Correlation Similarity Measure based Document Clustering with Directed Ridge Regression

R. Nagaraj*1 and V. Thiagarasu2

1Karpagam University, Coimbatore, India; nagukasc@gmail.com
2Computer Science, Gobi Arts & Science College, Gobichettipalayam, India; profdravt@gmail.com

Abstract

Correlation Preserving Indexing (CPI) can discover the intrinsic structures implanted in high-dimensional document space. To predict the result of one variable based on another variable is not suitable for all the situations since two variable prediction problems take places. In this paper, Directed Ridge Regression is introduced to predict two or more variables which are highly correlated in high dimensional document space. Directed Ridge Regression is a statistical technique to estimate the relationship among the variables based on the Eigen values to find the similarity between the documents. The directed ridge estimator alters the diagonal elements of the Eigen values. The objective of the Directed Ridge Regression is to achieve efficient document clustering in similarity measure. Experimental results shows that compared to Correlation Preserving Indexing, the Directed Ridge Regression achieves efficient document clustering.

Keywords: Correlation Similarity Measure, Directed Ridge Regression, Document Clustering, Latent Semantic Indexing

1. Introduction

Document clustering is nothing but grouping the related documents into clustering. The document clustering is a considerable process in data mining and it is a general method which is used in extracting the topic, automatic document association and information retrieval. In document clustering, the Euclidean distance measure is most comprehensively utilized. Spectral clustering approaches are used for low computation cost where the documents are estimated into a low-dimensional semantic space and for identifying the document clusters the conventional clustering method is used. The latent semantic indexing is one of the proficient techniques in spectral clustering. The main intent of this method is to find the best subspace approximation to the inventive document space by diminishing the global reconstruction error. The Locality preserving indexing (LPI) technique is a spectral clustering approach which is based on graph separation theory1. BY using a weighted function, this method focuses on detaining the similarity structure of the documents1. It does not correct the drawback of Euclidean distance.

The correlation preserving indexing is a method which is suggested to deal with the difficulty of similarities between the documents. This Correlation preserving Indexing method primarily considers the manifold structure which is entrenched in the similarities between the documents. The intent of the Correlation preserving Indexing is to achieve a most sympathetic semantic subspace by concomitantly growing the associations between the documents in the local patches and to reduce the associations between the documents outside these patches. This method is similarity measure which is focused on identifying the intrinsic formation between close by documents. CPI can proficiently recognize the intrinsic semantic structure of the high-dimensional document space.

In the correlation preserving indexing there is a two variable problem. In the correlation method based on one variable one result can be predicted based on another variable. This becomes not suitable for every variable, and hence the efficient clustering cannot be achieved in this method. To address this trouble Directed Ridge Regression (DRR) is applied in order to bring the relationships

*Author for correspondence
among the variables. It identifies the similarity between the documents by evaluation of relationship among the variables. In the regression analysis specifically Directed Ridge Regression (DRR) is used in which effective document clustering is accomplished.

2. Literature Review

A document separation method was proposed in order to accomplish the high correctness and capability for computation of the number of clusters in the set of documents\(^2\). A clustering is achieved by characterizing the document by a term-frequency vector with its magnitude to represent the number of times a specific word occurred in the document. A more affluent feature set is used to disseminate each and every document for precisely clustering the documents. The Gaussian Mixture Model together with the Expectation-Maximization approach is used to cluster more number of documents.

A new technique is called Locality preserving Indexing is utilized for document clustering\(^3\). The locality preserving Indexing (LPI) is a method in which the documents can be anticipated into a lower-dimensional semantic space. The documents associated to the identical semantics are close up to each other. This method investigates to identify both the numerical and discriminating structures of the document space. The main intent of the Locality Preserving Indexing is to discover the global Euclidean structure. It can have more discerning power. The derivation of the LPI is accomplished by determining the most favorable linear estimation to the Eigen functions of the Laplace Beltrami operator on the document. The second order derivatives of the functions on the manifolds are utilized by the Laplace Beltrami operator. It estimates the softness of the functions. Resulting to that, it can determine the nonlinear manifold configuration to some extent.

Distributional Clustering was clarified for the document classification. Data clustering is a complicated process in information processing and pattern recognition\(^4\). The dispute is both theoretical and computational. Impulsively, to cluster a dataset, the intention is to partition it into subsets such that points in the same subset are more “related” to each other than to points in other subsets. The usual clustering algorithms depend on choosing a similarity measure between data points and a “correct” clustering result can be dependent on an appropriate choice of a similarity measure. Based on the allocation of the class labels related with each word, this method clusters words into groups. The words are into groups predominantly for the document classification. In the document organization the focus point is class label. By allocating the class labels related with the words, measure word similarity. Distributional Clustering will cautiously cluster words that are investigative of more than one class.

For the correlation similarity determination a discriminant learning process is utilized\(^5\). This is one of the methods that explore for a global linear transformation to utilize the correlation of samples from the same class and diminish the correlation of samples from different classes in the transformed space. This method merges the similarity measure learning with the supervised discriminant learning in a very simple way to determine the nonlinear information instead of only linear information. In this method based on the description of within-class correlation and between-class correlation, the best alteration can be required to extend the dissimilarity between them, which is in conformance with good classification performance practically. The difference can be virtually calculated as a standard for the classification performance.

Semi-supervised learning that is based on a Gaussian random field model\(^6\). In a weighted graph the vertices represented the labeled and unlabeled data, with edge weights encoding the similarity between instances. After that the learning problem is formulated in terms of Gaussian random field on this graph, where the mean of the field is differentiate in terms of harmonic functions, and is proficiently attained using matrix methods or belief propagation. In this method, the solution is exclusively based on the structure of the data manifold, which is consequent from data features. In practice, however, this derived manifold structure may be inadequate for precise classification. The fully probabilistic framework is intimately related to Gaussian process classification, and this connection proposes principled ways of incorporating class priors and learning hyper parameters.

The metric learning algorithm for text documents was introduced\(^7\). Based on the exploiting the inverse volume of a given data set of points, this method selecting a metric from a parametric family. The captured metric is local, thus detaining local variations within the space and is characterize on the entire embedding space. A group of metric candidates is represented as a parametric family of transformations. Based on the some performance
criteria, the obtained metric is chosen. By examining the application of the metric learning techniques in the context of classification of text documents and images, the experimental results for text classification was provided.

For a given set of documents, a document clustering method was suggested based on the non-negative factorization of the term document matrix \( X \). This method is used to compute the latent semantic space whereas every axis confines the base topic of particular documents. Also, each and every document is disseminated as an additive grouping of the base topics. In every document the cluster membership can be effortlessly determined by classifying the base topic with which the document has the main projection value. This method is based on the singular vector decomposition and the associated spectral clustering methods in that the latent semantic space derived by NMF does not need to be orthogonal, and that each document is definite to take only non-negative values in all the latent semantic directions.

Spatial data mining is the proficient and effectual for the clustering approaches \(^8\). This method is used to recognize the relations and distinctiveness that may endure completely in spatial databases. The main intent of the spatial data mining is to mechanize such a knowledge detection process. The major role is to remove the interesting spatial patterns and features, detaining essential associations between spatial and non-spatial data, presenting data reliability succinctly and at higher hypothetical levels, helping to rearrange spatial databases to comprise data semantics and also attain better performance. However the quality of the results formed by both algorithms relies quite critically on the appropriateness of the hierarchy to the given data.

A pattern clustering methods from a statistical pattern recognition point of view, with a objective of providing useful advice and references to fundamental ideas available to the broad community of clustering practitioners \(^9\). The taxonomy of clustering techniques was introduced, and recognizes cross-cutting themes and recent advances. Therefore, the trouble is little prior information about the data, and it is possible to make a few postulations about the data as possible. Clustering is used in several fields to explain process for grouping of unlabeled data. The objective of this concept is to review the core models and methods in the large subset of cluster analysis with its roots in statistics and decision theory.

