International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Document Retrieval and Cluster Based Indexing using Rider Spider Monkey Optimization Algorithm

Madhulika Yarlagadda, K. Gangadhara Rao, A. Srikrishna

Abstract: ***Document retrieval process is more significant in the field of research community for retrieving the highly-relevant documents that fit for the user query. Even though various document retrieval methods are introduced, retrieving the exact document based on the indexing is a quite challenging task in the document retrieval framework. Thus, an effective document retrieval algorithm named Rider Spider Monkey Optimization Algorithm (RSOA) is proposed in this research. Initially, the documents are pre-processed by the stop word elimination and the stemming process, and the features are extracted to find the key words of the documents by applying the Term Frequency-Inverse Document Frequency (TF-IDF). The selected keywords are passed into the cluster-based indexing phase, where the cluster centroids are identified by using the proposed Rider Spider Monkey Optimization Algorithm. Moreover the query matching is carried out at two levels, at first, the query is forwarded and is matched to the entire cluster centroid to find the appropriate centroid. At the second level; the user query is matched based on the records present inside the matched centroid. Moreover, the query matching is progressed using the distance measure by the Bhattacharya distance to retrieve the documents. The performance is analyzed using the metrics, namely precision, F-measure, and recall and accuracy with the values of 90.141%, 91.876%, 91.178%, and 91.202%, respectively using 20 news group dataset .***

Keywords: cluster based indexing, cluster centroid, stop word removal, holoentropy, Rider spider monkey optimization.

I. INTRODUCTION

Retrieval of documents is the process of identifying the relevant records from the group of information document with respect to the user query [3]. However, the data structure otherwise called as an index that is assembled over the set of records in the domain, so that, the user can retrieve the best document that fits the user query. Different application uses different searching process, while the string pattern in the generic solution for the document retrieval problem [6].

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Smt. Madhulika Yarlagadda, Assistant Professor, Department of Information Technology, R.V.R & J.C College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India.
Dr. K. Gangadhara Rao, Professor, Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.
Dr. A. Srikrishna, Professor and head, Department of Information and Technology, RVR & JC College of engineering, Chowdavaram, Guntur, Andhra Pradesh, India.

Most of the traditional methods first scan all the documents and calculates a score for each cluster document based on the query of the user.
The exact relevant documents are selected by applying the ranking function, and the document with the maximum score is retrieved at first. The document retrieval process holds the polynomial time complexity while answering the queries for the large amount of documents.

One of the effective methods to enhance the retrieval performance in the information retrieval process is the cluster based method. However, the pre-processing step is used to retrieve the documents and the selected documents are partitioned to clusters based on the similarity of the documents. The cluster based method identifies the cluster based on the query of the user and from the cluster it retrieves the matched document, as is does not scan the whole document [8]. In order to access the important records from the domain, the user is required to provide the correct keyword of the content. Therefore, the efficiency and the effectiveness in retrieving the documents may get pretentious, as the system does not have the knowledge of user regarding the searched records [2].

Various data mining methods are utilized to simplify the information retrieval complexity by developing the valuable data from the collection of records. However, the entire documents are partitioned to different groups, and hence the documents are searched in each group by using some computing tools [3]. The clusters are grouped into disjoint clusters through k-means approach, where in each cluster group there exist the relevant documents [3]. The classical methods used in document retrieval process scan all the available documents and calculates the score between the user query and the documents. The relevant documents are placed by using the ranking function, which contains the polynomial complexity [3]. Due to the enormous growth of internet, the documents are stored in an electronic format or in the form of hard copy [16]. Effort has been taken for the integrated document retrieval and the document filling system, which is used in various aspects, like document reproduction, categorization, classification, document retrieval, and document storage. The overview of the documents is captured and the logical structures are retrieved by using the structure model. The conceptual structure and the domain information, such as organizational activities, and the document usage are captured by using the conceptual model [2]. Pseudo relevance feedback mechanism (PRF) is used to make high representation in terms of short query and also assure to enhance the retrieval performance.

PRF mechanism assures that the documents retrieved in the top ranked lists are equivalent to the user query and therefore these documents are used in the form of feedback to retrieve the appropriate information [1] [9]. These methods are most effective in the document retrieval process by enhancing the searching efficiency [1].

Due to the existence of the limited keyword query, the PRF method is not reliable in retrieving the top ranked information for the very short query. Since, the off-topic information is extracted by using the feedback process, which further integrates the retrieved documents with the original query. The appropriate feedback for the documents are effectively selected by assigning the weights to each information document, placing the feedback terms for various documents using the weight, and to design the language model for eliminating the effect of distraction caused by the observations [1] [10] [11]. Such methods are most effective in focusing and retrieving the feedback based documents, which are may be negative or positive feedback, and hence it attained better document retrieval performance based on the query relevant to the user. In certain cases, the negative feedback may present in the better information document set [12]. The learned negative feedback mechanism affects the irrelevant documents, as it holds the negative topics, and hence they are unseen [13]. In the negative feedback approach, the language model was introduced to search the space for retrieving the top-ranked information. The effectiveness of document retrieval is enhanced by combining the negative with the positive model. In general, the irrelevant documents attain better retrieval performance rather than the relevant documents. Since, the negative document contains the positive terms, the term distribution affects the PRF methods [1] [14] [15]. Several algorithms, like Frequent itemset based hierarchical clustering (FIHC) [18], Lazy associative tag recommender (LATR) [20], Topic document clustering (TDC) [19], Frequent itemset mining (FIM) [21] [22], and Hierarchical frequent term based clustering (HFTC) [17] are utilized to select the frequent itemset [8].

The primary intention of this research is to design and develop an effective cluster based indexing scheme for information retrieval. The proposed document retrieval process involves five phases, as pre-processing, feature extraction, keyword selection, indexing module, and query matching. Initially, the data taken from the database is pre-processed through stop word removal and stemming process. The features are extracted from the pre-processed documents to find the keywords of the documents using TF-IDF. Then, the key terms are selected by the holoentropy and are further fed into the cluster based indexing phase, where the proposed RSOA optimization is progressed to find the cluster centroid. The query matching is progressed at two levels, at first the query is matched with the cluster centroid and at second, the query is send and is matched in the entire records of the cluster centroid to retrieve the matched document.

