Abstract-
Efficient Document clustering can be performed based on the term level, sentence level and concept level techniques in the high dimensional document space. Most of the existing techniques have problems such as two-variable problem, high computational time and low similarity relatedness which reduces the clustering efficiency. To overcome the existing drawbacks, a hybrid clustering algorithm called Semantically Document Clustering algorithm is proposed in this paper. The Semantically Document Clustering algorithm is developed by combining the features of Directed Ridge Regression (DRR), Fuzzy relational Hierarchical clustering (FHC) and Conceptual clustering methods presented in our previous researches. The proposed Semantically Document Clustering algorithm utilizes the semantic weight of terms related to the concepts from Wikipedia and Word Net to categorize the texts in the documents. Then the similarity between the sentences is calculated by using the Jiang and Conrath measure which considers the concept weight and the similarity measure for effective clustering. The direct ridge regression is applied to build a Laplacian matrix and the diagonal elements of the normalized Laplacian matrix are varied to solve the two-variable problem. Then the fuzzy hierarchical rules are employed to classify the rows of the normalized Laplacian matrix into classes for calculating the membership for the observations and the center vectors. Thus the term relatedness, sentence relatedness and concept relatedness can be calculated and the documents can be clustered efficiently. Experiment results also show that the proposed hybrid approach Semantically Document Clustering method provides more accurate document clustering than the state-of-the-art clustering methods.

Keywords: Directed Ridge Regression, Fuzzy relational Hierarchical clustering, Conceptual clustering

Introduction
Document clustering is considered to be the fundamental procedure in grouping the unsupervised documents for effective applications in text mining and information retrieval. The clustering of documents is a much needed process in machine learning approaches, medical practices and artificial intelligence techniques. Document clustering is normally a centralized process that makes use of the descriptors and descriptor extraction for efficient performance. Based on various distance measures, a number of methods have been proposed in the recent years to handle document clustering. A typical and widely used distance measure is the Euclidean distance. The k-means method is one of the methods that use the Euclidean distance, which minimizes the sum of the squared Euclidean distance between the data points and their corresponding cluster centers. Since the document space is always of high dimensionality, it is preferable to find a low-dimensional representation of the documents to reduce computation complexity. Low computation cost is achieved in spectral clustering methods, in which the documents are first projected into a low-dimensional semantic space and then a traditional clustering algorithm is applied to finding document clusters. Latent semantic indexing (LSI) and Locality preserving indexing (LPI) are some of the most commonly used methods for document clustering. Correlation preserving indexing (CPI) is an efficient clustering method which explicitly considers the manifold structure embedded in the similarities between the documents [1]. The similarity-measure based CPI method focuses on detecting the intrinsic structure between nearby documents rather than on detecting the intrinsic structure between widely separated documents. Since the intrinsic semantic structure of the document space is often embedded in the similarities between the documents, CPI can effectively detect the intrinsic semantic structure of the high-dimensional document space. But the approach has two-variable problem due to the use of many variables in the high-dimensional document space reduces the performance. Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (FRECCA) [2] is an effective sentence level clustering approach that operates on the basis of Expectation-Maximization framework and was capable of identifying overlapping clusters of semantically related sentences. But the approach suffers from the serious problem of time complexity. Multi-layer classification (MLCLA) framework has been primarily used for categorizing the texts in the documents [3]. Using this property of MLCLA framework, the document clustering can also be performed. MLCLA framework performs layer-by-layer clustering with three classifiers. Two classifiers are implemented at the syntactic level and semantic level with the third classifier combining the outputs of the first two classifiers to classify the documents. The approach utilizes the Wikipedia concepts related to terms in syntactic level to represent document in semantic level and improves the clustering accuracy.

