An Improved Similarity Matching based Clustering Framework for Short and Sentence Level Text

M. John Basha[1], K.P. Kaliyamurthie[2]
[1]Department of CSE, P.T.R College of Engineering & Technology, Madurai, Tamil Nadu 625008, India
[2]Department of CSE, Bharath University, Chennai - 600073, Tamil Nadu, India

Article Info

Article history:

Received Aug 19, 2016
Revised Oct 21, 2016
Accepted Nov 5, 2016

Keyword:

Fuzzy c-means (FCM)
Mean square error (MSE)
Pairwise random swap (PRS)
Particle swarm optimization (PSO)
Peak signal noise ratio (PSNR)
Text clustering

ABSTRACT

Text clustering plays a key role in navigation and browsing process. For an efficient text clustering, the large amount of information is grouped into meaningful clusters. Multiple text clustering techniques do not address the issues such as, high time and space complexity, inability to understand the relational and contextual attributes of the word, less robustness, risks related to privacy exposure, etc. To address these issues, an efficient text based clustering framework is proposed. The Reuters dataset is chosen as the input dataset. Once the input dataset is preprocessed, the similarity between the words are computed using the cosine similarity. The similarities between the components are compared and the vector data is created. From the vector data the clustering particle is computed. To optimize the clustering results, mutation is applied to the vector data. The performance the proposed text based clustering framework is analyzed using the metrics such as Mean Square Error (MSE), Peak Signal Noise Ratio (PSNR) and Processing time. From the experimental results, it is found that, the proposed text based clustering framework produced optimal MSE, PSNR and processing time when compared to the existing Fuzzy C-Means (FCM) and Pairwise Random Swap (PRS) methods.

Corresponding Author:

M. John Basha,
AP & Head,
Department of CSE,
P.T.R College of Engineering & Technology, Madurai, Tamil Nadu 625008, India.
Email:

1. INTRODUCTION

Text clustering is the process of managing the large amount of digitally stored electronic data. The high volume of data is used for the data analysis, classification and retrieval techniques. In case of the prototype based clustering, the sequence of the prototypes are used for finding the best fit data with the unknown structures. To represent the clusters in k-means, only a single prototype is used. Multiple real applications use the prototype based clustering because, it provides less computational and memory space. Several other methods have been developed, which are based on stochastic global optimization such as simulated annealing and genetic algorithms. But these methods provide a high time complexity. Clustering algorithm and cluster validity are the commonly used correlated parts in the cluster analysis. Generally, to prevent the initialization problems, the k-means algorithm is executed many times with different parameters. The optimal solution is provided as the result. The quality of the clustering is computed using cost function. Thecategorization of the dataset depends on the cost function. The clustering methods are classified as density-based methods, graph based methods, grid based methods and methods for high dimensional space data. The major issues in the existing clustering algorithms include, high processing time consumption and high complexity.

Granados, et al [1] proposed a B Orjae technique to cluster the texts based on the string compression. It effectively computed the distortion present in the information of the text. When the proposed technique was applied for the structural datasets, the structure of the dataset was completely destroyed. *Lee, et al* [2] suggested a fuzzy based method to classify the text present in multi category document. A fuzzy relevance measure was used to convert the high dimensional document into a low dimensional document. The proposed clustering technique splitted the relevance space into multiple sub regions. The individual sub regions were then combined to create the individual category. The suggested clustering method provided optimal performance and speed than the other text clustering techniques. *Wei, et al* [3] proposed a lexical chain based wordnet. It used theontology hierarchical structure to determine the similarity between the terms of the words. The lexical chain was used toobtain the semantic relationship of the words present in the text. When compared to the classical methods, the proposed method increased the performance.