The contribution of this research is demonstrated as follows:

- The features are extracted from the documents by using TF-IDF such that suitable keywords are selected based on the holoentropy.
- The cluster based indexing is performed using the proposed RSOA optimization and the suitable centroids are identified through which the record matching is achieved.

The rest of this paper is organized as: the literature survey of the existing methods is elaborated in section 2. Section 3 explains the proposed spider monkey optimization algorithm for information retrieval, and the results along with the analysis are elaborated in section 4. Finally, section 5 concludes the paper.

II. MOTIVATION

The motivation of the proposed cluster based indexing is explained in this section, which involves various existing document retrieval methods.

A. Literature survey

Different existing document retrieval techniques are surveyed. Hao S *et* al. [1] introduced an expectation maximization algorithm to enhance the retrieval performance of the documents. The negative and the positive feedback were integrated by using the query language model to achieve valuable in document retrieval. However, it achieves better average precision in the entire document process, but the weights were not updated for each retrieval query. Kayest M. and Jain S.K [2] developed monarch butterfly algorithm for the information retrieval framework. Here, the documents were pre-processed by the stop word and the stemming process. The selective keywords were identified based on the holo-entropy concept to achieve the document retrieval. Even though it achieves better retrieval performance, it failed to use the recommender system. Djenouri Y *et al.* [3] developed Bees swarm optimization algorithm to achieve the document retrieval process. The collected documents were grouped into clusters based on the type of document. From each cluster group, the frequent item sets were extracted effectively. However, it enhances the quality of the documents, but the running time is competitively high. Farhi S.H and Boughaci D [4] developed a stochastic method for extracting the sub-graphs for retrieving the documents. The subgraph frequent set and the query size were utilized to minimize the index size of the documents. In order to achieve the best retrieval process, the query size was equivalent to the index size. It achieves better search space and attained effective quality solution, but the running time is very high. Biswas S *et al.* [5] introduced a linear space index method for the document retrieval process. The most relevant records were extracted through the query time using the linear data structure. Moreover, it attained better retrieval performance, but the forbidden patterns were not extracted. Ferrada H. and Navarro G [6] introduced a lempel-ziv based retrieval method to retrieve the top documents. The index used in the document retrieval processes were build in the form of arrays and suffix trees.

This method attained effective fast search structure, but the time used in the retrieval process is high. Wang N *et al.* [7] developed a hierarchical based retrieval method for document collection. Here, the documents were encrypted and the retrieval features were constructed based on the document attributes. It enhances the searching efficiency by using the parallel computing process.

However it was not suitable for the large sized documents. Djenouri Y *et al.* [8] developed a cluster based approach for document retrieval. The clusters were extracted effectively by combining the frequent itemset with the k-means clustering scheme. The patterns were selected from each cluster group to retrieve the relevant documents based on the user query. It attained better quality of documents, but it was not suitable for the data types, like videos and images.

B. Challenges

- Retrieving the exactly right documents using the short query is a major challenging task in the document retrieval process, as the average person uses the short query to express and explain their expectation to retrieve the documents [1].
- Reducing the usage space while retrieving the documents based on the user query is a challenging task associated with the document retrieval process. Relevant documents are retrieved based on the selected keywords [2].
- To retrieve the top significant records using the knowledge discovery of the domain is a major challenge of document retrieval process. However, the score function ranks the relevant documents based on the patterns [6].
- Due to high dimensionality, the execution time required by the data set is high for dealing with the large sized documents, which is a major challenge of the document retrieval process [8].
- Protecting the data confidentiality while searching the documents and enhancing the searching efficiency for the large sized document is challenging task in the document based retrieval mechanism [7].

III. PROPOSED CLUSTERING BASED DOCUMENT INDEXING USING RIDER SPIDER MONKEY OPTIMIZATION ALGORITHM

The document retrieval process gained more importance in synthesizing and browsing the information from various documents. An effective information retrieval mechanism using the optimization algorithm is proposed in this research. The proposed cluster indexing-based document retrieval involves five phases, as pre-processing, feature extraction, keyword selection, clustered indexing, as well as query matching. Initially, the records available in the database is forwarded into the pre-processing phase, where the documents are pre-processed using the two tasks, like stop word removal and stemming process. Later, the pre-processed records are fed to the feature extraction module, where the keywords of the documents are extracted through the TF-IDF process. The selective key terms are selected using the holoentropy process. The selective keywords are further processed by the proposed cluster-based indexing using RSOA optimization algorithm, which is developed by combining the Rider Optimization Algorithm (ROA) [23] with Spider Monkey Optimization (SMO) [24]. However, the proposed RSOA algorithm performs the clustering mechanism by finding the suitable centroids based on which the query matching is progressed. Matching the query is performed at two levels, at first the query is matched to the cluster centroid to identify the appropriate centroid. At the second level, the query is matched through the keywords of the optimal centroid matched at the first level matching and thus, it retrieves the best suitable document. Here, the matching is progressed using the distance measure based on the Bhattacharya distance. Figure 1 shows the block diagram of the proposed cluster indexing based document retrieval.

