ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 10, December 2013

An Optimized WebDocument Clustering Using Recurrent Set IGA & Confusion Matrix For Fact Retrieval

C. Josephine Christy[1], Dr. B. Nagarajan[2]

Research Scholar, Bharathiar University, Coimbatore, India [1]

Professor & Head, Department of Computer Applications, Bannari Amman Institute of Technology, Sathyamangalam, India[2]

Abstract: Initially the first phase derives the Genetic Algorithm for global clustering process to resolve the optimization solution in both clustering and feature selection. The second phase follows a concept of confusion matrix for derivative works and improved GA is included for the final classification. The third phase presents the optimization technique to evaluate the cluster optimality for proficient document clustering based on the optimized conceptual feature words. Final phase introduce a join approach to cluster the web pages which primarily finds the recurrent sets and then clusters the documents. These recurrent sets are generated by using recurrent pattern expansion technique. Then by applying Fuzzy K-Means algorithm on Optimized Web document clustering using Recurrent Set founds clusters having documents which are extremely related and have related features. Experimental results show that our approach is more efficient then the above two join approach and can handle more efficiently in robust nature. Performance evaluation show benefits in terms of cluster optimality, true negative rate and information retrieval on real and UCI repository bag of words dataset.

Keywords: Genetic Algorithm, Fuzzy K-Means Algorithm, Recurrent Pattern Expansion, Web document Clustering, Confusion Matrix, Optimization Technique, World Wide Web, Feature Selection

I. INTRODUCTION

Clustering is to segregate a group of text documents into several category groups. However the text document clustering is one of the essential tasks in text mining. The perception of association rule mining has been used for the computerization of document clustering. The fitness of a rule has been evaluated by its classification accuracy on a set of training examples.Second phase the feature selection is integrated with the global clustering process. Each rule can be represented by a string of bits. Third phase uses the cluster optimality for efficient document clustering based on the optimized conceptual feature words. Optimal weight is calculated for effective clustering of text document. The optimization technique considers conceptual weight for choosing the trait of the documents. The two fundamental methods of fuzzy clustering are fuzzy k-partitions clustering and another based on the fuzzy correspondence relations. There are two wide principles use for relationship examination. Recurrent Pattern expansions are a divide and conquer strategy that mines a complete set of frequent item sets without candidate generation. The final phase of the web document clustering id developed based on Recurrent Pattern Expansion and FKM that helps the search engine to retrieve relevant web documents needed for any user. Documents in the FKM are strongly correlated.

.

II. THE RESEARCH METHOD

[1] A. K. Santra, C. Josephine Christy and B. Nagarajan have proposed that cluster based niche memetic and genetic algorithm have been designed & implemented by optimizing feature selection of text in the document repository. [2] A. K. Santra and C. Josephine Christy have proposed that Genetic Algorithm and Confusion Matrix for Document Clustering have been designed & implemented by calculation Precision, True Positive Rate, True Negative Rate, False Positive Rate and False Negative Rate. [3] A. K. Santra and C. Josephine Christy have proposed that “An efficient document Clustering for Optimization Technique for Cluster Optimality” have been designed & implemented by calculation of Optimization Techniques.[4] Shady Shehata.,et.Al., 2010 proposes a

model which efficiently find significant matching concepts between documents is calculated based on a new concept-based similarity measure. The proposed similarity measure takes full advantage of using the concept analysis measures on the sentence, document, and corpus levels in calculating the similarity between documents. [5] Yun Yang., and Ke Chen., 2011 propose a novel weighted accord function guided by clustering validation criteria to settle initial partitions [6] Ninad Thakoor., and Jean Gao., 2011 addresses the data analysis problem using the branch and bound. [7] The heterogeneous documents are summarized and review of different document clustering are presented in a uniform manner by K.Sathiyakumari.,et.Al., 2011. [8] Swatantra kumar sahu.,et.Al, 2012 presents Document Clustering Approaches based on classification of Large data Sets . [9] K. Premalatha., and A.M. Natarajan., 2009 presents the document clustering based on Genetic algorithm with Simultaneous mutation operator and Ranked mutation rate. The mutation operation is significant to the success of genetic algorithms since it expands the search directions and avoids convergence to local optima. [10] Kiran G V R.,et.Al., 2010 propose a hierarchical clustering algorithm using closed frequent item sets that use wikipedia as an exterior knowledge to improve the document representation. [11] George Pallis.,et.Al., 2008 present a clustering-based pre fetching scheme where a graph-based clustering algorithm identifies clusters of ‘‘correlated’’ Web pages based on the users’ access patterns. Clustering-based pre fetching scheme can be incorporated easily into a Web proxy server, improving its performance. [12] Pushplata., and Mr. Ram Chatterjee.,2012 focuses on the STC algorithm applied on the search result documents which is stored in the dataset. It articulates the key requirements for web document clustering and clusters would be created on the full text of the web documents.

III. OPTIMIZED WEB DOCUMENT CLUSTERING USING RECURRENT SET (OWCRS) METHOD

The OWCRS follows four **different** phases to efficiently cluster the web documents. The first phase develops a derived genetic algorithm for performing the global clustering process. The derived genetic algorithm groups the documents into meaningful categories. Second phase pursue feature selection to efficiently use confusion matrix which contains information about actual and predicted classifications. Feature selection in the context of supervised learning adopts methods that are usually divided into two classes’ filters and wrappers.

Next phase is the cluster optimality for efficient document clustering based on the optimized conceptual feature words. Optimal weight is calculated for effective clustering of text document. The optimized technique transform a feature-represented document into a concept represented one. Final phase is the efficient web document clustering using join approach to cluster the web pages which primarily finds the recurrent sets and then clusters the documents. These recurrent sets are generated by using recurrent pattern expansion technique.

