ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2012

Validation of Document Clustering based on Purity and Entropy measures

M.Deepa[1], P. Revathy[2]

PG Student[1], Assistant Professor[2]

Rajalakshmi Engineering College, Thandalam

ABSTRACT— ***Document clustering aims to automatically group related document into clusters. If two documents are close to each other in the original document space, they are grouped into the same cluster. If the two documents are far away from each other in the original document space, they tend to be grouped into different cluster. The classical clustering algorithms assign each data to exactly one cluster, but fuzzy c-means allow data belong to different clusters. Fuzzy clustering is a powerful unsupervised method for the analysis of data. Cluster validity measure is useful in estimating the optimal number of clusters. Purity and Entropy are the validity measures used in this clustering.***

Keywords— Document **clustering, fuzzy c-means, validation, purity and entropy measures.**

I. INTRODUCTION

Document clustering is one of the crucial techniques for organizing documents in an unsupervised manner. It groups all documents so that the documents in the same group are more similar than ones in other groups. Hypothesis of cluster makes relevant documents tend to be more closely related to each other than to non-relevant document. It is a fundamental operation used in unsupervised document organization, automatic topic extraction, and information retrieval. Clustering involves dividing a set of objects into a specified number of clusters. The motivation behind clustering a set of data is to find inherent structure in the data and expose this structure as a set of groups. The data objects within each group should exhibit a large degree of similarity while the similarity among different clusters should be minimized. There are two major clustering techniques: "Partitioning" and "Hierarchical". Most document clustering algorithms can be classified into these two groups. Hierarchical clustering techniques produce a nested sequence of partition, with a single, all-inclusive cluster at the top and single clusters of individual points at the bottom.

The partitioning clustering techniques seeks to partition a collection of documents into a set of non-overlapping groups, so as to maximize the evaluation value of clustering. Although the hierarchical clustering technique is often portrayed as a better quality clustering approach, this technique does not contain any provision for the reallocation of entities, which may have been poorly classified in the early stages of the text analysis. Moreover, the time complexity of this approach is quadratic.

It has been recognized that the partitional clustering technique is well suited for clustering a large document dataset due to their relatively low computational requirements. The time complexity of the partitioning technique is almost linear, which makes it widely used.

A. Fuzzy c-means Algorithm

Traditional clustering approaches generate partitions; in a partition, each pattern belongs to one and only one cluster. Fuzzy clustering extends this notion to associate each pattern with every cluster using a membership function. The output of such algorithms is a clustering, but not a partition some times. Fuzzy clustering is a widely applied method for obtaining fuzzy models from data. It has been applied successfully in various fields including geographical surveying, finance or marketing. This method is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c} (\mu_{ij})^m \left\|x_i - v_j\right\|^2$$

where m is any real number greater than 1, $||x_i - v_j||'$ is the Euclidean distance between i^{th} data and j^{th} cluster center. µij is the degree of membership of xi in the cluster j, xi is the ith of d-dimensional measured data, vj is the dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership µij and the cluster centers vj.

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, \ldots . c$$

B.Validation measures

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2012

The purpose of clustering is to determine the intrinsic grouping in a set of unlabeled data, where the objects in each group are indistinguishable under some criterion of similarity. Clustering is an unsupervised classification process fundamental to data mining (one of the most important tasks in data analysis). It has applications in several fields like bioinformatics, web data analysis, text mining and scientific data exploration. Clustering refers to unsupervised learning and, for that reason it has no a priori data set information. However, to get good results, the clustering algorithm depends on input parameters like the optimal number of clusters. Currently, cluster validity indexes research has drawn attention as a means to give a solution. Many different cluster validity methods have been proposed without a priori class information. Clustering validation is a technique to find a set of clusters that best fits natural partitions (number of clusters) without any class information. Generally speaking, there are two types of clustering techniques, which are based on external criteria and internal criteria. External validation: Based on previous knowledge about data. Internal validation: Based on the information intrinsic to the data alone. The above two types of cluster validation to determine the correct number of groups from a data set, one option is to use external validation indexes for which a priori knowledge of dataset information is required.