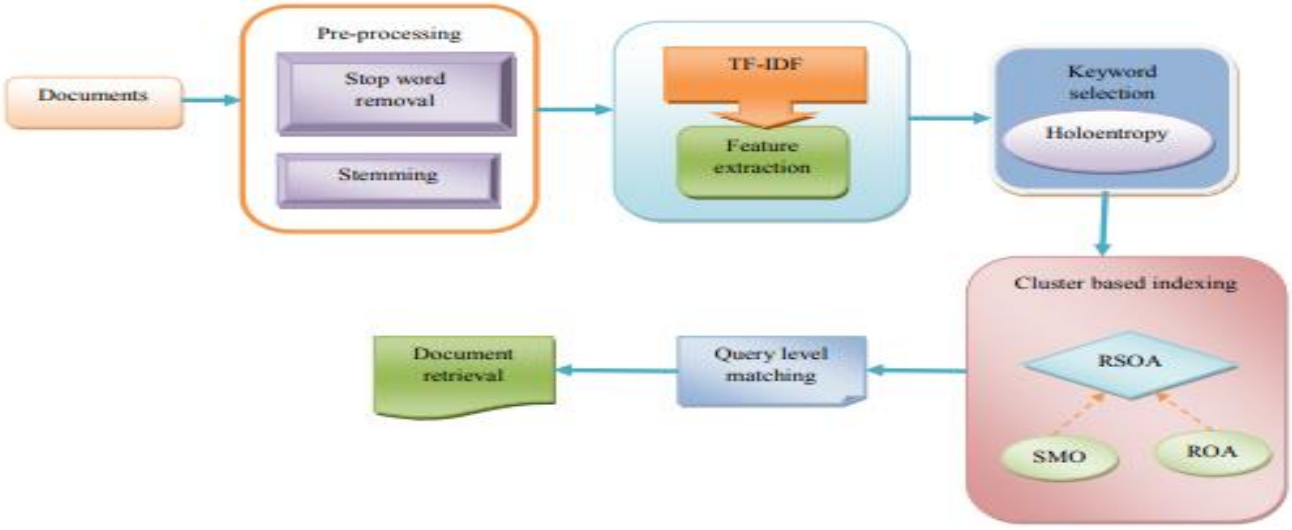

Figure 1. Block diagram of the proposed spider monkey optimization algorithm

A. Pre-processing

The documents present inside the database is forwarded into the pre-processing phase. In preprocessing step, the stop word removal and stemming process are carried out by the documents. Hence, the pre-processed documents are used for the further process in the document retrieval approach for the effective retrieval of the documents.

B. Feature extraction from the pre-processed documents

The features are extracted for the individual pre-processed documents acquired from the input database. The Term Frequency (TF) defines the count of occasion for each word present in the record, which determines the document measure of the term. Number of documents which exist under consideration is called corpus. The term importance of the document is computed using the weight allocated in the specific document. The document vector using the term frequency is expressed as,

$$A_m = \frac{d_m}{\max(d_m)} \tag{1}$$

Here, d_m denotes the TF of m^{th} term, and m varies as $m = 1,2,3.....,n$. The TF for the long-sized documents are reduced by dividing the TF with the total words present in the record as given by,

$$A_m = \frac{d_m}{size(a)} \tag{2}$$

where, d_m represents the TF of document a, and $size(a)$ denotes the raw frequencies in the document. Moreover, TF is the local measure and IDF is the global measure of the term importance in the document. It shows a small value for the repeated words, and the large values for the unique words. Hence, IDF is defined as,

$$B(b_m) = \log \frac{E}{c_m} \tag{3}$$

where, E represents the documents present in the dataset, and c_m denotes the document frequency. Hence, TF-IDF is computed using the below equation as,

$$D = A \times B \tag{4}$$

where, D denotes the TF-IDF, A indicates the term TF, and B indicates the term IDF. Thus, the individual document is represented as TF and IDF, which is further subjected to the keyword selection phenomenon.

C. Keyword selection

The keyword is selected from the extracted features based on the holoentropy process [25]. The best keywords are selected using the holoentropy, and for each attribute the holoentropy is calculated as,

$$X(e) = P.S(e) \tag{5}$$

where, e denotes the attribute vector, P be the weight value, and S be the entropy measure. The feature attribute with the maximum score is selected as the best keyword.

D. Centroid-based indexing for document retrieval using proposed RSOA

The selected keywords are subjected to the indexing process, which is a significant step for enabling the effective retrieval of the relevant documents. The centroid-based indexing scheme is used to generate the cluster centroids using the proposed RSOA, which can be further used to perform the query matching in order to retrieve the highly relevant documents.

D.(a). Solution encoding

The solution encoding is the representation of the solution vector, which presents the optimal centroids obtained using the proposed optimization algorithm. Here, each solution vector specifies the centroids in the document retrieval framework.

D.(b). Fitness function

Fitness function or the objective function aims to find the optimal solutions, which is based on the distance between the individual point with respect to the centroid. Hence, the fitness function is computed as,

$$F = \sum_{y=1}^{P} \sum_{z=1}^{Q} \left\| f_{yz}^{select}, S_a \right\| \tag{6}$$

where, f_{yz}^{select} represents the data point, and S_a denotes the a^{th} centroid.

D.(c). Proposed Rider Spider Optimization algorithm

The behavior of foraging in the spider monkey is depends on the structure of fission fusion, which is used in this research for retrieving the documents based on the user query. The foraging behavior of the swarm is used to reduce the social competition between the group members by partitioning the swarm into different subgroups to search their food. Here, the global leader is the female swarm, which guides the group and is the most dependable leader in penetrating the source of food. When the leader does not receive enough food sources, then it partitioned the group into various smaller subgroups, where each subgroup forages independently. The local leader leads the subgroup and makes a decision to plan for the foraging route. Depend on the availability of the food, the subgroup members communicated within and outside of the group to maintain the boundaries. The foraging behavior of the spider monkey involves four steps. At first they started to search their food; and compute the distance between the sources of food. Next, the members change their location through the distance of the sources of food. In third step, the local leader changes the best position inside the group. At the final step, the global leader changes its best position and partition the group into the smaller subgroups. The procedure involved in the proposed RSOA optimization algorithm for document retrieval is elaborated as follows:

The proposed optimization algorithm contains six different phases based on the iterative process, as local leader, local leader learning, local leader decision, global leader, global leader learning, and global leader decision phase.

a) Population initialization

The initial population of the spider monkeys is uniformly distributed, where each monkey has N dimensional vector. Each monkey is correspond to the solution in the optimization problem, where each spider monkey is represented as,

$$G_{m,n} = G_{\min n} + Y(0,1) \times (G_{\max n} - G_{\min n}) \quad (7)$$

where, G_m denotes the m^{th} spider monkey, and nth dimension, in the population initialization, $G_{\min n}$ and $G_{\max n}$ are the bounds of G_m, and $Y(0,1)$ is the random number lies in the interval of $[0,1]$.

b) Local leader phase

Here, each spider monkey changes their present location depends on the information of local leader. The fitness value for the new changed position is computed, and if the fitness measure of the updated position is greater than the previous position; then the spider monkey updates its location through the new position. Hence, the update equation of the location is expressed as,

$$G_{newmn} = G_{mn} + Y(0,1) \times (I_{in} - G_{mn}) + Y(-1,1) \times (G_{fn} - G_{mn}) \quad (8)$$

where, G_{mn} denotes the n^{th} dimension of m^{th} spider monkey, I_{in} denotes the n^{th} size of local leader, G_{fn} indicates the n^{th} size of f^{th} spider monkey. Let us assume $G = M$ in the above equation as,

$$M_{mn}(b+1) = M_{mn}(b) + Y(0,1) \times (I_{in} - M_{mn}(b)) + Y(-1,1) \times (M_{fn}(b) - M_{mn}) \quad (9)$$

$$M_{mn}(b+1) = M_{mn}(b)[1 - Y(0,1) - Y(-1,1)] + Y(0,1) I_{in} + Y(-1,1) M_{fn}(b) \quad (10)$$