Unsupervised learning was suggested for distributed clustering \(^10\). Unsupervised learning is an method of learning, in which the instances are automatically placed into meaningful groups based on their similarity. The intention is to make inferences about its cluster structure, it is necessary to examine whether the data set demonstrates a clustering tendency. Due to the improper measurement, there may be errors in the collected data set. In order to overcome this problem, take an unseen instance, eliminating the value of one of its attributes and then trying to classify it. The missing attribute is calculated to be the same as the value of the attribute on the closest matching instance. This value can be then compared to the actual value of the removed attribute and so can be moderated to be correct or not. This process is repetitive for each attribute. In order to give the average prediction accuracy, the number of attributes properly predicted is then totaled up and divided by the number of attributes.

### 3. Document Clustering Approaches

#### 3.1 Correlation Preserving Indexing

Correlation Preserving Indexing is one of the imperative techniques for document clustering. In particular, this method recognizes the manifold structure which is entrenched in the similarities between the set of documents. The semantic constitution is frequently inherent in large number of documents. In the document clustering the most momentous procedure is to determine the intrinsic constitution of the document space. Although the manifold structure is usually entrenched in the similarities between the documents, Correlation is a similarity measure which is significant for confining the manifold structure embedded in the high-dimensional document space. The main purpose of the method is to establish an optimal semantic subspace by concomitantly make use of the associations between the documents in the local patches and to reduce the associations between the documents outside these patches.

The correlation between the two vectors \( u \) and \( v \) is as followed as,

\[
\text{Corr}(u, v) = \langle \frac{u^T v}{\|u\| \|v\|} \rangle
\]

The correlation corresponds to an angle \( \theta \) such that \( \cos \theta = \text{Corr}(u, v) \). The higher the value of \( \text{Corr}(u, v) \), the stronger the association between the two vectors \( u \) and \( v \).
1. 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. In the original document space, if the two documents are close to each other then it grouped into the same cluster.

A2. In the original document space, if the two documents are far away from each other then it grouped into the different clusters.

\[ y_i \in \text{denotes the low-dimensional illustration of the} \]
\[ \text{i^{th} document } x_i \in \text{in the semantic subspace, in which } i = 1, 2, \ldots, n. \] Then the above postulation (A1) and (A2) can be described as,

\[
\text{Max} \sum \sum \text{Corr}\left(y_i, y_j\right)
\]

\[
\text{Min} \sum \sum \text{Corr}\left(y_i, y_j\right)
\]

Correspondingly, where \( N \left(x_i\right) \) denotes the group of nearest neighbors of \( x_i \). The optimization of (2) and (3) is analogous to the succeeding metric learning:

\[
d(x, y) = \alpha \ast \cos(x, y)
\]

In which \( d(x, y) \) represents the similarity between the documents \( x \) and \( y \), \( \alpha \) denotes the whether \( x \) and \( y \) are the nearest neighbors of each other.

3.2 Clustering Algorithm based on CPI

Given the set of documents \( x_1, x_2, \ldots, x_n \in IR^n \). The document matrix is represented by \( X \). Based on correlation preserving indexing for document clustering, the algorithm is followed as,

1. Construct the local neighbor patch, and calculate the matrices \( M_s \) and \( M_T \).

The matrices \( M_s \) and \( M_T \) are defined as,

\[
M_T = \sum \sum \left(x_i, x_j^T\right)
\]

\[
M_s = \sum \sum \sum \left(x_i, x_j^T\right)
\]

It is simple to confirm that the matrix \( M_s \) is semi positive definite. Whereas the documents are expected in the low dimensional semantic subspace in which the correlations between the document points among the nearest neighbors are preserved, this principle can be called as “correlation preserving indexing.”

2. The document vectors are assigned into the singular value decomposition subspace by throw away the zero singular values. The representation of the singular value decomposition is as \( X = U \Sigma V^T \). All the zero singular values in \( \Sigma \) are eliminated. The vectors in \( U \) and \( V \) related to the zero singular values have been eliminated. Therefore the document vectors in the SVD subspace can be accomplished by \( \hat{X} = U^T X \).

3. Compute CPI Projection. Based on the multipliers \( \lambda_1, \lambda_2, \ldots, \lambda_n \) is obtained. one can compute the matrix \( M = \lambda_0 M_T + \lambda_1 x_1 x_1^T + \ldots + \lambda_n x_n x_n^T \). Let \( W_{CPI} \) be the solution of the generalized Eigen value problem \( M, W = \lambda MW \). Then, the low dimensional representation of the document can be computed by,

\[
y = W_{CPI}^T \hat{X} = W^T X
\]

where, \( W = UW_{CPI} \) is the transformation matrix.

4. In the correlation preserving indexing semantic subspace to cluster the documents. While the documents were projected on the unit hyper sphere, the inner product is a natural calculation of similarity. A separation \( \left\{ \pi_j \right\}_{j=1} \) of the document can be examined by utilizing the maximization of the subsequent objection function:

\[
Q\left(\left\{ \pi_j \right\}_{j=1} \right) = \frac{1}{2} \sum \sum x^T \pi_j
\]

with \( \pi_j = \frac{m_j}{\left|m_j\right|} \) where \( m_j \) represents the mean of the document vectors enclosed in the cluster \( \pi_j \).

3.3 Directed Ridge Regression

Directed ridge regression is one of the efficient methods that give relationship among the several variables. This method brings the similarity between the documents by measuring the relationship among the 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, one can recognize how the characteristic value of the dependent variable modifies when any one of the independent variables is dissimilar, whereas
the other independent variables are held permanent. The effectiveness of the document clustering is achieved by using the concept of directed ridge regression analysis. In this method it alters only the diagonal elements corresponding to the small Eigen values.

3.4 Clustering Algorithm based on DRR

The set of the given documents \( x_1, x_2, \ldots, x_n \in R^n \). Let \( X \) indicates the document matrix. For document clustering based on DRR, the algorithm can be summarizing as follows:

1. Generate the local neighbor patch, and estimate the matrices \( M_s \) and \( M_t \).
2. The document vectors are assigned into the singular value decomposition subspace by throw away the zero singular values. The representation of the singular value decomposition is as \( X = USVT \). All the zero singular values in \( \Sigma \) are eliminated. The vectors in \( U \) and \( V \) related to the zero singular values have been eliminated. Therefore the document vectors in the SVD subspace can be accomplished by \( \hat{X} = UTX \).
3. After that compute the directed ridge estimator based on the relationship between the Eigen values \( U \) and variance \( \alpha_i \). The computation of directed ridge estimator,

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

\[
W = \hat{\alpha}(dk)^{+}\hat{X}
\]

where, \( K \) is the diagonal matrix.

4. Compute CPI Projection. Based on the multipliers \( \lambda_1, \lambda_2, \ldots, \lambda_n \) is obtained, one can compute the matrix

\[
M = \lambda_0M_T + \lambda_i x_i x_i^T + \ldots + \lambda_n x_n x_n^T.
\]

Let \( W_{CPI} \) be the solution of the generalized eigenvalue problem \( M \hat{W} = \lambda \hat{W} \). Then, the low dimensional representation of the document can be computed by

\[
Y = W_{CPI}^T \hat{X} = W^T \hat{X}
\]

where, \( W = UW_{CPI} \) is the transformation matrix.

5. 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 \( \left\{ \Pi_j \right\}_{j=1}^k \) of the document is used with the maximization of the following objection function:

\[
Q\left(\left\{ \Pi_j \right\}_{j=1}^k \right) = \sum_{j=1}^k \sum_{x \in \pi_j} x^T c_j
\]

with \( c_j = c = \frac{m_j}{\| m_j \|} \), where \( m_j \) is the mean of the document vectors contained in the cluster \( \Pi_j \).

4. Experimental results

The Correlation Preserving Indexing and Directed Ridge Regression are compared. The experimental results are attained when the number of nearest neighbors is set to seven or eight. The corresponding results of the correlation preserving indexing and the directed ridge regression are measured for both accuracy and normalized mutual information and the details are shown in Figures 1 and 2.

4.1 Accuracy

The accuracy metric (AC) is used to measure the clustering performance. The AC metric is defined as

\[
\text{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{FP}}
\]
where \( r_i \) is the cluster label attained by the algorithm, \( s_i \) is label offered by the corpus, \( \delta(x, y) \) is the delta function that equals one if \( x = y \) and equals zero otherwise, \( \text{map}(r_i) \) is permutation mapping function that maps cluster label \( r_i \) to the equivalent label from the data corpus.

