An Efficient Document Clustering by Optimization Technique for Cluster Optimality

A.K.Santra
Dean,
Care School of computer Applications,

C. Josephine Christy
Research Scholar,
Bharathiar University ,Coimbatore-638401

ABSTRACT

Document clustering grows to be a very famous technique with the popularity of the web which also indicates that quick and best clustering technique acts as an important issue. Document clustering is about identifying semantically interconnected groups from formless collection of text documents. Feature Selection is significant for clustering process because number of the isolated or redundant feature should misguide the clustering results. Existing work presented improved Niching memetic algorithm and improved Genetic algorithm (GA) for feature selection. To attain more perfect document clustering, more instructive features including optimal conceptual weight are essential.

In this paper, the proposed work presents the optimization technique to evaluate the cluster optimality for efficient document clustering based on the optimized conceptual feature words.

The conceptual words (similarity words) are extracted from the featured words by using feature selection process. The important of cluster words are identified by the optimal conceptual word weight values. Experiments are carried out to evaluate the proposed optimization technique for efficient document clustering in terms of Conceptual word weight, Number of conceptual words and optimal conceptual word weight.

Keywords

Document clustering, Conceptual words, cluster optimality

1. INTRODUCTION

Present document clustering techniques tend to ignore some foremost aspects that significantly limit their practical applicability. Clustering of Text document is generally noticed as a significant method, which conveys one obviously defined outcome, which requires being optimal in several manner. However, runs opposing to the truth that different people consists fairly different requirements with consider to clustering of texts because they can examine the similar documents from entirely different viewpoints. Thus, what are needed are document clustering techniques that offer various subjective viewpoints.

Text document clustering is one of the essential tasks in text mining. Clustering is to segregate a group of text documents into several category groups. Consequently that document in the similar category group illustrates the identical topic.

There are numerous uses of clustering in valid applications, for example, collecting the Web search results and classifying digital documents. Contrasting clustering structured data, clustering text data admits lot of new challenges. Among others, the amount of text data, dimensionality, sparsity and difficult semantics are the most essential ones. This distinctiveness of text data has need of clustering techniques to be scalable to great and high dimensional data, and able to manage sparsity and semantics. Most of the previous text document clustering methods use clustering techniques based only on term strength and document frequency in which single terms are employed as features for indicating the documents and they are handled independently.

The trouble of document clustering is commonly defined as follows: Set of text documents are partitioned into a prearranged or a routinely derived number of clusters, such that the documents allocated to each cluster are more related to each other than the documents assigned to dissimilar clusters. In other words, the documents in one cluster distribute the identical topic, and the documents in dissimilar clusters stand for different topics. In recent times, a flurry of activity has there in the text document clustering domain. All clustering technique studies in a somewhat special method and initiates biases. Clustering technique will frequently perform better in a particular field. Additionally, analysis of the resultant clusters will be complicated or even totally insignificant. This builds for a motivating and dynamic field of investigate.

In this paper, proposed optimization technique evaluates 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 proposal optimization technique considers conceptual weight for choosing the trait of the documents. Clustering text documents into category groups is a significant process in indexing, retrieval, management and mining of rich text data on the Web or in shared information systems. Among others, the demanding problems of text clustering are large volume, huge dimensionality and difficult semantics. For these problems, the proposed optimization technique has offered a proficient result which is scalable clustering with Conceptual optimal weight.

2. LITERATURE REVIEW

A. K. Santra, C. Josephine Christy and B.Nagarajan [1] have proposed that cluster based niche memetic and genetic algorithm have been designed & implemented by optimizing feature selection of text in the document repository. The contribution of genetic algorithm works with an evaluation of fitness function. Accuracy can be calculated through the document clustering. A. K. Santra, C. Josephine Christy [2] have proposed the application of real data sets with an abundance of irrelevant or redundant features. Improved GA relies on confusion matrices and uses the F-measure as the fitness function. In this case, identifying a relevant subset that adequately captures the underlying structure in the data can be particularly useful. Numerous clustering techniques have been presented in the literature. Clustering techniques are mostly classified into hierarchical and partitioning methods [3]. In hierarchical clustering technique, data objects are grouped into a tree of clusters. These methods have further been categorized into agglomerative and disruptive hierarchical clustering based on whether the hierarchical disintegration is created in a bottom-up or top-down manner [4]. K-means and its departure are the most familiar partitioning mechanism. All clustering techniques based on frequencies of terms and resemblances of data points experience from the same mathematical possessions of the fundamental spaces [5].