The spider optimization is modified using bypass rider equation of ROA, which is expressed as,

$$M_{mn}(b+1) = \lambda [M_{\upsilon n}(b) * \mu(n) + M_{\omega n}(b) * [1 - \mu(n)]] \quad (11)$$

Let us assume $\upsilon = \omega = m$,

$$M_{mn}(b+1) = \lambda [M_{mn}(b) * \mu(n) + M_{mn}(b) * [1 - \mu(n)]] \quad (12)$$

$$M_{mn}(b+1) = M_{mn}(b)[\lambda * \mu(n) + \lambda * [1 - \mu(n)]] \quad (13)$$

$$M_{mn}(b) = \frac{M_{mn}(b+1)}{\lambda * \mu(n) + \lambda * [1 - \mu(n)]} \quad (14)$$

Substitute the Eq. (14) in Eq. (10),

$$M_{mn}(b+1) = \frac{M_{mn}(b+1)}{\lambda * \mu(n) + \lambda * [1 - \mu(n)]} [1 - Y(0,1) - Y(-1,1)] + Y(0,1) I_{in} + Y(-1,1) M_{fn}(b) \quad (15)$$

$$M_{mn}(b+1) - M_{mn}(b+1) \frac{[1 - Y(0,1) - Y(-1,1)]}{\lambda * \mu(n) + \lambda * [1 - \mu(n)]} = Y(0,1) I_{in} + Y(-1,1) M_{fn}(b) \quad (16)$$

$$M_{mn}(b+1) \left[1 - \frac{[1 - Y(0,1) - Y(-1,1)]}{\lambda * \mu(n) + \lambda * [1 - \mu(n)]} \right] = Y(0,1) I_{in} + Y(-1,1) M_{fn}(b) \quad (17)$$

$$M_{mn}(b+1) \left[\frac{\lambda [\mu(n) + [1 - \mu(n)]] - 1 + Y(0,1) + Y(-1,1)}{\lambda [\mu(n) + [1 - \mu(n)]]} \right] = Y(0,1) I_{in} + Y(-1,1) M_{fn}(b) \quad (18)$$

$$M_{mn}(b+1) \left[\frac{\lambda [\mu(n) + 1 - \mu(n)] - 1 + Y(0,1) + Y(-1,1)}{\lambda [\mu(n) + 1 - \mu(n)]} \right] = Y(0,1) I_{in} + Y(-1,1) M_{fn}(b) \quad (19)$$

$$M_{mn}(b+1) \left[\frac{\lambda - 1 + Y(0,1) + Y(-1,1)}{\lambda} \right] = Y(0,1) I_{in} + Y(-1,1) M_{fn}(b) \quad (20)$$

$$M_{mn}(b+1) = \frac{\lambda}{\lambda - 1 + Y(0,1) + Y(-1,1)} [Y(0,1) I_{in} + Y(-1,1) M_{fn}(b)] \quad (21)$$

where, λ is the random number which varies from 0 to 1, and I_{in} denotes the i^{th} local group leader in n^{th} dimension.

c) Global leader phase

Once the local leader module is completed, then the global leader starts. Here, the spider monkey changes its location based on the knowledge of local group member, as well as the global leader. Hence, the updated equation of the global leader phase is expressed as,

$$G_{newmn} = G_{mn}(0,1) \times (V_n - G_{mn}) + Y(-1,1) \times (G_{fn} - G_{mn}) \quad (22)$$

where, V_n denotes the global leader position of n^{th} dimension.

d) Global leader learning phase

The selection of greedy scheme is integrated into the members for updating the location of the global leader. Moreover, the spider monkey having the best value is elected as the global leader among the population.

e) Local leader learning phase

The selection of greedy scheme is applied into the group to update the location of the local leader. The spider monkey having the best value is elected as the local leader of the population.

f) Local leader decision phase

If any of the local leader never change its position, then the members present in the group is updated by using the combined information or by the random initialization.

$$G_{newmn} = G_{mn} + Y(0,1) \times (V_n - G_{mn}) + Y(0,1) \times (G_{mn} - I_{in}) \quad (23)$$

g) Global leader decision phase

Here, the location of the global leader is clearly monitored, if the global leader never updates its position, next global leader partition the group into sub-groups. Table 1 illustrates the pseudo code of the proposed RSOA algorithm.

Table 1. Pseudo code of the RSOA algorithm

Sl. No	RSOA algorithm
1	Population initialization
2	Fitness calculation
3	Greedy selection is used to elect the local leader and the global leader
4	While(termination criteria is not satisfied)
5	Do
6	Generate the new position of the members using local leader experience, group member experience, and self experience.
7	Greedy selection is used between the new position and the existing position depends on the fitness value for selecting the best one.
8	Compute the probability function for the group members
9	Generate a new position for the group members through which the knowledge of global leader, group members, and self experience.
10	The greedy framework is utilized to the groups for updating the position of global leader and the local leader.
11	If the local group leader never change the position, then all the members in the group are re-directed for foraging.
12	When the global leader is not updated its position, then it partition the groups into various smaller sized subgroups.
13	End while

Thus, the output from the RSOA is the clusters, which is represented as,

$$C = C_l ;\ [1 \le l \le m] \tag{24}$$

There are a total of m number of clusters that are stored in such a way that whenever a new query arrives, the query is coordinated through the cluster centroids for retrieving the relevant documents.