### 4.2 Normalized Mutual Information

The normalized mutual information (MI) is defined as

\[
\overline{\text{MI}}_{(C,C')} = \frac{\text{MI}(C, C')}{\max(H(C), H(C'))}
\]

where \( C \) is the set of clusters provided by the document corpus and \( C' \) is the set of clusters obtained by the algorithm. \( H(C) \) and \( H(C') \) are the entropies of \( C \) and \( C' \) respectively. \( \text{MI} \) is the mutual information corresponding to the matrices \( C \) and \( C' \)

\[
\text{MI}(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \log \frac{p(c_i, c'_j)}{p(c_i) p(c'_j)}
\]

Here \( p(c_i) \) is the probability that a document arbitrarily chosen from the corpus belongs to the clusters \( c_i \) and \( p(c_i, c'_j) \) is the joint probability that the arbitrarily chosen document belongs to the clusters \( c_i \) and \( c'_j \) at the same time. It is simple to ensure that \( \overline{\text{MI}}(C, C') \) takes zero when the two sets are independent, and takes one when the two sets are equal.

### 5. Conclusion

Correlation preserving indexing is one of the innovative approaches for document clustering. In this method the manifold structure is entrenched in the similarities between the documents. Proposed system introduces the similarity between the documents by measuring the relationship among the variables. A directed ridge regression method is one of the effective methods in which capable to predict two or more variables which are highly correlated. An efficient document clustering can be attained by using the directed ridge regression method. Hierarchical fuzzy document clustering can be achieved using a similarity measure of the vectors representing documents.

### 6. References

Fuzzy Relational Spectral Clustering Method for Document Clustering

R.Nagaraj¹, Dr.V.Thiagarasu²

¹Research Scholar, Department of Computer Science, Karpagam University, Coimbatore, Tamil Nadu, India
²Associate Professor in Computer Science, Gobi Arts and Science College, Gobichettipalayam, Tamil Nadu, 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 unified framework for probabilistic model-based clustering, which permits one to understand and compare a vast range of model-based partitioning 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,

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

The correlation corresponds to an angle $\theta$ such that $\cos \theta = \text{Corr} (u, v)$. 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 and 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 \( i = 1, 2, \ldots, n \). Then the above assumption (A1) and (A2) can be followed as,

\[
\begin{align*}
\text{Max} & \sum_i \sum_{x \in M(x_i)} \text{Corr}(y_{i,x}y_j) \\
\text{Min} & \sum_i \sum_{x \in N(x_i)} \text{Corr}(y_{i,x}y_j)
\end{align*}
\]

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:

\[
d(x, y) = \alpha \cdot \cos(x, y)
\]

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

**Clustering Algorithm Based on CPI**

For a set of documents \( x_1, x_2, \ldots, x_n \in IR^d \). 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,

   \[
   \begin{align*}
   M_T & = \sum_i \sum_{x \in \nu(x)} \{ (x_i, x_j) \}, \\
   M_S & = \sum_i \sum_{x \in N(x_i)} \{ (x_i, x_j) \}
   \end{align*}
   \]

   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 in 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 \( \Sigma \) 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,

\[
\hat{X} = U^T X
\]

3. The CPI projection is computed. Based on the multipliers \( \lambda_1, \lambda_2, \ldots, \lambda_n \) is obtained. One can compute the matrix \( M = \lambda_1 M_T + \lambda_2 x_1 x_1^T + \cdots + \lambda_n x_n x_n^T \). Let \( W_{CPI} \) be the solution of the generalized Eigen value problem \( M y \propto \lambda M W \). Then, the low dimensional representation of the document can be computed by,

\[
Y = W_{CPI}^T \hat{X} = W^T X
\]

\( W = U W_{CPI} \) 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:

\[
Q(\{\pi_j\}_{j=1}^k) = \sum_j \sum_{x \in \pi_j} x^T c_j
\]

**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 \( \Sigma \). The computation of directed ridge estimator,

\[
\hat{\beta}(d(k)) = (\Lambda + d(k)I)^{-1} \hat{X} U
\]

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.,
\[
v_{ij} = \max_{x \in \mathcal{F}_i} \text{sim}(w_j, x).
\]
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:
\[
\text{Sim}(w_1, w_2) = \frac{1}{IC(w_1) + IC(w_2) - 2 \times IC(\text{LCS}(w_1, w_2))}
\]
Where LCS \( \text{LCS}(w_1, w_2) \) 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 \( x_1, x_2, \ldots, x_n \in \mathbb{R} \).

1. Given a set of data points require to partition into k clusters \( X = \{X_1, X_2, \ldots, X_n\} \)
2. Find the similarity values by using the \( s_{ij} = \exp \left( -\frac{1}{2\sigma^2} d^2(x_i, x_j) \right) \) where \( i = 1, 2, \ldots, N; j = 1, 2, \ldots, N \) where \( s_{ij} \) is the similarity between the sentences \( i \) and \( j \).
3. Form the similarity matrix \( W \) is defined as:
\[
W_{ij} = \exp \left( -\frac{1}{2\sigma^2} d^2(x_i, x_j) \right)
\]
4. To construct the Laplacian matrix:
\[
L = D - W
\]
5. 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:
\[
U = [u_1, \ldots, u_k] \in \mathbb{R}^{n \times k}
\]
6. Form the matrix \( Y \) from \( U \) by normalizing each of \( U \)'s rows to have unit length:
\[
Y_{ij} = \frac{U_{ij}}{\left( \sum_{j=1}^{n} U_{ij}^2 \right)^{\frac{1}{2}}}
\]
7. Treat each row of \( Y \) as a point in \( \mathbb{R}^k \) and classify them into \( k \) classes through the fuzzy relational algorithm.
8. Initialize membership value \( U = [u_{ij}] \) matrix: \( U^T(\Theta) \)
9. At \( K \)-step: calculate the centers vectors \( C^{\Theta} = \{c_j\} \) with \( U^{\Theta} \)
\[
c_j = \frac{\sum_{i=1}^{n} u_{ij}^m \cdot x_i}{\sum_{i=1}^{n} u_{ij}^m}
\]
10. If \( ||U^T((k+1)) - U^T(k)|| < \epsilon \) then STOP; otherwise return to step 2.
11. Assign the original points \( x_i \) 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, 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.
\[
\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}
\]

**Recall**
Recall value is calculated is based on the retrieval of information at true positive prediction, false negative.
\[
\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}
\]

The Figure 1 shows that when compared to CPI and DRR the precision is improved in SCFSC.
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 F1 score reaches its best value at 1 and worst score at 0.

\[
F_{\text{measure}} = \frac{2 \times \text{Precision} \times \text{recall}}{(\text{Precision} + \text{recall})}
\]

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,

\[
\text{Accuracy} = \frac{\text{True positive + True negative}}{\text{True positive + True negative + False positive + False negative}}
\]

The corresponding results of 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**


A novel semantic level text classification by combining NLP and Thesaurus concepts

R. Nagaraj¹, Dr. V. Thiagarasu², P. Vijayakumar³
Research Scholar, Karpatham University, Coimbatore, India
Associate Professor of Computer Science, Gobi Arts and Science College, Gobichettipalayam, India
MPhil Scholar, Kaamadhenu Arts and Science College, Sathyamangalam, India

Abstract: Text categorization (also known as text classification or topic spotting) is the task of automatically sorting a set of documents into categories from a predefined set. Automated text classification is attractive because it frees organizations from the need of manually organizing document bases, but it can be too expensive or simply not feasible given the time constraints of the application or the number of documents involved. In the previous approaches only the Wikipedia concepts related to terms in syntactic level are used to represent document in semantic level. This paper proposes a new approach to represent semantic level with the use of Word Net. The semantic weight of terms related to the concepts from Wikipedia and Word Net are used to represent semantic information. The semantic vector space model of terms by combining the Word Net and Wikipedia is being further improved the classification accuracy of the Text classification. Because of, two different concept extractor are gives the concepts related to the terms in the syntactic level o find the better concept vector space for documents. So we obtain the improved classification by using this approach. In this study the classification framework are presented. In classification framework, the primary information is effectively kept and the noise is reduced by compressing the original information, so that this framework can guarantee the quality of the input of all classifiers. This proposed method can help to further improve the performance of classification framework by introducing Wikipedia with Word Net. We find that the proposed approach result in a high classification accuracy.

