ISSN:0975-9646

Fuzzy Relational Spectral Clustering Method for Document Clustering

R.Nagaraj[1], Dr.V.Thiagarasu[2]

[1]*Research Scholar,Department of Computer Science,Karpagam University, Coimbatore, TamilNadu, India*

[2]*Associate Professor in Computer Science, Gobi Arts and Science College, Gobichettipalayam, TamilNadu, India*

Abstract **Correlation Preserving Indexing is a spectral clustering method which discovers intrinsic structures embedded in high-dimensional document space. But the problem is to predict the result of one variable based on another variable is not suitable for all the situations. So, the directed Ridge regression is used which computes the relationship among the variables based on the Eigen values to identify the similarity between the documents. But in these two methods the similarity is identified by taking terms. So, there is high computation and less clustering efficiency. Further to improve the cluster efficiency, in this manuscript an innovative technique is introduced which is called Sentence level document clustering in fuzzy relational spectral clustering (SCFSC). The spectral fuzzy can better handle clusters with a complex, nonlinear geometric structure and it does not need prior information on the number of clusters. In this method the similarity between the sentences are measured by using the standard similarity measure. By using the fuzzy relational spectral clustering the efficient clustering is achieved. An experimental results show that the proposed system achieves high clustering efficiency and less computation.**

Keywords: **Document Clustering, correlation measure, Directed Ridge Regression, Fuzzy spectral clustering**

I. INTRODUCTION

The main intent of the document clustering is to automatically group related documents into clusters. It is important tasks in machine learning and artificial Intelligence [1] [2] [3]. K-Means method [4] is one of the clustering methods which use the Euclidean distance to measure the similarity. Latent semantic indexing (LSI) [5] is one of the effectual spectral clustering methods which intend to identify the best subspace approximation to the original document space by decreasing the global construction error. Locality preserving indexing (LPI) [6] method is a different spectral clustering method which is based on graph partitioning theory. It applies a weighted function to each pair wise distance attempting to focus on capturing the similarity structure.

A document clustering method is called correlation preserving indexing is used which particularly considers the manifold structure embedded in the similarities between the documents [7]. The intent of correlation preserving indexing is to identify an optimal semantic subspace by concurrently maximizing the correlations between the documents in the local patches and minimizing the correlations between the documents outside these patches. But the drawback in this method is to predict the result of one variable based on another variable is not applicable for all the situations since two variable prediction problems takes place [8]. So, the directed ridge regression method is used in order to bring the relationships among the variables. This method achieves the similarity between the documents by measuring the relationship among the variables. These two methods achieve the similarity by taking terms. So, there is high computation complexity and less clustering efficiency.

In order to improve the cluster efficiency, a Sentence level document clustering in fuzzy relational spectral clustering (SCFSC) is introduced. In this method firstly to find the similarity between the sentences by using the Jiang and Conrath measure. The sentence similarity measure relies on a word-to-word semantic similarity measure. Then construct the similarity matrix for the data sets. Form the Laplacian matrix and compute the Eigen values and the Eigen vectors of the Laplacian matrix. Map each point to a lower-dimensional representation based on one or more eigenvectors. Then initialize the membership values and calculate the center vectors. At every iteration the membership value is updated according to the similarity. Finally, assign points to two or more classes based on the representation.

II. RELATED WORK

A Latent Semantic Analysis was suggested for automatic indexing and retrieval. The particular "latent semantic indexing" (LSI) analysis tried to utilize singular-value decomposition [9]. A large matrix of term document association data is taken and create a "semantic" space where in terms and documents that are directly associated are placed near one another. Singular-value decomposition permits the arrangement of the space to replicate the main associative patterns in the data, and avoid the smaller, less important influences. But the limitation in this method is not effective for the document clustering.

A spherical K-means method was proposed to cluster the documents in high-dimensional document space [10]. In this work, concept decompositions are introduced to estimate the matrix of document vectors. These decompositions are attained by taking the least-squares approximation onto the linear subspace covered by all the concept vectors. Then the concept vectors are localized in the word space, are sparse, and inclined towards orthonormality. The clustering is accomplished by measuring similarity between the words in the entire set of documents.