E. Query level matching to retrieve the highly-relevant documents

Once the indexing phase is progressed using the proposed RSOA algorithm for finding the suitable centroids then, the resulted centroids are further used to perform the query matching process. In the proposed approach, the query level matching is carried out at two levels. In the first level matching, the query send by the user is matched with the centroid with respect to the Bhattacharya distance and the appropriate centroids are retrieved based on the smallest value of the Bhattacharya distance measure. Once the centroid with the smallest distance is retrieved, the related documents are retrieved with the individual records present at the centroid. Thus, at the second level, the user query is coordinated with the records of the optimal centroid, which is retrieved at the first level matching such that it returns the most suitable document that is based on the smallest value of the Bhattacharya distance. The Bhattacharya distance measure is formulated as,

$$V(p,q) = \frac{1}{4}\ln\left(\frac{1}{4}\left(\frac{H_p^2}{H_q^2}+\frac{H_q^2}{H_p^2}+2\right)\right)+\frac{1}{4}\left(\frac{\left(L_p-L_q\right)^2}{L_p^2+L_q^2}\right) \tag{25}$$

where, $V(p,q)$ represents the Bhattacharyya distance between p and q, H_p^2 represents the variance of p^{th} distribution, H_q^2 represents the variance of q^{th} distribution, L_p indicates the mean of p^{th} distribution, L_q indicates the mean of the q^{th} distribution and p and q are the distributions.

IV. RESULTS AND DISCUSSION

This section describes the results and discussion of the developed cluster based indexing using RSOA algorithm for document retrieval process.

A. Experimental setup

The experimentation of the proposed algorithm is performed in the MATLAB tool using the 20 newsgroup database [26] and reuter database [27]. The 20 newsgroup dataset contains nearly 20, 000 newsgroup records, which are partitioned to 20 newsgroups. It is collected from the learning to filter netnews paper. Moreover, the 20 newsgroup dataset is mainly used in the text applications, like text clustering and text classification.

B. Evaluation metrics

a) Precision: It is termed as the ratio of observed positive measure with respect to the total positive measure.

$$Z = \frac{K}{K+L} \tag{26}$$

where, Z represents the precision, K denotes the true positive, and L denotes the false positive.

b) Recall: It is defines as the ratio of predicted positive measure with respect to the total measure.

$$Q = \frac{K}{K+V} \tag{27}$$

where, Q denotes the recall, and V represents the true negative.

$$F-measure = \frac{2(Q*Z)}{Q+Z} \tag{28}$$

C. Comparative methods

The performance of the proposed RSOA algorithm is analyzed using the existing methods, like ICIR (Intelligent Clusterbased Information Retrieval) [8], Positive and Negative Feedbacks with Single negative Model strategy (PNFSM) + Positive and Negative Feedbacks with Multiple negative Models strategy (PNFMM) [1], Hierarchical Attribute-based Encryption Scheme (H-ABE) [7], and Spider Monkey optimization (SMO) [24].

D. Comparative analysis

This section discusses the comparative analysis using the metrics, like accuracy, recall, F-measure, and precision by varying the cluster size.

a) Analysis using 20 newsgroup dataset

The comparative analysis based on the cluster size with respect to the metrics, like accuracy, F-measure, recall, and precision is elaborated in this section. Figure 2 a) depicts the analysis of accuracy based on the cluster size. When the cluster size is 8, the existing methods, like ICIR, PNFSM+PNFMM, SMO, and H-ABE attained the accuracy as 92.553%, 91.883%, 88.702%, and 84.326%, whereas, the proposed RSOA obtained better accuracy of 93.867%. Figure 2 b) depicts the analysis of F-measure based on the cluster size. When cluster size=12, the F-measure obtained by the proposed RSOA is 93.708%, however, the percentage of improvement of the proposed RSOA with respect to the existing methods, namely ICIR, PNFSM+PNFMM, SMO, and H-ABE is reported as 2%, 6%, 8%, and 19%, respectively.

Figure 2 c) depicts the analysis of precision based on the cluster size. When cluster size=20, the precision obtained by the existing methods, such as ICIR, PNFSM+PNFMM, SMO, and H-ABE is 87.381%, 86.040%, 85.357%, and 80.547%, while the proposed RSOA attained better precision of 90.141%, respectively. Figure 2 d) depicts the analysis of recall based on the cluster size. When cluster size=4, the existing methods, like ICIR, PNFSM+PNFMM, SMO, and H-ABE attained the recall as 95.656%, 93.239%, 91.266%, and 91.173, while the proposed RSOA obtained better recall of 96.697%, respectively.

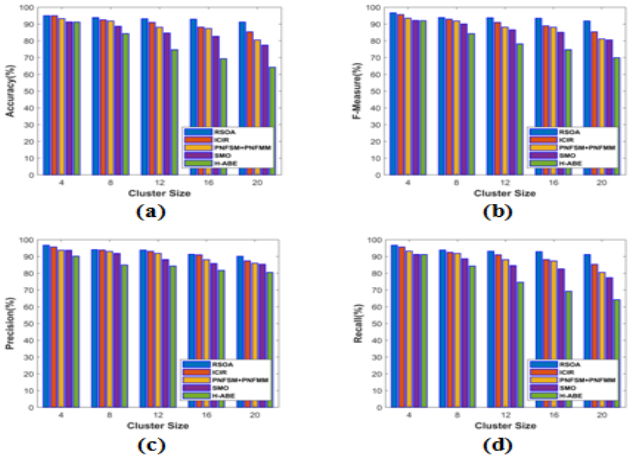

Figure 2. Analysis based on the cluster size, a) accuracy, b) F-measure, c) precision, d) recall

b) Analysis using Reuter dataset

The comparative analysis based on the cluster size with respect to the metrics, like accuracy, F-measure, recall, and precision is elaborated in this section. Figure 3 a) depicts the analysis of accuracy based on the cluster size. When the cluster size is 8, the existing methods, like ICIR, PNFSM+PNFMM, SMO, and H-ABE attained the accuracy as 91.997%, 66.032%, 65.825%, and 60%, while, the proposed RSOA obtained better accuracy of 94%. Figure 3 b) depicts the analysis of F-measure based on the cluster size. When cluster size=12, the F-measure obtained by the proposed RSOA is 91.158%, the percentage of improvement of the proposed RSOA with respect to the existing methods, namely ICIR, PNFSM+PNFMM, SMO, and H-ABE is reported as 3%, 36%, 39%, and 51%, respectively.

Figure 3 c) depicts the analysis of precision based on the cluster size. When cluster size=20, the precision obtained by the existing methods, such as ICIR, PNFSM+PNFMM, SMO, and H-ABE is 80.647%, 73.962%, 68.710%, and 66.495%, while the proposed RSOA attained better precision of 85.477%, respectively. Figure 3 d) depicts the analysis of recall based on the cluster size. When cluster size=4, the existing methods, like ICIR, PNFSM+PNFMM, SMO, and H-ABE attained the recall as 95.780%, 68.375%, 67.919%, and 63.575%, while the proposed RSOA obtained better recall of 96.839%, respectively.

