IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

Document Clustering Analysis Based on Hybrid Clustering Algorithm

Neha Garg[1], R.K. Gupta[2]

Department of CSE and IT, Madhav Institute of Technology and Science, Gwalior, India[1, 2]

Abstract: In today's era of World Wide Web, there is a tremendous proliferation in the amount of digitized text documents. As there is huge collection of documents on the web, there is a need of grouping the set of documents into clusters. Document clustering plays an important role in effectively navigating and organizing the documents. The k-means clustering algorithm is the most commonly document clustering algorithm, it takes less computation time than a matrix-based clustering algorithm. The major problem with this algorithm is that it is quite sensitive to selection of initial cluster centroids. This article proposed a hybrid Genetic K-means clustering algorithm that improves the quality of clusters. Further, author has also performs a comparisons of hybrid algorithm and k-means algorithm on two different text document dataset. The experimental results show that the proposed method is more effective and converge to more accurate clusters than previous method.

Keywords: Document Clustering, Cosine Similarity, k-means, Genetic Algorithm, Purity measure.

I. INTRODUCTION

Due to the rapid advancement of smart technologies in World Wide Web, the volume of digitized text documents has been increasing rapidly. Clustering plays an important role for organizing such massive document collection returned by search engines into meaningful clusters. The purpose of clustering is to introduce an order in a collection of documents by grouping or classification [3]. There are various clustering algorithms are available to cluster text documents.

The k-means clustering algorithm is one of the most widely used clustering algorithm that arrange the documents in order such that a document is close to its related document on the basis of similarity measure. It is simple to understand and computationally efficient. However, in this algorithm there is a need to define the number of clusters in advance which is difficult to set. Also the final results of the algorithm are sensitive to the selection of initial cluster centroids and may converge to local optima [8]. There are various methods have proposed by the authors for improving the performance of the k-means algorithm.

Yogesh kumar et al. [9] uses the feature of genetic algorithm and Discrete Differential Evolution (DDE) for text document clustering. K-means is a popular method but its results are heavily depends on initial selection of cluster centroids. So, for the refinement of the initial cluster seeds, the author uses the combination of GA and DDE which gives better results in less iteration. Experimental results of Reuters-21578 text dataset show that the proposed algorithm gives better results than the GA and DDE.

Vikas Kumar Sihag et al. [10] define a graph based method for computing the initial cluster centroids for k-means clustering algorithm. In this method, nodes and edge of the graph represent documents and value of similarity between documents respectively. This method first finds the edge that has lowest weight from others and deletes that edge from network for detecting a community structure then, computing the centrality of each node on the basis of cohesiveness and dissimilarity. The node that has high centrality is the initial seeds for the k-means clustering. Experiment results of the F-measure values illustrates that the graph based method performs better in terms of accuracy as compared to the existing method.Later, in 2014, author also defined a new method for the selection of k clusters by the modified spectral bisection. A result obtained by this method is given to a genetic algorithm. Final results show that the new method is better than the popular k-means method [11].

K. Premalatha et al. [13] presents a document clustering based on genetic with Simultaneous and Ranked Mutation. Simultaneous mutation means that genetic algorithm uses different mutation operators producing next generation and the mutation ratio depends on respective offspring that produces. Rank mutation means that the mutation depends on fitness rank of chromosomes of earlier population. Experimental results demonstrate that the proposed method performs better than simple GA and k-means method. Harish verma et al. [15] defines a modified genetic algorithm by defining new method for the initial population selection not adopting the random population. The creation of initial population is based on the similarity between the documents computedby sum of squared distances.

An entropy result shows that the performance of the new method gets improved.

The further section of this paper is organized as follows: Section II is devoted to methodology. Section III defines a standard k-means clustering algorithm. The hybrid genetic

DOI 10.17148/IJARCCE.2016.54190

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

k-means (GA-KM) clustering algorithm is described in Section IV. Section V discusses the experimental results by comparing the performance of GA-KM algorithm with k-means algorithm. Section VI concludes the paper.

II. METHODOLOGY

The Figure 1 depicts the stages of the process of clustering which starts with collecting text documents from various sources then preprocessing is done to extract the useful patterns. The preprocessed documents are represented as Vector space model, it is necessary to generate the document vector before applying any clustering algorithm on text dataset.

