Comparison Between K-Means and Genetic Algorithm in Text Document Clustering

Divyashree G
Assistant Professor
Dept of Information Science & Engineering
Sapthagiri College of Engineering
Bangalore, India

Gayathri Rayar
Assistant Professor
Dept of Information Science & Engineering
Sapthagiri College of Engineering
Bangalore, India

***Abstract*— In today's business practice through the means of IT, there are different data sources for a particular object. More precisely it can be said that the documents are large. Specific text document clustering method based on text mining concept can help to analyze and monitor the data sources. The prime objective of optimization of clustering algorithms is to achieve high intra-cluster similarity (i.e. *documents within a cluster should be similar*) and low inter-cluster similarity (i.e. *documents from different clusters should be dissimilar*). The clustering is the core technology into machine learning, pattern recognition, image analysis and information retrieval system based application. The existing works such as cosine similarity, Jaccard, Pearson Coefficient and K-Means algorithm will be optimized by using the genetic algorithm in the proposed work. The performance metric such as purity, entropy and F-measure will be evaluated for the K-means clustering algorithm and genetic algorithm and the final result is expected to posses' higher score of purity, lower score of entropy and maximized F-measure value.**

Keywords - Genetic Algorithm, K-Means Clustering Algorithm, Similarity Measures, Cosine Similarity, Jaccrd Coefficient, Pearson Coefficient.

I. INTRODUCTION

Data mining is known as Knowledge Discovery in Databases (KDD).Data mining is a process of analyzing large databases to find patterns that are valid, useful, and understandable. The valid means holds the new data with some certainty and useful means data mining should be able to act on the terms in the document finally, the understandable means humans should be able to read/identify the pattern. Data mining performed with large data, heterogeneous machine learning, statistics, artificial intelligence, databases and visualization.

Text mining is a part of data mining its aim is to extract high-quality information from the given text. The extraction of high quality information can be done through statistical pattern learning. text mining includes information retrieval, lexical analysis, pattern recognition, information extraction, data mining techniques, association analysis, visualization, and predictive analytics.

Cluster analysis or clustering is the process of grouping a set of objects in such a way that objects which are more similar are grouped under single clusters and the objects which are not similar are grouped under other clusters. It is the main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Document clustering, also known as text clustering is a technique, used to group the documents automatically. It is useful in organizing the documents based on indexing, that is helpful for sorting and quick searching.

Example:

- Apple can be clustered into the categories of fruits, as well as mobile companies
- An email received by a company, with subject line containing problem, can be parsed into a separate folder, and to be addressed by the customer care.

Document clustering is a document organization, extraction of the terms and fast information retrieval or filtering. Document clustering uses the concept of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Clustering is an example of unsupervised classification [1]. Classification of the data is done based on the procedure that assigns data objects to a set of classes. Unsupervised means clustering does not depend on any predefined classes and data training examples while classifying the data objects. Clustering is a crucial area of research, which finds applications in many fields including bioinformatics, pattern recognition, image processing, marketing, data mining, economics, etc. An example of document clustering is web document clustering for data searching by users. The application of document clustering can be classified into two types, online and offline. Online applications are constrained by efficiency problems when compared offline applications.

Document Clustering is a challenging task and it is being studied from many decades but still it is far from a trivial and solved problem. The some of the challenges among them are:

1. Selecting appropriate features of the documents that should be used for clustering.
2. Selecting an appropriate similarity measure between the documents.
3. Selecting an appropriate clustering method utilizing the above similarity measure.
4. Implementing the clustering algorithm in an efficient way that makes it feasible in terms of required memory and CPU resources.
5. Finding ways of assessing the quality of the performed clustering.

Text document clustering plays an important role in providing intuitive navigation and browsing mechanisms by organizing large amounts of information into a small number of meaningful clusters. However, the bag of words representation used for these clustering methods is often unsatisfactory as it ignores relationships between important terms that do not co-occur literally. However, in spite of a long tradition of research in similarity-based text document retrieval there is no clustering method that could function as a best to this end. The reason to stem from the fact that text document clustering must simultaneously deal with quite a number of problems [3]:

1. Problem of efficiency: Text document clustering must be efficient because it should be able to do clustering on ad-hoc collections of documents, e.g. ones found by a search engine through keyword search.

2. Problem of effectiveness: Text document clustering must be effective, i.e., it should relate documents that talk about the same or a similar domain.

3. Problem of explanatory power: Text document clustering should be able to explain to the user why a particular result was constructed. Lack of understandability may pose a much bigger threat to the success of an application that employs text document clustering than a few percentage points decrease in accuracy.