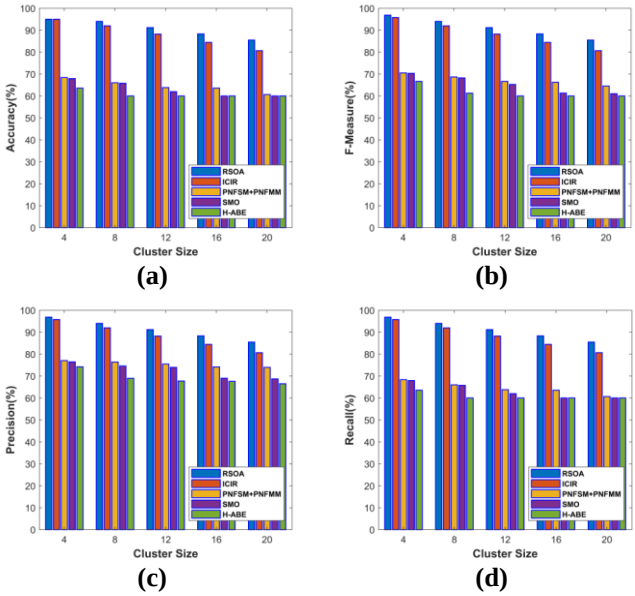

Figure 3. Analysis based on the cluster size, a) accuracy, b) F-measure, c) precision, d) recall

E. Performance analysis

This section discusses the performance analysis using the metrics, like accuracy, recall, F-measure, and precision by varying the population size.

a) Analysis using 20 newsgroup dataset

The performance analysis based on the cluster size with respect to the metrics, like accuracy, F-measure, recall, and precision is elaborated in this section. Figure 4 a) illustrates the analysis of accuracy based on the cluster size.

When cluster size=12, the accuracy obtained by the RSOA with population size=10 as 85.18%, RSOA with population size=20 as 87.77%, RSOA with population size=30 as 91.03%, RSOA with population size=40 as 93.94%, RSOA with population size=50 as 95%, respectively. Thus, it is very clear, that increasing the cluster size decreases the accuracy rate, respectively. Figure 4 b) depicts the analysis of F-measure based on the cluster size.

When cluster size=20, the F-measure obtained by the RSOA with population size=10 as 75.744%, RSOA with population size=20 as 80.238%, RSOA with population size=30 as 85.353%, RSOA with population size=40 as 90.154%, RSOA with population size=50 as 95.019%, respectively.

Figure 4 c) shows the analysis of precision based on the cluster size. When cluster size= 16, the precision obtained by the RSOA with population size=10 as 85.185%, RSOA with population size=20 as 87.779%, RSOA with population size=30 as 91.03%, RSOA with population size=40 as 93.943%, RSOA with population size=50 as 96%, respectively. It is observed that, when cluster size increases, the precision value will get decreases. Figure 4 d) depicts the analysis of recall based on the cluster size. When cluster size =4, the recall obtained by the RSOA with population size=10 as 94%, RSOA with population size=20 as 95%, RSOA with population size=30 as 96%, RSOA with population size=40 as 96%, RSOA with population size=50 as 96%. It shows that maximizing the population size increases the value of recall rate.

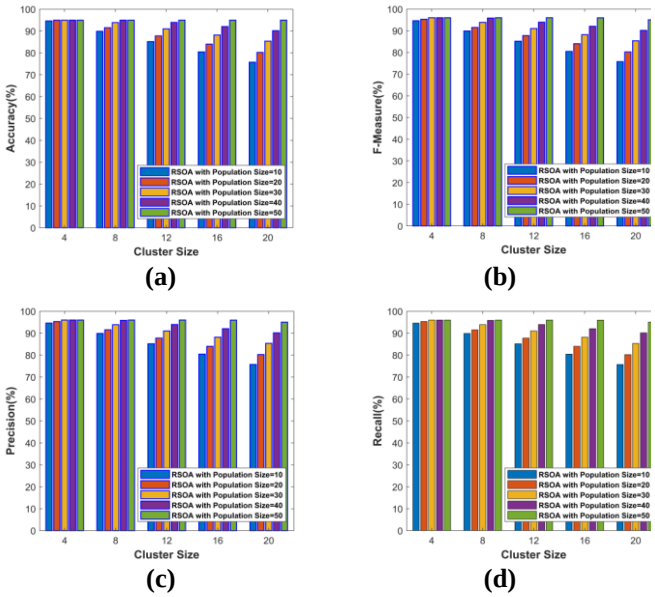

Figure 4. Performance analysis based on the cluster size, a) accuracy, b) F-measure, c) precision, and d) recall

b) Analysis using Reuter dataset

The performance analysis based on the cluster size with respect to the metrics, like accuracy, F-measure, recall, and precision is elaborated in this section. Figure 5 a) depicts the analysis of accuracy based on the cluster size. When cluster size=12, the accuracy obtained by the RSOA with population size=10 as 85%, RSOA with population size=20 as 88%, RSOA with population size=30 as 91%, RSOA with population size=40 as 94%, RSOA with population size=50 as 95%, respectively. When the cluster size is 20, the accuracy obtained by the RSOA with population size=10 as 75%, RSOA with population size=20 as 80%, RSOA with population size=30 as 85%, RSOA with population size=40 as 90%, RSOA with population size=50 as 95%, respectively. Thus, it is very clear, that maximizing the population size significantly increases the accuracy rate. Figure 5 b) shows the analysis of F-measure based on the cluster size. When cluster size=12, the F-measure obtained by the RSOA with population size=10 as 85%, RSOA with population size=20 as 88%, RSOA with population size=30 as 91%, RSOA with population size=40 as 94%, RSOA with population size=50 as 96%, respectively. When cluster size=20, the F-measure obtained by the RSOA with population size=10 as 75%, RSOA with population size=20 as 80%, RSOA with population size=30 as 85%, RSOA with population size=40 as 90%, RSOA with population size=50 as 95%, respectively.