The comparative study of generative models is proposed for document clustering. In the similarity based method, the average similarities is maximized within clusters and the average similarities is minimized between the clusters. The Model-based approaches attempt to learn generative models from the documents, in which every model representing one particular document group [11]. A united framework for probabilistic model-based clustering, which permits one to understand and compare a vast range of model-based partitional clustering methods using a common viewpoint that centers around two steps. The first step is called re-estimation step and the second step is called data re-assignment step. This two-step view facilitates one to effortlessly combine different models with different assignment strategies.

A model-based clustering is used with balance-constrained method to provide accuracy in the clustering methods. In model based clustering, a unifying bipartite graph view is presented [12]. Then a two-step iterative maximum-likelihood optimization process is presented and examined for hard, model-based clustering. A complete balanced sample assignment sub-problem is formulated to solve by using a greedy heuristic at each iteration of the process. A balance-constrained method is used in the sample assignment step instead of a maximum-likelihood assignment. A proficient iterative bi-partitioning heuristic is developed to minimize the computational complexity of this step and make the balanced sample assignment algorithm scalable to large datasets.

A kernel and spectral approaches are suggested for effectual clustering [13]. These two approaches are able to produce nonlinear separating hyper surfaces between clusters. The main idea of both approaches lies in their ability to construct an adjacency structure between data avoiding to handle with a prefixed shape of clusters. These approaches have a slight similarity with hierarchical methods in the use of an adjacency structure with the main difference in the philosophy of the grouping procedure. The limitation of this method is the clustering efficiency is less.

The spectral clustering method is suggested which utilizes the top Eigen vectors of a matrix which can be derived from the distance between the points [14]. One line of analysis makes the link to spectral graph partitioning in which the second Eigen vector of a graph's Laplacian is utilized to define a semi-optimal cut. The Eigen vector is seen as a solving a relaxation of an NP-hard discrete graph partitioning problem and it can be shown that cuts based on the second Eigen vector give a guaranteed approximation to the optimal cut. By building a weighted graph, this can be extended to clustering in that the nodes corresponds to data points and edges are associated to the distance between the points.

K-Means or clustering is essentially a partitioning method applied to examine data and treats annotations of the data as objects based on locations and distance between various input data points [15]. By using this method, partitioning the objects into equally exclusive clusters is done by it in such a fashion that objects within each cluster remain as close as possible to each other but as far as possible from objects in other clusters. Each cluster is represented by its centre point i.e. centroid. In most of the times the distances used in clustering do not actually represent the spatial distances. The K-Means clustering algorithm finds the desired number of distinct clusters and their centroids. A centroid is the point whose co-ordinates are attained by means of computing the average of each of the co-ordinates of the points of samples assigned to the clusters. The limitation in this project is a labeled dataset as training data and practically classification of labeled data is generally very difficult as well as expensive.

The Sentence similarity measures are an important technique in Webpage interval for enhancing retrieval effectiveness, where titles are used to symbolize documents in the named page discovering task [16]. The key idea of Sentence similarity measure is to compute the similarity between very short texts, primarily of sentence length. It presents an algorithm that takes account of semantic information and word order information implied in the sentences. By using information from a structured lexical database and from corpus statistics, the semantic similarity of two sentences is calculated. The use of a lexical database enables our method to model human common sense knowledge and the incorporation of corpus statistics allows our method to be adaptable to different domains. This method can be used in a various types of applications that include text knowledge representation and discovery. But the problem is the disambiguate word sense is not considered. So, that the clustering efficiency is less.

III. DOCUMENT CLUSTERING BASED ON CPI

Correlation Preserving Indexing is one of the techniques for document clustering which particularly considers the manifold structure embedded in the similarities between the documents. The main objective is to discover a best semantic subspace by concurrently maximizing the associations between the documents in the local patches and minimizing the associations between the documents outside these patches. This method is dissimilar from LSI and LPI that are based on a dissimilarity measure and are focused on detecting the intrinsic structure between widely separated documents rather than on detecting the intrinsic structure between nearby documents. Based on CPI method the similarity-measure generally focuses on identifying the intrinsic structures between nearby documents rather than on detecting the intrinsic structure between widely separated documents. Correlation Preserving Indexing can proficiently detect the intrinsic semantic structure of the high-dimensional document space.