Keywords: Text classification, vector space model, Wikipedia, Word Net.

1. Introduction

Data mining is the way to help organization make full use of the data stored in their databases and when it comes to decision making, this is true in all fields and all different types of organizations. Data mining is the task of discovering interesting and hidden patterns from large amounts of data where the data can be stored in databases, data warehouses, OLAP (online analytical process) or other repository information. It is also defined as knowledge discovery in databases (KDD).

Text mining, roughly equivalent to text analytics and it refers to the process of deriving high-quality information from text. Types of text mining tasks include text clustering, text categorization, concept/entity extraction, sentiment analysis, production of granular taxonomies, document summarization, and entity relation modeling. Text analysis involves information retrieval, to study word frequency distributions lexical analysis, pattern recognition, information extraction, tagging/annotation, data mining techniques including link and association analysis, predictive analytics and visualization. The main goal is, to turn text into data for analysis, through application of natural language processing (NLP) and analytical methods. Text analytics software can help by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analyzed with traditional data mining techniques. An organization can successfully use text analytics to gain insight into content-specific values such as intensity, sentiment, emotion, and relevance with an iterative approach. Reason for that text analytics technology is still considered to be an emerging technology.

Apart from manual classification and hand-crafted rules, there is a third approach to text classification called machine learning-based text classification. In ML (machine learning), the set of rules or, more commonly, the decision criterion of the text classifier, is acquired automatically from training data. This technique is also called statistical text classification if the learning method is statistical. In statistical text classification, we need a number of good example documents (or training documents) for each class. The necessary for manual classification is not eliminated because the training documents come from a person who has labeled them - where labeling refers to the process of annotating each document. But labeling is arguably an easier task than writing rules. Mostly anybody can look at a document and decide whether or not it is related to China. Formerly such labeling is already implicitly part of an existing workflow. For instance, you may go through the news articles returned by a standing query each morning and give relevance feedback by moving the relevant articles to a special folder like multicore-processors.
II. Related Works

VSM (Vector Space Model) is the most popular document representation model for text clustering, classification and information retrieval. In early literature, term-based VSM, representing one document as a term vector, was widely used. The weight of each term in a document is usually measured via two schemes: Binary (1 for term appearing in the document, 0 for not) and Term Frequency-Inverse Document Frequency (TF-IDF). However, both approaches only contain the literal information in document. Some methods were proposed to mine the underlying semantic structure in textual data, such as Latent Dirichlet Allocation (LDA) in Blei, Ng, & Jordan[1] and Latent Semantic Indexing (LSI) in Deerwester, Dumais, Landauer, Furnas, & Harshman[2]; Hotho et al. [3] took the synonyms in Wordnet of each term as the related concepts. Although empirical results have shown this method was efficient in some cases, Wordnet is manually built and its coverage is far too restricted. Thus, many researches began to make use of Wikipedia, the largest electronic encyclopedia to date. Nouali & Blache [4]. To some extent, these methods make up for the shortage of term-based VSM, but they cannot discover as much semantic information as described in text data only by analyzing syntactic information via statistic methods. Syed, Finin, and Joshi[5] was interested in finding semantically related concepts which were also common to a set of documents.

In Wang, Hu, Zeng, Chen, and Chen [6] constructed an informative thesaurus from Wikipedia so that the synonymy, polysemy, hyponymy, and associative relations between concepts can be explicitly derived. But they rely on an exact phrase matching strategy while this strategy is limited by the terms appearing in the documents and the coverage of Wikipedia concepts or article titles. Concept similarity matrix was measured by taking account of synonyms, hyponyms and associative concepts in Wikipedia. However, these methods do not use the contextual semantic relatedness to change the concept weight. In the existing paper, concept weight is effected by the semantic relatedness between concept and the given document, which is equal to the average relatedness between concept and other concepts (contextual concepts) within the document. Here, the semantic relatedness measure between concepts also adopted link-based concept relatedness method Milne and Witten [7], Medelyan, Witten, and Milne [8]. In Huang et al. [9] compared three models (concept-based VSM, Term + Concept VSM and Replaced VSM) with term-based VSM. In the experiments, they used the WordNet and Wikipedia as the background knowledge bases respectively. Experimental results showed that Term + Concept VSM usually can improve successfully the performance in text clustering and concept-based VSM did not perform better than term-based VSM in most cases. These observations gave us a hint: concept-based VSM can supply more information for discriminating documents, but only using concepts cannot represent document sufficiently. Concept mapping could result in loss of information or addition of noise. It is necessary to include both term and concept in representation documents. In order to make use of term and concept information in text classification and clustering tasks, an alternative method is to liner combining the similarity values which are calculated based on term-based VSM and concept-based VSM respectively. However, as shown in the literatures, this method depends on the input parameters.

In Hu et al.[12] built document-concept matrix through exact-match and relatedness-match which requires to compute the tf-idf value of term in the whole Wikipedia article collection. In Gabrilovich and Markovitch [13], [14],[15] used machine learning techniques to map document to the most relevant concepts in ODP or Wikipedia by comparing the textual overlap between each document and article. However, its feature generation procedure requires high processing efforts, because each document needs to be scanned multiple times. Besides, it produced too many Wikipedia concepts for each document and filtering step further increases the processing time. Besides identifying the related concepts, weighting the concepts is also a vital technology to build concept-based VSM. It is time consuming. Banerjee, Ramanathan, and Gupta [16] treated the entire document as query strings to Wikipedia and associate the document with the top articles in the returned result list. Due to the limited background knowledge and concept mapping technology, extracted concepts might not contain the term information exactly and completely. Many Researchers began to use both term and concept information to represent document, for instance, Term + Concept VSM and Replaced VSM. The Replaced VSM represents document with concepts and terms which do not have any related concept in knowledge base of Wang et al.[17].
2.1 Introduction to Word Net

The lexical database Word Net is particularly well suited for similarity measures, because it organizes nouns and verbs into hierarchies of is–a relations. In version 2.0, there are nine noun hierarchies that include 80,000 concepts, and 554 verb hierarchies that are made up of 13,500 concepts. Is–a relations in WordNet do not cross part of speech boundaries, so Word Net–based similarity measures are limited to making judgments between noun pairs (e.g., cat and dog) and verb pairs (e.g., run and walk). While WordNet includes adjectives and adverbs, these are not organized into is–a hierarchies so similarity measures cannot be applied. However, concepts can be related in many ways beyond being similar to each other. For example, a wheel is a part of a car, night is the conflicting to day, snow is made up of water, a knife is used to cut bread, and so forth. As such Word- Net provides additional (non–hierarchical) relations such as has–part, is–made–of, is–an–attribute–of, etc. In addition, each concept (or word sense) is described by a short written definition or gloss. Measures of relatedness are based on these additional sources of information, and as such can be applied to a wider range of concept pairs. For example, they can cross part of speech boundaries and assess the degree to which the verb murder and the noun gun are related. They can even measure the relatedness of concepts that do not reside in any is–a hierarch, such as the adjectives violent and harmful.

As Pucher [18] has shown different Word Net- based measures and contexts are best for word prediction in conversational speech. The LESK measure performs best for nouns using the noun-context. The LDC measure performs best for verbs and adjectives using a mixed word-context. In Demetriou et al.,[19] generated N-best lists from phoneme confusion data acquired from a speech recognizer, and a pronunciation lexicon. Then sentence hypotheses of varying Word-Error-Rate (WER) were generated based on sentences from different genres from the British National Corpus (BNC). It was shown by them that the semantic model can improve recognition, where the amount of improvement varies with context length and sentence length. Thereby it was shown that these models can make use of long-term information. Most of the work dealing with relatedness and similarity measures has been developed using WordNet. While WordNet represents a well structured taxonomy organized in a meaningful way, questions arise about the need for a larger coverage. E.g., WordNet 2.1 does not include information about named entities such as Condoleezza Rice, Salvador Allende or The Rolling Stones as well as specialized concepts such as exocytosis or P450.