4. Problem of user interaction and subjectivity: Applications that employ text document clustering must be able to involve the user. The results should be focusing one's attention on particularly relevant subjects. For example a search for "health" might turn up food-related issues that a user might want to explore in details relevant for him, such as "fruits", "meat", "vegetables" and others.

The rest of the paper describes as follows: Literature Survey describes some of the current knowledge related to the text document clustering as well as theoretical and methodological contributions to a clustering method, the Methodology brief about the systematic, theoretical analysis of the methods applied to a document clustering, the architecture of the system will brief the overall work of the system. Then the is described in the document pre-processing module. The similarity measures are calculated for the documents are explained in the similarity measure module. brief about the document clustering by using the K-Means algorithm Chapter7 tells the evaluation metrics used in the project and also hoe they are used and tabulate the results obtained. Chapter 8 explains the global genetic clustering algorithm works for text document clustering. The testing of these modules is shown in the chapter 9. The results are shown and written in the Chapter 10.

II. RELATED WORK

A. Background

Cluster analysis is the basic data analysis tools in data mining. Cluster analysis can be used as a standalone data mining tool for the data distribution and data preprocessing step for data mining algorithms operating on the clustering of the documents. Clustering algorithms are used to organize data, categorize data, for data compression and model construction. Clustering is a huge area of research, which finds applications in many fields including bioinformatics, pattern recognition, image processing, marketing, data mining, economics etc.

B. Clustering algorithms

Clustering is a group of set of objects and finds the relationship between the objects. Hence the clustering algorithm and similarity measures are used for the text document clustering. Data clustering algorithms can be mainly classified into following categories [2]:

1. Partition algorithms

A partition clustering algorithm partition the data set into desired number of sets in a single step. Partition clustering algorithm [13] splits the data points into k partition, where each partition represents a cluster. The partition is done based on certain objective function. The partition algorithm approaches require selecting a value for the desired number of clusters to be generated. Some of the examples for the partition clustering algorithms methods are k-means and a variant of k-means-bisecting, k-medoids.

2. Hierarchical algorithms

A hierarchical clustering algorithm divides the given data set into smaller subsets in hierarchical fashion. A hierarchical method [13] creates a hierarchical decomposition of the given set of data objects. Here tree of clusters called as dendrograms is built. Every cluster node contains child clusters, sibling clusters partition the points covered by their common parent. In hierarchical clustering we assign each item to a cluster such that if we have N items then we have N clusters. Find closest pair of clusters and merge them into single cluster. Compute distance between new cluster and each of old clusters. We have to repeat these steps until all items are clustered into K no. of clusters. It is of two types:

i. Agglomerative (bottom up)

Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. It starts by letting each object form its own cluster and iteratively merges cluster into larger and larger clusters, until all the objects are in a single cluster or certain termination condition is satisfied. The single cluster becomes the hierarchy's root. For the merging step, it finds the two clusters that are closest to each other, and combines the two to form one cluster.

ii. Divisive (top down)

A top-down clustering method is not a preferred method. It works same as the agglomerative clustering algorithm but in the reverse direction. This method starts with a single cluster containing all data, and then successively splits resulting clusters until only clusters of individual objects remain.

K-Means algorithm has biggest advantage of clustering large data sets and its performance increases as number of clusters increases. But its use is limited to numeric values. Therefore Agglomerative and Divisive Hierarchical algorithm was adopted for categorical data, but due to its complexity a new approach for assigning rank value to each categorical attribute using K- means can be used in which categorical data is first converted into numeric by assigning rank. Hence performance of K- mean algorithm is better than Hierarchical Clustering Algorithm.

3. Density-based methods

Density-based clustering methods will group the data objects based on the arbitrary shapes. Clustering is done based on criteria of density such as density connected points and based on an explicitly constructed density function. The popular density based clustering methods are DBSCAN and its extension, OPTICS and DENCLUE

4. Grid-based methods

Grid-based clustering methods use multiresolution grid structure to cluster the data objects. The benefit of this method is its speed in processing time. Some examples include STING, Wave Cluster.

5. Model-based methods

Model-based methods use a model for each cluster and determine the fit of the data to the given model. It is also used to automatically determine the number of clusters. Expectation-Maximization, COBWEB and SOM (Self-Organizing Map) are typical examples of model-based methods.

