Prof. A.D. Thakare, Dr. C.A. Dhote / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue6, November- December 2012, pp.1312-1316

# Virtual Center Based Algorithms for Information Retrieval

**Prof. A.D. Thakare*, Dr. C.A. Dhote****
*(Department of Computer Engineering, University of Pune, India
** (Department of Computer sc. & Engineering, Amravati University, India

## ABSTRACT

**This work suggests a novel evolutionary approach for post clustering retrieval. A novel matching function called virtual center based Matching function(VCF) and VCF based Genetic Algorithm(VCGA) are proposed to improve the retrieval performance. VCF is based on finding the virtual center from the set of centroids present in clustering space and the retrieval is done using VCGA. It is expected that the retrieval performance will be enhanced since number of cycles will be reduced and a single access to disk makes the entire class available when user's query is matched with document description. The performance of these algorithms will be compared with the existing matching function for information retrieval.**

**This paper is presented as a work-in-progress and extension to our previous research paper in which the idea of GA based model for clustering and retrieval was proposed [2]. The experimentation has done up to preprocessing of the documents and matching function algorithm is proposed. Findings regarding the feasibility and utility of the proposed approach will be presented as well as suggestions for follow-on research will be taken. This research proposes the feasibility and utility of post-clustering retrieval by using evolutionary approach.**

***Keywords*** **- Clustering, Information Retrieval(IR), Genetic algorithm(GA), Matching Function, Virtual Center Function(VCF)**

## I. INTRODUCTION

Data mining is component of the knowledge discovery in databases process concerned with the algorithmic means by which patterns are extracted and enumerated from data. This knowledge discovery process has several steps .One of the important steps is to clustering the data[4].

The fundamental concept of clustering is the grouping together of similar data items into clusters. A more elaborate definition, for example, "These clusters should reflect some mechanism at work in the domain from which instances or data points are drawn, a mechanism that causes some instances to bear a stronger resemblance to one another than they do to the remaining instances." The goal is to partition X into K groups Ck such every data that belong to the same group are more alike" than data in different groups. Each of the K groups is called a cluster. Many subspace clustering algorithms fail to yield good cluster quality because they do not employ an efficient search strategy [4].

The nature of the clustering problem is such that the ideal approach is equivalent to finding the global solution of a non-linear optimization problem. This is usually a difficult task to achieve. As a matter of fact, this is an np hard problem, e.g. a problem that cannot be solved in polynomial time

Genetic Algorithms are robust in searching a multidimensional space. It is a search algorithm developed from the biologically natural selection and evolution mechanism. Because of its abilities of self-adaptation and self-organization, it is widely used to solve some complicated optimization problems. Using GA problems are solved by an evolutionary process resulting in a best solution and solution is evolved [1].

Research in IR can be categorized into three categories probabilistic IR, knowledge based IR, learning systems based IR [6]. The information retrieval efficiency measures from recall and precision by using certain Matching Functions. Research work is going on in the field of information Retrieval to make it effective. The main issues are obtain more pages relevant to the users query, optimize the search time [9]. Genetic algorithms are stochastic search algorithm which tries to optimize the solution. In the field of Information Retrieval, GA has been used widely to optimize the query and obtain a relevant set of pages. GAs has been applied to solve some of the information retrieval problems. The problem areas include Genetic mining, Agents for Internet Search, Query Formulation, Query optimization, document Indexing, Ranking, Document clustering, Rough Sets.

## II. CLUSTERING TECHNIQUE:

Some of the issues from research studies in clustering are taken as challenges for this work. They are, first the clustering algorithms require certain parameters for clustering such as the no. of clusters, second cluster shapes, orientation of cluster etc. Due to a bad choice of initial cluster centers often clustering algorithms converges to local optimum. This results in the poor quality clusters and third Reducing the dimension of clustering space.

The GA-clustering uses searching capability of GA's for the purpose of appropriately determining a fixed number k of cluster centers in N, thereby suitably clustering the set of n unlabeled points[11]. The clustering metric that has been adopted was the sum of the Euclidean distances of the points from their respective cluster centers. The chromosomes

which were represented as strings of real numbers, encode the centers of a fixed number of clusters.

## III. INFORMATION RETRIEVAL:

From the Information Retrieval research studies, some of the issues are taken as a challenges which affects the performance and these are, first the performance of IR system completely depends on efficiency of Matching functions second, less work has been done in writing the new matching function and mostly existing matching functions are used to improve the performance third, previous attempts have done retrieval from population containing documents and fourth, previous attempts at using GA's have concentrated on modifying document representations or modifying query representations. Also, applicability of GA adapt various matching functions.