2.2 Introduction to Wikipedia

In the English version, as of 14 February 2006, contains 971,518 articles with 18.4 million internal hyperlinks, thus providing a large coverage knowledge resource developed by a large community, which is very attractive for information extraction applications [20]. Also, it provides also taxonomy by means of its categories: articles can be assigned one or more categories, which are further categorized to provide a category tree. In practice, the taxonomy is not designed as a strict hierarchy or tree of categories, but allows multiple categorization schemes to co-exist simultaneously. As of January 2006, 94% of the articles have been categorized into 91,502 categories. The strength of Wikipedia lies in its size, which could be used to overcome current knowledge bases' limited coverage and scalability issues. Such size represents on the other hand a challenge: the search space in the Wikipedia category graph is very large in terms of depth, branching factor and multiple inheritance relations, which creates problems related to finding efficient mining methods.

In addition, the category relations in Wikipedia cannot only be interpreted as corresponding to is–a links in taxonomy since they denote meronymic relations as well. As an example, the Wikipedia page for the Nigerian musician Fela Kuti belongs not only to the categories MUSICAL ACTIVISTS and SAXOPHONISTS (is–a) but also to the 1938 BIRTHS (has-property) [21]. This is due to the fact that, rather than being a well-structured taxonomy, the Wikipedia category tree is an example of a folksonomy, namely a collaborative tagging system that enables the users to categorize the content of the encyclopedic entries. Folksonomies as such do not strive for correct conceptualization in contrast to systematically engineered ontologies. They rather achieve it by collaborative approximation.

III. Semantic Level Text Classification By Thesaurus Concepts

Two-level Representation Model (2RM) that represents syntactic information and semantic information with two levels. Term-based VSM and tf-idf weighting scheme are used in syntactic level to record the syntactic information. Semantic level consists of Wikipedia concepts related to the terms in the syntactic level. These two levels are connected via the semantic correlation between terms and their relevant concepts. The key technique to build 2RM model is to construct the semantic level 2RM represents document in a two-level vector space containing syntactic (term) and semantic (related concept) information respectively.
3.1 Two-level representation model

In this section, Two-level Representation Model (2RM) that represents syntactic information and semantic information with two levels. Term-based VSM and tf-idf weighting scheme are used in syntactic level to record the syntactic information. Semantic level consists of Wikipedia concepts related to the terms in the syntactic level. These two levels are connected via the semantic correlation between terms and their relevant concepts. The key technique to build 2RM model is to construct the semantic level. In this paper, a context-based method is proposed to find the most relevant concept for each term based on the document structure information (e.g., document-paragraph) and Wikipedia link structure.

The semantic relatedness between term and its candidate concepts in a given document is computed according to the context information as follows (1).

$$\text{Rel}(t, c_i|d_j) = \frac{1}{|T| - 1} \sum_{|c_i|} \sum \text{SIM}(c_i, c_j)$$  \hspace{1cm} (1)

where $T$ is the term set of the $j$th document $d_j$, $t_i$ is a term in $d_j$ except for $t$ and $c_{s1}$ is the candidate concept set related to term $t$. $\text{SIM}(c_i, c_j)$ is the semantic relatedness between two concepts, which is calculated with the Wikipedia hyperlinks

$$\text{SIM}(c_i, c_j) = 1 - \frac{\log \max(|A||B|) - \log(|A||B|)}{\log(|W|) - \log(\min(|A||B|))}$$  \hspace{1cm} (2)

where $A$ and $B$ are the sets of all articles that link to concepts $c_i$ and $c_j$, respectively, and $W$ is the set of all articles in Wikipedia. The equation (2) is based on term occurrences on Wikipedia-pages. Pages that contain both terms indicate relatedness, while pages with only one of the terms suggest the opposite. Higher value of $\text{Rel}(t, c_i|d_j)$ means that concept $c_i$ is more semantically related to term $t$, because $c_i$ is much more similar to the relevant concepts of other terms in $d_j$ (such terms are the context of term $t$). The concepts with highest relatedness will be used to properly build the concept vector in semantic level, i.e., each term will be finally mapped into its most related concept. Based on $\text{Rel}(t, c_i|d_j)$ and term’s weight $w(t_i, d_j)$, the concept’s weight is defined as their weighted sum as follows (3).

$$W(c_i, d_j) = \sum w(t_i, d_j) * \text{Rel}(t, c_i|d_j)$$  \hspace{1cm} (3)

Different terms may be mapped to a same concept, and some term such as “dealt” has no concept in Wikipedia. Because of these many-to-one mapping, the synonym information can be considered in our proposed 2RM model. In order to deal with the second situation, some terms do not have related concept, a multi-layer classification framework is designed, to make use of term and concept information during the classification processing.

3.2 Multi-layer classification framework

In this step presents constructing the MLCLA framework. MLCLA framework includes two classification procedures in low layer; they can be implemented in series or parallel. When running in series; two data matrices based on different representation levels syntactic and semantic) can be loaded one by one. Therefore, the required memory space depends on the larger matrix plus compressed representation matrix, rather than the summation of term-based matrix and concept-based matrix. On the other hand, when running in parallel, two classifiers in low layer can be built at the same time, and the classifier in high layer is very fast on the basis of low dimension compression space. Now further analyze the time complexity of MLCLA. $N$ represents the number of documents, $M$ denotes the number of terms or concepts in document collection, $K$ is the number of classes and $m$ is the average number of terms or concepts in one document. The low layer of MLCLA includes two classification procedures, based on syntactic level and semantic level respectively.

In the low layer, the first classifier is trained and tested using the documents which are represented by term-based VSM, i.e., the syntactic information in 2RM model. According to the truth labels of training set and the predicted labels of test set of the first classifier, the center of each class can be determined by averaging the document vectors belonging to that class.

$$Z_k = \frac{\sum d_j}{|c_k|}$$  \hspace{1cm} (4)

where $|c_k|$ is the number of documents in the $k$th class $c_k$. Based on the class centers, each document can be represented with a $K$ dimension compressed vector $[S_{j1}, ..., S_{jk}]$ ($K$ equals to the number of classes) where the value of the $k$th element is the similarity between document and the $k$th class center.

$$S_{jk} = \frac{|d_j| * Z_k}{|d_j||Z_k|}$$  \hspace{1cm} (5)

Similarly, the second classifier is applied on the concept-based VSM, i.e., the semantic information in 2RM model, to get the second $K$ dimension compressed vector $[S'_{j1}, ..., S'_{jk}]$ for each document. Then, two $K$-dimension compressed vectors are combined as follows (6).

$$d_j = [S_{j1}, ..., S_{jk}, S'_{j1}, ..., S'_{jk}]$$  \hspace{1cm} (6)

$S_{jk}$ is the similarity between the $j$th document represented in syntactic level of the 2RM model and the $k$th class center obtained by the first classifier. $S'_{jk}$ is the similarity between the $j$th document represented in
A novel semantic level text classification by combining NLP and Thesaurus concepts

In MLCLA framework, the primary information is effectively kept and the noise is reduced by compressing the original information, so that MLCLA can guarantee the quality of the input of all classifiers. Thus we believe the final classification performance would be improved. Because MLCLA framework includes two classification procedures in low layer, they can be implemented in series or parallel. When running in series, two data matrices based on different representation levels (syntactic and semantic) can be loaded one by one. Therefore, the required memory space depends on the larger matrix plus compressed representation matrix, rather than the summation of term-based matrix and concept-based matrix. On the other hand, when running in parallel, two classifiers in low layer can be built at the same time, and the classifier in high layer is very fast on the basis of low dimension compression space.