Another option is to use internal validity indexes which do not require a priori information from dataset. Internal validation includes Bic index, Silhouette index, Davies-Bouldin index (DB), Calinski-Harabasz index, Dunn index, NIVA index. External validation includes F-measure, NMI measure, Purity, Entropy. The Bayesian information criterion (BIC) is devised to avoid over fitting, and is defined as: $BIC = -\ln(L) + vln(n)$.The Dunn index measures the ratio between the smallest cluster distance and the largest intra-cluster in a partitioning. DB measures the average similarity between each cluster and the one that most resembles it. The SD index is defined based on the concepts of the average scattering for clustering and total separation among clusters. The index PS uses non metric distance based on the concept of point symmetry, and measures the total average symmetry with respect to the cluster center .Silhouette clustering structure quality; taking into account group compactness, separation between groups. The external measures include Entropy, Purity, NMI Measure and F-Measure. To compute purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by N. Formally:

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N}\sum_{k}\max_{j}|\omega_k \cap c_j|$$

where $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ is the set of clusters and $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$ is the set of classes. We interpret ω_k as the set of documents in ω_k and c_j as the set of documents. Bad clusterings have purity values close to 0, a perfect clustering has a purity of 1. H is entropy as defined as follows

$$H(\Omega) = -\sum_{k} P(\omega_k)\log P(\omega_k) = -\sum_{k}\frac{|\omega_k|}{N}\log\frac{|\omega_k|}{N}$$

If the purity is high and the entropy value is less, then its tend to be a good clustering. Otherwise it's a bad clustering. The rest of the paper is organized as follows. In section 2, we briefly discuss related work on document clustering. Our work is described in section 3. Section 4 concludes the paper.

II. RELATED WORK

K-means algorithm is the most commonly used partitional clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time. The major problem with this algorithm is that it is sensitive to the selection of the initial partition and may converge to local optima. This view offers us a chance to apply PSO optimal algorithm on the clustering solution. Contrary to the localized searching in the K-means algorithm, the PSO clustering algorithm performs a globalized search in the entire solution space.

Utilizing the PSO algorithm's optimal ability, if given enough time, the PSO clustering algorithm we proposed could generate more compact clustering results from the document datasets than the traditional K-means clustering algorithm. However, in order to cluster the large document datasets, PSO requires much more iteration to converge to the optima than the K-mean algorithm does. The computation requirement for clustering extremely huge document datasets is still high. In terms of execution time, the K-means algorithm is the most efficient for the large data set.

The K-means algorithm [1] tends to converge faster than the PSO, but it usually only finds the local maximum. The fuzzy clustering in contrast to the usual (crisp) methods does not provide hard clusters, but returns a degree of membership of each object to all the clusters. The interpretation of these degrees is then left to the user that can apply some kind of a thresholding to generate hard clusters or use these soft degrees directly. Hard c-means is better known as k-means and in general this is not a fuzzy algorithm. However, its overall structure is the basis for all the others methods. The HCM algorithm has a tendency to get stuck in a local minimum, which makes it necessary to conduct several runs of the algorithm with different initializations. Then the best result out of many clusterings can be chosen based on the objective function value. Although often desirable, the relative property of the probabilistic membership degrees[2] can be misleading.

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2012

Hard c-means it is quite insensitive to its initialization and it is not likely to get stuck in an undesired local minimum of its objective function in practice. The sample weighting clustering, one of the important works is automatic assignment weights to the clustering samples. In this work, the weights are assigned to the samples according to the importance of samples. In general, the citing relationship between academic documents could indicate the authority degree of a document approximately, which also denotes the structure information in document set. The paper calculates the simplified Page Rank value of a document according to the citing relationship. Different samples or objects should play different roles in clustering process. There are still some unsolved problems of the sample weighting clustering algorithm [3]. How to transform structure information into the weight of samples? Hence, it is very useful to give the appropriate sample weighting in cluster.

The application of the algorithms above is limited for that they need users or heuristic principle to weight samples. Many well known clustering algorithms deal with documents as bag of words and ignore the important relationships between words like synonyms. The proposed FCDC algorithm utilizes the semantic relationship between words to create concepts. The special feature of proposed FCDC algorithm is: it treats the documents as set of related words instead of bag of words. Different words shares the same meanings are known as synonyms. Set of these different words that have same meaning is known as concept. So whether document share the same frequent concept or not is used as the measurement of their closeness [4].

