Document Clustering for Effective Information Retrieval System using Genetic Algorithm

V. Raja Manickam[1] Dr. A. Nagarajan[2]
[1]Research Scholar [2]Assistant Professor
[1,2]Department of Computer Applications
[1,2]Alagappa University – Karaikudi India

Abstract— Document clustering is a significant domain of interest in the field of document summarization. K-means clustering is one of the methods used for clustering documents. These methods suffer from issues and challenges like accuracy and time complexity. Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. The searches of the information retrieval can be based on metadata, full-text or other content-based indexing. In clustering system it can be very useful in web search for grouping search results into closely related sets of documents. It can improve the similarity search on information retrieval. To overcome these limitations a new genetic algorithm based document clustering method have been proposed in this research work. This work also proposed the Boolean operator based information retrieval scheme to find out the particular query raised by the user. This research work the document clustering is performed by 20Newsgroups document dataset. The objective of this research work is to cluster the document by using genetic algorithm and retrieved the user query by based on Boolean operator information retrieval system. This proposed method is implemented and evaluated by various quality measures like confidence value and collective strength. The experimental analysis in this proposed methodology provides the better time complexity, memory utilization and CPU utilization compared with various existing methods.

Key words: Genetic Algorithm, Information Retrieval, 20Newsgroups Documents, Document Clustering

I. INTRODUCTION

The tremendous growth of scientific databases put a lot of challenges before the researches to extract useful information from them using traditional data base techniques [4]. Hence effective mining methods are essential to discover the implicit information from huge databases. Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning statistics, and database systems. The goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use [5]. Data mining is ready for application in the business community because it is supported by three technologies that are as follows:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Mining information and knowledge from large databases has been recognized by many researchers as a key research topic in database systems and machine learning and by many industrial companies as an important area with an opportunity of major revenues [1]. Researchers in many different fields have shown great interest in data mining. Several emerging applications in information providing services, such as data warehousing and on-line services over the Internet, also call for various data mining techniques to better understand user behavior, to improve the service provided, and to increase the business opportunities. In response to such a demand, this thesis is to provide a survey, from a database researcher's point of view, on the data mining techniques developed recently. A classification of the available data mining techniques based on the neural networks genetic algorithm that can be implemented in this thesis on applicative data mining systems [3].

In information retrieval system it is the activity of obtaining information resources relevant to an information need from a collection of information resources. The searches of the information retrieval can be based on metadata, full-text or other content-based indexing.

- For clustering: To propose a family of Genetic Algorithm namely genetic algorithm based clustering technique is used to cluster the documents which are related to the users needed database. The quality and effectiveness of this MDGA based clustering method is also assessed the effectiveness of the quality measures.
- After the document clustering is completed, the user enters the needed query. The query related appropriate document is retrieved by using Boolean operator based information retrieval system.

II. RELATED WORK

The following papers motivated to do the research work,

In 2015, Kedar B. Sawant (1) reviews some existing methods for selecting the number of clusters as well as selecting initial centroid points. The author says, this overview of the existing methods of choosing the value of K i.e. the number of clusters along with new method to select the initial centroid points for the K-means algorithm has been proposed in their paper along with the modified K-Means algorithm to overcome the deficiency of the classical K-means clustering algorithm. The author followed by a proposed method for selecting the initial centroid points and the modified K-mean algorithm which will reduce the number of iterations and improves the elapsed time. The new method in this paper is closely related to the approach of K-means clustering because it takes into account information reflecting the performance of the algorithm. The improved version of the algorithm uses a systematic way to find initial centroid points which reduces the number of dataset scans and will produce better accuracy in less number of iteration with the traditional algorithm.

In 2015, Ahmed M. Fahim [2] presents a new method namely enhanced DBSCAN which clusters spatial databases that contain clusters of varying densities

effectively. The idea is to allow varied values for the Eps parameter according to the local density of the starting point in each cluster. The clustering process starts from the highest local density point towards the lowest local density one. And the value of Eps varies according to the local density of the initial point in current cluster. For each value of Eps, DBSCAN is adopted to make sure that all density reachable points with respect to current Eps are clustered. In this paper, the author has introduced a simple idea to improve the results of DBSCAN algorithm by detecting clusters with variance in density without requiring the separation between clusters. The experimental results in this paper showed the efficiency of this method.