In this paper, to improve the efficiency of document clustering, a hybrid document clustering approach called Semantically Document Clustering method is proposed. Semantically Document Clustering

Semantically Document Clustering Using Contextual Similarities

R.Nagaraj, Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, Tamilnadu- 641021

X.Agnise Kalarani, Professor, Department of Computer Applications, Karpagam Academy of Higher Education, Coimbatore, Tamilnadu- 641021
method is developed by combining the efficient document clustering methods proposed in previous works. The methods Document Clustering based on Direct Ridge Regression (DCDRR), Sentence level document clustering in fuzzy relational hierarchical clustering (SCFH) and Conceptual clustering methods are utilized for developing the proposed method. The DCDRR approach identifies the similarity between the documents by evaluation of relationship among the variables and thus avoids the two-variable problem in the high dimensional document space. The SCFHC approach can better handle clusters with a complex, nonlinear geometric structure and it does not need prior information on the number of clusters and hence can improve the clustering performance. The Conceptual clustering approach is utilized to improve clustering based on semantic relatedness between the concepts using Wikipedia and Word Net. Thus by combining the features of these three approaches, Semantically Document Clustering method can provide efficient and accurate document clustering.

The remainder of the paper is organized as follows: Section II describes the previous researches that were performed to achieve the target of efficient document clustering. Section III describes the methodologies used in this paper and the detailed explanation of the proposed techniques. Section IV represents the experimental results conducted to evaluate the performance of the techniques. Section V represents the conclusion of the research.

Related Works
Soumi Ghosh et al [4] proposed Cluster analysis technique for classifying data sets. K-Means 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. 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. The K-Means clustering algorithm finds the desired number of distinct clusters and their centroids. The limitation of this algorithm is a labeled dataset as training data and practically classification of labeled data is generally very difficult as well as expensive.

An additive spectral method for fuzzy clustering is proposed by Boris Mirkina et al [5] to find similarity between the topics. The method operates on a clustering model which is an extension of the spectral decomposition of a square matrix. The clusters are extracted one by one that makes the spectral approach quite natural. The key feature of Fuzzy Additive Spectral clustering method aside from the relational clustering approaches is that the cluster membership values directly contribute to the similarity, in an additive way, according to model. This comes with the price of imposing another novel feature, the clusters intensity, to account for the similarity index scale. This somewhat blurs the meaning of a fuzzy membership value as proportion or probability which must never exceed the unity.

Xinlei Chen et al [6] presented the Landmark based Spectral Clustering approach. The previous methods usually forfeit quite lot information of the original data, resulting that the deprivation of performance. In order to resolve the huge scale clustering problems, the Landmark based Spectral Clustering is suggested. Predominantly, selected the representative data points as the landmarks and symbolizes the original data points as the linear combinations of these landmarks. With the landmark-based representation, the spectral embedding of the data can then be proficiently computed. This method scales linearly with the size of the problem. But the main disadvantage is this method is that it is difficult to implement.

Ruizhang Huang et al [7] proposed an approach called Dirichlet Process Mixture Model for Feature Partition (DPMFP) to discover the latent cluster structure based on the DPM model without requiring the number of clusters as input. This approach enables the detection of number of clusters which can be utilized for clustering documents at a less time. Document features are grouped into two groups of words and the variational inference algorithm is employed to analyze the document collection structure and the document partition words to improve the accuracy of document clustering.

Yuan Ling et al [8] presented Multi-View Nonnegative Matrix Factorization (NMF) for clustering the documents based on the medication symptom words. The application of this approach in clustering the clinical documents helps in better medication. The approach initially constructs an integrating system for extracting medication names and symptom names from clinical notes and then utilizes multi-view NMF for clustering the words into meaningful clusters based on the sample-feature matrices.

Guoyu Tang et al [9] suggested the use statistical word sense to cluster the cross lingual documents. The documents are clustered by employing the concept of sense-based vector space model leverages on a sense-based latent Dirichlet allocation. This approach groups the documents based on the cross lingual word sense score and utilizes the latent Dirichlet allocation to cluster the word senses. The proposed approach however suffers from the drawback of low accuracy related to the concepts.