C. *Clustering applications*

Clustering is a major tool in a number of applications in many fields of business and science. Hereby, following information will summarize the basic directions in which clustering are used.

a. Finding Similar Documents: This feature is often used when the user has spotted one "good" document in a search result and wants more-like-this. The interesting property here is that clustering is able to discover documents that are conceptually alike in contrast to search-based approaches that are only able to discover whether the documents share many of the same words.

b. Organizing Large Document Collections: Document retrieval focuses on finding documents relevant to a particular query, but it fails to solve the problem of making sense of a large number of uncategorized documents. The challenge here is to organize these documents in a taxonomy identical to the one humans would create given enough time and use it as a browsing interface to the original collection of documents.

c. Duplicate Content Detection: In many applications there is a need to find duplicates or near-duplicates in a large number of documents. Clustering is employed for plagiarism detection, grouping of related news stories and to reorder search results rankings (to assure higher diversity among the topmost documents). Note that in such applications the description of clusters is rarely needed.

d. Recommendation System: In this application a user is recommended articles based on the articles the user has already read. Clustering of the articles makes it possible in real time and improves the quality a lot.

e. Search Optimization: Clustering helps a lot in improving the quality and efficiency of search engines as the user query can be first compared to the clusters instead of comparing it directly to the documents and the search results can also be arranged easily.

D. Prior Studies

This section describes abstraction of few of the related work, in terms of the problem taken in hand, their approach and finding.

Mohit Sharma and Pranjal Singh "Text Document Clustering and Similarity Measures" [16] Clustering is an important technique which organizes a large number of objects into small number of coherent groups. It leads to efficient and effective use of these documents for information retrieval. Clustering algorithms require a similarity metric to identify how the two different documents are related/similar to each other. This difference is often measured by some distance measure such as Cosine similarity, jaccard and others. In the work, the well known five different distance measures are used and compare their performance on datasets using k-means clustering algorithm. But, this work has a lack of efficient feature selection and representing the terms.

Anna Huang, "Similarity Measures for Text Document Clustering" [14] Clustering is a useful technique that groups a large quantity of unordered text documents into a small number of meaningful and coherent clusters. Partitional clustering algorithms have been identified more a more suitable than the hierarchical clustering algorithm schemes for clustering the large datasets. The different types of similarity measures have been used for clustering the data, such as euclidean distance measure, cosine similarity, and relative entropy. In the paper, different similarities measures are used to compared and analyze their effectiveness by using the similarity measures in partitional clustering for text document datasets. The work utilize the standard K-means clustering algorithm and report the results on seven text document datasets and five distance/similarity measures that have been most commonly used in text clustering. In the observed work there are three components that affect the final results, they are representation of the terms, distance or similarity measures, and the clustering algorithm itself.

Sanjivani Tushar Deokar, "Text Documents clustering using K Means Algorithm" [15] The paper discussed the implementation of K-Means clustering algorithm for clustering unstructured text documents, beginning with the preprocessing of unstructured text and reaching the resulting set of clusters. K-Means algorithms have been applied to text clustering in a straight forward way. The preprocessing steps use the calculated TF-IDF value and the similarity measure it used as cosine similarity. Here, the k-means algorithm using a set of points in m-dimensional vector space for text clustering. It is easier and less time consuming to find documents from a large collection when the collection is ordered or classified by group or category. The author has informed that the work can be further improved by using similarity measure of documents which would ultimately provide better clusters for a given set of documents.

E. *Gaps identified*

- The work has a lack of efficient feature selection and
  representing the terms [16].
- In the text document clustering work there are three components that affect the final results, they are representation of the terms, distance or similarity measures, and the clustering algorithm itself [14].
- In the documents clustering using k means algorithm work can be further improved by using similarity measure of documents which would ultimately provide better clusters for a given set of documents and the performance of the algorithm can still be improved by blending it with any standard optimization technique [15].

III. OVERVIEW OF THE PROPOSED SYSTEM

A. *Methodology of the proposed system*

Figure 1.3 shows the overall methodology of the proposed efficient document clustering system on centralized system.

The different steps in the proposed methodology are the following,

- Only standard text document data are selected and upload to each of the system for clustering.
- The collected text documents are pre-processed using different techniques such as removal of stop words and stemming the words in each system. Stop words are the non-descriptive words such as a, and, are, then, what, is, the, do etc. are removed from the each of the document. Stemming the words means words with different endings will be mapped into a single word Ex: production, produce, product, produces will be mapped to the stem "produc" to reduce the appearance of same words with different forms. After pre-processing steps are performed now, the data are ready to calculate the similarity measures.
- Initially calculate the term frequency (TF) and inverse document frequency (IDF) for each document. Term frequency tells occurrence of the word in the document and inverse document frequency represents the weight of word means to tell how important the word in the document.
- The similarity measures such as Cosine similarity, Jaccard coefficient and Pearson correlation coefficient is applied to the pre-processed document collection in the system.

- Apply the K-Means clustering algorithm to each of the calculated similarity values at each local system. K-Means algorithm produces the number of clusters based on the similarity value and sends the clustered data to the global system.
- The clustered data are received at the global system and apply Genetic algorithm for the document clustering. Genetic algorithm apply the step by step process and perform the clustering operation on the text documents
- Performance factors/metrics such as purity, entropy and F-Measure are used to analyze the proposed document clustering. Quality of the cluster is calculated by purity and entropy whereas the accuracy of the cluster is calculated by F-Measure.