III. RESEARCH CONTRIBUTIONS

A. Document Clustering

Document clustering involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster (2). Document clustering is generally considered to be a centralized process. Examples of document clustering include web document clustering for search users.

The application of document clustering can be categorized to two types, online and offline. Online applications are usually constrained by efficiency problems when compared to offline applications.

Document clustering often takes the following steps:

1) Tokenization

Tokenization is the process of parsing text data into smaller units (tokens) such as words and phrases.

2) Stemming and lemmatization

Different tokens might carry out similar information (e.g. tokenization and tokenizing). And we can avoid calculating similar information repeatedly by reducing all tokens to its base form using various stemming and lemmatization dictionaries.

3) Removing stop words and punctuation

Some tokens are less important than others. For instance, common words such as "the" might not be very helpful for revealing the essential characteristics of a text. So usually it is a good idea to eliminate stop words and punctuation marks before doing further analysis.

4) Computing term frequencies or TF-IDF

After pre-processing the text data, we can then proceed to generate features. For document clustering, one of the most common ways to generate features for a document is to calculate the term frequencies of all its tokens. Although not perfect, these frequencies can usually provide some clues about the topic of the document. And sometimes it is also useful to weight the term frequencies by the inverse document frequencies.

5) Clustering

We can then cluster different documents based on the features we have generated. See the algorithm section in cluster analysis for different types of clustering methods.

6) Evaluation and Visualization

Finally, the clustering models can be assessed by various metrics. And it is sometimes helpful to visualize the results by plotting the clusters into low (two) dimensional space.

B. Genetic Algorithm

In the field of artificial intelligence, a genetic algorithm (GA) is a search heuristic that mimics the process of natural selection. This heuristic (also sometimes called a metaheuristic) is routinely used to generate useful solutions to optimization and search problems [3].

A typical genetic algorithm requires:

– A genetic representation of the solution domain,
– A fitness function to evaluate the solution domain.

Basically the Genetic Algorithm Operations include[2]:

1) Fitness operator: Metrics to measure scheduler performance for each chromosome in the problem, and calculate the values for each chromosome.
2) Selection operator: Select two chromosomes with the highest quality values from the population, that couple to produce two offspring. Copies individual strings according to their objective function values [3]. The objective function is based on a biological measure of fitness, which is to be maximized. This is how the natural selection is simulated.
3) Crossover operator: Exchanges two subparts of the selected chromosomes, the position of the subparts selected randomly. It uses the population of strings at the current generation as a "mating pool" from which to select a random pair of string and randomly combines the value in these two strings to form a new string [7]. This process is repeated to form a new population of strings.
4) Mutation operator: Randomly changes the allele value in some location [8]. This operator is used to perform an occasional random alteration of each string. It has a low probability of alteration.

Genetic algorithm is an exploration algorithm that is based on the process of natural selection. Genetic algorithm differ from traditional search algorithms [5] in the following ways,

- Genetic algorithms use a set of parameters rather than a rules, Prioritization, Genetic single parameters, and each parameter is coded as a string.
- Genetic algorithms start with a population of points (strings) rather than a single point.
- Genetic algorithms use only function values, not derivative or additional information.
- Genetic algorithms use probabilistic transaction rules rather than the deterministic rules.

C. Boolean Model

Retrieval systems based on Boolean logic have long served as the cornerstone of the commercial document retrieval system market and remain very important because of the relative simplicity of the query language and the ease with which it can be understood and implemented. The most common use for a Boolean expression is to state what characteristics must be present in material to be retrieved in a system that retrieves and presents to users bibliographic records or full-text [2]. A second use of Boolean expressions, likely to increase in importance over the next decade, is in rules incorporated into document and email filtering systems. Boolean expressions typically use three operators: AND, OR, and NOT.