The correlation between the two vectors *u* and *v* is as followed as,

$$Corr(u,v) = \frac{u^T v}{\sqrt{u^T u}\sqrt{v^T v}} = < \frac{u}{\|u\|}, \frac{v}{\|v\|} \quad (1) > \ (1)$$

The correlation corresponds to an angle θ such that Cos θ = Corr (u, v). The higher the value of Corr (u, v), the stronger the association between the two vectors u and v.

Online document clustering intend to group documents into clusters, in which the unsupervised learning is converted into semi-supervised learning by using the following information.

A1. If the two documents are close to each other then it grouped into the same cluster.

A2. If the two documents are far away from each other then it grouped into the different clusters.

$y_i \in Y$ is the low-dimensional representation of the i th document $x_i \in$ X in the semantic subspace, where i= 1,2,…n. Then the above assumption (A1) and (A2) can be followed as,

$$Max \sum_i \sum_{x_j \in N(x_i)} [Corr(y_i] y_j) \quad (2)$$

$$Min \sum_i \sum_{x_{j \in N(x_j)}} Corr(y_i, y_j) \quad (3)$$

respectively, where N (x_i) denotes the set of nearest neighbors of x_i. The optimization of (2) and (3) is equivalent to the following metric learning:

$$d(x, y) = \alpha * cos(x, y) \quad (4)$$

Where d(x, y) indicates the similarity between the documents x and y, α corresponds to whether x and y are the nearest neighbors of each other.

Clustering Algorithm Based on CPI

For a set of documents $x_1, x_2, \ldots . x_n \in IR^n$. Let X denotes the document matrix. For document clustering based on CPI, the algorithm can be summarized as follows:

1. Construct the local neighbor patch, and compute the matrices M_S and M_T

The matrices M_S and M_T are defined as,

$$M_T = \sum_i \sum_j [(x_i] x_j^T), \quad (5)$$

$$M_S = \sum_i \sum_{x_j \in N(x_i)} [(x_i] x_j^T) \quad (6)$$

It is simple to confirm that the matrix M_T is semi positive definite. While the documents are anticipated in the low dimensional semantic subspace where the correlations between the document points among the nearest neighbors are preserved.

2. Allocate the document vectors into the SVD subspace by throwing away the zero singular values. The singular value decomposition of X can be written as $X = U \Sigma V^T$. The all zero singular values in Σ have been eliminated. Therefore, the vectors in U and V that correspond to these zero singular values have been eliminated as well. Therefore the document vectors in the SVD subspace can be attained by,

$$\hat{X} = U^T X \quad (7)$$

3. The CPI projection is computed. Based on the multipliers $\lambda_0, \lambda_1, \ldots, \lambda_n$ is obtained. One can compute the matrix $M = \lambda_0^* M_T + \lambda_1^* x_1 x_1^T + \cdots \lambda_n^* x_n x_n^T$. Let W_{CPI} be the solution of the generalized Eigen value problem $M_S W = \lambda M W$. Then, the low dimensional representation of the document can be computed by,

$$Y = W_{CPI}^T \hat{X} = W^T X \quad (8)$$

$W = UW_{CPI}$ denotes the transformation matrix.

4. Cluster the documents in the CPI semantic subspace. Since the documents were anticipated on the unit hyper sphere, the inner product is a natural measure of similarity. A partitioning $\{\Pi_j\}_{j=1}^k$ of the document can be searched using the maximization of the following objection function:

$$\{Q(\{\pi_{(j)}\}_{j=1}^k) = \sum_{j=1}^k \sum_{x \in \pi_j} x^T c_j \quad (8)$$

Directed Ridge regression

In the directed ridge regression is one of the effective methods for the document clustering. It gives the relationship among the several variables. The value of regression analysis as a numerical tool may be extensively diminished when the set of independent variables are approximately collinear. Usually in the directed ridge regression method, to identify how the typical value of the dependent variable changes when any one of the independent variables is different, while the other independent variables are held fixed. The directed ridge estimator based on the relationship between the Eigen values U and variance $\widehat{\alpha_i}$. The computation of directed ridge estimator,

$$\hat{\alpha} (dk)^{(0)} = (\Lambda + KI)^{-1} \hat{X} \, U$$

$$W = \hat{\alpha} (dk) * \hat{X} \quad (9)$$

Where K is the diagonal matrix.