A. Derived Genetic Algorithm System

The OWCRS method clusters the group of documents into effectiveness of document clustering process. It is carried over by labeling the clusters with the relevant keywords or phrases.Selecting key phrases according to the document is done by the key phrase strategies. Usually a complete list of phrases has been shared by two documents in a cluster. These phrases are assigned scores depends on the set of features evaluated from the matching process in OWCRS method. The statuses of the candidate phrases are assigned in descending order with the top L phrases are assigned as a label for the cluster.

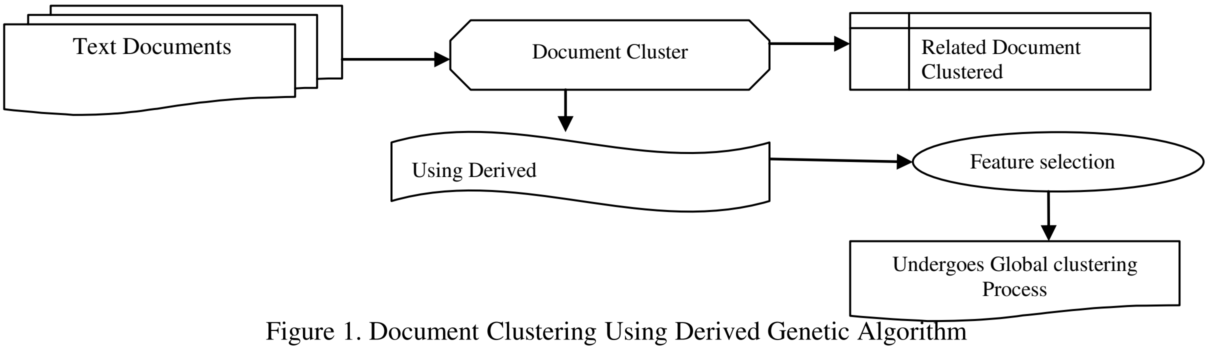
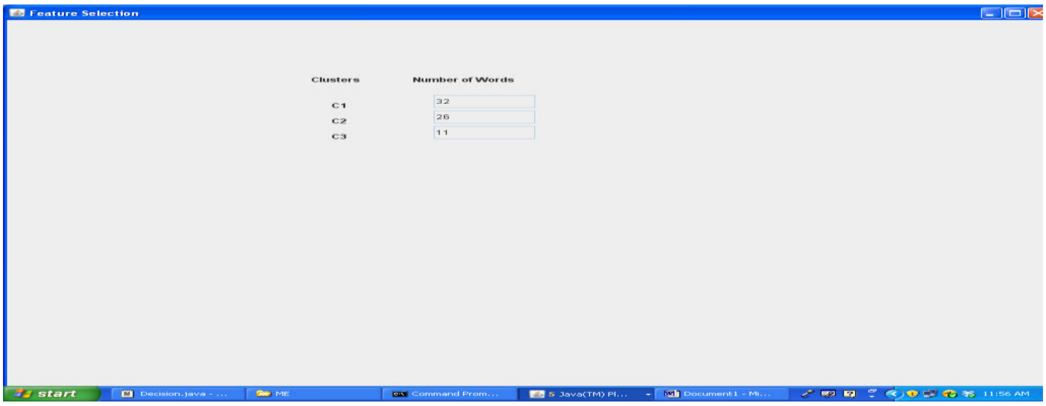

Text Documents
Document Cluster
Related Document Clustered
Using Derived
Feature selection
Undergoes Global clustering Process

Figure 1. Document Clustering Using Derived Genetic Algorithm

The above Figure 1 describes the clustering process using the genetic algorithm. It separates the common themes hidden within the quantity of document clusters. The document clustering application is to cluster the documents into knowledge maps. This application is capable for obtaining successive knowledge retrieval. The text documents are clustered depending on the related features using the Derived Genetic Algorithm technique.

A.1.1 Feature Selection Using Derived GA

To improve the efficiency of clustering process in derived Genetic Algorithm of OWCRS method, not only reduces the size of extracted feature but also reduced the potential inequality of the original feature set. The efficiency and effectiveness of clustering is also improved. Optimized Web document Clustering using Recurrent Set (OWCRS) method is to implement a derived Genetic Algorithm (GA) and it has been combined with optimization of crossover, selection, mutation to preserve the population diversity during the global clustering process and feature selection.



Clusters	Number of Words
C1	32
C2	26
C3	11

Figure 2. Feature Selection Words

The above screenshots displays the count of words after the clustering process. The three different clusters C1, C2, C3 are done for the real datasets (i.e.) sports related data and related features are grouped together for efficient global clustering process.

A.1.2 Derived Genetic Algorithm

A derived genetic algorithms fit in to the larger class of algorithms which produce solutions by natural evolution such as mutation, selection, and crossover of OWCRS technique. The flow of the algorithm is given as follows:

Input: Web Document set WDS, number of generations ‘n’
Output: Best classifier over WDS

```
Step 1: Create a population 'NewPop'
Step 2: Select the group of individuals (chromosome) randomly
Step 3: Evaluate the fitness of each chromosome in 'NewPop'
Step 4:  Optimize selection of individual is done by
Step 4.1: For all members in NewPop
Step 4.2: Evaluate P(i) in every i in NewPop
Step 4.3: Based on p(i) and fitness of I, select two i1, i2 from NewPop
Step 5: While size (newPop) < size (oldPop)
Step 5.1: Select par1 and par2 in oldPop
Step 5.2: Generate n1, n2 through crossover optimization (parent1, parent2)
Step 6: Compute optimize mutation
Step 6.1: n1 = mut(n1) and n2 = mut(n2)
Step 6.2: add n1 and n2 to newPop
Step 7: For choosing the best classifier,
Step 7.1: Compute total time taken to evaluate the global clustering process by
Step 7.2: Total time = $\Sigma^{n}_{i=0}$ time of task $_{i}$
Step 7.3: Select the best chromosome C based on total time;
Step 8: Classifier associated with K.
```

The above Algorithm takes document set WDS as input and provides a best classifier over the document set in OWCRS technique. NewPop selects the group of individuals randomly and evaluate the fitness of each chromosome. Optimize the selection based on the NewPop and OldPop of the system to generate n1, n2 through crossover optimization. Compute optimize mutation for choosing the best classifier. The classifier is associated with K values.