In this paper we used IR system based on Boolean model that was built and implemented by 20Newsgroups [6]

data. The processing of the entire system was conducted as following:

1) Select the highest 15 terms frequency from the top 10 documents retrieved using the 20Newsgroups data set.
2) Construct various Queries from the selected terms.
3) Represent queries as a tree and calculate the fitness function which is either precision or recall for each query.
4) Select the best two queries.
5) Perform Crossover (one point crossover is used).
6) Perform Mutation (three different mutation techniques are used; for more details see the next section).
7) Update Population by replacing the new two queries with the worst two queries of the 10 Queries selected in step 2.
8) Go to step 3.

In order to use GA a set of parameters must be determined, these parameters are:

- Number of generation: the number of iteration can be determined by predefined scale of accepted error, or can be defined before the GA starts. In this paper the number of iterations used is 3303 iterations.
- Fitness Function Operator: Fitness function is a performance measure or reward function which evaluates how each solution is good. In this paper precision and recall are used as two fitness functions.
- Selection operator: In this research we used a single point crossover strategy with crossover probability Pc =0.8. The best two individuals with best fitness values are chosen from a population, and represented as trees. When the one point crossover is applied (i.e. if Random number < Probability of crossover) the two trees will exchange sub tree between them.
- Mutation operators: In this experiment the mutation operator works as the most important operator for the learning of query. Each node from the new off springs may be mutated; that depends on mutation probability (pm=0.2), this probability of mutation.

Different types of mutations are used in this research as follows:

- Mutation on Boolean operator: randomly exchanging one operator to another.
- Mutation on term node (leaf node): in Boolean model one term is selected randomly from the offspring and replace by any other one from the terms in a given collection of documents. But in fuzzy model the term is not replaced, only the term weight is changed.
- Mutation by inserting or deleting operator between two nodes in the off springs.

D. Algorithm steps for Proposed Document Clustering Method

The algorithm steps is as follows,

1) Step 1: text[]= sentence from document
2) Step 2: words[]= words of text[]
3) Step 3: Compare words[] with stop words[]
4) Step 4: Remove matching words
5) Step 5: for i = 1 to n stem (word i).
6) Step 6: Create A= adjacency matrix.
7) Step 7: Create D= Diagonal matrix.
8) Step 8: Find C= D – A.
9) Step 9: Find eigen values and eigen vectors of C.
10) Step 10: Select λ d = m / c ; m= total number of edges and c= total vertex degree.
11) Step 11: Select eigen value near to λ d .
12) Step 12: Repeat steps 6 to 13 until no change in sign.
13) Step 13: k= no of clusters

E. Algorithm steps for Proposed Genetic Algorithm Method

The algorithm steps is as follows,

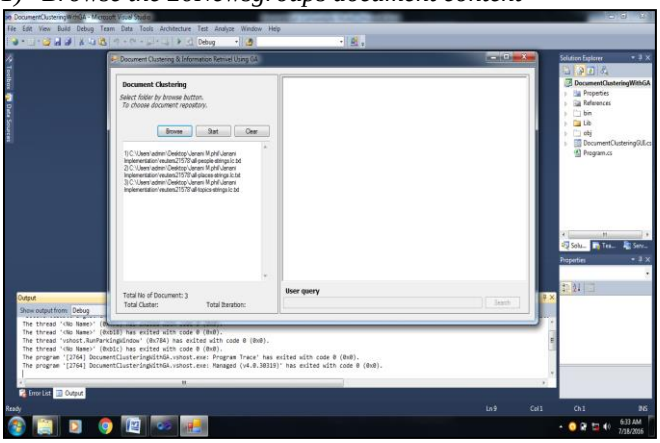
1) Step 1: Generate an initial, random population of chromosomes (The chromosome is the header portion which stores the number of clusters and the number of elements in each cluster)
2) Step 2: Test the fitness of each chromosome in the population.
3) Step 3: Select parents as the fit members of the population.
4) Step 4: Reproduce from selected parents to produce a new population.
5) Step 5: Mutate according to some probability.
6) Step 6: Test the fitness of each chromosome in the new population.
7) Step 7: Evaluation
8) Step 8: Iterate steps 3 to 7 until termination criterion is met.