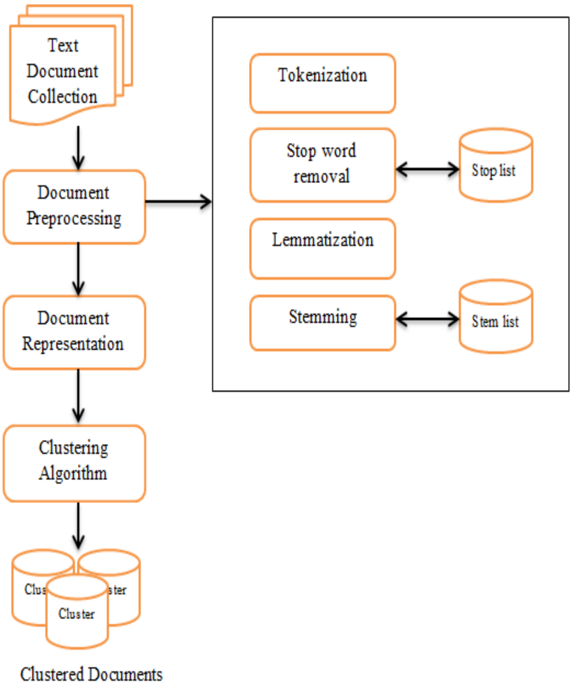

Fig. 1 The stages of the process of Clustering

Preprocessing of Text Document
Preprocessing is a critical step in text mining, used for extracting the interesting knowledge from unstructured text data. There are various preprocessing methods such as tokenization, stop words removal, lemmatization, stemming etc. are defined to extract the useful patterns from the text documents because the text data contains some special formats like date formats, number formats and the most common words which comes in every text documents that unlikely to help text mining such as prepositions, articles, and pro-nouns needs to be removed from the documents.

A. Text Document Representation

In the vector space model a document d is represented as d= {wt_1, wt_2,….,wt_n} where t_1,t_2,….t_n denotes the distinct terms present in a document d and w_n denotes the weight of term t_n .

The term weight is determined using the Term Frequency-Inverse Document Frequency (TF-IDF).
Weight of term t in document d is computed as:

$$W_{td}= tf_{td} * idf_{td} = tf_{td} * \log (N/df_{td}) \qquad (1)$$

Where tf_{td} denotes the number of occurrences of term t appears in document d.

df_{td} denotes the number of documents in which term t appears.

N is the total number of documents.

B. Similarity Measure

Before clustering of documents, similarity of documents must be determined in a cluster analysis.There are various similarity measures available to compute the similarity of two documents such as Euclidean distance, Cosine similarity, Manhattan distance, Jaccard coefficient etc. Among these measures the Cosine similarity measure as the best similarity measure for k-means document clustering [2] has been used. If there are two vectors d1and d2 then the Cosine similarity can be calculated as:

$$\text{Cos}\ (d1, d2) = \frac{d1.d2}{|d1||d2|} \qquad (2)$$

where · denotes the dot product and | | denotes the length of the vector.

III. STANDARD K-MEANS CLUSTERING ALGORITHM

The k-means algorithm [12] is a centroid based partitioning clustering method. The process of k-means starts by initializing the k cluster centroids by selecting k documents from the collection. For each of the remaining documents, find the cluster whose centroid is most similar on the basis of similarity measure. For every cluster, recompute the cluster centroid based on the mean value of the current member documents. This process is continued until there is no change in the clusters of documents [1]. Although the k-means method is simple to understand and it is quite sensitive to initial selection of cluster centroids. There are several methods of creating the initial clusters centroids [3]:

In first method, select the k documents randomly as the initial cluster seeds. In second method, instead of picking the cluster centroids randomly, pick the most likely documents for the initial centroids. In third method,create the samples of documents as the k initial clusters and then uses the various optimization techniques to find the optimal cluster centroids.

Algorithm k-means (k, D)

```
1.   Select k documents from the document collection D to form k initial cluster centroids.
2.   Repeat
2.1  Reassign every document to the cluster whose centroid is most similar.
2.2  Update the cluster centroid by calculating the mean value of documents for each cluster.
3.   Until, no changes.
```

DOI 10.17148/IJARCCE.2016.54190

IV. HYBIRD GENETIC -K-MEANS (GA-KM) CLUSTERING ALGORITHM

A Genetic Algorithm (GA) is a search technique used in computing to find the approximate solutions to optimization and search problems [14]. The searching capability of GA has been used for the purpose of determining appropriate cluster centroids. An implementation of genetic algorithm begins with the population of chromosomes. An objective and fitness function associated with each chromosome is used to guide the selection of chromosomes which are used to generate the new candidate solution through crossover and mutation. Crossover generates new chromosomes by combining sections of two or more selected parents. After that mutation is applied on these individuals to yield a new population. The process of selection, crossover and mutation is repeated until the highest ranking fitness is reached. The basic steps of GA are defined as follows:

A. Individual Representation

The chromosome data structure stores the entire population in a matrix of size N*L where Nis the number of individuals in the population and Lis the length of the genotypic representation of individuals. Initially create k arrays of documents of size (n/k) where n is the total number of documents and k is the number of clusters which represent k initial clusters. Now, each individual is encoded with real numbers containing k cluster centroids of the initial clusters.