IV. DOCUMENT CLUSTERING BASED ON SCFSC

SCFSC is used to cluster the documents at the sentence level. Sentence clustering intends at grouping sentences with similar meanings into clusters. Generally, vector similarity measures, such as cosine, are used to define the level of similarity over bag-of-words encoding of the sentences. The Spectral clustering method uses eigenvectors of matrices constructed using measures of similarity between the data points. Fuzzy clustering algorithms assign, for each observation in the data set, degrees of membership to the different clusters that provide information about the uncertainty of the clustering assignments. That is, when an observation belongs to a cluster, it tends to have a high degree of membership to that cluster and low degrees of membership to the remaining clusters. The two sentences being compared and represented in a reduced vector space of dimension n. n is denoted as the number of distinct nonstop words appearing in the two sentences. Semantic vectors, V_1 and V_2 represents the sentences S_1 and S_2 in this reduced vector space are first constructed. The elements of V_i are determined as follows: Let v_{ij} be the jth element of V_i and let w_j be the word corresponding to dimension j in the reduced vector space. Consider the two cases, depending on whether w_j appears in S_i :

Case 1: If w_j appears in S_i , set v_{ij} equal to 1.

Case 2: If w_j does not appear in S_i, compute a word-to-word semantic similarity score between w_j and each nonstopword in S_i and set v_{ij} to the highest of the similarity scores, i.e., $v_{ij} \arg = \max_{x \epsilon \{s_i\}} sim(w_j, x)$. Once V_1 and V_2 have been determined, the semantic similarity between S_1 and S_2 can be defined using a standard measure of similarity. The sentence similarity measure relies on a word-to-word semantic similarity measure. The Jiang and Conrath measure is based on the idea that the degree to which two words are similar is proportional to the amount of information they share. The similarity between words w_1 and w_2 is defined as:

$$Sim(w_1, w_2) = \frac{1}{IC(w_1) + IC(w_2) - 2 \times IC(LCS(w_1, w_2))}$$

Where LCS $(w_1 1, w_1 2)$ is the word that is the deepest common ancestor of w_1 and w_2, IC (w) is the information content of word w, and defined as IC(w) = -log P(w), where P(w) is the probability that word w appears in a large corpus.

Algorithm:

Input: Given a set of documents $x_1, x_2, \ldots x_n \epsilon R$.

Given a set of data points require to partition into k clusters X= $\{X_1 1, X_1 2, \ldots X_1 n\}$

1. Find the similarity values by using the $s_{ij}\ where\ i = 1,2 \ldots N; j = 1,2, \ldots N$ where s_{ij} is the similarity between the sentences i and j.
2. Form the similarity matrix W is defined as: $W_{ij} = \exp\left(-\frac{1}{2\sigma^2} d^2(x_i, x_j)\right)$
3. To construct the Laplacian matrix: $L_{NCut} = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}$
4. Find the k first eigenvectors of L (chosen to be orthogonal to each other in the case of repeated eigenvalues), and form the matrix U by stacking the eigenvectors in columns: U= $[[u]_1, \ldots u_k] \epsilon R^{nxk}$
5. Form the matrix Y from U by normalizing each of U's rows to have unit length: $Y_{ij} = \frac{U_{ij}}{\left[\sum_{j=1}^{n} [U_{ij}^2]\right]^{\frac{1}{2}}}$
6. Treat each row of Y as a point in R^k and classify them into k classes through the fuzzy relational algorithm.
7. Initialize membership value U = $[u_1 ij]\ matrix, U^{\dagger}((0))$
8. At K-step: calculate the centers vectors $C^{(k)} = [c_j] with\ U^{(k)}$ $c_j = \frac{\sum_{i=1}^{N} u_{ij}^m . x_i}{\sum_{i=1}^{N} u_{ij}^m}$
9. Update $U^{(k)}, U^{(k+1)}$ $u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}$
10. If $\|U^{\dagger}((k+1)) - U^{\dagger}((k))\| < \epsilon$ then STOP: otherwise return to step 2.
11. Assign the original points xi to cluster j if and only if row i of the matrix Y was assigned to cluster j.

V. EXPERIMENTAL RESULTS

SCFSC is compared Correlation preserving indexing (CPI) and Directed ridge regression (DRR). The Reuters- data set is the most widely used data set for text categorization research and it contains over 21 thousand documents from over 600 categories, even though most categories contain few documents. Almost half the documents are labeled, and the remainder is unlabeled. Of the labeled documents, roughly 17 percent belong to more than one class. A subset is chosen that includes 1,833 documents, each labeled as belonging to one of 10 different classes. The number of documents in each of the 10 classes is, correspondingly, 355, 334, 259, 211, 156, 135, 114, 99, 97, and 73. Similarity values were calculated using cosine similarity. The results are compared in terms of precision, Recall, F-measure and accuracy.