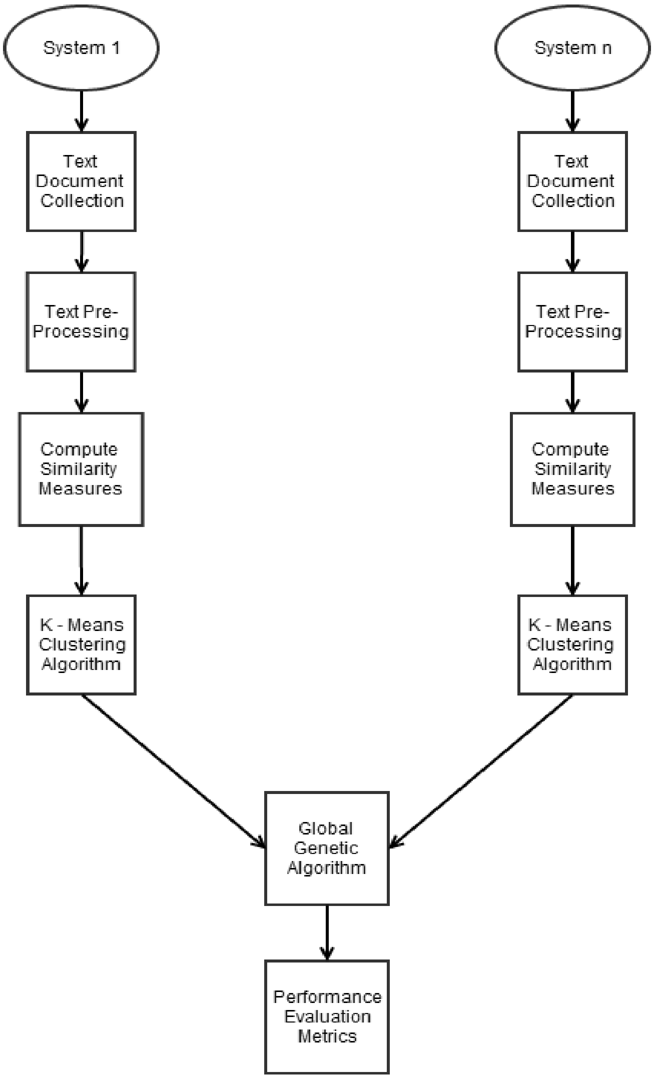
System 1
System n
Text Document Collection
Text Document Collection
Text Pre-Processing
Text Pre-Processing
Compute Similarity Measures
Compute Similarity Measures
K - Means Clustering Algorithm
K - Means Clustering Algorithm
Global Genetic Algorithm
Performance Evaluation Metrics

Figure 1.3: Flow Diagram for the document clustering

B. *Architecture of the system*

The proposed system architecture where it can be seen that after collecting the text document, it is subjected for pre-processing which will usually include truncation of stop-words, stemming of words, and followed by filtering process. The pre-processed text document then subjected to calculate the similarity measures such as Cosine similarity, Jaccard co-efficient and pearson correlation coefficient. The system also considers design of clustering where K-Means algorithm at the local system and genetic algorithm at the local system will be implemented. Finally, after document clustering, the final parameters such as purity, entropy and F- Measure of performance analysis will be extracted to identify the efficient of the proposed architecture.

The architecture of the proposed framework for efficient document clustering on centralized system is as shown in Figure 3.2