IV. A Novel Semantic Level Text Classification By Combining NLP And Thesaurus Concepts

In this section introduces a measure of relateness based on formulation of information content, which is a value that is assigned to each concept in a hierarchy based on evidence found in a corpus. Before describing this measure of relatedness we first introduce the notion of information content, which is simply a measure of the specificity of a concept. A concept with a high information content is very specific to a particular topic, when concepts with lower information content are associated with more general and less specific concepts. Thus, carving fork has a high information content while entity has low information content.

Information content of a concept is estimated by counting the frequency of that concept in a large corpus and thereby determining its probability via a maximum likelihood estimate. According to this, the negative log of this probability determines the information content of the concept(7):

\[
IC(\text{concept}) = -\log(P(\text{concept}))
\]  

(7)

If sense-tagged text is available, we can be attained the frequency counts of concepts directly, since each concept will be associated with a unique sense. If sense-tagged text is not available it will be necessary to adopt an alternative counting scheme. In this technique counting the number of occurrences of a word type in a corpus, and then by using the number of different concepts/senses associated with that word, dividing that count. This value is then assigned to each concept.

Fig 1. Overall architecture diagram

For example, suppose that the word type bank occurs 20 times in a corpus and there are two concepts associated with this type in the hierarchy, one for river bank and the other for financial bank. Each of these concepts would receive a count of 10. If the occurrences of bank were sense tagged then the relevant counts could simply be assigned to the appropriate concept. In this method we choose to assign the total count to all the concepts and not divide by the number of possible concepts. Thus we would assign 20 to river bank and financial bank in the example above. This decision was based on the observation that by distributing the
frequency count over all the concepts associated with a word type we effectively assign a higher relative frequency to those words having fewer senses. This would lead us to estimate higher probability and therefore assign a lower value of information content to such concepts.

Regardless of how they are counted, the frequency of a concept includes the frequency of all its subordinate concepts since the count we add to a concept is added to its subsuming concept as well. Note that the counts of more specific concepts are added to the more general concepts, which is not from the more general to specific. Thus, counts of more specific concepts percolate up to the top of the hierarchy, incrementing the counts of the more general concepts as they proceed upward. As a result, concepts that are higher up in the hierarchy will have higher counts than those at lower more specific levels and have higher probabilities associated with them. Such high probability concepts will have low values of information content since they are associated with more general concepts.

This measure of semantic similarity uses the information content of concepts along with their positions in the noun is – a hierarchies of Word Net to compute a value for the semantic relatedness of the concepts. The principle idea behind this measure of semantic relatedness is that two concepts are semantically related proportional to the amount of information they share in common. The quantity of information common to two concepts is determined by the information content of the lowest concepts in the hierarchy that subsumes both the concepts. This concepts is known as the lowest common subsumer of the two concepts. Thus, the measure of similarity is defined as follows (8):

$$\text{SIM}_{res}(c_1, c_2) = IC(lcs(c_1, c_2))$$ (8)

We note that this measure does not consider the information content of the concept themselves, nor does it directly consider the path length.

$$\text{SIM}^*(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log(P(c))]$$ (9)

Where S(c_i, c_k) is the set of concepts that subsume both c_i and c_k. Notice that although similarity is computed by considering all upper bounds for the two concepts but the information measure has the effect of identifying minimal upper bounds, because no class is less informative than its superordinates.

The semantic relatedness between term and its candidate concepts in a given document is computed according to the context information as follows (10).

$$\text{Rel}(t, c_j(d)) = \frac{1}{1 - \frac{1}{|T|} \sum \frac{1}{|S|} \sum \text{SIM}(c_i, c_k) \text{SIM}^*(c_i, c_k)}$$ (10)

where T is the term set of the jth document d_j, t_i is a term in d_j except for t and c_0 is the candidate concept set related to term t_i. \(\text{SIM}^*(c_i, c_k)\) is the semantic relatedness between two concepts, which is calculated with the Wikipedia hyperlinks and \(\text{SIM}^*(c_i, c_k)\) is the semantic relatedness between two terms, which is calculated with the WordNet hyperlinks.

$$\text{SIM}(c_i, c_k) = 1 - \frac{\log(|A|) - \log(|B|)}{\log(|W|) - \log(\min(|A|,|B|))}$$ (11)

where A and B are the sets of all articles that link to concepts c_i and c_k respectively, and W is the set of all articles in Wikipedia. The equation (11) is based on term occurrences on Wikipedia-pages. Pages that contain both terms indicate relatedness, while pages with only one of the terms suggest the opposite. Higher value of \(\text{Rel}(t, c_j(d))\) means that concept c_i is more semantically related to term t, because c_i is much more similar to the relevant concepts of other terms in d_j (such terms are the context of term t). The concepts with highest relatedness will be used to properly build the concept vector in semantic level, i.e., each term will be finally mapped into its most related concept. Based on \(\text{Rel}(t, c_j(d))\) and term’s weight w(t_i, d_j), the concept’s weight is defined as their weighted sum as follows (12).

$$W(c_i, d_j) = \sum w(t_i, d_j) * \text{Rel}(t_i, c_i)$$ (12)

Different terms may be mapped to a same concept, and some term such as “dealt” has no concept in Wikipedia. Because of these many-to-one mapping, the synonym information can be considered in our proposed 2RM model. In order to deal with the second situation, some terms do not have related concept, a multi-layer classification framework is designed, to make use of term and concept information during the classification processing.

4.1 Classification framework

In this step we are constructing the classification framework. This classification framework includes two classification procedures in low layer; they can be implemented in series or parallel. When running in series; two data matrices based on different representation levels syntactic and semantic) can be loaded one by one. Therefore, the required memory space depends on the larger matrix plus compressed representation matrix, rather than the summation of term-based matrix and concept-based matrix. On the other hand, when running in parallel, two classifiers in low layer can be built at the same time, and the classifier in high layer is very fast on the basis of low dimension compression space. Now we further analyze the time complexity of classification framework. N represents the number of documents, M denotes the number of terms or concepts in document collection, Q is the number of classes and m is the average number of terms or concepts in one document. The
low layer of this framework includes two classification procedures, based on syntactic level and semantic level respectively.

In the low layer, the first classifier is trained and tested using the documents which are represented by term-based VSM, i.e., the syntactic information in 2RM model. According to the truth labels of training set and the predicted labels of test set of the first classifier, the center of each class can be determined by averaging the document vectors belonging to this class.

\[ Z_k = \frac{\sum_j d_j}{|c_k|} \]  

(13)

where \(|c_k|\) is the number of documents in the kth class \(c_k\). Based on the class centers, each document can be represented with a K dimension compressed vector \([S_{j1}, \ldots, S_{jk}]\) (K equals to the number of classes) where the value of the kth element is the similarity between document and the kth class center.

\[ S_{jk} = \frac{d_j \cdot Z_k}{||d_j|| \cdot ||Z_k||} \]  

(14)

Similarly, the second classifier is applied on the concept-based VSM, i.e., the semantic information in 2RM model, to get the second K dimension compressed vector \([S'_{j1}, \ldots, S'_{jk}]\) for each document. Then, two K-dimension compressed vectors are combined as follows (15).

\[ S_{jk} = [S_{j1}, \ldots, S_{jk}, S'_{j1}, \ldots, S'_{jk}] \]  

(15)

\(S_{jk}\) is the similarity between the jth document represented in syntactic level of the 2RM model and the kth class center obtained by the first classifier. \(S'_{jk}\) is the similarity between the jth document represented in semantic level of the 2RM model and the kth class center obtained by the second classifier. This combined document representation will be the input of the third classifier in the high layer of classification framework.

In classification framework, the primary information is effectively kept and the noise is reduced by compressing the original information, so that this framework can guarantee the quality of the input of all classifiers. Thus we believe the final classification performance would be improved. Because classification framework includes two classification procedures in low layer, they can be implemented in series or parallel. When running in series, two data matrices based on different representation levels (syntactic and semantic) can be loaded one by one. Therefore, the required memory space depends on the larger matrix plus compressed representation matrix, rather than the summation of term-based matrix and concept-based matrix. On the other hand, when running in parallel, two classifiers in low layer can be built at the same time, and the classifier in high layer is very fast on the basis of low dimension compression space.