Precision

Precision value is computed is based on the retrieval of information at true positive prediction, false positive.

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

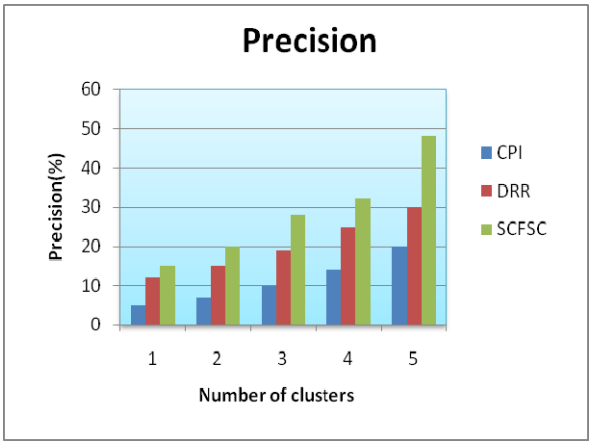

Fig 1. Precision

The Figure 1 shows that when compared to CPI and DRR the precision is improved in SCFSC.

Recall

Recall value is calculated is based on the retrieval of information at true positive prediction, false negative.

$$Recall = \frac{True\ Positive}{(True\ positive + False\ negative)}$$

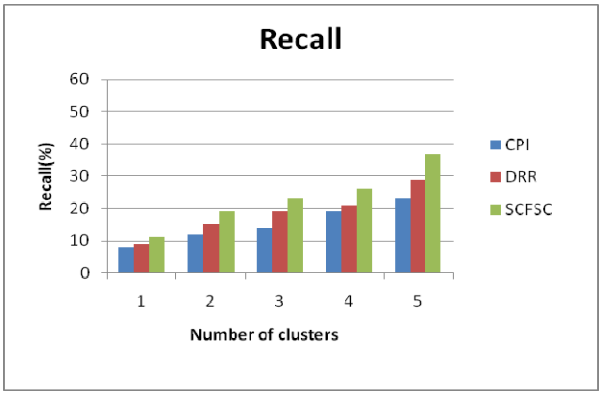

Fig 2. Recall

The corresponding results of the CPI, DRR and (SCFSC) are measured for Recall. Figure 2 shows that when compared to CPI and DRR the recall is improved in SCFSC.

F-Measure

F-measure is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct results divided by the number of all returned results and r is the number of correct results divided by the number of results that should have been returned. The F -Measure score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0.

$$F - measure = \frac{2.Precision.recall}{(precision + recall)}$$

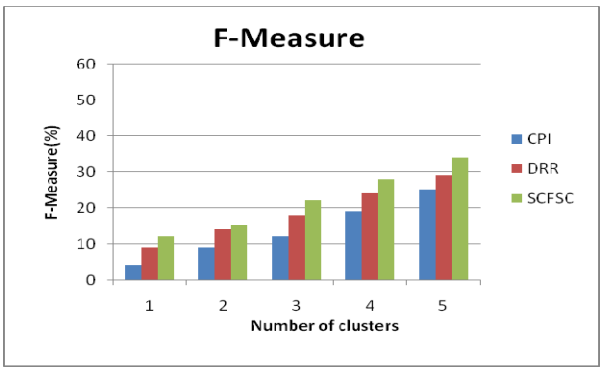

Fig 3. F-Measure

The corresponding results of CPI, DRR and SCFSC are measured for F-Measure. Figure 3 clearly shows that when compared to CPI and DRR the F-Measure is improved in SCFSC.