Figure 5 c) shows the analysis of precision based on the cluster size. When cluster size= 16, the precision obtained by the RSOA with population size=10 as 80%, RSOA with population size=20 as 84%, RSOA with population size=30 as 88%, RSOA with population size=40 as 92%, RSOA with population size=50 as 96%, respectively. When cluster size= 20, the precision obtained by the RSOA with population size=10 as 75%, RSOA with population size=20 as 80%, RSOA with population size=30 as 85%, RSOA with population size=40 as 90%, RSOA with population size=50 as 95%, respectively. It is observed that, when cluster size increases, the precision value will decrease. Figure 5 d) depicts the analysis of recall based on the cluster size. When cluster size =4, the recall obtained by the RSOA with population size=10 as 94%, RSOA with population size=20 as 95%, RSOA with population size=30 as 96%, RSOA with population size=40 as 96%, RSOA with population size=50 as 96%. When cluster size =12, the recall obtained by the RSOA with population size=10 as 85%, RSOA with population size=20 as 88%, RSOA with population size=30 as 91%, RSOA with population size=40 as 94%, RSOA with population size=50 as 96%.

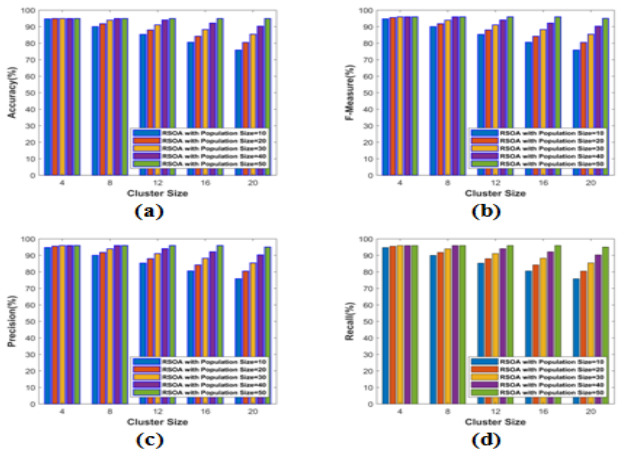

Figure 5. Performance analysis based on the cluster size, a) accuracy, b) F-measure, c) precision, and d) recall

F. Representation of word cloud

This section describes the word cloud representation. Figure 6 a) shows the word cloud 1, figure 6 b) shows the word cloud 2, and figure 6 c) depicts the word cloud 3, respectively.

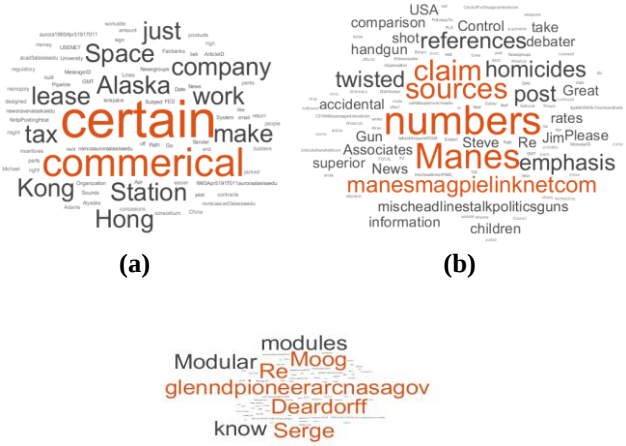

Figure 6. a) word cloud 1, b)word cloud 2, c) word cloud 3

G. Comparative discussion

This section elaborates the comparative discussion of the proposed RSOA algorithm based on the cluster size. Table 2 illustrates the comparative discussion based on the cluster size using 20 news group dataset. When the cluster size is 20, the precision obtained by the proposed RSOA is 90.141%, while the existing methods, like ICIR, PNFSM+PNFMM, SMO, and H-ABE obtained the precision as 87.381%, 86.04%, 85.357%, and 80.547%, respectively. For 20 cluster size, the F-measure obtained by the existing methods, like ICIR, PNFSM+PNFMM, SMO, and H-ABE is 85.353%, 81.150%, 80.544%, and 69.948%, while the proposed attained the F-measure as 91.876%, respectively. It is clearly depicted that the proposed RSOA attained better performance based on the cluster size.

Table 2. Comparative discussion made using 20 news group dataset

Methods	Metrics			
	Precision (%)	*F-measure (%)*	*Accuracy (%)*	*Recall (%)*
RSOA	90.141	91.876	91.202	91.178
ICIR	87.381	85.353	85.356	85.350
PNFSM+PNFMM	86.040	81.150	80.547	80.542
SMO	85.357	80.544	77.451	77.379
H-ABE	80.547	69.948	64.257	64.228

Table 3 illustrates the comparative discussion based on the cluster size using Reuter dataset. When cluster size=20, the accuracy obtained by the existing methods, like ICIR, PNFSM+PNFMM, SMO, and H-ABE is 80.647%, 60.677%, 60%, and 60%, while the proposed RSOA obtained better accuracy of 85.477%, respectively.

Table 3. Comparative discussion made using Reuter dataset

Methods	Metrics			
	Accuracy (%)	*F-measure (%)*	*Precision (%)*	*Recall (%)*
RSOA	85.477	85.476	85.477	85.476
ICIR	80.647	80.646	80.647	80.647
PNFSM+PNFMM	60.677	64.537	73.962	60.631
SMO	60	61.04	68.710	60
H-ABE	60	60	66.495	60

V. CONCLUSION

The Clustering-based indexing scheme named Rider Spider Monkey optimization algorithm is proposed in this research for document retrieval and bi-level matching is performed using the Bhattacharya distance. Moreover, the documents are pre-processed through the stop word elimination and the stemming process. The features are extracted by applying the TF-IDF and the suitable keywords are selected using the holoentropy. However, the indexing is performed using the proposed RSOA, which is the integration of the ROA and SMO. The cluster centroids are identified and the query matching is performed based on the cluster centroid. The query matching is carried out at two levels using the Bhattacharya distance to retrieve the matched documents. The implementation of the proposed algorithm is done and analyzed based on the performance measures. Moreover the proposed clustering-based indexing scheme attained better performance measures of precision, F-measure, recall, and accuracy to be 90.141%, 91.876%, 91.178%, and 91.202%, respectively using 20 news group dataset. In future, the performance of the document retrieval process may be based on additional features extracted from the database.