When the cluster size increases in large size then there is a problem in maintaining the quality of clusters. FIHC has a higher probability of grouping unrelated documents into the same cluster. The k-medoids technique, the basic strategy of K-Medoids clustering Algorithms is to find k clusters in n objects by first arbitrarily finding a representative object (the Medoids) for each cluster. Each remaining object is clustered with the medoids to which it is the most similar. K-Medoids method uses representative objects as reference points instead of taking the mean value of the objects in each cluster.

The algorithm takes the input parameter k, the number of clusters to be partitioned among a set of n objects. After the creation of data points, the user has to specify the number of clusters. In the output window, the data points in each cluster are displayed by different colors and the execution time is calculated in milliseconds. It starts with the Normal distribution of input data points first followed by Uniform distribution. The total elapsed time and the execution time for each cluster to each run are calculated in milliseconds. The time taken for execution of the algorithm varies from one run to another run and also it differs from one computer to another computer. The number of data points is the size of the cluster [5]. When the number of clusters increases, in turn will increase the execution time for both the algorithms. Hence, the time taken for execution depends on the number of clusters and the number of data points chosen by the user.

The drawback is the performance of the k-medoids clustering algorithm is relatively less. Frequent Item set Hierarchical Clustering (FIHC) [6] is a novel data mining algorithm for hierarchical grouping of text documents. The approach does not give reliable clustering results when the number of frequent sets of terms is large. In this paper we propose WDC (Word sets based Clustering), an efficient clustering algorithm based closed words sets. WDC uses a hierarchical approach to cluster text documents having common words. WDC [7] found scalable, effective and efficient when compared with existing clustering algorithms like K-means and its variants. Although standard clustering techniques such as k-means can be applied to document clustering, they usually do not satisfy the special requirements for clustering documents: high dimensionality, high volume of data, ease for browsing, and meaningful cluster labels. It initially discovers a set of tightly relevant keyword clusters that are disposed throughout the feature space of the collection of documents, and further cluster the documents into document clusters by using these keyword clusters.

The similarities between the representative keywords and the documents of the corpus for document categorization. All of these methods have a major problem that the high dimension of the feature space results in high computation complexity and space needs. This is because the native feature space consists of the unique terms in documents, which can be tens or hundreds of thousands of terms for even a moderate-sized text collection. The DCF (Description Comes First) method can generate document clustering description. For the clustering description base on DCF is generate before document clustering, there is semantic interval' between clustering description and cluster central vector. So, it contradicts to the intuition of 'First clustering, second description', and decreases the readability of clustering description.

A method based on combination strategy, i.e. combination of the DCF and DCL[8] (Description Comes Last) is proposed to solve the problem of the weak readability of clustering. labeling a clustered set of documents is an important and challenging work in document clustering applications. And it is also one of the difficult problems of search results clustering. The traditional algorithm of document clustering can cluster the documents collection, but it cannot give concept description to the clustered results.. For both the algorithms, a set of n data points are given in a two-dimensional space and an integer K (the number of clusters) and the problem is to determine a set of n points in the given space called centers, so as to minimize the mean squared distance from each data

point to its nearest center. The performance of the algorithms is investigated during different execution of the program for the given input data points. Based on

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2012

experimental results the algorithms are compared regarding their clustering quality and their performance, which depends on the time complexity between the various numbers of clusters chosen by the end user. clustering results by comparing them with others generated by other clustering algorithms, or by the same algorithm using different parameters. measure is often called the squared-error distortion and this type of clustering falls into the general category of variance based clustering[9].

III. PROPOSED METHOD

In the proposed system, the partition based clustering algorithm called fuzzy c-means algorithm is used.FCM is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters. FCM starts with an initial guess for the cluster centers, which are intended to mark the mean location of each cluster. The initial guess for these cluster centers is most likely incorrect. Additionally, FCM assigns every data point a membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point.

The processing steps applied to Document clustering are given in Figure 1.

