J. Appl. Math. & Informatics Vol. **41***(2023), No. 6, pp. 1327 - 1339*
https://doi.org/10.14317/jami.2023.1327

AN EFFICIENT DENSITY BASED ANT COLONY APPROACH ON WEB DOCUMENT CLUSTERING

M. REKA

Abstract. World Wide Web (WWW) use has been increasing recently due to users needing more information. Lately, there has been a growing trend in the document information available to end users through the internet. The web's document search process is essential to find relevant documents for user queries.As the number of general web pages increases, it becomes increasingly challenging for users to find records that are appropriate to their interests. However, using existing Document Information Retrieval (DIR) approaches is time-consuming for large document collections. To alleviate the problem, this novel presents Spatial Clustering Ranking Pattern (SCRP) based Density Ant Colony Information Retrieval (DACIR) for user queries based DIR. The proposed first stage is the Term Frequency Weight (TFW) technique to identify the query weightage-based frequency. Based on the weight score, they are grouped and ranked using the proposed Spatial Clustering Ranking Pattern (SCRP) technique. Finally, based on ranking, select the most relevant information retrieves the document using DACIR algorithm.The proposed method outperforms traditional information retrieval methods regarding the quality of returned objects while performing significantly better in run time.

AMS Mathematics Subject Classification : 65H05, 65F10, 91C20, 62G07, 08A02.
Key words and phrases : Document information retrieval, web, ranking, score, clustering, user query.

1. Introduction

In recent years, there has been an information overload due to the rapid spread of the internet. To process this information, use the get web document information task to retrieve documents that best match the user's query.Today, search engines are handy tools for collecting data and extracting knowledge from the internet.However, even the most popular search engines often return result sets that contain many pages that are not useful to users. With the number of

Received April 19, 2023. Revised August 16, 2023. Accepted September 26, 2023.

text documents on the internet, finding information of value to a particular user can be difficult.

The first Ant Colony Optimization (ACO) algorithm has been applied to the Traveling Salesman Problem (TSP) [21,22]. The ACO algorithm belongs to the natural class of problem solving techniques which is initially inspired by the efficiency of real ants as they find their fastest path back to their nest when sourcing for food.

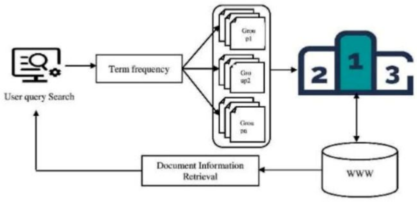

Figure 1. Compact of Document Information Retrieval (DIR)

To tackle the above problems, figure 1 illustrates the compact of Document Information Retrieval (DIR) based on clustering and ranking. First, the user searches the document in search engine, and then our proposed technique analysis the term frequency. After that, based on frequency weight, they are grouped into clusters (group-1, group-2…group-n). After that, ranking the document by grouping information. Finally, the proposed technique retrieves the document in World Wide Web (WWW) for user queries.

In this paper is novel developed a Spatial Clustering Ranking Pattern (SCRP) based Density Ant Colony Information Retrieval (DACIR) algorithm to retrieve the relevant document. The proposed Density Ant Colony Information Retrieval (DACIR) approach improves to information retrieval by using clustering and closed-term frequent item set mining to extract documents helpful to web users. Results show that the proposed method outperforms other state-of-the-art DIR methods in terms of running time and quality of the retrieved document. In section 2 discussed for preliminaries based on my study. A detailed description of the proposed methodology was implemented in Section 3 and followed by results and discussion were described in Section 4. Finally, this paper concluded in Section 5.

2. Preliminaries

J. Gong [1], a new method of network shared resources of deep data mining, is developed by the author to improve operations by deep data mining information shared by the network based on K-means clustering. J. Chiang et al. (2015), the author proposes a fuzzy clustering algorithm. This method is developed to detect meanings in information on the internet and to locate fuzzy linguistic topologies. C. C. Yang et al.[3], the author, researched the density-based clustering algorithm and proposed a clustering technique that measures website opinions based on distance. J. Wang et al. [4], the author describes a new approach: Multi-view Random-walk Graph regularization Low-Rank Representation (MRGLRR).D. Bollegala et al. [5], the author introduces an approach for calculating page counts and semantic similarity extracted from website search engines. This method calculates various word co-occurrences using page count. And extracting and synthesizing lexical patterns from text fragments. It is used to identify numerous semantic relations between words.