Lin Yue et al [10] presented a clustering approach called fuzzy document clustering based on the domain-specified ontology. In this approach, first a domain-specific ontology is constructed to provide the controlled vocabulary for feature selection. Then with the vector space model (VSM), singular value decomposition (SVD) is performed to translate all of the term-document vectors into a concept space. The fuzzy value of each feature is estimated and correlation between two terms is considered for effective clustering of the documents with the domain specific clusters.

Yinglong Ma et al [11] presented a three phase approach for document clustering based on topic significance degree. In the first phase, the best topic model is determined by analyzing the significance degree and then a formal concept about significance degree of topics is discovered by LDA method. In the second phase, the initial clustering centers are selected by using the k-means algorithm and in the third phase the k-means algorithm is used in the clustering centers for document clustering. Chun-Ling Chen et al [12] presented the Fuzzy-based Multi-label Document Clustering (FMDC)
approach along with the Word Net semantic terms to provide efficient document clustering.

**Methodologies**

Document clustering can be achieved with high efficiency by including the terms relatedness, sentence relatedness and concept relatedness. Hence the proposed hybrid approach called as Semantically Document Clustering includes the terms, sentences and concepts for efficient document clustering by combining the Direct Ridge Regression, Sentence clustering based Fuzzy relational Hierarchical clustering and Conceptual clustering methods. The Semantically Document Clustering method can be developed only by understanding DCDRR, SCFHC and Conceptual clustering approaches. Hence the three previously proposed techniques and the need for developing these techniques are discussed in this section.

The similarity based clustering methods have been largely utilized in the recent time to provide efficient document clustering. Correlation Preserving Indexing (CPI) method deals with highly unsupervised documents that can be clustered using the similarity measure of the documents. Though CPI clustering is very effective, it has the two-variable problem. When one variable is used in high dimensional space it provides results but when more than one variable is used, CPI does not produce efficient results. To overcome the two-variable problem, the Direct Ridge Regression (DRR) is utilized which efficiently provides relationships between many variables. The Document Clustering based on Direct Ridge Regression (DCDRR) approach is presented to provide better document clustering overcoming the two-variable problem. DCDRR determines the similarity between the documents by measuring the relationship among the variables. The regression analysis value as a numerical tool may be extensively diminished when the set of independent variables are approximately collinear. Using DRR, when one of the independent variables is not similar, the dependent variables in the diagonal elements are changed. This approach improves the document clustering as the dependent variables are utilized for similarity measurement.

Similarly, Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (FRECCA) is an efficient sentence level clustering algorithm inspired from the mixture model approach. The problem with the approach is the time complexity in executing the cluster membership process. Hence, Sentence level document clustering in fuzzy relational hierarchical clustering (SCFHC) is an innovative technique that is proposed to cluster the documents at the sentence level without much time complexity. SCFHC groups the sentences with similar meanings into clusters from which the vector similarity measures, such as cosine, are determined to define the level of similarity over bag-of-words encoding of the sentences. The SCFHC assigns each observation with degrees of membership with the uncertainty information of the clustering assignments. When observations belong to a cluster, it assigns high degree of membership to that cluster and low degrees of membership to the remaining clusters.

Multi-layer Classification (MLCLA) framework is one of the concept based clustering approach that uses the Wikipedia concepts related to terms in syntactic level to represent document in semantic level so that the documents can be clustered efficiently. But the performance of the MLCLA framework can be further improved by modifying the approach and now the approach called Conceptual clustering is proposed with the inclusion of Word Net concepts. This approach uses the concepts of Wikipedia and Word Net improves the semantic relatedness and thus the document clustering efficiency can be improved. In this paper, the Semantically Document Clustering is developed by hybrid combining DCDRR, SCFHC and Conceptual clustering algorithms, utilizes the term, concept and sentence informations. The semantic relatedness among the terms, concepts and sentences is calculated such that semantic weight can be estimated. The semantic weight of terms related to the concepts from Wikipedia and Word Net are used to represent semantic information. The directed ridge regression measure is used to find relationship between the variables of each document and then the fuzzy relational hierarchical clustering is used to cluster the documents. The Semantically Document Clustering includes semantic concepts to select the most appropriate concept for a term in a given document. The proposed clustering approach can be utilized for improving the accuracy of document clustering.