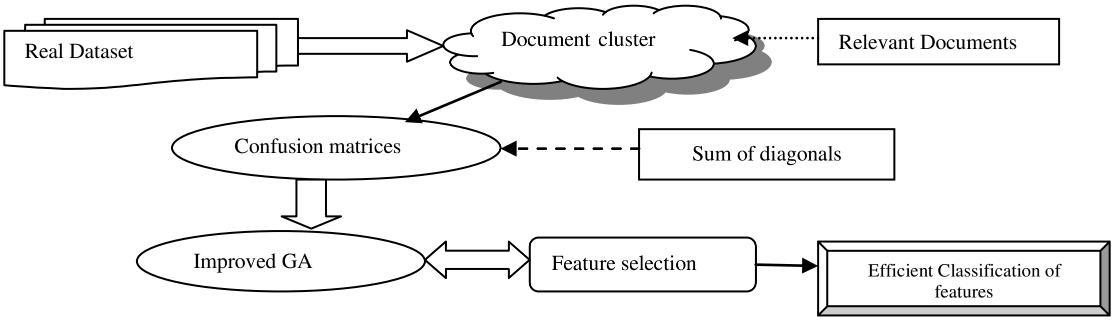


Figure 3. Confusion Matrix for Derived IGA work of OWCRS technique

The above fig illustrates discerns general themes hidden within the corpus. Applications of document clustering go beyond organizing document collections into knowledge maps. This can facilitate subsequent knowledge retrievals and accesses. Optimized Web document Clustering using Recurrent Set (OWCRS) method uses the confusion matrix with the sum of diagonals to improve the Genetic Algorithm.

B.2.1 Confusion matrix

A confusion matrix of OWCRS technique contains information about actual and predicted method done by a classification system. Performance of such systems is normally evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

		Predicted	
		Negative	Positive
Actual	Negative	W	x
	Positive	Y	z

The entries in the confusion matrix are described below as,

- W is the number of correct predictions that an instance is negative,
- x is the number of incorrect predictions that an instance is positive,
- Y is the number of incorrect of predictions that an instance negative, and
- z is the number of correct predictions that an instance is positive.

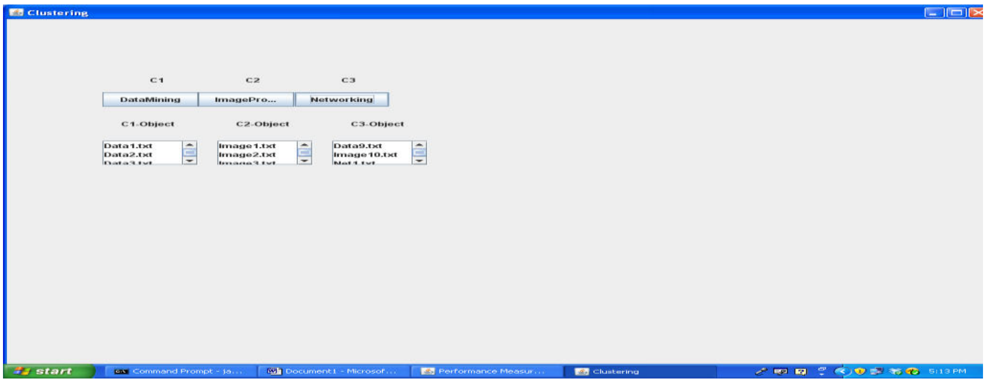


Figure 3. Confusion Matrix Classification

The confusion matrix follows a sum of diagonals approach to classify the documents in the real dataset. The Cluster C1 contains the Data mining documents, cluster C2 includes image processing documents and cluster C3 contains the networking documents.

B.2.2 Improved Genetic Algorithm

An Improved Genetic Algorithm (GA) is a search heuristic which involves inheritance, mutation, selection, and crossover in OWCRS technique. The flow of the algorithm is described below:

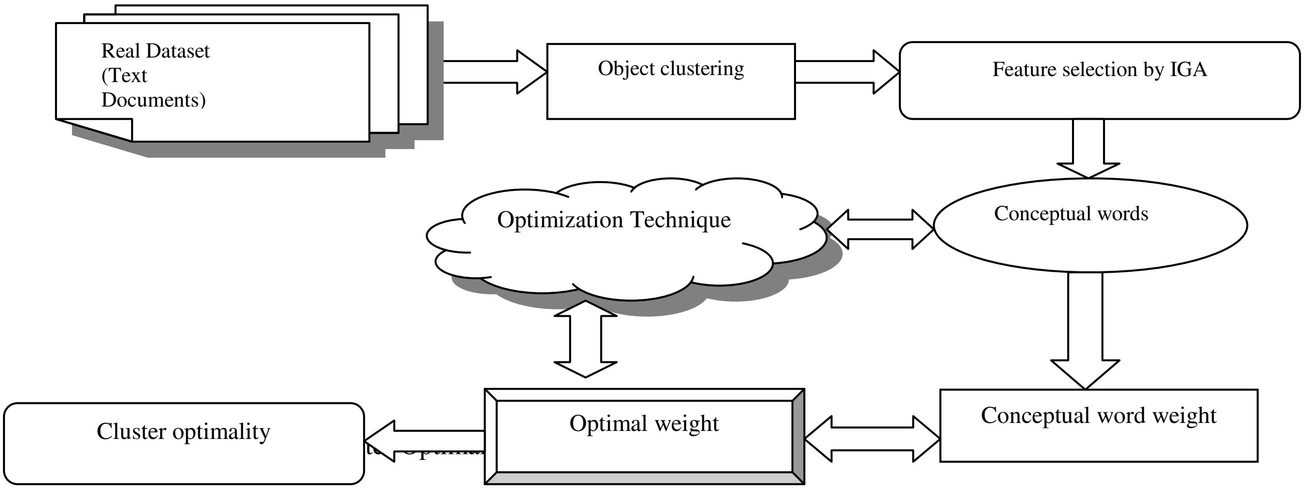
Input: Web Document set WDS, number of generations 'n'
Output: Best classifier over WDS in IGA

```
Step 1: Evaluate the sets of candidate positive and negative terms
Step2: Repeat Step 3 to Step 6 from section 3.1.2 algorithm
Step 3: If oldPop = newPop
    Step 3.1: Select the best chromosome K in oldPop;
    Step 3.2 Eliminate redundancies from K;
Step 4: Classifier associated with K.
```