Y. Xu et al. [6] the author presented a new method called BioRank based on gene rankings and annotated gene sets, which can be used to assess the similarity between cells. B. Zhang et al. [7], the author proposes a Rank after Clustering (RaC) algorithm for opinion leaders handling social media information. G. Kang et al. [8], the author proposes that service clustering of websites can be done using documents and structural information to describe service relationships from networks. S. Shehata et al. [9], the author, defines a new mining model to analyze concept-based terms such as resolution, registration, and corpus level. D. Huang et al. [10] propose a novel ensemble clustering method based on rapid cluster propagation through random walks to propagate similarity. Computes the weights of the cluster graph vertices and their edges using the Jaccard coefficients for each cluster.

T. B. Mudiyanselage et al. [11], the author, defines a new approach to the graph-based analysis of websites to reduce the problem's dimensionality. This method is used to preserve the original high-dimensionality of the physical object of the feature space. It calculates the feature selection of websites' features based on solid inference. B. Xu et al. [13], the author developed a method called structural similarity (SSIM) to measure the relationships between clusters. It proposes an SR-SSIM method based on Band selection (BS). A state-of-the-art BS method tests the classification of different data sets. A. M. Sheri et al. [14], the author, builds an automatic consensus-building operation based on the text classifier. It generates basic information from two separate TIM documents to form primary clusters.

H. Luet al [15], The author proposed a method based on the standard feature mining algorithm called Intraframe Clustering and Interframe Association (ICIA). H. Qin et al. [16], the author proposes a Density-Based Clustering (DBC) framework based on a community detection algorithm to speed up carefully-designed community alignment techniques significantly. E. Uzun [17], Authors propose a new method called UzunExt. This method uses the string method and additional information to extract the content quickly.S. Miloudi et al. [18], the author proposes an efficient approach from multiple transactional databases to search for optimal clustering.L. Wang et al. [19], the authors propose a multi-cluster-based SBIR re-ranking method.

P. Li et al. [16], the author developed a framework called Neighborhood-Adaptive Multi-cluster Ranking (NAMR) to optimize the diversity of local structures. X. -F. Wang et al. [19], the author proposes a new density-based clustering framework. This is a concept-based data method to materialize information on cluster centres. Chiang.J et al. [2], The author proposes a discovering latent semantics in web documents using fuzzy clustering models. Kumar.S et al. [12], the author propose a A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem. Ravi.J [20], proposed a A robust measure of pair-wise distance estimation approach: RD-RANSAC in clustering method.

3. Proposed Methodology for DIR

This phase presents a Spatial Clustering Ranking Pattern (SCRP) based Density Ant Colony Information Retrieval (DACIR) algorithm to resolve DIR problems. Figure 2 defines the overview of the proposed SCRP-DACIR algorithm for DIR. The proposed methodology contains five stages there are i) preprocessing, ii) score calculation, iii) weight analysis, iv) clustering and ranking, and v) information retrieval.

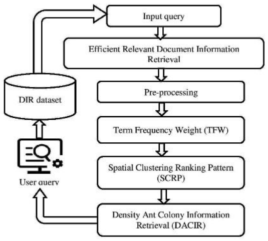
Input query
Efficient Relevant Document Information Retrieval
DIR dataset
Pre-processing
Term Frequency Weight (TFW)
User query
Spatial Clustering Ranking Pattern (SCRP)
Density Ant Colony Information Retrieval (DACIR)

Figure 2. Architecture diagram for SCRP-DACIR

First, we collect the user query data set from the online kaggle repository, then pre-process the dataset to structure the document. Then Term Frequency Weight (TFW) is used to calculate the query weightage-based frequency. Then Document Score Computing Rate (DSCR) technique analyses the query-sensitive importance score. Based on the query score, they are grouped and ranked using the proposed Spatial Clustering Ranking Pattern (SCRP) technique. Finally, based on ranking, select the most relevant information retrieves the document using the DACIR algorithm.