V. Performance Evaluation

5.1 Data set

The proposed representation model and classification framework were tested on three real data, 20Newsgroups, Reuters-21578 and Classic3. Six subsets were extracted from 20Newsgroups: 20NGDiff4, 20NG-Sim4, 20NG-Binary, 20NG-Multi5, 20NG-Multi10 and 20NG-Long.

Table 1: 20NewsGroup subsets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>20NG-Binary</td>
<td>talk.politics.mideast, talk.politics.misc</td>
</tr>
<tr>
<td>20NG-Multi5</td>
<td>comp.graphics, rec.motorcycles, rec.sport.baseball, sci.space, talk.politics.mideast</td>
</tr>
<tr>
<td>20NG-Multi10</td>
<td>alt.atheism, comp.sys.mac.hardware, misc.forsale, rec.autos, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, talk.politics.guns</td>
</tr>
<tr>
<td>20NG-Diff4</td>
<td>comp.graphics, rec.sport.bassball, sci.space, talk.politics.mideast</td>
</tr>
<tr>
<td>20NG-Sim4</td>
<td>comp.graphics, comp.os.ms-windows.misc, rec.autos, sci.electronics</td>
</tr>
<tr>
<td>20NG-Long</td>
<td>comp/, sc/, talk/</td>
</tr>
</tbody>
</table>

Tables 1 and 2 list the categories and the number of documents contained in these subsets. In this paper, 20NG-Long is a collection of long documents containing three categories “comp”, “sci” and “talk”. In each category, 70 documents with the most large size were extracted from the corresponding topic in 20Newsgroups (documents from topic “rec” were not included because there are few long documents in “rec/”). In 20NG-long, the minimal document’s size is 10 K, the maximal one is 158 KB and the average size is 29 KB. Another two data subsets were created from Reuters-21578: R-Min20Max200 and R-Top10. R-Min20Max200 consists of 25 categories with at least 20 and at most 200 documents, 1413 documents totally. In R-Top10, 10 largest categories were extracted from the original data set including 8023 documents. For Classic3, the whole dataset was used in the experiment.

www.iosrjournals.org 20 | Page
A novel semantic level text classification by combining NLP and Thesaurus concepts

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classes</th>
<th>Documents</th>
<th>Words</th>
<th>Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>20NG-Binary</td>
<td>2</td>
<td>500</td>
<td>3376</td>
<td>2987</td>
</tr>
<tr>
<td>20NG-Multi5</td>
<td>5</td>
<td>500</td>
<td>3310</td>
<td>2735</td>
</tr>
<tr>
<td>20NG-Multi10</td>
<td>10</td>
<td>500</td>
<td>3344</td>
<td>2772</td>
</tr>
<tr>
<td>20NG-Diff4</td>
<td>4</td>
<td>4000</td>
<td>5433</td>
<td>4362</td>
</tr>
<tr>
<td>20NG-Sim4</td>
<td>4</td>
<td>4000</td>
<td>4352</td>
<td>3502</td>
</tr>
<tr>
<td>20NG-Long</td>
<td>3</td>
<td>210</td>
<td>4244</td>
<td>3737</td>
</tr>
<tr>
<td>R-Min20Max200</td>
<td>25</td>
<td>1413</td>
<td>2904</td>
<td>2450</td>
</tr>
<tr>
<td>R-Top10</td>
<td>10</td>
<td>8023</td>
<td>5146</td>
<td>4109</td>
</tr>
<tr>
<td>Classic3</td>
<td>3</td>
<td>3891</td>
<td>4745</td>
<td>3737</td>
</tr>
</tbody>
</table>

Table 2: Data set summary

For our experiment, we consider the two subsets of 20Newsgroups datasets, one subset created from Reuters-21578 and Classic3. We are taking 20NG-Multi10 subset and 20NG-Sim4 subset which is extracted from the 20Newsgroups. 20NG-Multi10 dataset consists of totally 500 documents and 10 classes. In 20NG-Multi10, 3344 words and 2772 concepts are there. As well as 20NG-Sim4 consists of 4000 documents and 4 classes. 4352 words and 3502 concepts are there in this dataset. From the Reuters-21578, we are taking the R-Top 10 subset which consists of 10 classes and 8023 documents. In this dataset, 4109 concepts and 5146 words are there. For the Classic 3, 3 classes and 3891 documents are present. This dataset consists of 4745 words and 3737 concepts. In this paper, we only consider the single-label documents. Wikipedia and WordNet are used as background knowledge. Wikipedia contains 2,388,612 articles (i.e., concepts) and 8,339,823 anchors in English and WordNet in version 2.0, there are nine noun hierarchies that include 80,000 concepts, and 554 verb hierarchies that are made by 13,500 concepts.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training set size of dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>20NG-Multi10</td>
<td>0.92 0.999 0.903 0.895 0.883 0.88</td>
</tr>
<tr>
<td>20NG-Sim4</td>
<td>0.91 0.88 0.873 0.861 0.85 0.85</td>
</tr>
<tr>
<td>R-Top10</td>
<td>0.92 0.85 0.79 0.79 0.69 0.73</td>
</tr>
<tr>
<td>Classic3</td>
<td>0.98 0.93 0.88 0.86 0.8 0.8</td>
</tr>
</tbody>
</table>

Table 3: F-measure result from the experiment of the dataset

From Table 2 we can see the number of words and concepts extracted from each data set. The words were extracted by preprocessing steps, selecting only alphabetical sequences, stems them, removing stop words and filtering them by the document frequency. Then, we determined the Wikipedia and WordNet concepts for these words in each document via the method (Note: once a word was stemmed, its original form was used to correctly identify relevant Wikipedia and WordNet concept). Table 2 shows that the number of distinct concepts appearing in a data set is usually lower than the number of words. Meanwhile, parts of words (about 10 percent) do not have relevant concepts. The main one is to test the performance of proposed 2RM model and classification framework on real datasets by comparing with various flat document representation models plus basic classification algorithm (e.g., SVM or KNN). Table 3 shows the F-measure result from the experiment of the dataset. It shows the F-measure is improved in the proposed system compared to the existing system.

5.2 Accuracy comparison

In this section, performance is evaluated in terms of accuracy. In this graph, we have taken the parameters called accuracy and training set size of four datasets namely 20NG-Multi10, 20NG-Sim4, R-Top10 and Classic3. It helps to analyze the existing system and proposed combining technique. Accuracy can be calculated from the formula given as follows

\[
\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}}
\]
A novel semantic level text classification by combining NLP and Thesaurus concepts

Fig 2. Accuracy comparison for 20NG-Multi10

Fig 3. Accuracy comparison for 20NG-Sim4

Fig 4. Accuracy comparison for R-Top10

Fig 5. Accuracy comparison for Classic3
The Accuracy parameter will be the Y axis and training set size of dataset will be the X axis. Then we compare the accuracy performance. From this graph we identify that the accuracy of the proposed system is higher than the existing system. From this we easily understood the proposed system has more effective than exiting one.

5.3 Precision comparison

In this section performance is evaluated in terms of precision. Graph gives the precision comparison between the existing and proposed. It can be defined as

\[
\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}
\]

![Fig 6. Precision comparison for 20NG-Multi10](image)

![Fig 7. Precision comparison for 20NG-Sim4](image)

![Fig 8. Precision comparison for R-Top10](image)
A novel semantic level text classification by combining NLP and Thesaurus concepts

Fig 9. Precision comparison for Classic3

In the graph X-axis will be training set size of dataset of four dataset namely 20NG-Multi10, 20NG-Sim4, R-Top10 and Classic3 and Y-axis will be precision parameter. In the data sets also our proposed system has more precision compare to existing system. From this graph, proposed paper has effective in precision parameter.

5.4 F-measure comparison

F-measure distinguishes the correct classification of document labels within different classes. In essence, it assesses the effectiveness of the algorithm on a single class, and the higher it is, the better is the clustering. It is defined as follows:

\[ F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \]

then,

\[ F(i,j) = 2 \times \frac{P(i) \times R(i)}{P(i) + R(i)} \]

\[ F_c = \frac{\sum_{i \in C} (F(i))}{|C|} \]

where for every class i is associated a cluster j which has the highest F-measure, \( F_c \) represents the overall F-measure that is the weighted average of the F-measure for each class i and \(|i|\) is the size of the class.