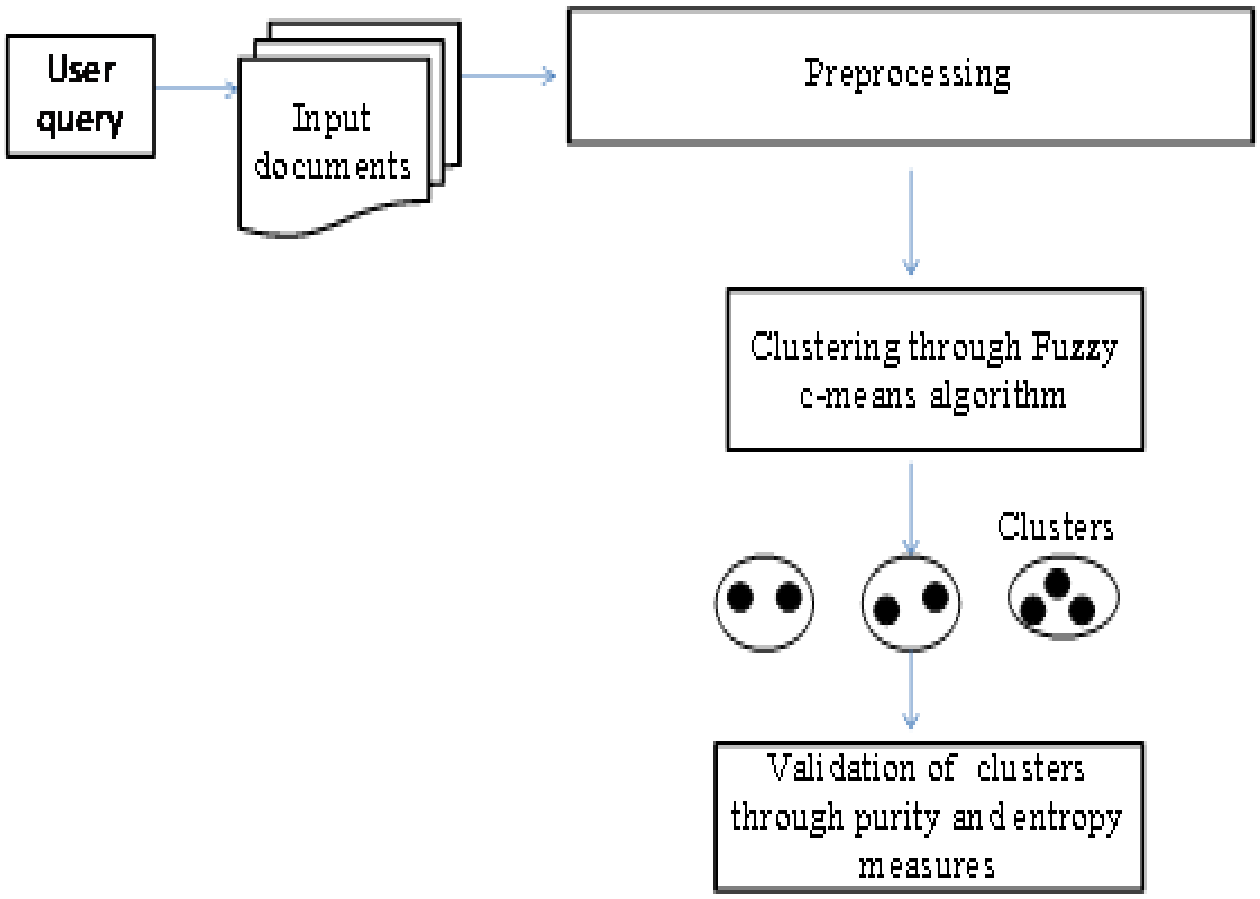

Fig. 1 Processing steps.

FCM iteratively moves the cluster centers to the right location within a data set. This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster center weighted by that data point's membership grade.

.