**Algorithm:** Semantically Document Clustering

**Input:** Set of documents \( x_1, x_2, \ldots, x_n \in R \).

**Output:** Clustered documents

1. Begin
2. Determine concept weight
   
   \[
   W(c_i, d_j) = \sum w(t_k, d_j) * Rel(t_k, c_i | d_j) \tag{1}
   \]
   
   // Where \( Rel(t, s_i | d_j) \) = semantic relatedness between term and sentence, \( \sum w(t_k, d_j) \) = term’s weight

3. Define measure of similarity
   
   \[
   SIMres(c_1, c_2) = IC(\text{ics}(c_1, c_2)) \tag{2}
   \]

4. Compute semantic relatedness between terms and its candidates
   
   \[
   Rel(t, c_i | d_j) = \frac{1}{|T| - \sum \frac{1}{|CS_i|} \sum SIM(c_i, c_k)} \tag{3}
   \]
   
   // where \( T \) is the term set of the \( j^{th} \) document \( d_j \) except for \( t \) and \( CS_i \) is the candidate concept set related to term \( t \), \( SIM(c_i, c_k) \) is the semantic relatedness between two terms

5. Similarity between the sentences is find by using Jiang and Conrath measure.
   
   \[
   Sim(w_{1,2}) = \frac{1}{IC(w_1) + IC(w_2) - 2 * IC(AGCS(w_1, w_2))} \tag{4}
   \]

6. Sentence weight is computed as
   
   \[
   W^*(s_i, d_j) = \sum w(t_k, d_j) * Rel(t_k, s_i | d_j) \tag{6}
   \]
Center of each class can be determined by averaging the
document vectors belonging to this class

\[ S_{jk} = \frac{d_{jk}}{||z_{j}||} \]  

(7)

8. Similarity between document and the \( k \)th class center is
determined as,

\[ u_{i} = \frac{d_{ik}}{||z_{i}||} \]  

9. Compute directed ridge estimator

\[ \hat{\alpha} (dk) = (\Lambda + KI)^{-1} \hat{X} U \]  

Hence \( W = \hat{\alpha} (dk)^n \)  

(8)

10. Construct Laplacian matrix

\[ L_{NCut} = s^{-1/2}(s - W)s^{-1/2} \]  

(9)

11. Find eigenvectors of Laplacian matrix and form the matrix

\[ U = [u_{1}, \ldots, u_{k}] \in R^{n \times k} \]  

(10)

12. Form Y matrix from U by normalizing each of U’s rows to
have unit length

\[ Y_{ij} = \frac{u_{ij}}{\sqrt{u_{i1}^2 + \ldots + u_{in}^2}} \]  

(11)

13. Classify each row of Y into k classes through the fuzzy
relational algorithm.

14. Initialize membership value of U

\[ U = [\mu_{ij}] \text{ matrix}, U^{(0)} \]  

(12)

15. Calculate the centers vectors

\[ C_{j} = \frac{\sum_{i=1}^{n} \mu_{ij} y_{i}}{\sum_{i=1}^{n} \mu_{ij}^2} \]  

(13)

16. Update membership value

17. Assign the original points to cluster j if row of the matrix Y
was assigned to cluster j.

18. Compute Semantic relatedness between two concepts

\[ SIM(c_{i}, c_{k}) = 1 - \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))} \]  

(14)

where A and B are the sets of all articles that link to concepts \( c_{i} \)
and \( c_{k} \) and W is the set of all articles.