The above algorithm describes better classification of documents using the Improved Genetic Algorithm in OWCRS technique. The OWCRS evaluates the set of candidate positive and negative terms. NewPop selects the group of individuals randomly and evaluate the fitness of each chromosome. Optimize the selection based on the NewPop and OldPop of the system to generate n1, n2 through crossover optimization. Compute optimize mutation for choosing the best classifier. If oldPop = newPop then selects the best chromosome K in oldPop and removes the redundancies for better classification.

C.3 Cluster Optimality of OWCRS technique

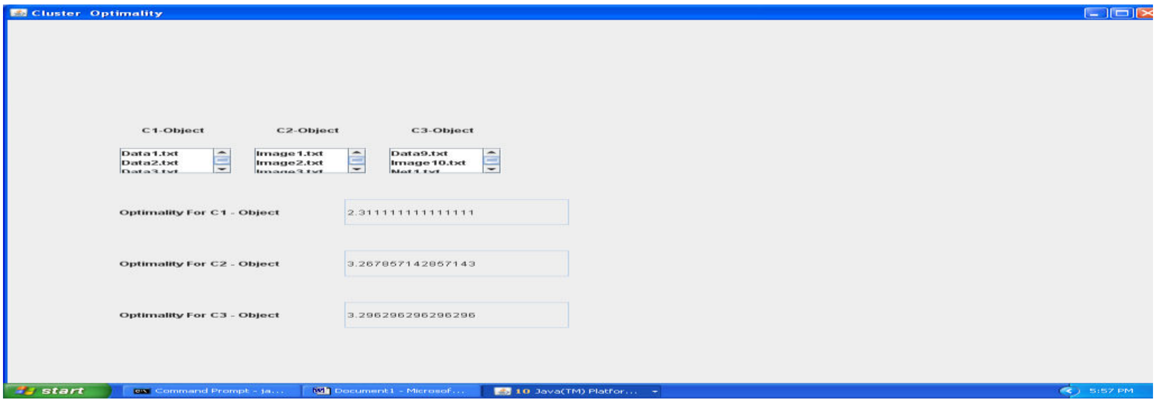
The third phase of OWCRS technique is designed to perform optimization clustering process based on the conceptual optimal weight. The real dataset are used for the evaluation of the cluster optimality while clustering the text documents. Then conceptual optimal weight is calculated based on optimal weight document clustering. The concept weight is also called the Semantic weight. The following figure shows the overview of the proposed system architecture.



The above Figure describes the optimization technique to evaluate the cluster optimality. Then the next step is to identify the featured words. Feature selection is important for clustering efficiency and effectiveness because it not only condenses the size of the extracted feature set but also reduces any potential biases embedded in the original feature set. Previous research commonly has employed feature selection metrics such as TF (term frequency), TF×IDF (term frequency × inverse document frequency), and their hybrids. The feature selection process is carried out using improved Genetic Algorithm of OWCRS technique.

C.3.1 Conceptual Optimal Weight of OWCRS technique

After the preprocessing step alters the text objects, the system converts the attributes to numeric one and employs the weighted vector to signify the text objects in OWCRS. The new concept weighting processes assist in mining exact and functional information from the data, thus minimizing the curse of dimension problem in feature weighting of OWCRS technique.



Cluster Optimality

C1-Object: Data1.txt, Data2.txt, Data3.txt

C2-Object: Image1.txt, Image2.txt, Image3.txt

C3-Object: Data9.txt, Image10.txt, Net1.txt

Optimality For C1 - Object: 2.311111111111111

Optimality For C2 - Object: 3.267857142857143

Optimality For C3 - Object: 3.296296296296296

Figure 5. Cluster Optimality

The above Fig describes the cluster optimality using the optimization technique. The Cluster C3 produces the most optimal result when compared to the other two cluster group using the OWCRS optimization technique. The significance in a text for conceptual optimal weight in which the semantics of document are calculated. The optimization system of OWCRS technique formulates the following assumptions:

a. Number of times the words present in the document is most probably the characteristic words.
b. Word length also distresses the significance of words.
c. If probabilities of that word are high, then word obtains additional weight.
d. One word may be the characteristic word even if it doesn't appear in the document.

The combination of above illustrated four assumptions leads to the weighting structure with cluster optimality of OWCRS technique. This optimization technique provide more accurate outcome course of its conceptual hierarchy.

C.3.2 Calculating Optimal weight

The calculation of words optimal weight using Term Frequency, Inverse Document Frequency (TF-IDF) formula in the document clustering is demonstrated. The optimization technique of OWCRS also considers the conceptual word and conceptual word weight. OWCRS technique optimal weight is calculated by the conceptual word weight and the number of conceptual words to be presented. Formula for calculating optimal weight is given below.

$$OptimalWeight = \frac{\sum_{i=1}^{n} Conceptual\,WordWeight}{No.OfConceptualSimilarityWords} \text{--->1}$$

Finally, the optimization technique of OWCRS ranks the weights and chooses the keywords that have with optimal weight. After choosing the concepts the optimization technique represents each document as a concept vector space.