In [6], proposed several local search operations to effectively design an MA for simultaneous clustering and feature selection which incorporate local searches with traditional GAs, have been proposed and applied successfully to solve a wide variety of optimization problems. These studies [7] show that pure GAs are not well suited to fine tuning structures in complex search spaces and that hybridization with other techniques can greatly improve their efficiency. In [8], have proposed about data clustering is a common technique for statistical data analysis and has been used in a variety of engineering and scientific disciplines such as biology (genome data). In [7] have proposed the purity of a cluster represents the fraction of the cluster corresponding to the largest class of documents assigned to that cluster; thus, the purity of the cluster [9].

In the scheme for feature selection prepared illustrate feature selection for an unsupervised learning task, specifically conceptual clustering [10]. They communicate a sequential feature selection approach based on an existing COBWEB conceptual clustering technique. In their estimation they demonstrate that feature selection considerably expands the consequences of COBWEB [11]. The disadvantage that Devaney and Ram face, however, is that COBWEB is not scalable like K-Means. In [12], identifies the semantic relations using the ontology. The ontology is used represent the term and concept relationship [13]. The synonym, meronym and hypernym relationships are represented in the ontology. The concept weights are estimated with reference to the ontology [14]. The concept weight is used for the clustering process. In [15], the concept weight is highlight for complex semantic problem area. But in this proposed system the concept weight is calculating for curse of dimensions problem. Consequently, the proposed system derives the high level requirement for text clustering approaches that they either rely on concept weight.

3. OPTIMIZATION TECHNIQUE FOR EFFICIENT DOCUMENT CLUSTERING

The proposed optimization technique is designed to perform clustering process based on the conceptual optimal weight. The proposed technique can transform a feature-represented document into a concept represented one. Therefore, the target document corpus will be clustered in accordance with the concepts representing individual document, and thus, achieve the proceeding of document clustering at the conceptual level. The system uses the text documents for the clustering process. Initially document preprocessing is done. 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 (i.e., non-trimmed) 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 Niching memetic algorithm and improved GA algorithm.

Then conceptual optimal weight is calculated and based on this optimal weight document clustering is performed. The concept weight is also called the Semantic weight. The following figure shows the overview of the proposed system

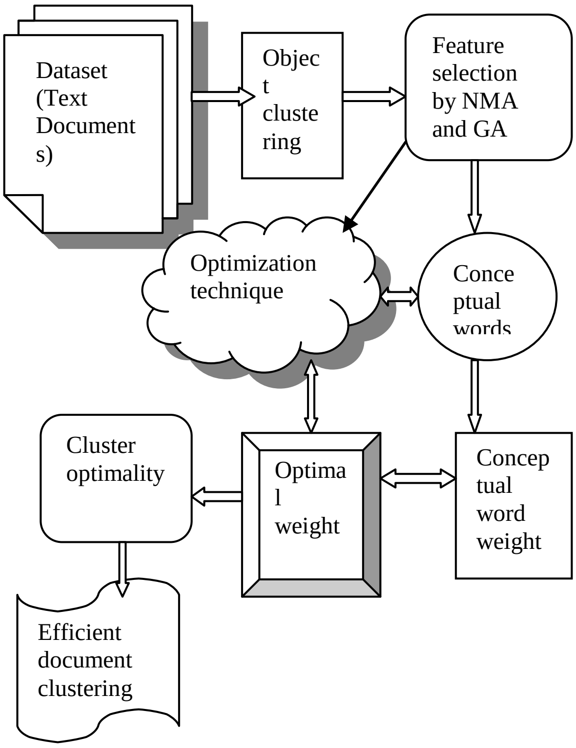


Figure 1: Optimization Technique for Efficient Document Clustering

3.1. CONCEPTUAL OPTIMAL WEIGHT

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. So each and every data point has particular measured value. Our objective is somewhat different from preceding approaches. Optimization technique also observe how new concept weighting processes assist in mining exact and functional information from the data, thus minimizing the curse of dimension problem in feature weighting. Furthermore to deal with the concern of text clustering, an appropriate method for calculating and selecting the feature selection is proposed. Various words have diverse significance in a text, thus a significant indicator for conceptual optimal weight gives to the semantics of document is calculated. When designing the method of calculating the optimal weights, the proposed optimization system formulates the following assumptions:

a. Number of times the words present in the document, most probably it is the characteristic words (if the number of times word occurrence is high then the word frequency will be high).
b. Word length also distresses the significance of words.