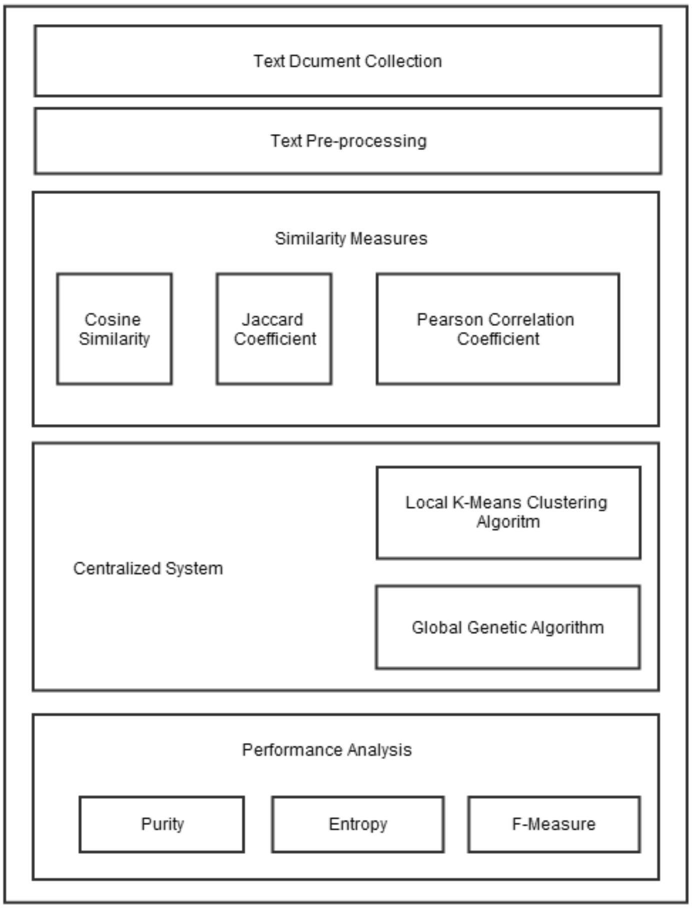

Figure 3.2: Architecture of the proposed system

The architecture of the document clustering consists of following five modules

1. Document Pre-processing Module
2. Similarity Measures Module
3. Local document clustering Module
4. Global document clustering Module
5. Evaluation Metric Module

1. *Document Pre-processing Module*

The text document pre-processing can be done by the following process shown in the figure 4.1.

Text document collection: Selecting and accessing the data from the system to perform the document clustering. The collected document should be in the .txt format.

Text Document Preprocessing: Initially the collected text documents are composed of a lot of elements or the words. Preprocessing requires the reduction in the document contents.

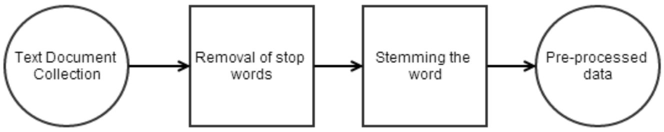

Figure 4.1: Text document pre-processing

Text document Collection includes the processing of data like indexing, filtering etc which are used to collect the documents that need to be clustered, index them to store and retrieve in a better way, and filter them to remove the extra data, for example, stop words.

Removal of Stop Word: Stop words are the words that are non-descriptive for the topic of a document such as a, and, are, then, what, is, the, do etc. It is the frequently occurring words that are not searchable. This is done to improve the speed and memory consumption of the application. There are standard stop word lists available but in most of the applications these are modified depending on the quality of the dataset. The syntax to remove the stop words is

Stemming the words means words with different endings will be mapped into a single word or Stemming is the process of reducing words to their stem or root form. For example ‘cook’, ’cooking’, ‘cooked’ are all forms of the same word used in different constraint and but for measuring similarity these should be considered same and production, produce, product, produces will be mapped to the stem “produc”.

Preprocessed data preprocessing consists of steps that take as input a plain text document and output a set of tokens to be included in the vector model.

After pre-processing, the pre-processed data is subjected to calculate the tf–idf (term frequency–inverse document frequency), is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining.

TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

TF (t) = (Number of times term t appears in a document) /
(Total number of terms in the document)

IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

IDF (t) = loge (Total number of documents / Number of
documents with term t in it)

2. *Similarity Measures Module*

A similarity measure or similarity function is a real valued function that quantifies the similarity between two objects. The similarity measure gives the degree up to which each objects are close to or separate from each other. This module performs the calculation of the Cosine similarity, Jaccard coefficient and Pearson correlation coefficient

A variety of similarity or distance measures have been proposed and widely applied, such as the cosine similarity, Jaccard coefficient and Pearson correlation coefficient.

1. *Cosine Similarity*

The similarity of two documents corresponds to the correlation between the vectors, where the documents are represented as term vectors. This is quantified as the cosine of the angle between vectors, which is called the cosine similarity. An important property of the cosine similarity is its independence of document length .The result of the cosine similarity lies between 0-1. If the cosine similarity between the two documents is 1 then, the documents are similar. If the cosine similarity between the two documents is 0 then, the documents are not similar. The mathematical formula to calculate the cosine similarity is given as shown in equation

$$\cos(\theta) = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

Where, A- Term count value in the document1
B- Term count value in the document2

2. *Jaccard Coefficient*

The Jaccard coefficient, also known as Tanimoto coefficient, measures similarity as the intersection divided by the union of the objects. For text document, the Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms. The value of the Jaccard coefficient exists in the range of 0-1.The value 1 means the objects are similar and the value 0 means the documents are different. The formal definition is in the equation

$$J(A,B) = \frac{A \cap B}{A \cup B}$$

Where, A- Term count value in the document1
B- Term count value in the document2

3. Pearson correlation coefficient

Pearson's correlation coefficient is another measure of the extent to which two vectors are related. There are different forms of the Pearson correlation coefficient formula. The value of this measure lies between the 0 to 1. The value is 1 when the number of terms present in the document1 is equal to the number of terms present in the document2. Given the term set T = {t1, . . . , tm}, a commonly used in the form is equation