3.1 Pre-processing

Initially, the need for preprocessing of the collected dataset from the Kaggle repository because real-world data was inconsistent, noisy and missing information owing to the large data document size. In this stage, method reads the web document dataset that holds the data point initialization. Then each data term performs Tokenization, Stemming, Remove Stop Words, and the process verifies to all the dimensions. To stop words are often used in a document but are considered inappropriate (E.g. a, is, an, to etc.). These phrases did not do much to explain the meaning of text and should be eliminated. Stemming is connecting words to their original stems, root or base.

Algorithm

```
Input:Set of documents$s_d = d_1, d_2 \ldots d_n$
Output:Preprocessed document $P_d$
Begin
    Import set of documents $s_d$
    Read Document
    For each $d_i$ in $s_d$
    Compute stop word $(w_s)$ removal in the document
    While ($s_d[i]$==null)
                        $d_i \leftarrow s_d[i]$
    If ($d_i == w_s[]$)
    Remove stop words
    Else
    Check next term
    End if
    End while
    Compute stemming $T_s$ in the document
    While ($s_d[i]$==null)
                        $d_i \leftarrow s_d[i]$
    If ($d_i == R_s(T_s)$)
    Identify the root of the information
    End if
    End while
    Return $P_d$ ←optimized document
    End for
End
```

The proposed preprocessing steps to successfully remove stop words and stemming in the collected dataset. The above algorithm reads a set of documents $d_1, d_2 \ldots d_n$ to analyze the stemming T_s and stop words w_s in document. This process reduces the dimensionality of the data.Let us assume R_s denotes the root suffix.

3.2 Term frequency weight (TFW)

This stage analyses the user query weight using Term Frequency Weight (TFW) from the processed documentP_d. This method finds word frequency and counts the number of instances of individual words.This step uses the inverse document frequency to calculate the weight of each term.

$$Q_l(S, s_d) = \frac{(P_d L)^T (P_d S)}{||P_d L||\ .||P_d S||}$$

The above expression identifies the inverse term frequency $IT_{frequency}$ indocumentsN_d, and P_d is a preprocessed document dataset. Hence the term weight $W_s$calculated described the following expression,

$$W_s = (T_{frequency}) * (IT_{frequency})$$

The proposed method proficiently identifies the term frequency weight from the preprocessed dataset in this phase.

3.3 Spatial Clustering Ranking Pattern (SCRP)

The proposed Spatial Clustering Ranking Pattern (SCRP)algorithm is applied for query ranking from Term Frequency Weight (TFW) in this stage. This technique, spatial clustering, can be defined as organizing data or information into groups.The documents are sorted based on the relevance scores obtained by the user queries.This method calculated two document similarities and sensitive ranking scores.Consider set ofterms$M = \{M_1, M_2, \ldots M_n\}$and objects$O = \{O_1, O_2 \ldots O_n\}$. Every object O_i in terms in$M(O_i \subset M, \forall \in [1 \ldots n])$.

$$S_t\left(d_a, d_b\right) = \frac{\sum\left(W_{sa} * W_{sb}\right)}{\sqrt{{W_{sa}}^2 * {W_{sb}}^2}}$$

The above equation is used to find the sensitive term in the dataset based on user interest U_i and cluster centroidC_c.

$$R_k = \sum R\left(S_s\right) * \beta$$

The above equation is used to identify the ranking R_k based on sensitive information in the document set S_s, and β is a positive constant value. This section efficiently groups and ranks the document.Clustering similar text documents to find documents efficiently.