19. Cluster documents based on term relatedness, sentence
relatedness and concept relatedness.

20. End

Description:
In Semantically Document Clustering method, the documents are
clustered based on terms, sentences and concepts. Initially, the
concept weight for each document is determined. Then the
similarity measure and the semantic relatedness between terms
and its candidates are calculated using the given formulas. The
similarity between the sentences is calculated by using Jiang and
Conrath measure in which the concept weight and similarity
measure play a crucial role in document clustering. The sentence
weight of each document is calculated with which the class center
is determined. The similarity between the document and the class
center using determined on which the DRR is applied. The direct
ridge estimator is computed to build an \( m \times n \) Laplacian matrix.
The Eigen vectors of the Laplacian matrix are determined and the
normalized matrix is build. The diagonal elements are changed

\[ \hat{\alpha} (dk)^n \]  

Performance Evaluation
Experiments are performed on real time datasets 20Newsgroups and Reuters-21578. The three methods DCDRR, SCFHC and Conceptual clustering are
compared along with the proposed Semantically Document Clustering method to determine the most
efficient clustering technique in terms of accuracy, precision, recall and f-measure.

i. Clustering Accuracy

![Figure 1. Accuracy Comparison](image)

The methods DCDRR, SCFHC, Conceptual clustering and Semantically Document Clustering are compared in
terms of cluster accuracy shown in Figure 1. In the x-axis, the number of clusters is taken while along the y-axis,
the clustering accuracy in percentage is taken. When the number of clusters is 5, the clustering accuracy
of DCDRR is 70%, FRECCA is 73%, and Conceptual clustering is 82% while that of Semantically Document
Clustering is 92%. From the comparison graphs, the Semantically Document Clustering method is found to be
the efficient technique with high clustering accuracy.

ii. Precision

Figure 2 shows the comparison of the four methods such as DCDRR, SCFHC, Conceptual Clustering method and
Semantically Document Clustering in terms of precision...
rate. If the numbers of clusters is 5, the precision value is 90% in Semantically Document Clustering, for Conceptual Clustering 80%, for SCFHC 73% and for DCDRR 66%. This clearly shows the precision value is increases in the Semantically Document Clustering when compared to the existing methods.

### iii. Recall
Figure 3 shows the comparison of the four methods such as DCDRR, SCFHC, Conceptual Clustering and Semantically Document Clustering in terms of recall. If there is five numbers of clusters, the recall value is 88% in the Semantically Document Clustering method, for Conceptual Clustering 77%, for SCFHC 72% and for DCDRR 62%. This clearly shows the recall value increases in the Semantically Document Clustering method when compared to the existing methods.

### iv. F-Measure
Figure 4 shows the comparison of the four methods such as DCDRR, SCFHC, Conceptual Clustering method and Semantically Document Clustering method in terms of F-Measure. If there is five numbers of clusters, the F-Measure is 90% in the Semantically Document Clustering method, for Conceptual Clustering method 80%, for SCFHC 72% and for DCDRR 69%. This clearly shows when the number of clusters is increased the F-Measure value is increases in the Semantically Document Clustering method when compared to the existing methods.

### Conclusion
Document clustering using term relatedness, sentence relatedness and concept relatedness is a most challenging but efficient procedure. The accuracy of clustering the documents can be improved considerably by including the terms, sentences and concepts together. Hence in this research work, a hybrid approach called Semantically Document Clustering method is proposed by combining the previously presented three effective clustering approaches DCDRR, SCFHC and Conceptual clustering methods. The three approaches are based on the term, sentences and concepts respectively and hence the proposed Semantically Document Clustering method can provide better clustering of documents along with avoiding the drawbacks such as multi-collinearity, two-variable problem and time complexity.

### References


[6] Xinlei Chen, and Deng Cai. "Large Scale Spectral Clustering with Landmark-Based Representation" In Association for the Advancement of Artificial Intelligence, 2011.