IV. RESULTS AND DISCUSSIONS

This section brings out the results obtained from the proposed algorithm. The input is getting from 20Newsgroups document data set. First of all this document is performed based on clustering processing. And then the genetic algorithm is processed to derive the processing. Finally the Boolean operator is performed by genetic algorithm for retrieve the appropriate query which was need by the user. In this research work the various types of documents in 20Newsgroups are tested using the proposed methodology. There are three different phases are need to provide the results in proposed technique, that phases are i) Document clustering ii) Genetic algorithm iii) Boolean operator based information retrieval. Based on these above mentioned procedure the experimental results are discussed below.

A. Implementation of Genetic Algorithm based document clustering with Information Retrieval

1) Browse the 20Newsgroups document content



Browse the 20Newsgroups document content

2) Start the session to perform the basic clustering function

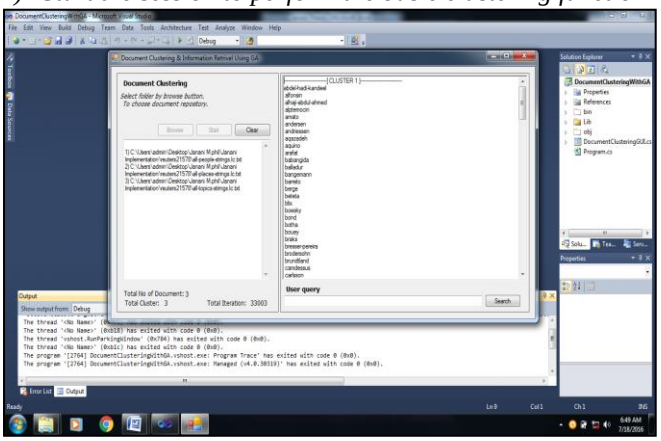


Fig. 2: Start the session to perform the basic clustering function

In this session the 20Newsgroups document is clustered and then the total iterations are calculated. The number of iterations in this session represents the 20Newsgroups document clustered iterations. The clustering process of this research work is based on genetic algorithm.

3) User query to retrieve the particular information

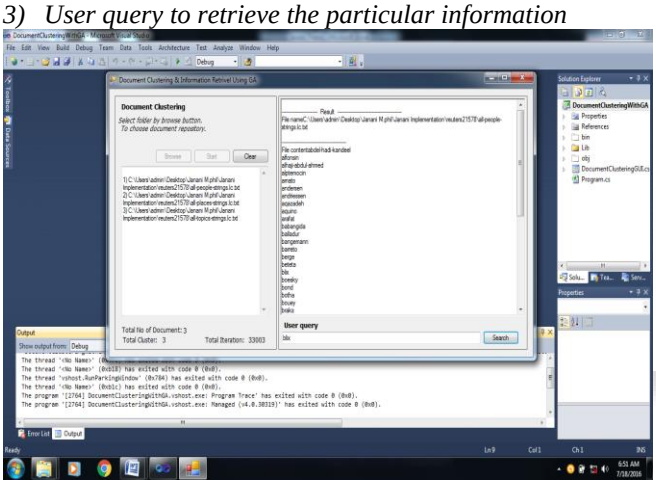


Fig. 3: User query to retrieve the particular information

In this session the user select a query which will he needed. The genetic algorithm based clustering is also presented to cluster the document. When the user enter a query, the Boolean operator based information retrieval is processed to find out the appropriate word, in this level the result provided by where the users query is have. In which particular document the user query is embedded is displayed on the result.

4) Next clusering document selection screen

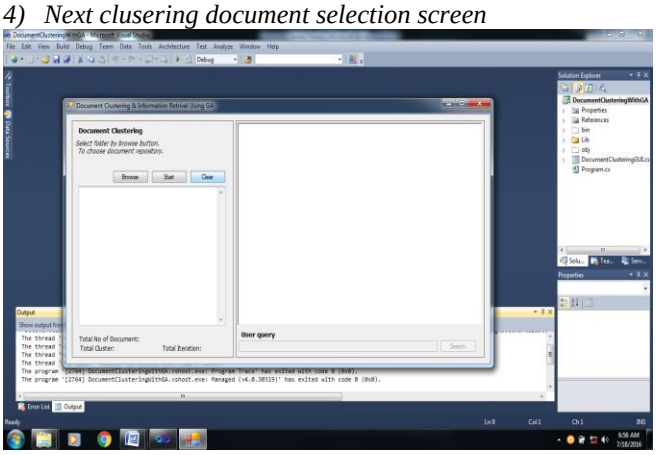


Fig. 4: Next clustering document selection screen

This is the clear function to use to select more and more documents to perform the genetic algorithm based document clustering with information retrieval. The entire processing of the implementation is completed on a small amount of time. So this is the main advantage in this research.

