named as Document
Graph. Similarity score among all nodes is represented in
table named as Relation Graph.

3.2 K Means Clustering
K Means is simple portioned algorithm can be used for
clustering large dataset on predefined attributes.
Algorithm for k-means
1. Decide on a value for k.
2. Initialize the k cluster nodes (randomly, if necessary).
3. Decide the class memberships of the N objects by assigning
them to the nearest cluster center.
4. Re-estimate the k cluster centers, by assuming the
memberships found above are correct.
5. If none of the N objects changed membership in the last
iteration, exit. Otherwise go to step 3.

3.3 Nearest Neighbor Clustering
Algorithm for Nearest Neighbor
1. Set i = 1 and k = 1. Assign pattern X_i to cluster C_i.
2. Set i = i + 1. Find nearest neighbor of X_i among the patterns
already assigned to clusters. Let d_m denote the distance from
X_i to its nearest neighbor. Suppose the nearest neighbor is in
cluster m
3. If d_m greater than or equal to t then assign X_i to C_m where t
is the threshold specified by the user. Otherwise set k = k+1
and assign X_i to a new cluster C_k
4. If every pattern has considered then exit. Otherwise go to
Step 2.

3.4 Agglomerative Hierarchical Clustering
Algorithm for Agglomerative Hierarchical
1. Start by assigning each node to a cluster, so that if you
have N nodes, you now have N clusters, each containing just
one node. Let the distances (similarities) between the clusters
the same as the distances (similarities) between the nodes they
contain
2. Find the closest (most similar) pair of clusters and merge
them into a single cluster, so that now you have one cluster
less. 3. Compute distances (similarities) between the new
cluster and each of the old clusters.
4. Repeat steps 2 and 3 until new clusters and old clusters are
different. Otherwise exit.

3.5 Performance Measurement and Trade-
off Criteria.
After getting the minimal cluster for the input query, these
three algorithms are compared with each other on the basis of
four parameters space utilization, computational cycles,
accuracy and redundancy. To show the trade of criteria, all
parameters result for a particular input query is shown.

4. IMPLEMENTATION
The clustering of input text document with three clustering
techniques namely Nearest Neighbor, Agglomerative
Hierarchical and K Means in order to analyze and compare
them on the basis of performance has been implemented. The
experimental results given below describe trade-off criteria
for choosing the algorithm for your application need..
Space utilization, computational cycles, accuracy and
redundancy are the parameters used to compare all
algorithms.
Relation graph (Weighted Document Graph) gives the similarity score between all nodes. The range of similarity score is 0 to 1. For example, in above figure Node 2 and 9 have the highest similarity score (one). That means node 2 and 9 are perfectly matched. Similarly, node 2 and 8 have lowest similarity score (zero). That means node 2 and 9 are completely distinct.

In Figure 3, graphical representation shows that Nearest Neighbor has highest space (5371 Kb) and time (8457 computational cycles) complexity. K-Means clustering technique has lowest time complexity (269 computational cycles) and moderate space complexity (2051 Kb). Agglomerative Hierarchical technique has moderate time complexity (6971 computational cycles) and lowest space complexity (1757 Kb).

**Figure 2: Relation Graph**

Resultant clusters with matching score are represented for all three clustering techniques. Resultant clusters are found out for at least 50% match with the given query. In the result ‘Match Found (%)’ column gives matching percentage of that particular query with each resultant cluster.

**Figure 3: Space Utilization and Computational Cycle Comparison**

**Figure 4: Comparative Results**

Trade-off criteria to choose a particular technique for an application depends on four parameters named as accuracy, redundancy, computational cycles and space complexity. Accuracy, computational cycles and space complexity are discussed in previous results. Here above three values with redundancy is shown. Redundancy is data other than input query is present in minimal cluster. Redundancy of Nearest Neighbor clustering technique is very low almost zeros. Redundancy of Agglomerative Hierarchical is very high and accuracy is moderate but space utilization and computational required low. Accuracy is moderate and redundancy is very high for K-Means clustering technique. But again it depends on master node selection and input query.

**Figure 5: Trade-off Criteria**
6. CONCLUSION AND FUTURE WORK

Performance based analysis and comparison of multi-algorithmic clustering techniques has been discussed. Four different parameters are considered for comparison names as space utilization, computational cycles, accuracy and redundancy. If your application needs exact data to mine, then Nearest Neighbor is best because it has lowest redundancy. Agglomerative hierarchical can be used if redundancy is not important factor. K-Means selection is depends on your application type.

The implemented system focuses only on three clustering algorithms. In future the model can be built in which any clustering algorithm can be compared with any existing or new clustering algorithm. Here I have considered input only as one text document. In future, it possible multiple file in multiple formats as a input.

7. REFERENCES