Peng, et al [4] proposed a novel CFu-tree based down-top incremental conceptual hierarchical text clustering approach for clustering the text in the document. The comparison variation (CV) criterion decided whether to merge or split the clusters. When compared to the existing K-Means algorithm, the proposed text clustering algorithm was efficient. *Yuan and Shi* [5] proposed a text clustering algorithm to prevent the issues in the division based clustering method. The complex features such as, synonym and co-occurring words were obtained from the multiple semantic information. Using the divide and conquer technique, the iteration ended with the expected cluster number. By dynamically updating the center number, optimal clustering results were obtained. *Bharthi, et al* [6] suggested a three-stage dimension reduction model to generate an informative feature subspace. The dimensions of the feature space were minimized. The total execution time for creating the cluster and creating the document cluster was significantly reduced. *Song, et al* [7] proposed a novel hybrid semantic similarity measure based fuzzy control Genetic Algorithm (GA) for clustering the documents. The Semantic Space Model (SSM) was used as the corpus-based method. The reduction in the dimensions of the SSM was used to obtain the true relationship between the documents. The thesaurus based method was combined with the SSM to provide the semantic similarity measure. When compared to the traditional GA, the proposed hybrid semantic strategy provided optimal performance. *Gong, et al* [8] proposed a validity index based method to address the issues of the adaptive feature selection for clustering the text stream. The threshold of the cluster valid index was used to reselect the features for creating a valid cluster. The quality of the proposed clustering algorithm was high. *Yao, et al* [9] proposed a k-means based Chinese text clustering algorithm to cluster the text. The average similarity parameter was used to obtain the similarity threshold value. Initially, the original cluster center that was above the threshold value was chosen as the candidate collection, then the cluster *Lin, et al* [10] proposed a novel similarity measure to compute the similarity between two documents. The proposed method considered the situations, such as, features in both the documents, features in only one document and features absent in both the documents. If both the documents had the features, the similarity between them was increased. If only one document had the features, then a fixed value was chosen as the similarity. If none of the documents had the features, the similarity value was found to be absent. When compared to the other measures, the proposed method produced optimal results.

Liu, et al [11] proposed a semantic tree based text clustering algorithm for clustering the parallel texts. The parallel algorithms were used to minimize the time complexity. It initiated the processes at the same time. The master process performed the data partitioning, information collection and clustering processes. The slave process calculated the word frequency. The proposed algorithm produced accurate results with less time complexity. *Li, et al* [12] suggested a Fuzzy Mahalanobis distances based text clustering algorithm to increase the precision and efficiency of the dataset. The proposed method was found to be more valid than the traditional fuzzy partitioning text clustering algorithms. *Nguyen, et al* [13] analyzed the quality issues of the clustering results. The extended Semantic Evaluation by Exploration (SEE) method was used to retrieve the INFONA documents. *Gao, et al* [14] proposed a genetic algorithm based text clustering. It integrated the latent semantic analysis. When compared to the single clustering method, the proposed clustering algorithm provided optimal clustering solutions. *Shi, et al* [15] proposed a patented text clustering algorithm named, Clustering by Genetic Algorithm Model (CGAM). The proposed model integrated the fitness function in the Genetic Algorithm (GA) and convergence criterion in the K-Means algorithm. When compared to the traditional GA and K-Means, the proposed algorithm obtained optimal clustering results. Summarization of documents based on the same topics play the major role in the quick understanding and creation of leagal judgements between the documents and topics. *Venkatesh et al* [16] utilized the hierarchical Latent Dirichlet Allocation (hLDA) using similarity measure between topics and documents and to find the summarization of each document using the same topics. The processing overhead is high due to the more number of pose taggers, processing tools and diverse choices of natural language processing scenarios in clustering algrotihms. *Bano et al* [17] created the large scale corpus with the annotation of disease names that train the probabilistic neural network model. They employed the context rank based hierarchical clustering method and optimal rule filtering algorithm to remove the unwanted special characters in the datasets.

The sentence clustering is used in multiple applications such as, classification and categorization of the documents, automatic summary generation, organizing the documents, etc. In text processing, the sentence clustering is used for the text mining process. The size of the cluster is unique for each cluster. The existing sentence clustering algorithms create multiple issues, such as, complexity, sensitivity, instability, etc. Compared to the sentence clustering, the clustering of the short texts are very difficult. As the short texts in the commercial products, new, FAQs and scientific abstracts are widely used by the users in real life, the clustering of the short texts demands focus. In this paper, the proposed text based clustering framework clusters the sentences as well as the short texts.

The proposed algorithm is executed till the duplicate clusters are removed. After the removal of the unwanted words, the proposed system checks all the words in the document for finding the exact word. The similarities between the sentences are used to find the ratio of the similarity of the words. Document clustering is an automatic analytic process that assigns documents to unknown categories. In this task, only the inherent structure of data is considered; therefore, it is more difficult than supervised text categorization because no information about correctly categorized examples is provided in advance. To overcome this difficulty, in this paper the CLUDIPSO based clustering is proposed. The key advantage of CLUDIPSO is the creation of real number vectors for each particle. The vectors represent the search space defined by the variables corresponding to the problem to solve.