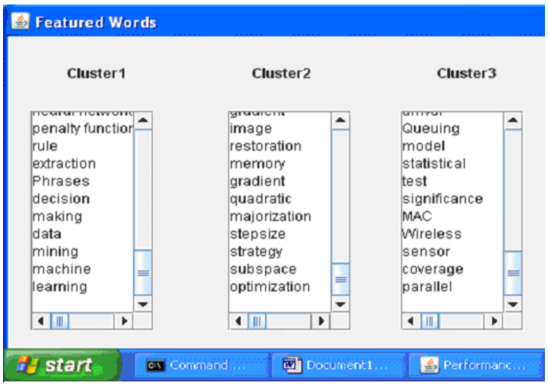
c. The word will obtain additional weight, if the probabilities of that word are high.
d. One word may be the characteristic word even if it doesn't appear in the document.

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

3.2. Calculating Optimal weight

Several researchers newly put their focus on calculating the words optimal weight using Term Frequency, Inverse Document Frequency (TF-IDF) formula in the document. But proposed optimization technique also considers the conceptual word and conceptual word weight. In this work, the optimal weight will be calculated by the conceptual word weight and the number of conceptual words to be presented. Formula for calculating optimal weight is given below.

$$OptimalWeight = \frac{\sum_{i=1}^{n} ConceptualWordWeight}{No.OfConceptualSimilarityWords} \text{ -------->} \ (1)$$



Cluster1
penalty functior
rule
extraction
Phrases
decision
making
data
mining
machine
learning

Cluster2
image
restoration
memory
gradient
quadratic
majorization
stepsize
strategy
subspace
optimization

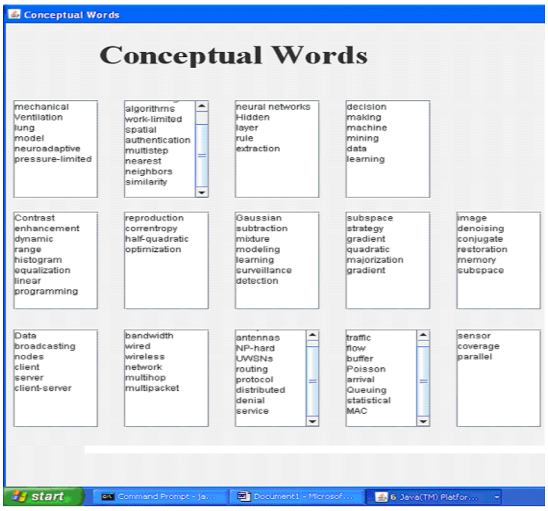
Cluster3
Queuing
model
statistical
test
significance
MAC
Wireless
sensor
coverage
parallel

Figure 2: Featured words Preliminarily, documents were subjected to the following pre-processing steps: (1) First, we removed all words occurring in a list of common stopwords, as well as punctuation marks and numbers; (2) then, we extracted all n-grams, defined as sequences of maximum three words consecutively occurring within a document (after stopword removal); (3) at this point we have randomly split the set of seen data into a training set (70%), on which to run the GA, and a validation set (30%), on which tuning the model parameters. We performed the split in such a way that each category was proportionally represented in both sets (stratified holdout). Then featured words are identified using improved Niching memetic and improved GA algorithm as shown in figure 2. Next from the each cluster, the conceptual words are discovered which is represented in the figure 3.

Finally, the optimization technique ranks the weights and chooses the keywords that have with optimal weight. After choosing the concepts the proposed optimization technique represents each document as a concept vector i.e., the concept-based document representation.