REFERENCES

1. S. Hao, C. Shi, Z. Niu, and L. Cao, "Modeling positive and negative feedback for improving document retrieval", *Expert Systems with Applications*, vol. 120, 2019, pp.253-261.
2. M. Kayest, and S.K. Jain, "Optimization driven cluster based Indexing and matching for the document retrieval", *Journal of King Saud University-Computer and Information Sciences*, 2019.
3. Y. Djenouri, A, Belhadi, and R. Belkebir, " Bees swarm optimization guided by data mining techniques for document information retrieval", *Expert Systems with Applications*, vol. 94, 2018,pp.126-136.
4. S.H. Farhi, and D.Boughaci, "Graph based model for information retrieval using a stochastic local search", *Pattern Recognition Letters*, vol. 105, 2018, pp.234-239.
5. S.Biswas, A. Ganguly, R. Shah and S.V. Thankachan, "Ranked document retrieval for multiple patterns", *Theoretical Computer Science*, vol. 746, 2018,pp.98-111.
6. H. Ferrada, and G. Navarro, "Lempel–Ziv compressed structures for document retrieval", *Information and Computation*, vol. 265, 2019, pp.1-25.

7. N. Wang, J. Fu, B.K. Bhargava and J. Zeng, "Efficient retrieval over documents encrypted by attributes in cloud computing", *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 10, 2018,pp.2653-2667.
8. Y. Djenouri, A. Belhadi, P. Fournier-Viger and J.C.W. Lin, "Fast and effective cluster-based information retrieval using frequent closed itemsets", *Information Sciences*, vol. 453, 2018, pp.154-167.
9. A. Keikha, F. Ensan, and E. Bagheri, "Query expansion using pseudo relevance feedback on Wikipedia", *Journal of Intelligent Information Systems*, vol. 50, no. 3, 2018,pp. 455–478.
10. B. He, and I. Ounis, "Finding good feedback documents", *In ACM conference on information and knowledge management, Hong Kong, China*, 2009,pp. 2011–2014.
11. Y. Zheng, and H. J. Xiangji, "A simple term frequency transformation model for effective pseudo relevance feedback", *In International ACM SIGIR conference on research and development in information retrieval, Gold Coast, Australia*, 2014,pp. 323–332.
12. M. Dehghani, H. Azarbonyad, J. Kamps, D. Hiemstra, & M. Marx, "Luhn revisited: Significant words language models", *In ACM international on conference on information and knowledge management, Pisa, Italy*, 2016, pp. 1301–1310.
13. Y. Lv, and C. X. Zhai, "Negative query generation: Bridging the gap between query likelihood retrieval models and relevance", *Information Retrieval Journal*, vol. 18, no. 4, 2015,pp. 359–378.
14. M. Karimzadehgan, and C. X. Zhai, "Improving retrieval accuracy of difficult queries through generalizing negative document language models", *ACM international conference on information and knowledge management, Glasgow, United Kingdom*, 2011,pp. 27–36.
15. Y. Ma, and H. Lin, "A multiple relevance feedback strategy with positive and negative models", *Plos One*, vol. 9, no. 8, 2014.
16. Q. Liu, P.A. Ng, "Document Processing and Retrieval: Text Processing", *Kluwer Academic Publishers, Norwell*, 1996.
17. F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering", *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002 pp. 436-442.
18. B.C. Fung, K. Wang, and M. Ester, "Hierarchical document clustering using frequent itemsets", *In Proceedings of the 2003 SIAM international conference on data mining, Society for Industrial and Applied Mathematics*, 2003, pp. 59-70.
19. H.Yu, D. Searsmith X. Li, and J. Han, "Scalable construction of topic directory with nonparametric closed termset mining", *IEEE, International Conference on Data Mining (ICDM'04)*, 2004, pp. 563-566.
20. G.V. Menezes, J.M. Almeida, F. Belém, M.A. Gonçalves, A. Lacerda, E.S. De Moura, G.L. Pappa, A. Veloso, and N.,Ziviani, "Demand-driven tag recommendation", *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2010,pp. 402-417.
21. M.J. Zaki, and C.J. Hsiao, "CHARM: An efficient algorithm for closed itemset mining", *In Proceedings of the 2002 SIAM international conference on data mining*, 2002,pp. 457-473.
22. P. Fournier-Viger, J.C.W. Lin, B.Vo, T.T. Chi, J. Zhang, and H.B. Le, "A survey of itemset mining", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 4, 2017,pp.1207.
23. D. Binu, and B.S. Kariyappa, "RideNN: A New Rider Optimization Algorithm-Based Neural Network for Fault Diagnosis in Analog Circuits", *IEEE Transactions on Instrumentation and Measurement*, 2018.
24. JC. Bansal, H. Sharma, S.S. Jadon, M. Clerc, "Spider Monkey Optimization algorithm for numerical optimization", *Memetic Computing*, vol.6, no. 1, 2014, pp.31–47.
25. V.M. Mane, and D.V. Jadhav, "Holoentropy enabled-decision tree for automatic classification of diabetic retinopathy using retinal fundus images", *Biomedical Engineering/Biomedizinische Technik*, vol. 62, no. 3, 2017,pp.321-332.
26. 20 Newsgroup database, "http://qwone.com/~jason/20Newsgroups/", accessed on May 2019.
27. Reuter database,"https://archive.ics.uci.edu/ml/machine-learningdatabases/reuters21578-mld/", accessed on May 2019.

AUTHORS PROFILE

Smt. Madhulika Yarlagadda, is an Assistant Professor in Information Technology department at R.V.R & J.C College of Engineering, Chowdavaram,Guntur,Andhrapradesh,India and is a research scholar at Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India. Her research interests are Data Mining, Information Retrieval Systems, Web Mining, Algorithms. Email:madhulika.yarlagadda@gmail.com

Dr. K. Gangadhara Rao, is a professor in Department of Computer Science and Engineering at Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. His areas of interest include Data Mining, Cloud Computing, Computer Networks and Software Engineering. Email: kancherla123@gmail.com

Dr. A. Srikrishna, is professor and head of the Department of Information and Technology at RVR & JC College of engineering, Chowdavaram, Guntur, Andhra Pradesh, India. Her research interests are in Image processing, Computer vision, Information security, and algorithms
Email:atlurisrikrishna@gmail.com