The research areas in IR and various issues that can be solved using the optimization and searching technique of GA [9]. GA also deals with the different application domains in IR which are emerging research areas. Authors also discuss applicability of GA in different areas of IR such as genetic mining, query optimization, document clustering and query optimization etc. Using GA to improve retrieval performance [10] was proposed which shows method to be applicable to three well known document collections where more relevant documents are presented to the users in the genetic modification. Authors presented a new fitness function for approximate information retrieval which is very fast and very flexible than cosine similarity fitness function.

## IV. DOCUMENT PREPROCESSING

Vector space model is to represent each document as a vector of certain weighted word frequencies. The vector space model procedure can be divided in to three stages. The first stage is the document indexing where content bearing terms are extracted from the document text . The second stage is the weighting of the indexed terms to enhance retrieval of document relevant to the user. The last stage ranks the document with respect to the query according to a similarity measure.

The implementation part is concerned with the preprocessing part and the application of clustering algorithm and GA based approach is proposed. The classification of textual data requires pre-processing phase to be completed before applying the clustering methods. Vocabulary for preprocessing the text is defined manually. Using the vocabulary frequency count for each document is computed and weighted vector is generated. Documents are preprocessed using TFIDF format. Weighted vector is given as an input to the algorithm, which in turn gives clustered data.

4.1 Keyword Extraction :

In the keyword extraction process, the vocabulary is limited manually to avoid high dimensionality. The steps are followed while constructing vocabulary like, identify the keywords from documents, Suffix streaming, and finding least frequently and most frequently used words and removing them from list. Identify synonyms and place them under same category. Thus with the final vocabulary, frequency of occurrences of each keyword in document was computed. Now each word has to be weighted with respect to their power of discrimination. For weighing of these words Inverse Document Frequency (IDF) is used.

2.2 Model for Ranked Retrieval :

Various weighing schemes are used to discriminate one document from the other. In general this factor is called collection frequency document. Most of them, e.g. the inverse document frequency, assume that the importance of a term is proportional with the number of document the term appears in. The documents are converted into TFIDF format by using the formula TFIDF = tfij * log (N/n)

The experiment is conducted on sample text documents and weighted average of keywords for every document is calculated as shown in figure 1. For the first document the weighted average is zero, for first document it is 1.3, for second it is 2.2 and so on. Maximum value is for document 5 which describes the theme of the subject.

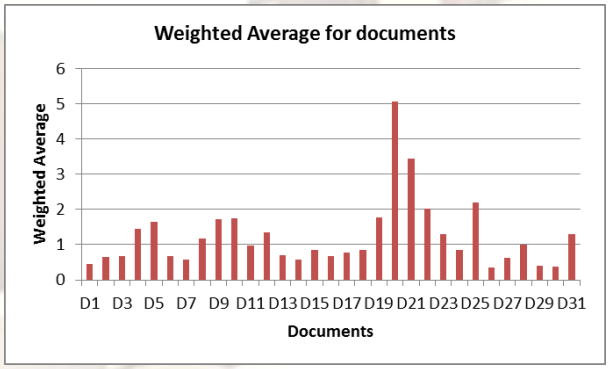


**Fig1:Documents with weighted average for keywords.**

## V. . CLUSTERING ALGORITHM

The results are tested by collecting the sample documents. Initially some sample documents were collected. The documents were already categorized semantically to different classes to compare the result of the k means algorithm with our own results. Weka tool is used for clustering algorithm. The document collection includes industry related documents which are taken as

sample documents. The result of algorithm is as shown in Table 1.:
Number of iterations: 3
Within cluster sum of squared errors: 18.64285815626974
Missing values globally replaced with mean/mode

| Document Weighted Index | Cluster 1 | Cluster 2 |
|---|---|---|
| 0.0167 | 0 | 0.0667 |
| 0.5083 | 0.3333 | 1.0333 |
| 3.2 | 3.2 | 3.2 |
| 3.0667 | 3.2 | 2.6667 |
| 2.2542 | 1.4278 | 4.7333 |
| 2.0042 | 2.4056 | 0.8 |
| 0.7375 | 0.7333 | 6.1167 |
| 0.0 | 0.8056 | 0.5333 |
| 0.9333 | 3.675 | 1.9111 |
| 3.675 | 0.5333 | 2.133 |

Table 1.: Sample cluster instances

Clustered Instances
0 18 ( 75%)
1 6 ( 25%)