Accuracy

Accuracy is computed as,

$$Accuracy = \frac{(True\ positive + True\ negative)}{(True\ positive + True\ negative + @False\ positive + False\ negative)}$$

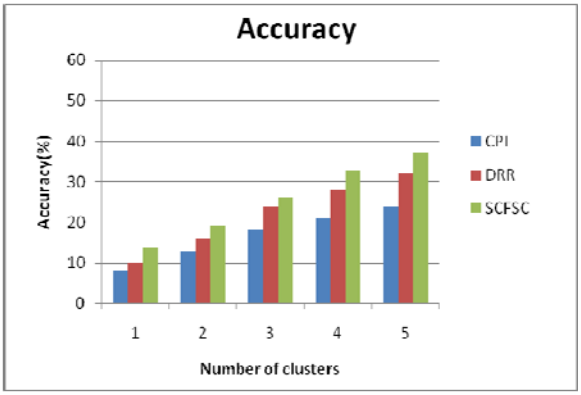

Fig 4. Accuracy

The corresponding results of the CPI, DRR and SCFSC are measured for Accuracy. Figure 4 clearly shows that when compared to the CPI and DRR, the accuracy is improved in SCFSC.

VI. CONCLUSION

Correlation preserving indexing is a method which computes the similarity based on correlation. In this method, concurrently exploits the correlation between the documents in the local patches and reduces the correlation between the documents in the outside patches. But due to the two variable problems, directed ridge regression is used in which the similarity between the documents is measured by computing the relationship among the variables based on the Eigen values. To increase the clustering efficiency, a Sentence level Document Clustering based on fuzzy relational model is used. Sentence clustering it is significant to cluster the sentence is probable to be associated to more than one theme or topic present within a document or set of documents. The Fuzzy based spectral Clustering provides the efficient document Clustering and reduces the computation time.

This can be further extended to develop a probabilistic based fuzzy relational clustering algorithm. This technique directs to a fuzzy partition of the fuzzy rules, for each cluster, which corresponds to a new set of fuzzy sub-systems.

REFERENCES

[1] R.T. Ng and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB), pp. 144-155, 1994.

[2] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.

[3] S. Kotsiantis and P. Pintelas, "Recent Advances in Clustering: A Brief Survey," WSEAS Trans. Information Science and Applications, vol. 1, no. 1, pp. 73-81, 2004.

[4] J.B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. Fifth Berkeley Symp. Math. Statistics and Probability, vol. 1, pp. 281-297, 1967.

[5] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman, "Indexing by Latent Semantic Analysis," J. Am. Soc. Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[6] D. Cai, X. He, and J. Han, "Document Clustering Using Locality Preserving Indexing," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 12, pp. 1624-1637, Dec. 2005.

[7] Taiping Zhang, Yuan Yan Tang, Bin Fang and Yong Xiang, "Document Clustering in Correlation Similarity Measure Space" proc. vol. 24, no. 6 ,pp .June 2012.

[8] M. El-Dereny and N. I. Rashwan, "Solving Multicollinearity Problem Using Ridge Regression Models," Int. J. Contemp. Math. Sciences, Vol. 6, 2011, no. 12, 585 – 600.
[9] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman, "Indexing by Latent Semantic Analysis," J. Am. Soc. Information Science, vol. 41, no. 6, pp. 391-407, 1990.
[10] I.S. Dhillon and D.M. Modha, "Concept Decompositions for Large Sparse Text Data Using Clustering," Machine Learning, vol. 42, no. 1, pp. 143-175, 2001.
[11] S. Zhong and J. Ghosh, "Generative Model-Based Document Clustering: A Comparative Study," Knowledge of Information System, vol. 8, no. 3, pp. 374-384, 2005.
[12] S. Zhong and J. Ghosh, "Scalable, Balanced Model-Based Clustering," Proc. Third SIAM Int'l Conf. Data Mining, pp. 71-82, 2003.
[13] Maurizio Filippone, Francesco Camastra, Francesco Masulli, Stefano Rovetta," A survey of kernel and spectral methods for clustering," Department of Computer and Information Science, University of Genova, and CNISM, Via Dodecaneso 35, I-16146 Genova.
[14] Andrew Y. Ng, Yair Weiss, Michael I. Jordan, "On Spectral Clustering: Analysis and an Algorithm,". In Proceedings of the 41st Annual symposium on Foundations of Computer Science 2002.
[15] Soumi Ghosh, Sanjay Kumar Dubey," Comparative Analysis of K-Means and Fuzzy C-Means Algorithms," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.
[16] Yuhua Li, David McLean, Zuhair A. Bandar," Sentence Similarity Based on Semantic Nets and Corpus Statistics," IEEE Transactions On Knowledge And Data Engineering, Vol. 18, No. 8, August 2006.