D.4 Join Approach of OWCRS technique

In vector space model of OWCRS technique, each document is defined as a multidimensional vector of keywords whose axis corresponds to the keyword. The keywords are extracted from the document and weight associated with each keyword determines the importance of the keyword in the document. Thus, a document is represented as, Ei = (v1i, v2i, v3i, v4i,…………….,vmi) where, vk,i is the weight of term k in document i indicating the significance and weight of the keyword words.

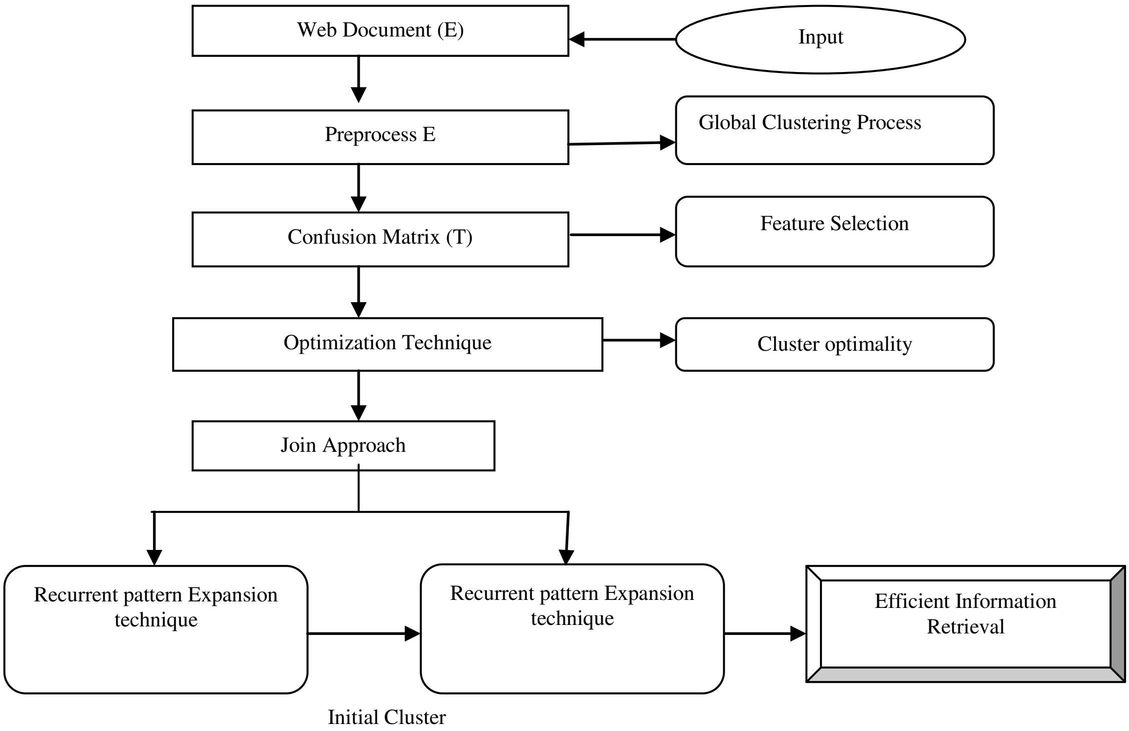


Figure 6: Optimized Web document Clustering using Recurrent Set Technique

The Optimized Web document Clustering using Recurrent Set Technique (OWCRS) takes the Web Document Set 'E' as the input then uses the preprocess technique to remove the full stops and the unwanted words and special characters. It selects the noun as the keyword from the 'E' and ignores verbs, adjectives, adverbs and pronounce. The confusion matrix 'T' uses the sum of diagonals to classify the features into the specific clusters. The matrix formed by the number of counting occurs on the each term of document 'E'. The optimization technique clusters the web documents optimality using the optimal weight calculation.

The join approach of OWCRS technique primarily finds the recurrent sets and then clusters the documents. Then by applying Fuzzy K-Means algorithm.

Input: Web Document set, E, to be clustered, Value of min_sup, to be used in RP-expansion, Value of 'n' fuzziness parameter.

Output: Confusion matrix 'T' which shows how much a document belongs to a cluster, Matrix containing cluster centers 'D'.

Step 1: Preprocessing of E Remove unwanted words such as verbs, adjectives, adverbs and articulate.
Step 2: Create Confusion matrix: Confusion matrix, 'T' created by counting the number of occurrences of each term in each document Ek.
Step 3: Opmization Technique Cluster optimality calculated by optimal weight
Step 4: Extraction of Join Approach Recurrent pattern Expansion approach is used to extract maximum recurrent sets of documents from the term confusion matrix T using the value of minimum support min_sup given as an input.

Step 5: Calculate initial cluster Initial cluster centroids, di, are calculated using the maximal recurrent sets. For each recurrent set fi present in F, cluster di is calculates as di = (E1 +E2 + … + Ei)/I where i is the number of documents in recurrent set ri. D = { di : di is the centroid vector for cluster k }.
Step 6: Calculation final clusters for fact retrieval
Final clusters are calculated using the cluster centroids D, and the 'm' fuzziness parameter

The above Step 1 to 6 follows a set of instructions and produces the efficient information retrieval using the Confusion matrix 'T'. The information is retrieved by calculating the minimum support value in OWCRS technique. The step by step processes are undergone to overcome all the risks and obtain efficient document clustering on the web.

D.4.1 Recurrent Pattern Expansion Approach

A transaction UCI Repository Bag of words dataset 'E' initially constructs the recurrent pattern. It collects the set of recurrent items (Ritems) and their support counts after scanning the bag of words dataset (E) at once. The root of the recurrent pattern tree is created and it is labeled as "null". For each transaction Qtrans in 'E' do the following task in OWCRS technique.