3.4 Density Ant Colony Information Retrieval (DACIR)

This section applied for Density Ant Colony Information Retrieval (DACIR) technique for web document retrieval from clustering ranking pattern. In this stage, efficiently retrieves the document-based ants' food foraging method.During the foraging period, the ant colony always finds the best route between the nest and the food. Because it emits pheromones on its way as it seeks it, allowing it to randomly choose a path and move forward when it encounters an intersection that it does not pass. It also occasionally releases pheromones depending on the length of its root.

$$\gamma^{i}_{x,y} = \begin{cases} \frac{1}{l_v} & i^t h \ ant \ location \ connection(x,y) \\ 0 & Otherwise \end{cases}$$

The above equation is used for every ant of tour$\gamma_{x,y}$.αRefers to the pheromone evaporation rate. Z is a constant value.

$$I_R = 0.9 \times \frac{fitness_i}{fitness_{max}} + 0.1$$

The above equation is used to efficiently retrieve the web documentI_R. In this phase, the proposed retrieve the web documents based clustering ranking pattern.

4. Experimental Result

In this phase, we present the details of the experimental result analysis performed and evaluate the performance of the proposed approach. Before illustrating experimental results, we briefly describe the simulation setup. Simulation parameters settings for DIR are defined in table 1.

Table 1: Simulation parameters settings for DIR

Parameter	**Values**
Dataset name	Centre for Inventions and Scientific Information (CISI)
Simulation tool	Anaconda
Language used	Python
Operating system	Windows 10
Number of users	250

The proposed implementation is evaluated on the Windows 10 operating system with 8GB RAM and Intel core i5 processor.

4.2 Performance Evolution

Here, the performance described by the proposed Spatial Clustering Ranking Pattern (SCRP) based Density Ant Colony Information Retrieval (DACIR) approach performance parameters is the accuracy of information retrieval, precision

(pr), recall (re), false rate (fr) and time complexity (tc)

$= \frac{\sum R_d}{\sum S_r} * 100$

Let's assume R_d denotes a set of retrieved documents d, and S_r is the number of succeeded document retrieval.
Recall(re)= $\frac{\sum D_r}{\sum R_D} * 100$

D_r is several appropriate documents retrieved, and $R_D$denotesrelevant documents in the dataset.
Precision(pr) = $\frac{\sum D_r}{\sum I_d} * 100$

Let us assume $I_d$is several retrieved documents

4.3 Information retrieval result

To illustrate the proposed Spatial Clustering Ranking Pattern (SCRP) based Density Ant Colony Information Retrieval (DACIR) approach compared with other algorithms, there are UzunExt(E. Uzun (2020)), K-Means Clustering (KMC) (J. Gong (2021)) and Density-Based Clustering (DBC)(H. Qin et al. (2022)).

Table 2: Analysis of classification query information retrieval performance

Query retrieval performance in %				
No users/ Methods	UzunExt	KMC	DBC	SCRP-DACIR
50	73.4	77.8	80.1	82.5
100	75.6	79.2	84.2	86.9
150	81.1	84.4	87.5	90.8
200	84.3	88.2	89.7	93.1
250	89.5	91.7	93.3	95.6

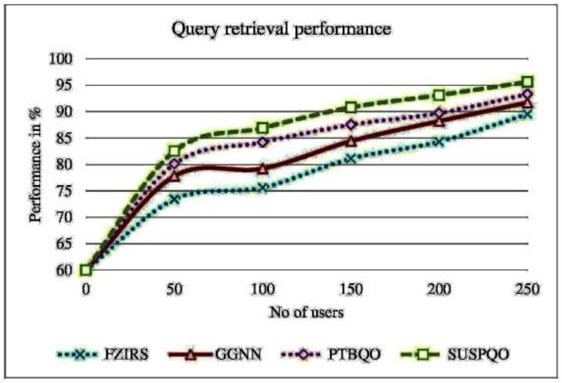

Figure 3. Impact of classification query information retrieval performance

Figure 3 and Table 2 denote the impact of query information retrieval performance based on the spider optimization algorithm and ranking-based search information. The proposed SCRP-DACIR method achieved 95.6% performance for 250 users. The result of existing methods is UzunExt achieved 89.5%, the KMC attained 91.7%, and the DBC attained 93.3% of retrieval performance for 250 users.