The remainder of the paper is systematized as follows: Section II describes the existing text clustering techniques. Section III illustrates the proposed text based clustering framework and section IV describes the performance results of the proposed technique. Section V illustrates the conclusion of this paper.

2. RESEARCH METHOD

It is composed of following processes to achieve the reduction in processing time and MSE.

a. Preprocessing
b. Similarity Computation
c. Vector Data Formation
d. Clustering particle
e. Mutation

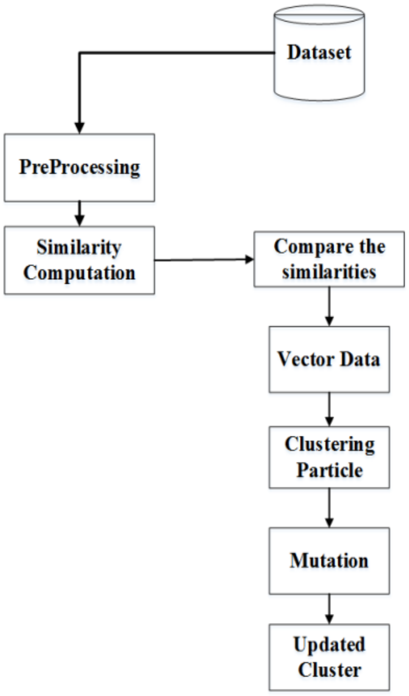

Figure 1. Overall Flow of the Proposed Text Based Clustering Framework

Figure 1 shows the workflow of the proposed text-based clustering framework that includes the sequential processes. Initially, the dataset is passed to the preprocessing block where the data warehouse entries and the table declaration are discussed in detail. Then, the cosine similarity between the topics and documents in the dataset is computed. Based on the similarity values, the vectors that represent the data are computed based on the ranking values. Then, the Particle Swarm Optimization (PSO) algorithm is used as the clustering tool to find the optimality.

2.1. Preprocessing

Each value of the input dataset is preprocessed and the resultant dataset is stored back in the database. The database load utility reads the user-provided data and stores them in the table. The input Reuter's dataset contains the user provided data. The database load utility supports four different formats of files.Before the data loading, the table must be defined. The data warehouse stores the private data and also makes the edge decision support system. The key aim of the data warehouse is to collect multiple information from various sources that follows different platform. The collected variable data are united for performing the business decisions.

2.2. Similarity Computation

Based on the input dataset, the similarity is calculated. Further deviation in the input dataset are also considered to refine the data in the files. In this paper, the vector similarity is accomplished using the cosine similarity.Cosine similarity is used to estimate the similarity between the vecors of an inner product space and measures the cosine of the angle between them.Cosine similarity is commonly used for the positive space whose outcome always lies between [0,1]. Cosine similarity is most suited for the high-dimensional positive spaces.Its merits are used in the field of data mining for measuring the cohesion between the clusters.The technique is also used to measure cohesion within clusters in the field of data mining.The Cosine Similarity of two vectors d1 and d2 is calculated as depicted in (1),

$$\cos(d1, d2) = dot\ (d1, d2)/\ \|d1\|\|d2\| \tag{1}$$

where,

$$dot\ (d1, d2) = d1[0] * d2[0] + d1[1] * d2[1]$$
$$\|d1\| = sqrt(d1[0]^2 +\ d1[1]^2$$

2.3. Vector Data Formation

The term vector is an algebraic model for representing text documents as vectors of identifiers. It is used in information filtering, information retrieval, indexing and relevancy rankings. Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero otherwise zero.

2.4. Clustering Particle

The clustering of the vectors is performed using the PSO algorithm. When the size and dimensionality of the dataset is large, the traditional PSO is not a best option, hence in this paper, a new version of the PSO named, CLUDIPSO is proposed. Three specific characteristics of the CLUIDIPSO makes it suitable for handling the larger datasets. The characteristics include, new representation of particles for reducing the dimensionality, reduce the computational time and increasing the speed of the silhouette computation. Based on the similarity distance threshold value, optimal clusters are generated.

2.5. Mutation

The mutation process is used to update the particle's position. The traditional PSO is used for solving only the continuous problems, but the proposed mutation process is not dependent on the position of the particles and further at each iteration, the position updating process is carried out in all the dimensions. To compute the dimension at which the particle is updated the following steps are performed.