4. EXPERIMENTAL EVALUATION OF OPTIMIZATION TECHNIQUE

The performance of our proposed optimization technique is evaluated in this section. The optimization technique is implemented in java platform to identify the optimal weight of each conceptual word. The optimization technique was tested with a file of 500 text documents. For each dataset, 30% of the documents are randomly selected as test documents, and the rest are used to create training sets as follows: γ percent of the documents from the positive class is first selected as the positive set P. The rest of the positive documents and negative documents are used as unlabeled set U. We range γ percent from 10%- 50% to create a wide range of scenarios.



Conceptual Words

mechanical
Ventilation
lung
model
neuroadaptive
pressure-limited

algorithms
work-limited
spatial
authentication
multistep
nearest
neighbors
similarity

neural networks
Hidden
layer
rule
extraction

decision
making
machine
mining
data
learning

Contrast
enhancement
dynamic
range
histogram
equalization
linear
programming

reproduction
correntropy
half-quadratic
optimization

Gaussian
subtraction
mixture
modeling
learning
surveillance
detection

subspace
strategy
gradient
quadratic
majorization
gradient

image
denoising
conjugate
restoration
memory
subspace

Data
broadcasting
nodes
client
server
client-server

bandwidth
wired
wireless
network
multihop
multipacket

antennas
NP-hard
UWSNs
routing
protocol
distributed
denial
service

traffic
flow
buffer
Poisson
arrival
Queuing
statistical
MAC

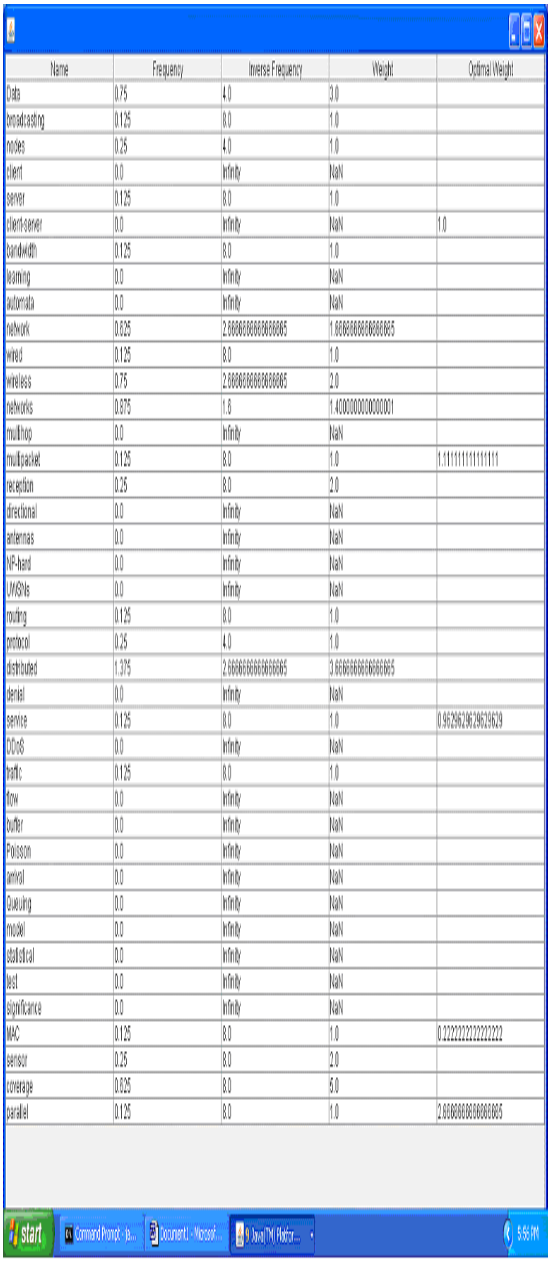
sensor
coverage
parallel

Figure 3: Conceptual words

Based on the term frequency and inverse document frequency, the term weight will be calculated. Then the optimal weight of each conceptual words are calculated by using the formula which is given as,