B. Fitness Computation

Fitness value is derived from the objective function through ranking method. Objective function value(s) are stored in a matrix of size N*O where N is the number of individuals in the population and O is the number of objectives to compute the fitness function. We use the fitness function for chromosome i as 1/DBj, where DBj is the well-known Davies-Bouldin index [4,5] computed for chromosome i.

C. Selection

Selection process selects the chromosomes from the mating pool with higher fitness value. In this article, Rank selection strategy is adopted, chromosomes are sorted according to their fitness value and chromosomes are selected for the crossover on the basis of their rank.

D. Crossover

Crossover is a probabilistic process that accepts the two parent chromosomes and creates the two child chromosomes for the next generation of solutions. The chromosome of length L and random crossover point is generated in the range of [1, L-1]. In this article, uniform crossover with crossover probability p_c is adopted.

E. Mutation

Each chromosomes undergoes mutation with a fixed mutation probability p_m, since floating point consideration is used therefore, generate a random number ∂ between [0,1] with uniform distribution. After mutation, the value x at gene position is modified by:

$$x = x \pm 2 * \partial * x, \qquad \text{for } x \neq 0 \qquad (3)$$

$$x = x \pm 2 * \partial, \qquad \text{for } x = 0 \qquad (4)$$

Algorithm: GA based refinement

```
1) Create k number of Arrays $A_1$, $A_{2,}$…$A_k$.
2) Move the documents from input array to Arrays $A_k$ until $S_c = \lceil n/k \rceil$
3) Repeat step 2 until all documents removes from input array.
4) Now, initialize each chromosome containing k centroids from Arrays $A_k$.
5) Test the fitness of each chromosome in the population.
6) While (Not convergence reached) do
i. Select parents as the most fit members of the population to reproduce;
ii. Perform crossover with rate $p_c$ over selected pair of parents.
iii. Perform mutation with rate $p_m$ and get the new population;
iv. Test the fitness of each chromosomes in the new population;
endwhile
1. Return with k cluster centroids vectors.
```

The hybrid genetic k-means (GA-KM) clustering algorithm first execute the Genetic algorithm (GA) to find the optimal cluster centroids and the results of the GA algorithm is then used as initial centroid vectors of the k-means (KM) algorithm which gives the final list of clusters as output.

V. EXPERIMENTAL RESULTS

Here we demonstrate the outcome of the experiments by applying the hybrid GA-KM algorithm and k-means algorithm. We took Reuter21578 and Cnae9 text dataset for comparing the performances of both the algorithms. There are various measures to validate the quality of clusters. These measures can be internal or external [6]. Internal measure permits us to compare the different clusters results without reference to outer information e.g. inter-cluster similarity and intra-clustersimilarity. External measure isevaluated the how wellthe clustering is workingby some a priori knowledge e.g. Purity, Entropy, F-measure etc.

Purity

Every cluster is appointed to a class which is most successive in the cluster and the accuracy of this assignment is evaluated by counting the number of correctly assigned documents [7]. Purity is calculated as:

$$\text{Purity} = \frac{1}{N} \sum_k \max_j \left| c_k \cap t_j \right| \qquad (5)$$

Where c_k be cluster k.

t_j is the classification which has maximum count in cluster c_k and

N is the number of documents.

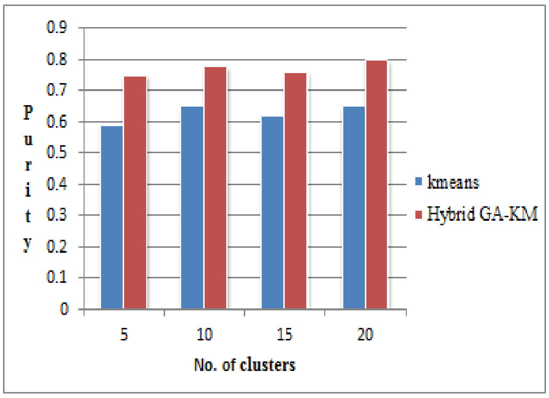

Fig. 2 Purity results of two different algorithms when run on Cnae9 dataset

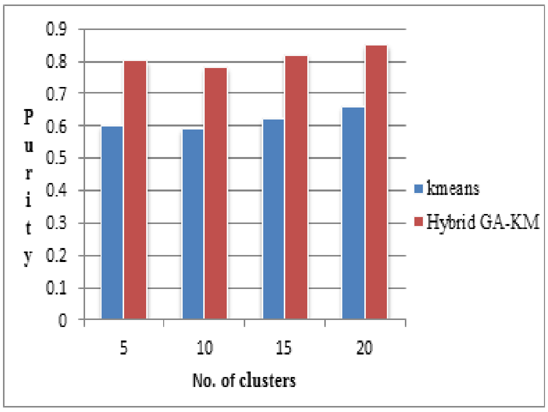

Fig. 3 Purity results of two different algorithms when run on Reuter21578dataset.