Select and sort the frequent items in Qtrans according to the order of Mfreq. Let the sorted frequent list in Qtrans be [d ǀ Dlist], where d is the first element and Dlist is the remaining list. The insert_tree is called as ([d ǀ Dlist], T) on OWCRS technique if T has a child M such that M.item-name=d.itemname, then increment M"s count by 1, else create a new node M, and let its count be 1, its parent link be linked to T, and its node-link to the nodes with the same item-name via the node-link structure of OWCRS technique. If Dlist is nonempty, call insert_tree (Dlist,M) recursively.

D.4.2 Fuzzy K-means Method (FKM)

Fuzzy K-means clustering on OWCRS technique allows each feature vector to belong to more than one cluster with different association degrees (between 0 and 1) and fuzzy limitations between clusters. In the present the optimal number of clusters is same as the number of recurrent item sets obtained using recurrent pattern expansion technique. FKM clustering is based on optimizing the objective function of OWCRS technique. FKM starts with arbitrary Confusion matrix 'T' and a fixed number of clusters. Number of columns and rows of the matrix 'T' depends on the documents and the number of clusters.

The recurrent sets generated are of regularity greater than the minimum support supplied by the user. In the generated recurrent sets of documents, the terms are taken to be the transactions and the documents are the items of the transactions in bad of words dataset. OWCRS technique helps for efficient information retrieval by deciding the number of clusters and also the centers of these clusters which is simply the centroid of the respective recurrent item-set.

IV. EXPERIMENTAL EVALUATION

Optimized Web document clustering using Recurrent Set (OWCRS) is implemented using JAVA platform to identify the web document cluster efficiency. The real dataset are used to preprocess, performs the global clustering process using the confusion matrix, cluster optimality value. Bag of words data set contains five text collections. For each text collection, D is the number of documents, W is the number of words in the vocabulary, and N is the total number of words in the collection (below, NNZ is the number of nonzero counts in the bag-of-words). After tokenization and elimination of stop words, the terms of unique words was truncated by only keeping words that occurred more than ten times. These data sets are ideal for clustering and topic modeling experiments.

The performance of the Optimized Web document Clustering using Recurrent Set (OWCRS) is measured in terms of

- True negative rate
- Accuracy

- Cluster optimality
- Information Retrieval Efficiency

V. RESULTS AND DISCUSSION

In this work, we efficiently evaluated the information retrieval from the web document clustering. The below table and graph describes the performance of the Optimized Web document Clustering using Recurrent Set (OWCRS) with Efficient Concept-Based Mining Model (CBMM), Derived Genetic Algorithm (DGAO) with optimization and Improved Genetic Algorithm (IGA) model.

A.1 Performance Result of True Negative Rate:

True Negative Rate (TNR) is defined as the quantity of negatives cases that were classified appropriately. It is calculated using the equation

$$TNR = a / a + b$$

Cluster Object	True Negative Rate			
	Proposed OWCRS	CBMM Model	DGAO Scheme	IGA Method
10	1.1	0.9	0.6	0.2
20	1.2	1.0	0.5	0.3
30	1.1	0.9	0.6	0.4
40	1.3	0.8	0.5	0.2
50	1.5	0.7	0.4	0.2
60	1.7	0.7	0.3	0.2
70	1.8	0.7	0.4	0.3
80	1.9	0.7	0.5	0.4

Table 1 Cluster Object vs. True Negative Rate

The above table describes the true negative rate based on the cluster object. Whenever the cluster object increases, the true negative rate will also be high. The true negative rate are evaluated in Optimized Web document Clustering using Recurrent Set (OWCRS), Efficient Concept-Based Mining Model (CBMM), Derived Genetic Algorithm (DGAO) with optimization and Improved Genetic Algorithm (IGA) model.

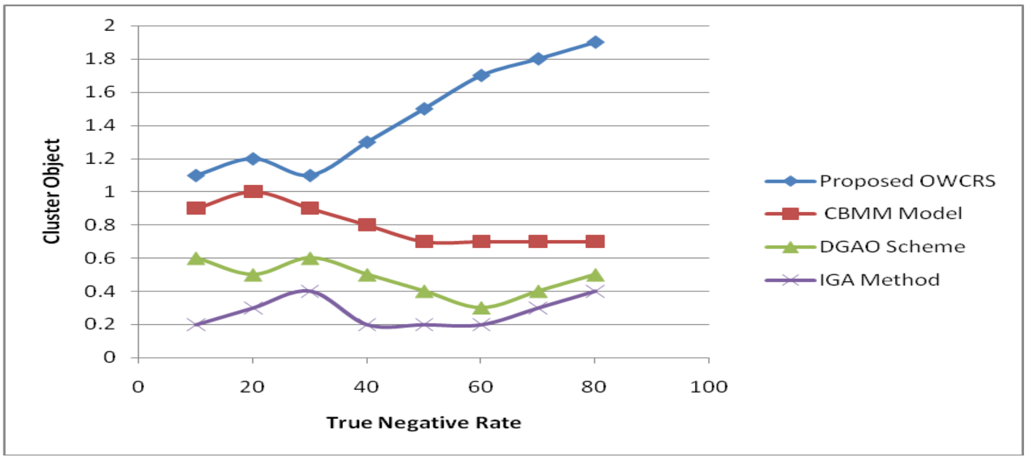


Figure 7: Cluster Object vs. True Negative Rate

The above Fig describes the true negative rate of the Optimized Web document Clustering using Recurrent Set method on real dataset. The documents are clustered and true negative rate increases in the Optimized Web document Clustering using Recurrent Set (OWCRS) when compared with Efficient Concept-Based Mining Model (CBMM), Derived Genetic Algorithm (DGAO) with optimization and Improved Genetic Algorithm (IGA) model. The variance of OWCRS is approximately 20-25% higher than in the other models.