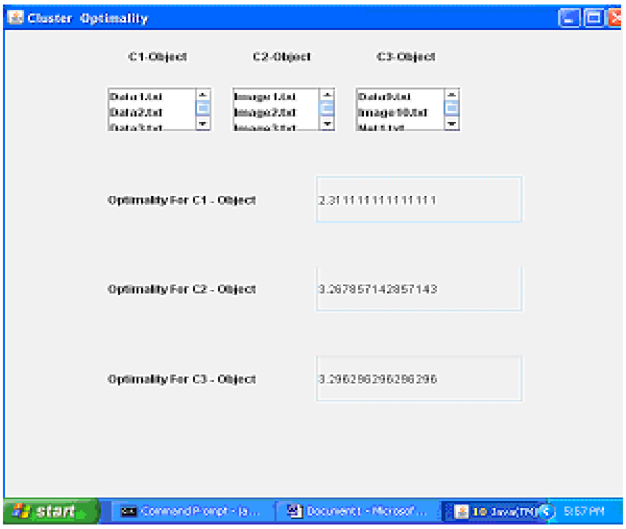
$$OptimalWeight = \frac{\sum_{i=1}^{n} ConceptualWordWeight}{No.OfConceptualSimilarityWords}$$

The optimal weight for each and every conceptual words and cluster optimality for each cluster, generated from our experiment is shown in figure 4.3 and figure 4.4 respectively.



Name	Frequency	Inverse Frequency	Weight	Optimal Weight
Data	0.75	4.0	3.0	
broadcasting	0.125	8.0	1.0	
nodes	0.25	4.0	1.0	
client	0.0	Infinity	NaN	
server	0.125	8.0	1.0	
client-server	0.0	Infinity	NaN	1.0
bandwidth	0.125	8.0	1.0	
learning	0.0	Infinity	NaN	
automata	0.0	Infinity	NaN	
network	0.625	2.6666666666666665	1.6666666666666665	
wired	0.125	8.0	1.0	
wireless	0.75	2.6666666666666665	2.0	
networks	0.875	1.6	1.4000000000000001	
multihop	0.0	Infinity	NaN	
multipacket	0.125	8.0	1.0	1.111111111111111
reception	0.25	8.0	2.0	
directional	0.0	Infinity	NaN	
antennas	0.0	Infinity	NaN	
NP-hard	0.0	Infinity	NaN	
UWSNs	0.0	Infinity	NaN	
routing	0.125	8.0	1.0	
protocol	0.25	4.0	1.0	
distributed	1.375	2.6666666666666665	3.6666666666666665	
denial	0.0	Infinity	NaN	
service	0.125	8.0	1.0	0.9629629629629629
DDoS	0.0	Infinity	NaN	
traffic	0.125	8.0	1.0	
flow	0.0	Infinity	NaN	
buffer	0.0	Infinity	NaN	
Poisson	0.0	Infinity	NaN	
arrival	0.0	Infinity	NaN	
Queuing	0.0	Infinity	NaN	
model	0.0	Infinity	NaN	
statistical	0.0	Infinity	NaN	
test	0.0	Infinity	NaN	
significance	0.0	Infinity	NaN	
MAC	0.125	8.0	1.0	0.2222222222222222
sensor	0.25	8.0	2.0	
coverage	0.625	8.0	5.0	
parallel	0.125	8.0	1.0	2.6666666666666665

Figure 4: Optimal Weight of conceptual words



Cluster Optimality

C1-Object	C2-Object	C3-Object
Data1.txt Data2.txt Data3.txt	Image1.txt Image2.txt Image3.txt	Data9.txt Image10.txt Net1.txt

Optimality For C1 - Object 2.3111111111111111

Optimality For C2 - Object 3.267857142857143

Optimality For C3 - Object 3.296296296296296

Figure 5: Cluster Optimality

5. RESULTS AND DISCUSSION ON OPTIMIZATION TECHNIQUE

The text documents are represented as unstructured data. It is very difficult to cluster text documents. The document clustering needs a preprocessing task to translate the unstructured data values into a structured one. The documents are huge dimensional data elements. Initially, the dimension is minimized using the stop word removal and stemming process. The Experimentation is conducted with 500 text documents collected from Goggle Search Engine which were used in the assessment. The datasets were taken as related topic of Data mining, Image processing and Networking. After preprocessing the system transformed a feature represented document into concept represented one. Therefore, the target document corpus will be clustered in accordance with the concept represented one and thus achieve the proceeding of document clustering at the conceptual level. Then the proposed technique anchors the analysis process. Finally, it is important to measure the efficiency of the proposed method. The proposed optimization technique is evaluated in terms of Conceptual word weight, Number of conceptual words and optimal conceptual word weight.