V. PERFORMANCE ANALYSIS

A. Comparison Analysis

The comparison table analysis contains 3 Existing algorithms with our new proposed algorithm. The chosen factor here to determine the performance is the algorithm's speed to cluster the document with various sizes of documents. Following parameter will be simulating at the time of results calculation.

– CPU Utilization: CPU utilization can be calculated by using,

CPU Utilization (%) = (Used CPU) / (Total CPU capability) * 100.

– Memory Utilization: Memory utilization can be calculated by using,

Memory Utilization (%) = (Used Memory) / (Total Available Memory) * 100

The comparison table for this analysis is given below:

Factors	Neepa shah et al. [9]	Proposed Method
Input	20Newsgroups	20Newsgroups
Memory	54.68	21.56
CPU usage	19	17

Table 1: Comparison Analysis of Proposed Algorithm

From the results calculation it's analyzed that we can increase performance parameters by using proposed encryption model as compare existing methods. Also, we can see that the existing methods. The proposed genetic algorithm based document clustering techniques are better than others as they have low memory utilization time and low CPU capability.

VI. CONCLUSION

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. The searches of the information retrieval can be based on metadata, full-text or other content-based indexing. In clustering system it can be very useful in web search for grouping search results into closely related sets of documents. In document clustering involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. It can improve the similarity search on information retrieval. In this research, a new genetic algorithm based document clustering is implemented to cluster the document. The selection of document in this work is 20Newsgroups document dataset. The information retrieval is also performed in this paper to provide the appropriate query which is needed by the user. The IR in this research is performed by using Boolean operator to retrieve the appropriate query. The main motivation in this research work is too discovered of high level prediction rules which are all related to the information retrieval. In this paper the fitness function is designed based on the two measures like all confidence and the collective strength.

The experimental analysis and comparative analysis provide the better and efficient results in this proposed

method compared with latest existing methods. The analysis evolution precedes the high quality of the proposed method, and the level of genetic algorithm based document clustering with information retrieval based on 98% of better query retrieval.

REFERENCES

[1] Kedar B. Sawant, "Efficient Determination of Clusters in K-Mean Algorithm Using Neighborhood Distance", International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015, PP 22-27.
[2] Ahmed M. Fahim, "A Clustering Algorithm for Discovering Varied Density Clusters", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 02 Issue: 08, Nov-2015. Pp.566-573.
[3] Alaa H. Ahmed and Wesam Ashour, "An Initialization Method for the K-means Algorithm using RNN and Coupling Degree", International Journal of Computer Applications (0975 – 8887), Volume 25– No.1, July 2011, pp. 1-6.
[4] Aly, A. A. Applying genetic algorithm in query improvement problem. International Journal Information Technologies and Knowledge, 7[1]: 309-316., 2007.
[5] Banjamin C.M Fung ,KeWang ,Martin Ester, " Hierarchical document clustering using Frequent itemsets", In proceedings of SIAM International Conference on Data mining 2003.
[6] M.P.S Bhatia, Deepika Khurana, "Analysis of Initial Centers for k-Means Clustering Algorithm", International Journal of Computer Applications", Volume 71– No.5, May 2013.
[7] G. Bordang and G. Pasi "Soft Clustering for Information retrieval application", WIREs on Data Mining and Knowledge Discovery, Vol.–1, No.- 2, pp 138-146, 2011.
[8] S. Chiang, S. C. Chu, Y. C. Hsin and M. H. Wang, "Genetic Distance Measure for K-Modes Algorithm", International Journal of Innovative Computing, Information and Control ICIC,ISSN 1349-4198, Volume 2, Number 1, pp.33-40 February 2006.
[9] Neepa shah et al. "Distributed document clustering using K-Means", IJARCSSE, Volume 4, issue 11, November 2014, pp. 24-29.