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}}$$

Where, x_i – Term count value of the document1
y_i – Term count value of the document2
$\bar{x}$ - Mean term count value of the document1
$\bar{y}$ - Mean term count value of the document2

3. Local document clustering Module

The local text document clustering is done by using the K-Means algorithm.K-Means is a simple yet very powerful algorithm for clustering data. It is a predictive algorithm for determine the clusters. There is a whole lot of research done around K-Means because it provides fast and reliable solutions for most practical applications. The idea is to put the data points in to a cluster with the smallest distance from the clusters mean to the data point.

The standard K-means algorithm will work as follows. For a given set of data objects D and a pre defined number of clusters k, k data objects are selected randomly to initialize k clusters, each one being the randomly selected centroid value of a cluster. The remaining objects are then assigned to the cluster represented by the nearest or most similar centroid value. Next, the new centroid value is re-computed for each cluster and in turn all documents are re-assigned based on the calculated new centroid value. This step iterates until a final and fixed solution is reached, where all data objects remain in the same cluster after an update of centroid value.

Some of the properties if the K-Means algorithms are

- There should be always K clusters.
- Each cluster should contain atleast one item in it.
- The clusters are non-hierarchical and they do not overlap each other.
- Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'centroid value' of clusters.

The k-means algorithm for the text document data is as explained below. Consider, each cluster's center is represented by the mean value of the objects in cluster.

```
INPUT: K - The number of clusters
       D - Set of documents similarity value
OUTPUT: A set of K- clusters
Begin
1. Choose k objects from D as initial cluster center.
2. Assign each object to the cluster based on the closest centroid.
3. Calculate the mean value of the objects for each cluster and update.
4. Repeat the step 2 and 3 until there is no change in the cluster center.
End.
```

4. Global document clustering Module

The global text document clustering is performed by using the genetic algorithm. The main objective of the genetic algorithm is to provide exact/perfect clustering of text documents by providing the high intra similarity and low inter similarity.

Genetic algorithm developed by Goldberg was inspired by Darwin's theory of evolution which states that the survival of an organism is affected by rule "the strongest species that survives". Darwin also stated that the survival of an organism can be maintained through the process of reproduction, crossover and mutation. Darwin's concept of evolution is then adapted to computational algorithm to find solution to a problem called objective function in natural fashion. A solution generated by genetic algorithm is called a chromosome, while collection of chromosome is referred as a population. A chromosome is composed from genes and its value can be either numerical, binary, symbols or characters depending on the problem want to be solved. These chromosomes will undergo a process called fitness function to measure the suitability of solution generated by GA with problem. Some chromosomes in population will mate through process called crossover thus producing new chromosomes named offspring which its genes composition are the combination of their parent. In a generation, a few chromosomes will also mutation in their gene. The number of chromosomes which will undergo crossover and mutation is controlled by crossover rate and mutation rate value. Chromosome in the population that will maintain for the next generation will be selected based on Darwinian evolution rule, the chromosome which has higher fitness value will have greater probability of being selected again in the next generation. After several generations, the chromosome value will converges to a certain value which is the best solution for the problem. The genetic algorithm is used in various application areas.

The operations that performs in the genetic algorithm contains are Selection, Crossover and Mutation

1.Selection: The candidate individuals are chosen from the population in the current generation based on their fitness. The individuals with higher fitness values are more likely to be selected as the individuals of population in the next generation.

2. Crossover: Crossover is a genetic operator that combines (mates) two chromosomes (parents) to produce a new chromosome (offspring). The idea behind crossover is that the new chromosome may be better than both of the parents if it takes the best characteristics from each of the parents. Crossover occurs during evolution according to a user-definable crossover probability. Consider the following 2 parents which have been selected for crossover. The "|" symbol indicates the randomly chosen crossover point.

Parent1:11001|010
Parent2:00100|111

After interchanging the parent chromosomes at the crossover point, the following offspring are produced:

Offspring1:11001|111
Offspring2:00100|010

3. Mutation: The mutation operator is applied to each bit of an individual with a probability of mutation rate. After mutation, a bit that was "0" changes to "1" and vice versa. In fact, it is possible that a regular node becomes a cluster head and a cluster head becomes a regular node. Individual before mutation: 0 1 1 1 0 0 1 1 0 1 0 individual after mutation: 0 1 1 0 0 0 1 1 0 1 0