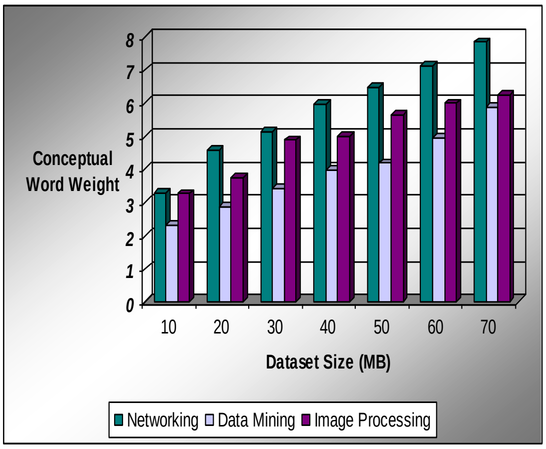


Figure 6: Conceptual word weight

Figure 6 shows the performance result of our proposed Optimization technique Based on the conceptual word weight.

The datasets are taken for our experiment from 10,20…70MB. For each dataset, Conceptual word weight will be calculated. The networking Dataset has the highest Conceptual word weight value compared with Data mining and Image processing.

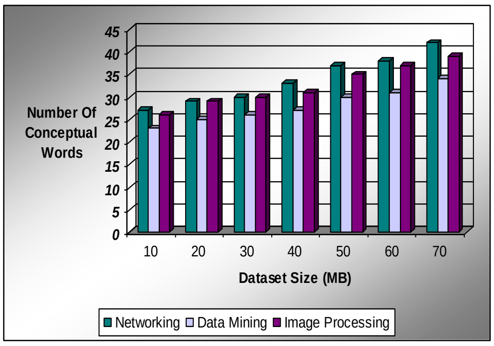


Figure 7: Number of Conceptual words

Figure 7 illustrates the experimental performance of proposed Optimization technique in terms of number of conceptual words. There are three types of datasets named as Networking, Data mining and Image processing, are taken for our experiment. The featured words to be taken as input in this process. From the feature selection, the conceptual words (similarity words) are discovered. Figure 5.2 shows the performance graph which demonstrates number of conceptual words that is extracted from the each datset.

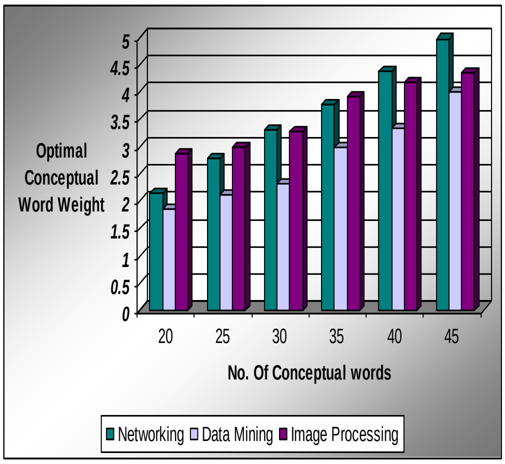


Figure 8: Optimal Conceptual word weight

Figure 8 gives the optimal conceptual word weight of proposed Optimization technique for number of conceptual words. The optimal conceptual word weight was calculated by the ratio of summation of each conceptual word weight to the number of conceptual similarity words. Optimal Conceptual word weight increases the accuracy of text document clustering.

6. CONCLUSION

In this paper, we have implemented a new mechanism, Optimization technique with cluster optimality to identify the optimal conceptual word weight for efficient clustering of text documents. When weighed by the concept, the clustering system can improve the accuracy and performance of text documents. The important of cluster words have been extracted by the optimal conceptual word weight values. The conceptual words have been identified through the feature selection process. For Feature selection, we used improved Niching memetic algorithm and improved Genetic algorithm (GA). From the optimized conceptual feature words, the cluster optimality has been calculated. Experimental results shows that the proposed optimization technique for efficient document clustering gives better result in terms of Conceptual word weight, Number of conceptual words and optimal conceptual word weight.