Recall performance in %				
Methods/no of users	UzunExt	KMC	DBC	SCRP-DACIR
50	71.1	76.5	78.6	80.4
100	73.8	78.4	81.2	83.5
150	78.6	81.9	85.3	87.3
200	83.2	85.8	88.1	90.5
250	86.9	88.3	90.4	92.1

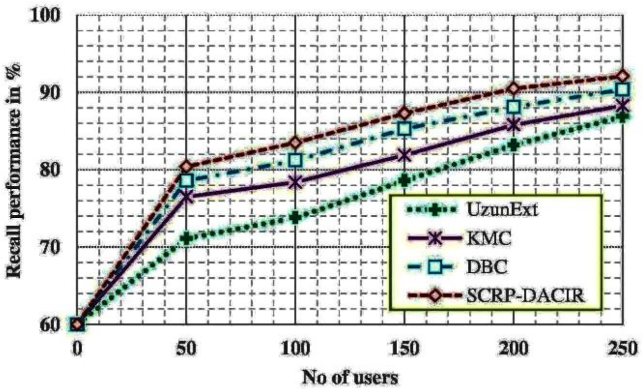

Figure 4. Impact of recall performance

Figure 4 and Table 3 illustrate the impact of query retrieval recall performance based on the spider optimization algorithm and ranking-based search information. The figure y-axis presents each method's performance gradually increased in percentage besides x-axis shows several web users, 50, 100, 150, 200 and 250. The proposed attained 92.1% of retrieval recall performance for 250 users; similarly, the previous techniques expressed that UzunExt has 86.9% of recall performance, KMC has 88.3% of recall performance, and DBC is 90.4% of recall performance.

Table 4: Impact of precision performance

Precision performance in %				
Methods/ No of users	UzunExt	KMC	DBC	SCRP- DACIR
50	72.6	77.2	79.6	81.1
100	74.9	79.5	82.4	84.2
150	79.3	82.4	86.1	88.9
200	84.5	86.1	89.7	91.5
250	87.2	89.6	91.3	93.4

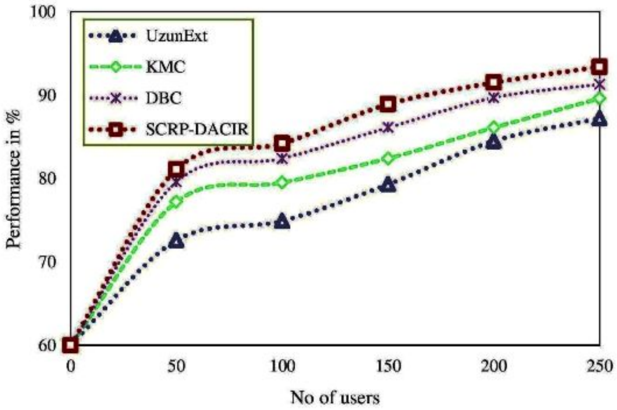

Figure 5. Impact of precision performance

Figure 5 and Table 4 describe the query retrieval precision performance vs the number of users like 50, 100, 150, 200 and 250. The proposed method efficiently identifies the correct information from users' need queries based on rank search information and spider optimization algorithm. Therefore the proposed SCRP-DACIR technique produces a better result than previous UzunExt, KMC, and DBC techniques.

Figure 6 explores the information retrieval false rate performance vs the number of users. The proposed method produced a 15.4% false rate performance for 250 users, the UzunExt method had 28.1%, the KMC technique had 24.7%, and the DBC technique had a 21.8% of false rate performance. However, the proposed SCRP-DACIR method produced less false rate performance than other methods.

Figure 7 shows the analysis of information query retrieval time complexity in milliseconds vs the number of users 50, 100, 150, 200 and 250. The proposed technique achieved 14.6ms, the UzunExt method produced 30.6ms, the KMC technique had 23.4ms, and the DBC technique produced 23.4ms time complexity performance.