A.2 Performance Result of Accuracy:

Accuracy is defined as a rate at which it retrieves (i.e.) fetching of the information from the clustered web documents.

$$\text{Accuracy} = \frac{\text{No. of correctly classified web documents}}{\text{No. of clustered documents}}$$

No. of iterations	Accuracy			
	Proposed OWCRS	CBMM Model	DGAO Scheme	IGA Method
25	99	85	80	75
50	98	85	79	75
75	98	87	81	77
100	99	89	81	78
125	98	90	82	78
150	98	89	82	79
175	97	87	83	79
200	95	85	83	78
225	96	84	81	75

Table 2 No. of iterations vs. Accuracy

It described the accuracy based on number of iterations of the system. The Optimized Web document clustering using Recurrent Set (OWCRS) is compared with Efficient Concept-Based Mining Model (CBMM), Derived Genetic Algorithm (DGAO) with optimization and Improved Genetic Algorithm (IGA) model for accuracy.

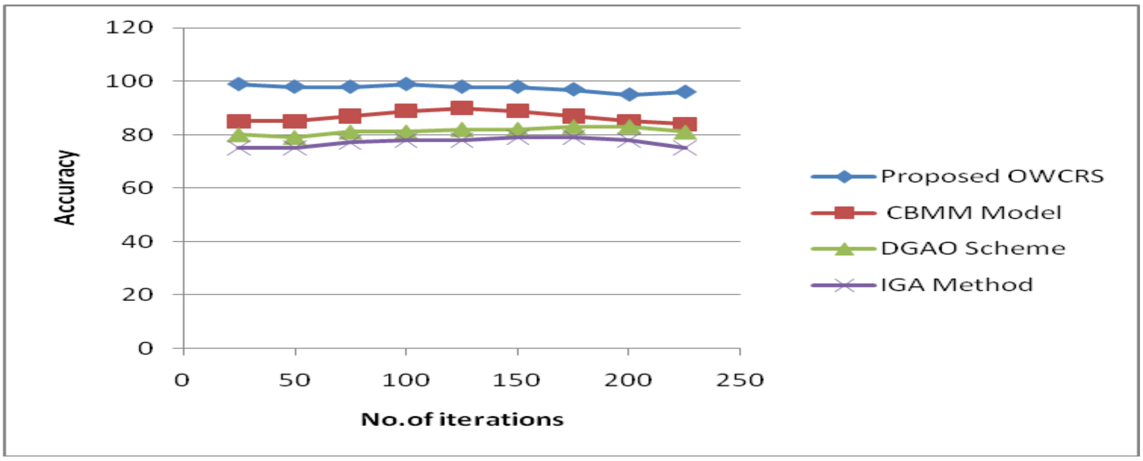


Fig 7 No. of iterations vs. Accuracy

The above figure describes the accuracy of OWCRS technique with Efficient Concept-Based Mining Model (CBMM), Derived Genetic Algorithm (DGAO) with optimization and Improved Genetic Algorithm (IGA) model on real dataset. The variance of the OWCRS accuracy in web document is approximately 95 – 98 % high when compared to all other schemes.

A.3 Performance Result of Cluster Optimality:

Cluster optimality is defined as the amount of time (i.e.) optimal to cluster the web documents. Cluster optimality provides an efficient clustering using the OWCRS technique.

$$OptimalWeight = \frac{\sum_{i=1}^{n} Conceptual\,WordWeight}{No.OfConceptualSimilarityWords} *100$$

No. of clusters	Cluster Optimality (%)			
	Proposed OWCRS	CBMM Model	DGAO Scheme	IGA Method
5	99	82	80	65
10	98	83	79	66
15	98	84	81	67
20	99	85	81	68
25	98	86	82	69
30	98	85	82	66
35	97	85	83	65
40	98	85	83	64

Table 3 No. of clusters vs. Cluster Optimality

The cluster optimality of OWCRS system is measured based on the number of clusters. The above table (table 5.3) describes the cluster optimality on Optimized Web document clustering using Recurrent Set (OWCRS), Efficient Concept-Based Mining Model (CBMM), and Derived Genetic Algorithm (DGAO) with optimization and Improved Genetic Algorithm (IGA) model. The OWCRS system uses the optimization technique with optimal weight calculation to find the optimal solution. It is measured in terms of percentage (%).

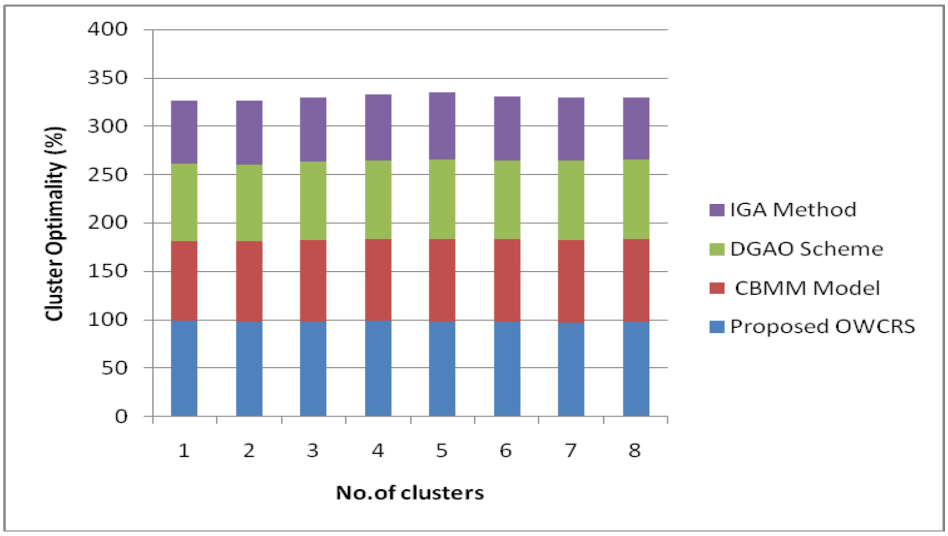


Figure 8. No. of clusters vs. Cluster Optimality

The comparison result provides the 85 – 95 % optimal result on real dataset when compared with the Efficient Concept-Based Mining Model (CBMM) and Derived Genetic Algorithm (DGAO) with optimization and Improved Genetic Algorithm (IGA) techniques.