8. 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] K. Deep and K. N. Das. Quadratic approximation based Hybrid Genetic Algorithm for Function Optimization. AMC, Elsevier, Vol. 203: 86-98, 2008.

[4] Sun Park, Dong Un An, Choi Im Cheon, "Document Clustering Method Using Weighted Semantic Features and Cluster Similarity," digitel, pp.185-187, 2010 Third IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning, 2010

[5] Wen-Hui Yang, Dao-Qing Dai, and Hong Yan, Fellow, IEEE,” feature extraction and uncorreleted discriminant analysis for high dimentional data”, IEEE transactions on knowledge and data engineering, vol. 20, no. 5, may 2008

[6] Yanjun Li, Congnan Luo,,” Text clustering with feature selection by using statistical data”, IEEE Transactions on Knowledge and Data Engineering, vol: 20 no:5, may 2008.

[7] Huan Liu, Senior Member, IEEE, and Lei Yu, Student Member, ieee,” Toward Integrating Feature Selectio Algorithms for Classification and Clustering”, ieee transactions on knowledge and data engineering, vol. 17, no. 4, April 2005

[8] C. Wei, C.S. Yang, H.W. Hsiao, T.H. Cheng, Combining preference- and content-based approaches for improving document clustering effectiveness, Information Processing & Management 42 (2) (2006) 350–372.

[9] Renchu Guan, Xiaohu Shi, Maurizio Marchese, Chen Yang, and Yanchun Liang, “Text Clustering with Seeds Affinity Propagation” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 4, APRIL 2011

[10] Y.J. Li, C. Luo, and S.M. Chung, “Text Clustering with Feature Selection by Using Statistical Data,” IEEE Trans. Knowledge and Data Eng., vol. 20, no. 5, pp. 641-652, May 2008.

[11] B.J. Frey and D. Dueck, “Non-Metric Affinity Propagation for Un- Supervised Image Categorization,” Proc. 11th IEEE Int’l Conf. Computer Vision (ICCV ’07), pp. 1-8, Oct. 2007.

[12] L.P. Jing, M.K. Ng, and J.Z. Huang, "An Entropy Weighting KMeans Algorithm for Subspace Clustering of High-Dimensional Sparse Data," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 8, pp. 1026-1041, Aug. 2007.

[13] Z.H. Zhou and M. Li, "Distributional Features for Text Categorization," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 3, pp. 428-442, Mar. 2009.

[14] F. Pan, X. Zhang, and W. Wang, "Crd: Fast Co-Clustering on Large Data Sets Utilizing Sampling-Based Matrix Decomposition," Proc. ACM SIGMOD, 2008.

[15] Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee, "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 3, MARCH 2011

8. AOTHURS PROFILE

A. K. Santra received the P. G. degree and Doctorate degree from I.I.T., Kharagpur in the year 1975 and 1981 respectively. He has got 20 years of Teaching Experience and 19 years of Industrial (Research) Experience. His area of interest includes Artificial Intelligence, Neural Networks, Process Modeling, Optimization and Control. He has got to his credit (i) 35 Technical Research Papers which are published in National / International Journals and Seminars of repute, (ii) 20 Research Projects have been completed in varied application areas, (iii) 2 Copy Rights for Software Development have been obtained in the area of Artificial Neural Networks (ANN) and (iv) he is the contributor of the book entitled "Mathematics and its Applications in Industry and Business", Narosa Publishing House, New Delhi. He is the recognized Supervisor for guiding Ph. D. / M. S. (By Research) Scholars of Anna University-Chennai, Anna University- Coimbatore, Bharathiyar University, Coimbatore and Mother Teresa University, Kodaikanal. Currently he is guiding 12 Ph. D. Research Scholars in the Department. He is a Life member of CSI and a Life member of ISTE

.

C. Josephine Christy received her M.Sc., M.Phil., M.B.A., from Bharathiar University, Coimbatore. Currently she is working as Asst.Professor in Bannari Amman Institute of Technology, Sathyamangalam. Her area of interest includes Text Mining, Web Mining. She presented 2 papers in International Journal, 2 papers international conferences and 6 papers in national Conferences. She is a Life member of Computer Society of India and a Life member of Indian Society for Technical Education

.