**The clustered instances in the figure represent that all the documents are divided into two groups.**

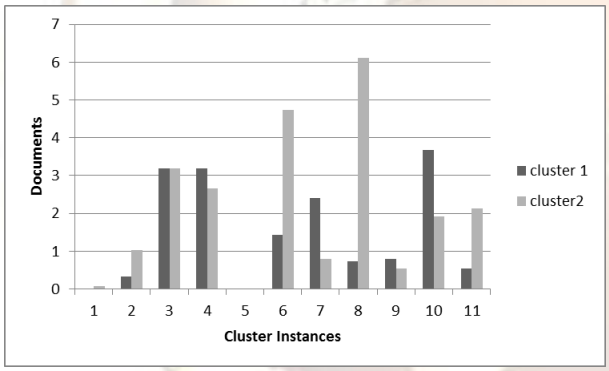


**Fig 2: Document Cluster instances.**

## VI. GA BASED MODEL

GA based clustering model was proposed in our earlier work [1], which is based on evolutionary approach for clustering & information retrieval.

### 5.1 GA Based Information Retrieval Model:

Fitness function is a performance measure or reward function which evaluate how good each solution be. The information retrieval problem is how to retrieve user required documents. Table of existing fitness functions which are used for information retrieval purpose. Suppose,X = (x1 , x2 , x3 ,….., xn) , | X | = number of terms occur in X , | *X* U*Y* | = number of terms occur in both X and Y [2].

Table: Fitness functions

| Similarity Measure Sim (X,Y) | Weighted Term Vectors |
|---|---|
| Dice coefficient | $\frac{2\sum_{i=1}^{t} x_i \cdot y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2}$ |
| Cosine coefficient | $\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{t} x_i^2 \cdot \sum_{i=1}^{t} y_i^2}}$ |
| Jaccard coefficient | $\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2 - \sum_{i=1}^{t} x_i \cdot y_i}$ |

The cosine function is mostly used to measure the similarity between two documents in the criterion function, but it may not work well when the clusters are not well separated. To solve this problem, the concepts of neighbors and link is applied to document clustering[6].
Modern information retrieval (IR) systems consist of many challenging components, e.g. clustering, summarization, etc. Nowadays, without browsing the whole volume of datasets, IR systems present users with clusters of documents they are interested in, and summarize each document briefly which facilitates the task of finding the desired documents[8]

5.2 VCF-Matching Function Algorithm :

**Assumption1:** Search space(population) contains ideal or good quality clusters and query vector contains more than one keywords.

1. Suppose Query vector is Q: a'i+b'j
2. Find the index of these keywords in the keyword list.
3. If present, for these two indices of each center take mean value of their coefficients.
4. Repeat step 3 for each cluster center.
5. Compare mean value for all cluster centers & choose maximum value.
6. Suppose V=ai+bj+ck+dl is the desired center, This can be obtained by considering the virtual center as Vv=ai+bj+0k+0l& desired center is obtained by taking the mean value (a+b)/2

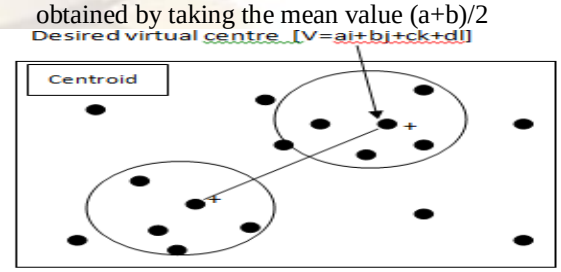


**Fig 3: Desired Virtual Center**

The advantage of Virtual Center Matching Function is that the Documents retrieved will be in the order such that they are more relevant to the keywords instead of that to center of cluster.

5.3 VCGA- GA Based IR Algorithm:

1.Generate the population of virtual centroids.

C={C1,C2,C3,…..Cn}

2. Apply fitness function(matching function) to each individual & generate fitness score.

3. Select best fit individuals (max. fitness score) & apply GA operator crossover.

4. Choose two centers randomly, perform crossover & generate two new individuals say c1' & c2'.

5. Generate new population.

6. Repeat steps 2 to 5 till termination criteria reached.

7. If {c1',c2',….cm'} is the set of desired centers then find the distance of each center from the centers in the original set & select minimum one.

8. If distance of c1' is minimum from center C5 then center C5 & all documents in its cluster will be retrieved to the user.

The flow propogation of the algorithm is shown in the fig.4

The advantages of GA based IR algorithm are :

1. Fast & Effective retrieval since number of cycles will be reduced. A single access to disk makes the entire class available.

2. Users query is matched with only cluster centers instead of with every document assuming that centers are the ideal representatives of the clusters.