D. Algorithm

The Algorithm for the genetic algorithm process is as follows

INPUT: k-means clustered data
OUTPUT: clusters of document
Begin
1. Determine the number of chromosomes, generation, and mutation rate and crossover rate value
2. Generate chromosome and the initialization value with a random value
3. Process steps 4-7 until the number of generations is met
4. Evaluation of fitness value of chromosomes by calculating objective function
5. Chromosomes selection
6. Crossover
7. Mutation
8. New Chromosomes (Offspring)
9. Solution (Best Chromosomes)
End

5. Evaluation Metric Module

In order to check the quality and accuracy of the clustering algorithm the proposed system uses the metrics such as purity, entropy and F-Measure. The purity and entropy measures are used to calculate the quality of the clusters whereas the F-Measure used to check the accuracy of the clustering operations. The evaluation metrics such as the purity, entropy and F-measure are explained below

1. **Purity:** The metric purity evaluates the consistency of a cluster that is the degree to which a cluster contains documents from a single category. If the purity value is one it contains documents from a single category therefore it is an ideal cluster. The purity value lies in the range of 0-1.The higher the purity value, better the quality of clusters. If the purity value is one it contains documents from a single category therefore it is an ideal cluster. The higher the purity value, better the quality of clusters. The formal definition of purity is as given below in the equation 8.1

$$P(Cj) = \frac{1}{nj} \max_h (n_j^h) \qquad (8.1)$$

Where, max $_h(n_j^h)$ - is the number of documents that are from the dominant category in cluster Cj and
(n_j^h) - represents the number of documents from cluster
Cj assigned to category h

2. **Entropy:** In general, is a measure of the number of specific ways in which a system is arranged. This measure evaluates the distribution of categories in a given cluster. The entropy results lies between the 0-1. If the entropy value is smaller, the quality of clusters is better. The mathematical formula for the entropy is given as below in the equation 8.2

$$E(C_i) = -\frac{1}{\log c} \sum_{h=i}^{n} \frac{n_i^h}{n_i} \log \frac{n_i^h}{n_i} \qquad (8.2)$$

Where, (n_i^h) - represents the number of documents in cluster Ci assigned to category h
n_i – represents the size of the cluster.

3. **F-Measure:** It is a combined value of Precision and Recall. The Precision and recall computed for each class and its weighted average gives the value of F-measure. The value of this metrics also lies between 0-1. More the F-measure more the accuracy. It is calculated as shown in the equation 8.3

F-Measure= 2 * Precision*Rec all/ (Precision + Recall) (8.3)

Precision and recall are the basic measures used in evaluating strategies in finding the F-Measure evaluation metric. Both precision and recall is applied to the collected documents and the documents can be assumed as either relevant or irrelevant data that is it measures the degree of relevancy.

Precision is the ratio of the number of relevant text documents retrieved to the total number of irrelevant and relevant documents retrieved. It is usually expressed as shown in the equation 8.4

Precision= A/ (A+C) (8.4)

Where, A – Number of relevant documents retrieved
C – Number of irrelevant documents retrieved

Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as shown in the 8.5

Recall = A/ (A+B) (8.5)

Where, A – Number of relevant documents retrieved
B – Number of irrelevant documents retrieved

IV. RESULTS AND DISSCUSSIONS

The proposed system framework for efficient document clustering on centralized system has used the two clustering algorithm for efficient text documents clustering. Initially in the local systems the proposed work uses the K-Means clustering algorithm and in the central global system the proposed work uses the genetic algorithm. The K-Means algorithm works based on the three similarity measures such as Cosine similarity, Jaccard coefficient and Pearson correlation coefficient. In the proposed system three performance metrics are used such as purity, entropy and F-Measure. Among them purity and entropy are used to evaluate the overall quality of the clusters and F-Measure is used to measure the accuracy of the clusters.

From the table 9.1, the conclusion of the proposed system is that, among the three similarity measures Jaccard and Pearson correlation coefficient measures generate more coherent clusters results than cosine similarity measure. Among the two clustering algorithm K-Means clustering algorithm and genetic algorithm, the genetic algorithm produces the good clusters than the K-Means.

Table 9.1: Final evaluated results of K-means and genetic algorithm

Evaluation Metrics	K- Means clustering algorithm			Genetic algorithm
	Cosine Similarity	Jaccard Coefficient	Pearson Correlation Coefficient	
Purity	0.62	0.86	0.86	1
Entropy	0.26	0.01	0.09	0
F-Measure	0.59	0.83	0.91	1

The figure 9.1 shows the results of evaluation metric calculated by using the purity, entropy and F-measure for the both K-means clustering algorithm and genetic algorithm. The graph shows that the genetic algorithm has the highest purity and F-measure value and lower entropy value so, the genetic algorithm will produce better cluster results than K-means clustering algorithm.