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2012

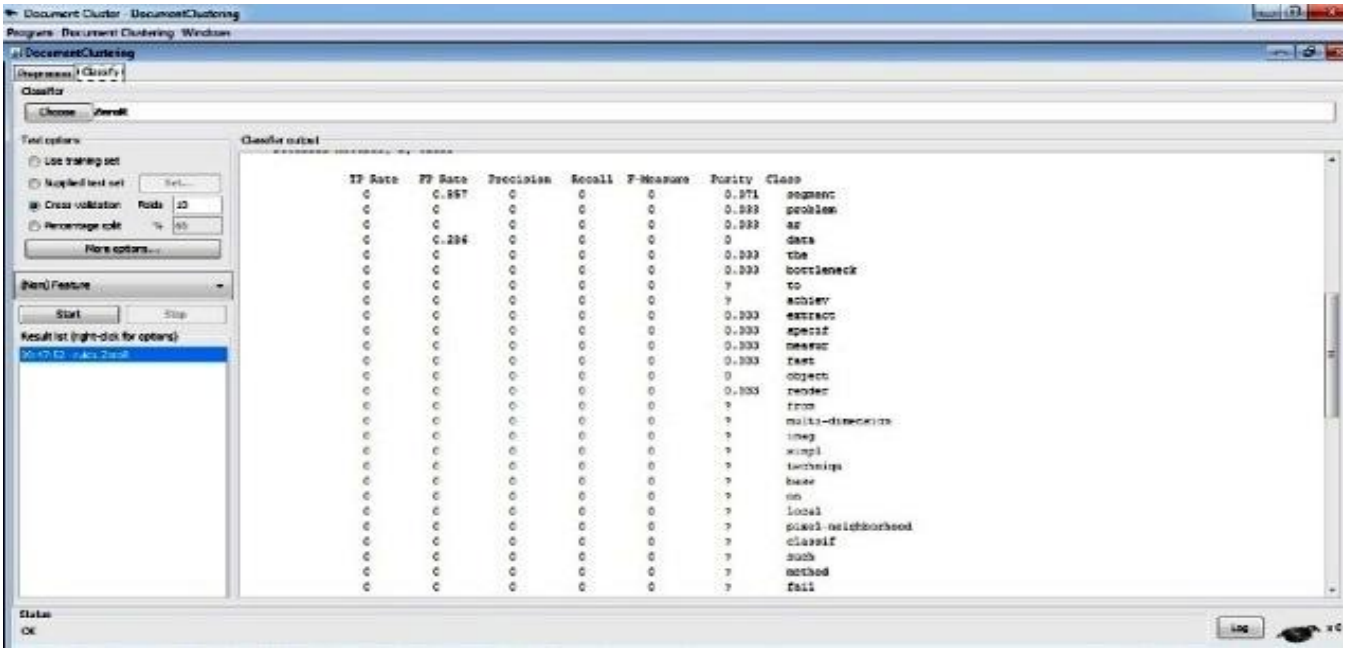

Fig 2. Evaluation of measures

The preprocess step is used to select the file with .arff extension. In the classify step the correctly classified instance and incorrectly classified instance are specified. The confusion matrix is formed ant the corresponding words purity and entropy value is calculated.