5.4 Expected Results from the proposed research:

The results are expected in two phases, in first phase, good quality clusters & each cluster is having a strong representative (center) which has capability to represent a set of documents and in second phase, a new matching function & IR algorithm will results in fast & effective retrieval of desired documents to the user.

## VII. CONCLUSION

The need to structure & learn vigorously growing amount of data has been a driving force for making clustering a highly active research area. Evolutionary algorithms are not specifically learning algorithm but they offer a powerful & domain independent search ability that can be used in many learning tasks. The proposed work will add to the existing matching functions and supplement to the existing methods that will certainly improve the performance of clustering process & Information Retrieval.

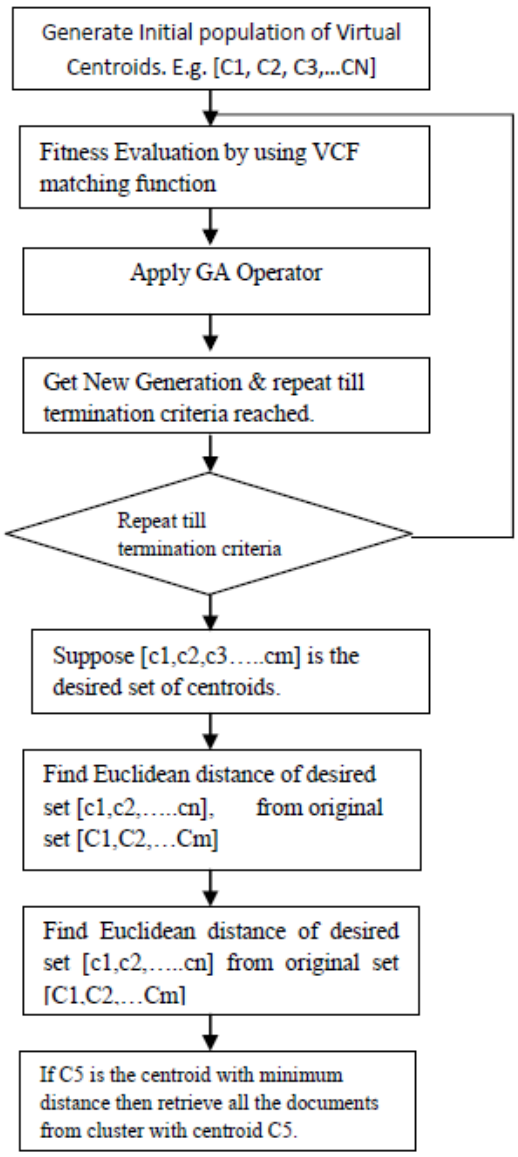


**Fig 4: Flow Propogation.**

## REFERENCES

[1] Wei Jian-Xiang, Lui Huai, su Xin- Ning, "Application of Genetic Algorithm in documentclustering"International onference on Information Technology and computer Science,2009,pp 145-148.

[2] Dr. C.A. Dhote, A. Thakare, "Evolutionary Approach for effective Clustering and IR", ICCTD-2011, Chengdu, China.

[3] Bangorn Klabbankoh, OuenPinngern, "Applied Genetic Algorithms In Information Retrieval"

[4] Yanping Lu, Shengrui Wang, Shaozi Li, change,Particle Swarm optimizer for variable weighing in clustering high-Dimensional data, Zhou January 2011, Machine Learning.

[5] Jiawei Han and Micheline kamber, " Data Mining" Concepts and techniques.

[6] Praveen Pathak Michael Gordon Weiguo Fan, Effective Information Retrieval using Genetic Algorithms based Matching Functions Adaptation Proceedings of the 33rdHawaii International Conference on System Sciences 2000

[7] Congnan Luo a, Yanjun Li b, Soon M. Chung c, " Text document clustering based on neighbors", Data &Knowledge Engineering 68 (2009) 1271–1288, Elsevier

**[8]** Wei Song, Lim Cheon Choi, Soon Cheol Park, Xiao Feng Ding, Fuzzy evolutionary optimization modeling and its Application to unsupervised categorization andextractive Summarization, August 2011,**Expert SystemswithApplication: An international Journal.**

**[9]** Philomina Simon, S. Siva Sathya "Genetic Algorithm For Information Retrieval", 978-1-4244-4711-4/2009 IEEE

[10] Ahmed A. A. Radwan, Bahgat A.Abdel latef, Abdel Mgeid A. Ali, and Osman Sadek "Using Genetic Algorithm to Improve Information Retrieval Systems",World Academy of Science, Engineering Technology,2006.

[11] Nicholas O. Andrews and Edward A. Fox, "RecentDevelopmentinDocument Clustering Techniques", Dept of Computer Science, Virgina Tech 2007.