Steps involved in the proposed CLUDIPSO based mutation	
Step 1:	All the dimensions of the velocity vector are normalized between [0,1] range.
Step 2:	Based on [18] the random number is calculated
Step 3:	All the dimensions that are above 'r' are chosen in the position vector and updated.
Step 4:	Updated Cluster is provided as the result.

3. RESULTS AND ANALYSIS

The performance of the proposed text based clustering algorithm is compared with the existing Fuzzy C-Means (FCM) and Pairwise Random Swap (PRS) clustering techniques for the metrics, such as,

a. Mean Square Error (MSE)
b. Processing Time
c. Peak Signal Noise Ratio

3.1. Mean Square Error (MSE)

The MSE is calculated using the equation (2),

$$f = \frac{1}{D}\sum_{j=1}^{D}\sum_{l=1}^{E}\left\|a_j - c_l\right\| \text{ ^}2\ K \quad (2)$$

Where, K indicates the indicator function. The D denotes the number of objects and the E denotes the number of clusters. Each object belongs to the cluster with the minimum Euclidean distance to the center centroid.

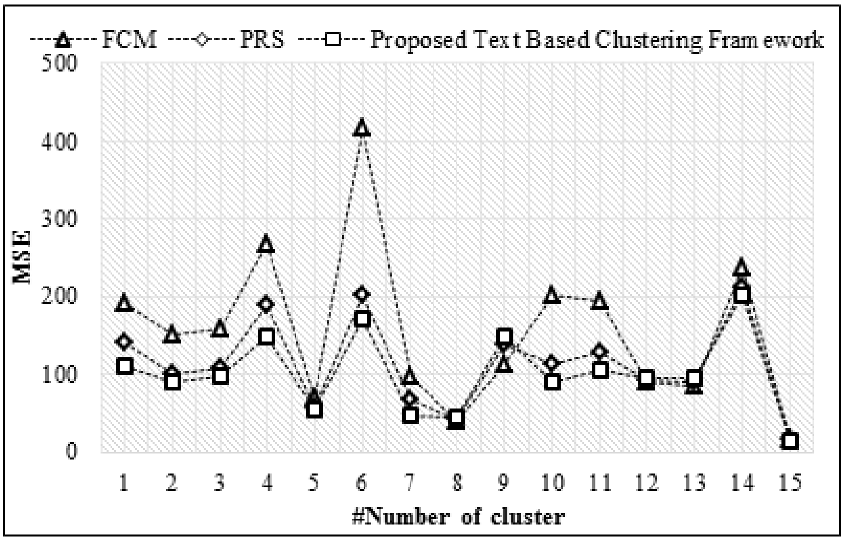

Figure 2. Comparison of the MSE for the proposed framework with the FCM and PRS

Figure 2 shows the comparative analysis of proposed text-based clustering with the existing FCM and PRS techniques regarding the MSE values. The effectiveness of any protocol proposed is determined with the minimum MSE values. The existing FCM and PRS provides the MSE values of 191 and 140 for single cluster. They provide 17 and 16 for 15 clusters. But, the optimal clustering-based similarity measurement in proposed text-based clustering reduces the values to 110 and 15 for single and 15 clusters respectively. The comparative analysis between the proposed TBC with the existing PSR (which provides minimum values) stated that the proposed TBC achieved the 21.42 and 6.67 % reduction in MSE compared to PSR for minimum and maximum clusters respectively.

3.2. Processing Time

The processing time of the proposed framework and the existing FCM and PRS is shown in Figure 3. From the figure it’s obvious that the proposed method produced optimal PSNR than the existing clustering techniques.

Figure 3 shows the comparative analysis of proposed text-based clustering with the existing FCM and PRS techniques regarding the processing time values. The effectiveness of any protocol proposed is determined with the minimum processing time. The processing time of the existing FCM and PRS are 73 and 17 secs for single cluster. They provide 36 and 34 secs for 15 clusters. But, the optimal clustering-based similarity measurement in proposed text-based clustering reduces the values to 10 and 24 secs for single and 15 clusters respectively. The comparative analysis between the proposed TBC with the existing PSR (which provides minimum values) stated that the proposed TBC achieved the 41.17 and 29.41 % reduction in processing time compared to PSR for minimum and maximum clusters respectively.