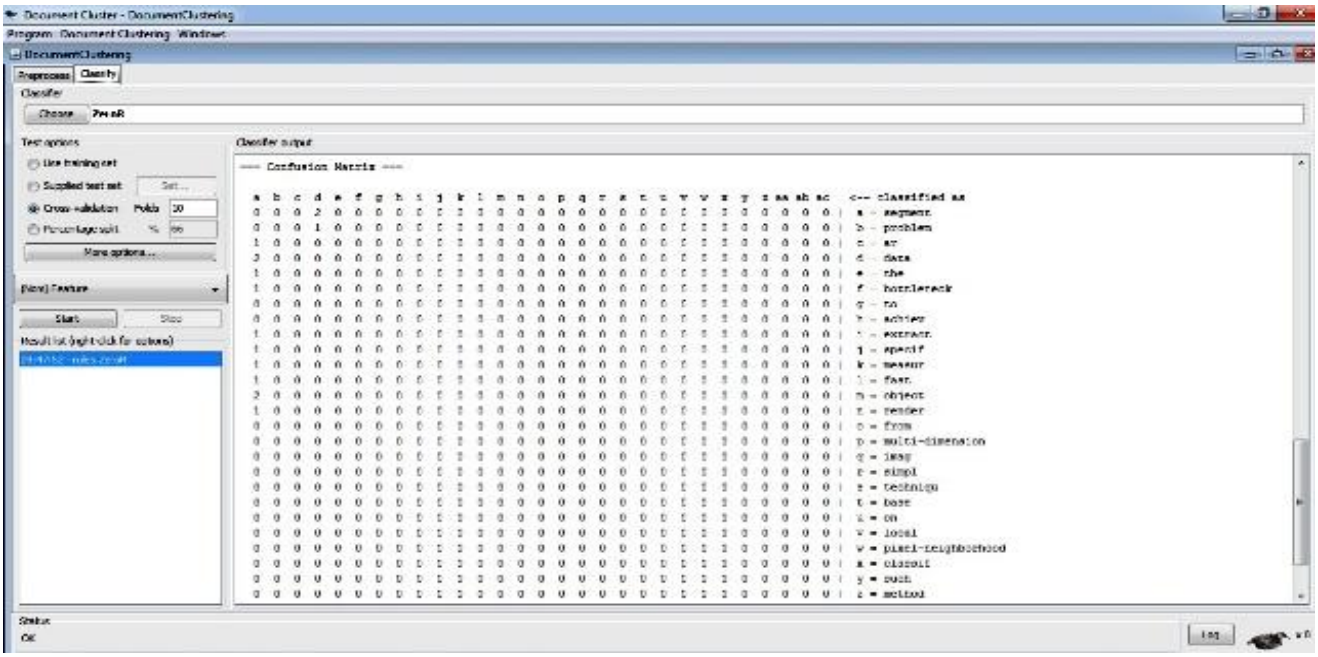

Fig 3. Confusion matrix

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2012

The command line function FCM outputs a list of cluster centers and several membership grades and a fuzzy inference system by creating membership functions to represent the fuzzy qualities of each cluster.

IV. CONCLUSION:

Fuzzy clustering is a powerful unsupervised method for the analysis of data and construction of models. Cluster validity measures are useful in estimating the optimal number of clusters because the number of cluster is not known a priori. The analysis of the fuzzy c-means algorithm provides better clustering of documents and it is observed that the quality of cluster is evaluated using the measures. Using these measures, the user can easily find the related documents in the file system.

REFERENCES

[1] Xiaohui Cui, Thomas E. Potok," Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm", Applied Software Engineering Research Group, Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6085, USA.2006

[2] Matjaž Juršič, Nada Lavrač," Fuzzy Clustering of Documents" Department of Knowledge Discovery, Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia.2008

[3] Chengzhi Zhang, Xinning Su, Dongmin Zhou," Document Clustering Using SaDlple Weighting" Nanjing University, Nanjing 210094, China.2007

[4] Rekha Baghel, Dr. Renu Dhir," A Frequent Concepts Based Document Clustering Algorithm" International Journal of Computer Applications (0975 – 8887) Volume 4 – No.5, July 2010

[5] T.Velmurugan, T.Santhanam," Clustering Algorithm for Statistically Distributed Data Points" Journal of theoretical and applied information technology, May 15, 2011, vol 27 no.1

[6] Anuj Sharma, Renu Dhir," A Wordsets Based Document Clustering Algorithm for Large Datasets", International Conference on Methods and Models in Computer Science, 2009

[7] Hsi-Cheng Chang',2, Chiun-Chieh Hsu2 and Yi-Wen DengZ," Unsupervised Document Clustering Based on Keyword Clusters", lntemational Symposium on Communications and Information Technologies 2004 (ISCIT 2004) Sappom, Japan, October 26- 29, 2004

[8] Chengzhi ZHANG," Document Clustering Description Based on Combination Strategy", Fourth International Conference on Innovative Computing, Information and Control,2009

[9] Jain, A.K., M.N. Murty and P.J. Flynn, Data Clustering: A Review, ACM Computing Surveys, Vol. 31, No. 3, Sep. 1999, pp. 264-323, DOI:10.1.1.18.2720&rep=rep1&type=pdf

Biography

Ms.Deepa.M has completed her B.Tech in Information Technology at C.S.I College of Engineering affiliated to Anna University, Chennai, India. Currently she is pursuing her M.E. in Computer Science and Engineering at Rajalakshmi College of Engineering, affiliated to Anna University, Chennai, India. Her area of interest includes Data mining.

Mrs.Revathy.P was awarded with Master of Engineering from Sathayabama University ,Chennai. Presently working as Assistant Professor in Rajalakshmi Engineering College, Chennai in the department of Computer Science & Engineering. Her Area of research interest includes Data Mining.

.