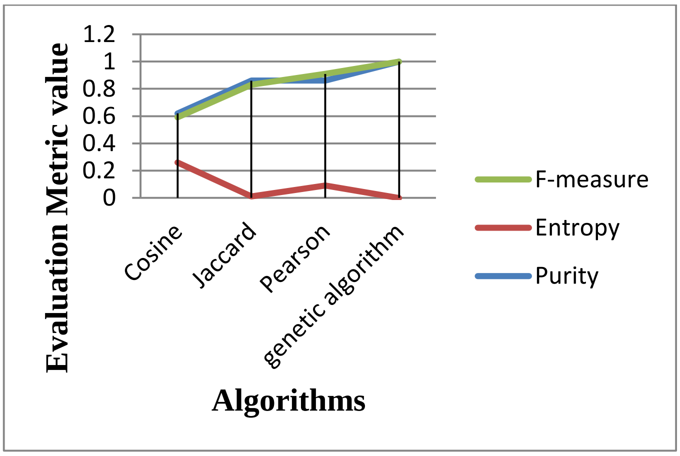

Figure 9.1: Graph showing the final results of K-means and genetic algorithm

V. CONCLUSION

The proposed framework for efficient document clustering on text documents and the two clustering algorithm such as K-Means and Genetic clustering algorithm were used. The K-Means algorithm was developed using Cosine similarity, Jaccard coefficient and Pearson correlation coefficient similarity metrics. The correctness of the algorithm was checked by taking the 10 text documents.

Genetic algorithm is used as a good clustering algorithm to perform clustering on the text documents. Hence, by using the genetic algorithm clustering results, the similarity measures are analyzed to encounter which is the best similarity measure.

By taking the 100 text documents, the effectiveness of the measures were evaluated and analyzed. The Jaccard coefficient and Pearson correlation coefficient measures generate more coherent clusters than the cosine similarity measures and genetic algorithm produces the good cluster result than K-Means clustering algorithm.

VI. FUTURE ENHANCEMENT

With rapid growth of IT environment, there will be a large number of documents has to be maintained. So, by using the large data sets as well as the different data sets the clustering can be performed.

The proposed system works only for the text document which is in the format .txt. So, this system is still being extended to work for the images also.

VII. REFERENCES

[1] Data Mining "Concepts and Techniques" Second Edition by Jiawei HanandMicheline Kamber University of Illinois at Urbana-Champaign.

[2] ManjotKaur, NavjotKaur, "Web Document Clustering Approaches Using K-Means Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013.

[3] Andreas Hotho, Steffen Staab, GerdStumme, "Text Clustering Based on Background Knowledge", *Technical Report, volume 425. University of Karlsruhe, Institute AIFB, (2003).*

[4] M. Eisenhardt, W. Muller, and A. Henrich, "Classifying documents by distributed P2P clustering." in INFORMATIK, 2003.

[5] S. Datta, C. R. Giannella, and H. Kargupta, "Kmeans Clustering over a Large, Dynamic Network," Proc. SIAM Int'l Conf. Data Mining (SDM), 2006.

[6] M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," Proc. KDD Workshop Text Mining, 2000.

[7] G. Forman and B. Zhang, "Distributed Data Clustering Can Be Efficient and Exact," SIGKDD Explorations Newsletter, vol. 2, no. 2, pp. 34-38, 2000.

[8] S. Datta, K. Bhaduri, C. R. Giannella, R. Wolff and H. Kargupta, " Distributed data mining in Peer-to-Peer network's", IEEE Internet Computing, vol.10 , no. 4, pp. 18-26, July 2006.

[9] I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan, "Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications," Proc. SIGCOMM, 2001.

[10] Neethi Narayanan, J.E.Judith, Dr.J.Jayakumari "Enhanced Distributed Document Clustering Algorithm Using Different Similarity Measures"Proc of 2013 IEEE Conf on ICT 2013.

[11] Denny Hermawanto,"Genetic Algorithm for Solving Simple Mathematical Equality Problem", Indonesian Institute of Sciences (LIPI), INDONESIA.

[12] A. K. Santra, C. Josephine Christy, "Algorithm and Confusion Matrix for Document Clustering" Dean, CARE School of Computer Applications, India.

[13] Aastha Joshi, Rajneet Kaur, "Comparative Study of Various Clustering Techniques in Data Mining" Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India.

[14] Anna Huang, "Similarity Measures for Text Document Clustering" The University of Waikato, Hamilton, New Zealand, April 2008.

[15] Sanjivani Tushar Deokar , "Text Documents clustering using K Means Algorithm" International Journal of Technology and Engineering Science [IJTES], July 2013.

[16] Mohit Sharma and Pranjal Singh "Text Document Clustering and Similarity Measures" IIT Kanpur, India, November 19, 2013