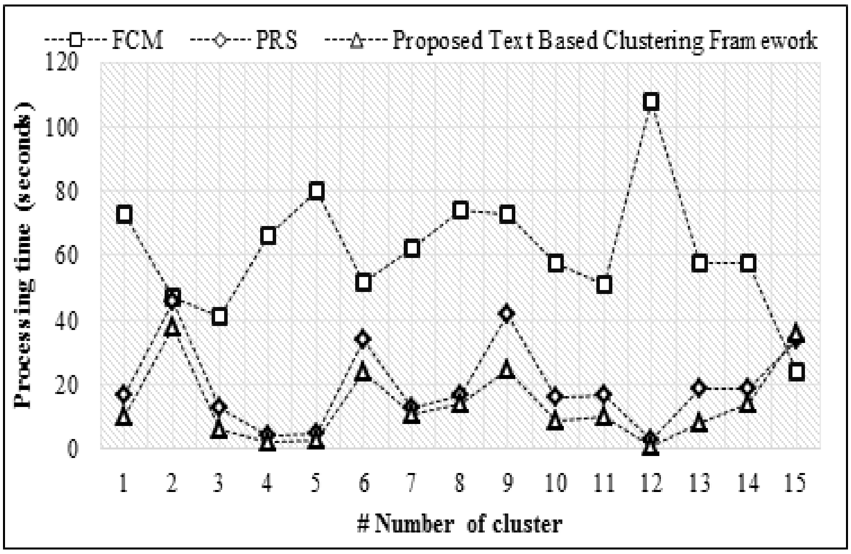

Figure 3. PSNR Comparison For the Proposed Method and Existing FCM And PRS Method

3.3. Peak Signal Noise Ratio (PSNR)

The PSNR is the ratioof the maximum possible value of the signal and the power of distorting noise that affects the quality of the representation it is calculated by the following equation,

$$PSNR = 20\log_{10}(\frac{MAX_F}{\sqrt{MSE}}) \quad (3)$$

Where,MAX_f denotes the maximum signal value that exists in the original data. Figure 4 shows the PSNR comparison for the proposed method and the existing FCM and PRS method.

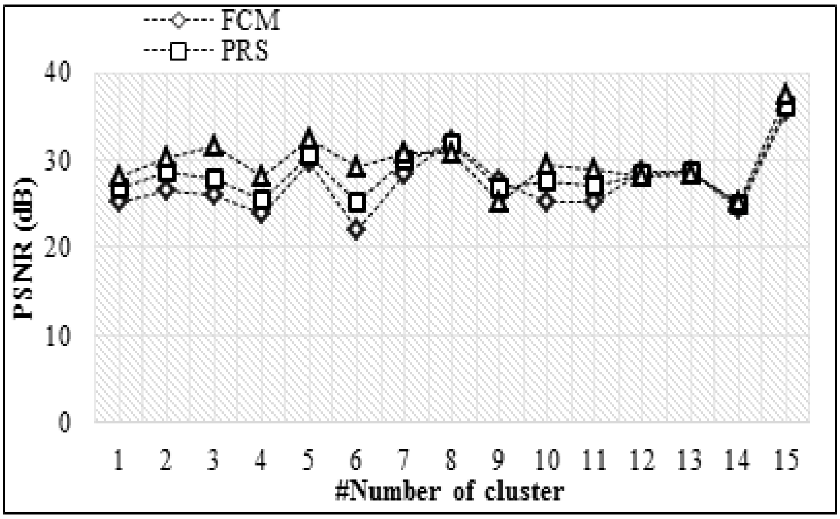

Figure 4. PSNR Comparison for the Proposed Method and Existing FCM and PRS Method

Figure 4 shows the comparative analysis of proposed text-based clustering with the existing FCM and PRS techniques regarding the PSNR values. The effectiveness of any protocol proposed is determined with the higher PSNR. The PSNR values for FCM and PRS are 25.3 and 26.7 dB for single cluster. They provide 35.7 and 36.1 dB for 15 clusters. But, the optimal clustering-based similarity measurement in proposed text-based clustering increases the values to 28.1 and 37.5 dB for single and 15 clusters respectively. The comparative analysis between the proposed TBC with the existing PSR (which provides maximum values) stated that the proposed TBC achieved the 4.98 and 3.73 % improvement in PSNR values compared to PSR for minimum and maximum clusters respectively.