A.4 Performance Result of Information Retrieval Efficiency:

Information Retrieval Efficiency is defined as the efficient way of retrieving the information and responding to the sender with accurate fact result.

$$\text{Efficiency} = \left\{ \frac{\text{Amount of Request send}}{\text{Response Time}} \right\} *100$$

No. of web Documents	Information Retrieval Efficiency			
	Proposed OWCRS	CBMM Model	DGAO Scheme	IGA Method
100	99.1	85.7	80.2	75.6
200	98.1	85.1	79.4	75.8
300	98.3	87.4	81.7	77.4
400	99.5	89.3	81.6	78.3

500	98.4	90.9	82.4	78.2
600	98.3	89.5	82.6	79.1
700	97.4	87.6	83.8	79.6

Table 4: **No. of web Documents vs. Information Retrieval Efficiency**

The above table 4 described the Information retrieval efficiency based on the number of web documents in OWCRS. The outcome of the Optimized Web document clustering using Recurrent Set on information retrieval efficiency is measured.

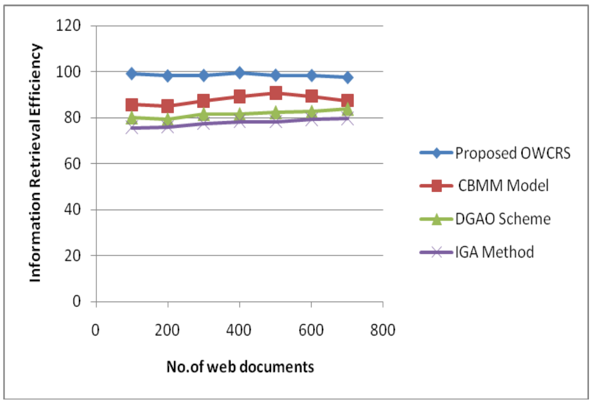


Fig 9 No. of web Documents vs. Information Retrieval Efficiency

The above Fig describes the information retrieval efficiency on the UCI bag of words dataset. The Optimized Web document clustering using Recurrent Set (OWCRS) method are compared with Efficient Concept-Based Mining Model (CBMM) and Derived Genetic Algorithm (DGAO) with optimization and Improved Genetic Algorithm (IGA) model provide the more efficient fact retrieval from the bag of words dataset.The experimental evaluation performed on real and bag of words dataset provides related documents in the same cluster so that searching of documents becomes more efficient.

VI. CONCLUSION

Optimized Web document Clustering using Recurrent Set (OWCRS) method is implemented in web document clustering for efficient fact retrieval. The last phase introduces a join approach to cluster the web pages which primarily finds the recurrent sets and then clusters the documents. Then by applying Fuzzy K-Means Algorithm (FKM) on OWCRS found clusters having documents which are extremely related and have related features. Experimental results show that our approach is more efficient then the above two join approach and can handle more efficiently in robust nature. Performance evaluation show benefits in terms of cluster optimality, true negative rate, information retrieval and 95 – 98 % more accurate information retrieval on real and UCI repository bag of words dataset. Future work would focus on civilizing the cluster sets by semantic based clustering and leveling the documents in each cluster using subject based modeling.

REFERENCES

[1] A.K. Santra, C. Josephine Christy and B.Nagarajan, " Cluster Based Hybrid Niche Memetic and Genetic Algorithm for Text Document Categorization", IJCSI, vol.8, Issue 5, no. 2,pp. 450-456, Sep 2011.

[2] A.K. Santra and C. Josephine Christy,"Genetic Algorithm and Confusion Matrix for Document Clustering" , IJCSI, vol.9, Issue 1, no. 2,pp. 322-328, Sep 2012.

[3] A.K. Santra and C. Josephine Christy,"An Efficient Document Clustering by Optimization technique for cluster Optimality, IJCA, vol.43, No. 16,pp. 15-20, April 2012.

[4]Shady Shehata., Fakhri Karray., and Mohamed S. Kamel., "An Efficient Concept-Based Mining Model for Enhancing Text Clustering," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 10, OCTOBER 2010

[5] Yun Yang., and Ke Chen., "Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 2, FEBRUARY 2011

[6] Ninad Thakoor., and Jean Gao., "Branch-and-Bound for Model Selection and Its Computational Complexity," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 5, MAY 2011

[7] K.Sathiyakumari., G.Manimekalai., V.Preamsudha., "A Survey on Various Approaches in Document Clustering.," International Journal Computer Technology Vol 2 (5), 1534-1539, 2011

[8] Swatantra kumar sahu., Neeraj Sahu., G.S.Thakur., "Classification of Document Clustering Approaches," International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 5, May 2012 ISSN: 2277, 2012

[9] K. Premalatha., A.M. Natarajan., "Genetic Algorithm for Document Clustering with Simultaneous and Ranked Mutation," Modern Applied Science, Vol 3, No:2, 2009

[10] Kiran G V R., Ravi Shankar., and Vikram Pudi., "Frequent Item set Based Hierarchical Document Clustering Using Wikipedia as External Knowledge," Springer-Verlag Berlin Heidelberg 2010

[11] George Pallis., Athena Vakali., Jaroslav Pokorny., "A clustering-based pre fetching scheme on a Web cache environment," Elsevier Journal, 2008

[12] Pushplata., Mr. Ram Chatterjee., "An Analytical Assessment on Document Clustering," International Jornal on Computer Network and Information Security, 2012.

BIOGRAPHY

B.Nagarajan received his Ph. D. degree in Pattern Recognition in the year 2010 from Anna University, Chennai. He has been in the teaching profession for more than a decade since 1997. His areas of academic interest are Image Processing, Pattern Recognition and Neural Networks. Also, he has published 14 papers in International Journals and presented 18 research papers in the National/International Conferences. He is the reviewer / editorial board member of 10 international journals from various countries like Singapore, Hong Kong, Korea, United States, Thailand, Romania and India.