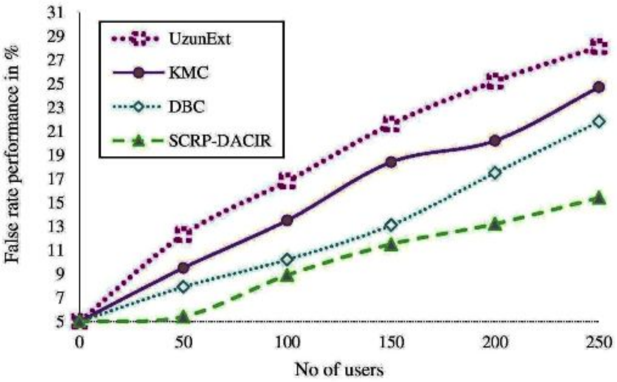

Figure 6. Impact of precision performance

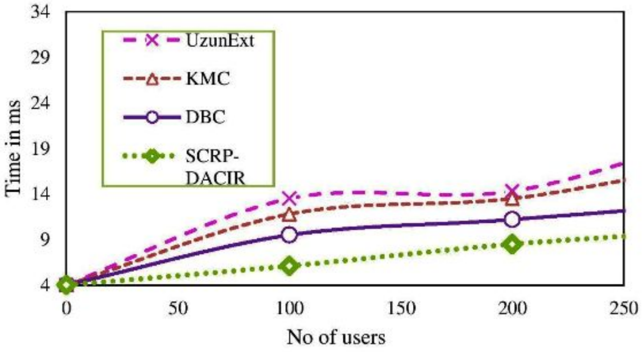

Figure 7. Impact of precision performance

5. Conclusion

A novel Spatial Clustering Ranking Pattern (SCRP) based Density Ant Colony Information Retrieval (DACIR) algorithm was proposed in this paper that uses frequent and efficient helpful extract patterns from the web.This proposed algorithm is the first process is preprocessing for the structured document. The document score identifies from the term frequency weight. The proposed SCRP grouped and ranked the document based on the score weight. Finally, the DACIR technique retrieves the relevant document based on user needs.The experiment result obtained 95.6% of DIR performance with minimum time. The results show that the proposed approach benefits from the extracted patterns and significantly improves the quality of retrieved documents. The proposed method is compared with several other state-of-the-art information retrieval methods on a benchmark dataset. The results show that the proposed method outperforms other methods, especially when dealing with multiple user queries.

Conflicts of interest : The author declares no conflict of interest.