4. CONCLUSION

Text clustering is the process of grouping the large amount of information into meaningful clusters. Existing FCM and PRS clustering techniques are used for clustering the texts in the document. But, these methods do not produce an optimal processing time, peak signal noise ratio and mean square error values, hence in this paper an efficient text based clustering framework is proposed to cluster the text documents that contains both the sentencesand short texts. Initially, the dataset is preprocessed to remove the noise, then the similarity between the words is calculated using the cosine similarity. Based on the computedsimilarity, the vector data is generated. The vector data is then clustered using the CLUIDIPSO technique. To optimize the clusters, mutation proess is deployed. The mutation process is repeated till an optimal cluster is obtained. The performance of the proposed text based clustering framework is compared with the existing FCM and PRS clustering methods. When compared to the existing methods, the proposed method reduced the processing time and MSN values and increased the PSNR value. Thus our text based clustering framework is proved to be better than the existing clustering FCM and PRS methods.

REFERENCES

[1] A. Granados, K. Koroutchev, and F. de Borja Rodriguez, "Discovering Data Set Nature through Algorithmic Clustering Based on String Compression", *IEEE Transactions on Knowledge and Data Engineering,* vol. 27, pp. 699-711, 2015.

[2] S.J. Lee and J.Y. Jiang, "Multilabel Text Categorization Based on Fuzzy Relevance Clustering", *IEEE Transactions on Fuzzy Systems,* vol. 22, pp. 1457-1471, 2014.

[3] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering using WordNet and lexical chains", *Expert Systems with Applications,* vol. 42, pp. 2264-2275, 2015.

[4] T. Peng and L. Liu, "A novel incremental conceptual hierarchical text clustering method using CFu-tree", *Applied Soft Computing,* vol. 27, pp. 269-278, 2015.

[5] M. Yuan and Y. Shi, "Text Clustering Based on a Divide and Merge Strategy", *Procedia Computer Science,* vol. 55, pp. 825-832, 2015.

[6] K.K. Bharti and P.K. Singh, "A three-stage unsupervised dimension reduction method for text clustering", *Journal of Computational Science,* vol. 5, pp. 156-169, 2014.

[7] W. Song, J.Z. Liang, and S.C. Park, "Fuzzy control GA with a novel hybrid semantic similarity strategy for text clustering", *Information Sciences,* vol. 273, pp. 156-170, 2014.

[8] L. Gong, J. Zeng, and S. Zhang, "Text stream clustering algorithm based on adaptive feature selection", *Expert Systems with Applications,* vol. 38, pp. 1393-1399, 2011.

[9] M. Yao, D. Pi, and X. Cong, "Chinese text clustering algorithm based k-means", *Physics Procedia,* vol. 33, pp. 301-307, 2012.

[10] Y.S. Lin, J.Y. Jiang, and S.J. Lee, "A similarity measure for text classification and clustering", *IEEE Transactions on Knowledge and Data Engineering,* vol. 26, pp. 1575-1590, 2014.

[11] G. Liu, Y. Wang, T. Zhao, and D. Li, "Research on the parallel text clustering algorithm based on the semantic tree", in *6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT)*, 2011, pp. 400-403.

[12] C. Li, Y. Tan, and J. Kong, "An Mahalanobis distances based text clustering algorithm", *International Conference on Automatic Control and Artificial Intelligence (ACAI 2012),* pp. 465 - 468, 2012.

[13] S.H. Nguyen, W. Swieboda, and H.S. Nguyen, "On semantic evaluation of text clustering algorithms", in *IEEE International Conference on Granular Computing (GrC),* 2014, pp. 224-229.

[14] M.T. Gao and B.J. Wang, "Text clustering ensemble based on genetic algorithms", in *International Conference on Systems and Informatics (ICSAI),* 2012, pp. 2329-2332.

[15] K. Shi and L. Li, "High performance genetic algorithm based text clustering using parts of speech and outlier elimination", *Applied Intelligence,* vol. 38, pp. 511-519, 2013.

[16] R.K. Venkatesh and N.I.E.M. India, "Legal Documents Clustering and Summarization using Hierarchical Latent Dirichlet Allocation", *IAES International Journal of Artificial Intelligence (IJ-AI),* vol. 2, pp. 27-35, 2013.

[17] S. Bano, K.L. University, K.R. Rao, and E. Sri Prakash College of, "Partial Context Similarity of Gene/Proteins in Leukemia Using Context Rank Based Hierarchical Clustering Algorithm", *International Journal of Electrical and Computer Engineering (IJECE),* vol. 5, pp. 483-490, 2015.

[18] X. Hu, R.C. Eberhart, and Y. Shi, "Swarm intelligence for permutation optimization: a case study of n-queens problem", in *Proceedings of the 2003 IEEE Swarm Intelligence Symposium, 2003. SIS'03.*, 2003, pp. 243-246.