The results obtained by applying both algorithms to above mentioned datasets are presented in Table 1. The performance of the clusters is evaluated by the Purity measure. From the results it is evident that Purity of clusters is higher for the hybrid algorithm as compared to the k-means algorithm, it means that proposed algorithm gives better results than the existing algorithm.

Table 1. Performance comparisons of k-means and hybrid GA-KM clustering algorithm.

		k-means	Hybrid GA-KM
	No.of Clusters	Purity	Purity
Cnae9 Dataset	5	0.59	0.75
	10	0.65	0.78
	15	0.62	0.76
	20	0.65	0.80
Reuter 21578 Dataset	5	0.60	0.80
	10	0.59	0.78
	15	0.62	0.82
	20	0.66	0.85

VI. CONCLUSION

In this paper, we have proposed our work by using the hybrid algorithm for improving theclustering performance in text mining. The proposed algorithm first executes the genetic algorithm for the refinement of the initial cluster centroids and the result of the genetic algorithm is given as initial seeds of the k-means algorithm. Our experimental results illustrate that hybrid GA-KM algorithm achieves better results than the k-means algorithm when applied to real datasets.

As a future work, would like to work more on this algorithm to make it fully automatic with the goal that it will require no parameter to be determined.

REFERENCES

[1]. J. Han, M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann, 2nd Ed., 2006.

[2]. S. Jaiganesh, P. Jaganathan,“An Appropriate Similarity Measure for K-Means Algorithm in Clustering Web Documents”, International Journal for Scientific Research & Development, Vol. 3, Issue 02, 2015.

[3]. M. konchady, “Text Mining Application Programming”, Programming Series, Charles River Media, 2006.

[4]. S. Bandyopadhyay, U. Mauilk, “Nonparametric genetic clustering: Comparison of validity indices”, IEEE Trans. System Man Cybern.-Part C Applications and Reviews, Vol. 31, pp. 120-125, 2001.

[5]. D.L. Davies, D.W. Bouldin, “A cluster separation measure”, IEEE Trans. Pattern Anal. Intell., Vol. 1, pp. 224-227, 1979.

[6]. M. Steinbach, G. Karypis, V. Kumar, “A Comparison of document clustering techniques”, Technical report, Department of Computer Science and Engineering, University of Minnesota, 2000.

[7]. S. Karol, V. Mangat, “Evaluation of text document clustering approach based on particle swarm optimization”, Central European Journal of Computer Science, Vol. 3, pp. 69-90, 2013.

[8]. S.Z. Selim, M.A. Ismail,“K-means type algorithms: A generalized convergence theorem and characterization of local optimality”, IEEE Trans. Pattern Anal. Mach. Intell. Vol. 6, Issue 01, pp. 81–87, 1984.

[9]. Y.K. Meena, Shashank, V.P. Singh, “Text document clustering using genetic algorithm and discrete differential evolution”, International Journal of Computer Applications, Vol. 43, No. 1, April 2012.

[10]. V.K. Sihag, S. Kumar, “Graph based Text Document Clustering by Detecting Initial Centroids for k-means”, International Journal of Computer Applications, Vol. 62, No. 19, Jan 2013.

[11]. P.M. Dhanya, M. Jathavedan, A. Sreekumar, “Implementation of Text clustering using Genetic Algorithm”, International Journal of Computer Science and Information Technologies, Vol. 5, 2014.

[12]. A.K. Jain, R.K. Dubes, “Algorithms for clustering data”, Prentice Hall, 1998.

[13]. K. Premalatha, A.M. Natarajan, “Genetic Algorithm for Documents Clustering with Simultaneous and Ranked Mutation”, Modern Applied Science, Vol. 3, No. 2, 2009.

[14]. J. Holland, “Adaptation In Natural and Artificial Systems”, University of Michigan Press, 1975.

[15]. H. Verma, E. Kandpal, B. Pandey, J. Dhar, “A Novel Document Clustering Algorithm Using Squared Distance Optimization Through Genetic Algorithms”, International Journal on Computer Science and Engineering, Vol. 2, No. 5, pp. 1875-1879, 2010.

DOI 10.17148/IJARCCE.2016.54190