Data availability : Not applicable

References

1. J. Gong, *In-depth Data Mining Method of Network Shared Resources Based on K-means Clustering*, 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), **7** (2021), 694-698.
2. J. Chiang, C.C.-H. Liu, Y.-H. Tsai and A. Kumar, *Discovering Latent Semantics in Web Documents Using Fuzzy Clustering*, IEEE Transactions on Fuzzy Systems **7** (2015), 2122-2134.
3. C.C. Yang and T.D. Ng., *Analyzing and Visualizing Web Opinion Development and Social Interactions With Density-Based Clustering*, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans **7** (2011), 1144-1155.
4. J. Wang, L. Wang H., J.-X. Liu, X.-Z. Kong and S.-J. Li, *Multi-View Random-Walk Graph Regularization Low-Rank Representation for Cancer Clustering and Differentially Expressed Gene Selection*, IEEE Journal of Biomedical and Health Informatics **7** (2022), 3578-3589.
5. D. Bollegala, Y. Matsuo and M. Ishizuka, *A Web Search Engine-Based Approach to Measure Semantic Similarity between Words*, IEEE Transactions on Knowledge and Data Engineering **7** (2011), 977-990.
6. Y. Xu, H.D. Li, Y. Pan, F. Luo, F.X. Wu and J. Wang, *A Gene Rank Based Approach for Single Cell Similarity Assessment and Clustering*, IEEE/ACM Transactions on Computational Biology and Bioinformatics **7** (2021), 431-442.
7. B. Zhang, Y. Bai, Q. Zhang, J. Lian and M. Li, *An Opinion-Leader Mining Method in Social Networks With a Phased-Clustering Perspective*, IEEE Access **7** (2020), 31539-31550.
8. G. Kang, J. Liu, Y. Xiao, Y. Cao, B. Cao and M. Shi, *Web Services Clustering via Exploring Unified Content and Structural Semantic Representation*, IEEE Transactions on Network and Service Management **7** (2022).
9. S. Shehata, F. Karray and M. Kamel, *An Efficient Concept-Based Mining Model for Enhancing Text Clustering*, IEEE Transactions on Knowledge and Data Engineering **7** (2010), 1360-1371.
10. D. Huang, C.D. Wang, H. Peng, J. Lai and C.K. Kwoh, *Enhanced Ensemble Clustering via Fast Propagation of Cluster-Wise Similarities*, IEEE Transactions on Systems, Man, and Cybernetics: Systems **7** (2021), 508-520.
11. T.B. Mudiyanselage and Y. Zhang, *Feature selection with graph mining technology*, Big Data Mining and Analytics **7** (2019), 73-82.
12. S. Kumar and M. Singh, *A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem*, Big Data Mining and Analytics **7** (2019), 240-247.
13. B. Xu, X. Li, W. Hou, Y. Wang and Y. Wei, *A Similarity-Based Ranking Method for Hyperspectral Band Selection*, IEEE Transactions on Geoscience and Remote Sensing **7** (2021), 9585-9599.
14. A.M. Sheri, M.A. Rafique, M.T. Hassan, K.N. Junejo and M. Jeon, *Boosting Discrimination Information Based Document Clustering Using Consensus and Classification*, IEEE Access **7** (2019), 78954-78962.
15. H. Lu, K. Gu, W. Lin and W. Zhang, *Object Tracking Based on Stable Feature Mining Using Intraframe Clustering and Interframe Association*, IEEE Access **7** (2017), 4690-4703.

16. H. Qin, H.R. Li, G. Wang, X. Huang, Y. Yuan and X.J. Yu, *Mining Stable Communities in Temporal Networks by Density-Based Clustering*, IEEE Transactions on Big Data **7** (2022), 671-684.
17. E. Uzun, *A Novel Web Scraping Approach Using the Additional Information Obtained from Web Pages*, IEEE Access **7** (2020), 61726-61740.
18. S. Miloudi, Y. Wang and W. Ding, *A Gradient-Based Clustering for Multi-Database Mining*, IEEE Access **7** (2021), 11144-11172.
19. L. Wang, X. Qian, X. Zhang and X. Hou, *Sketch-Based Image Retrieval With Multi-Clustering Re-Ranking*, IEEE Transactions on Circuits and Systems for Video Technology **7** (2020), 4929-4943.
20. J. Ravi, *A robust measure of pairwise distance estimation approach: RD-RANSAC*, International Journal of Statistics and Applied Mathematics **2(2)** (2017), 31-34.
21. M. Dorigo, V. Maniezzo, and A. Colorni, *Positive feedback as a search strategy Technical report 91-016*, Politecnico di milano, Dip. Elettronica, 1991.
22. M. Dorigo, V. Maniezzo, and A. Colorni, *The ant system: Optimization by a colony of cooperating agents*, IEEE Transactions on Systems, Man, and Cybernetics-Part B **26(1)** (1996), 29-42.

M. Reka is currently working as an Assistant Professor in the Department of Computer Applications, SONA College of Arts and Science, Salem, Tamilnadu, India. She is received the B.Sc. degree in Computer Science from Periyar University in 2002, M.C.A. degree from Periyar University in 2005, M.Phil. Degree in Computer Science from Periyar University in 2006 and a PhD degree from Anna University in 2017. She has more than 13 years of teaching experience in KSR Educational Insitution, Tiruchengode, Tamilnadu. Her research interests in Data Mining, Networking, Data Communications and Computer Graphics. She is a life member of ISTE (Indian Society for Technical Education).

Assistant Professor in the Department of Computer Applications, SONA College of Arts and Science, Salem, Tamilnadu, India.
e-mail: rekamca2010@gmail.com