Fig 10. F-measure comparison for 20NG-Multi10
A novel semantic level text classification by combining NLP and Thesaurus concepts

VI. Conclusion

In this work, we represent document as a two-level model with the aid of WordNet and Wikipedia. In the two-level representation model, one for term information, the other for concept information and these levels are connected by the semantic relatedness between terms and concepts. A context-based method is adopted to identify the relatedness between terms and concepts by utilizing the link structure among WordNet and Wikipedia articles, which is also used to select the most appropriate concept for a term in a given document. By combining the WordNet and Wikipedia, we get the improved performance of the existing paper. Based on the two-level representation model, we propose a classification framework to analyze text data. By introducing the combining technique of WordNet and Wikipedia accuracy rate will be increased and error rate is decreased observed from the experiment.

References

A novel semantic level text classification by combining NLP and Thesaurus concepts

Semantically Document Clustering Using Contextual Similarities

R.Nagaraj, Research Scholar, 
Department of Computer Science, 
Karpagam Academy of Higher Education, 
Coimbatore, Tamilnadu- 641021 
rnagaraj789@gmail.com

X.Agnise Kalarani, Professor, 
Department of Computer Applications, 
Karpagam Academy of Higher Education 
Coimbatore, Tamilnadu- 641021

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 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 (SCFHC) 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 similarities, 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
\[
W(c_i, d_j) = \sum w(t_k, d_j) \times Rel(t_k, c_i | d_j) \quad (1)
\]
// Where \(Rel(t, s | d_j)\) = semantic relatedness between term and sentence, \(\sum w(t_k, d_j) = \) term’s weight

3. Define measure of similarity
\[
SIMres(c_i, c_2) = IC(lcs(c_i, c_2)) \quad (2)
\]

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

5. Similarity between the sentences is find by using Jiang and Conrath measure,
\[
Sim(w_1, w_2) = \frac{1}{IC(|w_1| + IC(|w_2| - 2 \times IC(lcs(w_1, w_2))}} \quad (5)
\]
6. Sentence weight is computed as,
\[ W^*(d_j) = \sum w(t_k, d_j) \cdot Rel(t_k, s|d_j) \] (6)

7. Center of each class can be determined by averaging the document vectors belonging to this class
\[ S^k = \frac{d_k}{||d_k||} \] (7)

8. Similarity between document and the \( k \)th class center is determined as,
\[ c_k = \frac{d_k}{||d_k||} \] (7)

9. Compute directed ridge estimator
\[ \hat{\alpha}(d_k) = (\Lambda + KL)^{-1}X^T \] U. Hence \[ W = \hat{\alpha}(d_k)^* \] (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 \( U \) by stacking the eigenvectors in columns
\[ U = [u_1, \ldots, u_k] \in \mathbb{R}^{m \times k} \] (10)

12. Form \( Y \) matrix from \( U \) by normalizing each of \( U \)'s rows to have unit length
\[ Y_{ij} = \frac{u_{ij}}{\sum_j u_{ij}^2} \] (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
\[ c_j = \frac{\sum_{i=1}^{m} \mu_{ij}^m y_i}{\sum_{i=1}^{m} \mu_{ij}^m} \] (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
\[ SIM(c_i, c_k) = 1 - \frac{\log(\text{max}(|A|, |B|) - \log(\text{max}(|A|, |B|))}{\log(|W|) - \log(\text{min}(|A|, |B|))} \] (14)

\( \text{where A and B are the sets of all articles that link to concepts c_i and c_k and W is the set of all articles.} \)

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 variably to solve the two-variable problem. The fuzzy rules are applied to classify the rows of the normalized matrix into \( k \) classes. Using these classes, the membership of the observations is calculated along with the center vectors and updated periodically. The semantic relatedness of concepts is also calculated. Using the calculated term relatedness, sentence relatedness and concept relatedness, the clustering of the documents is performed efficiently.

**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-url)
**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.

![Precision Comparison](image)

**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.

![Recall Comparison](image)

**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.

![F-measure Comparison](image)

**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**

[1] Taiping Zhang, Yuan Yan Tang, Bin Fang, and Yong Xiang. "Document clustering in correlation similarity measure space." IEEE Transactions on Knowledge and Data


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


ENERGY EFFICIENT COOPERATIVE LOAD BALANCING AND DYNAMIC CHANNEL ALLOCATION MECHANISM IN MOBILE AD HOC NETWORKS

Nagaraj .R 1, Agnise Kala Rani .X2

ABSTRACT
An Energy Efficient -TRACE protocol (EE-TRACE) is introduced to avoid receiving a duplicate of the same packet in the mobile Adhoc network, and improve the energy efficiency. In EE-TRACE protocol, the IS packets include source ID and the packet sequence number. The informations in the IS packets are used to avoid receiving a duplicate of the same packet. An Efficient cluster head selection is done by using remaining battery and node stability. The proposed system introduced a novel MAC protocol, CDCA-TRACE, which combines dynamic channel allocation and cooperative load balancing algorithms to achieve efficient channel allocation and load balancing. The proposed method achieves high performance in terms of end to end delay, Throughput and energy consumption. The proposed EE-TRACE protocol is used to avoid duplicate of the same packet and improve energy efficiency.

Keyword : MH-TRACE protocol, Channel Allocation and load balancing.

I. INTRODUCTION
In an ad-hoc network, each node moves independently which are exchange their information with others using

1Ph.D Research Scholar, Karpagam Academy of Higher Education, Coimbatore – 641 021, Tamil Nadu. Email : nagukasc@gmail.com

2Professor, Department of Computer Applications, Karpagam Academy of Higher Education , Coimbatore-641 021, Email : agneskala72@gmail.com

multihop wireless links. Each node in the network acts as a router and forwarding the data packets to other nodes. Due to the node movement, topology of the network is varying dynamically. The MANETs have several salient characteristics such as Dynamic topologies, Bandwidth-constrained, Energy-constrained operation, Limited physical security etc [1].

For improving and automating the quality of service of the networks, proficient resource allocation methods are required. If traffic varies significantly, then resource allocated in the statical manner is insufficient or under-exploited [2]. It resource allocation is crucial for the medium access control (MAC) protocol of a MANET. It only manages bandwidth the utilization, does not adapted to dynamic environment [3] [4].

The MAC protocols for Adhoc network classified into two types [5]. They are coordinated and uncoordinated MAC protocols. In uncoordinated protocols, the nodes compete with each other to share the common channel. In uncoordinated protocols, the bandwidth is efficient when the network has low loads and the network load increases, their bandwidth efficiency is decreases.

For low network loads, these protocols are bandwidth efficient due to the lack of overhead. However, as the network load increases, their bandwidth efficiency decreases. Also, due to idle listening, these protocols are in general not energy efficient. On the other Resource allocation is carried out in a static manner on the hours to
month's scale of time in telecommunication networks. The Coordinated channel access methods offered quality of service (QoS), reduce energy dissipation, and increase throughput for large networks [6].

Vishnu Kumar Sharma et al. introduced a resource allocation technique to Mobile Ad Hoc Networks. In order to achieve quality of service the proposed system introduced an Agent based Bandwidth Reservation Routing Technique in Mobile Ad Hoc Networks. The mobile agents are forwarding the packets through minimum cost, congestion and bandwidth path. The every node status is collected in the network. The intermediate nodes are used to determine the available bandwidth on the link. After completing the new bottleneck bandwidth field updation the informations are feedback from destination to source [7]. The resource reservation technique is introduced to bandwidth reservation, if the available bandwidth is larger than the bottleneck bandwidth, then bandwidth reservation for the flow is executed. However it does not consider the load balancing.

Lin Gao et al. introduced a Multi-radio Channel Allocation scheme. In order to overcome the channel allocation problem the proposed system is introduced. The proposed min-max coalition-proof Nash equilibrium (MMCPNE) channel allocation scheme used to increase the achieved data rates of communication sessions. The min-max coalition-proof Nash equilibrium (MMCPNE) channel allocation scheme is executed in game [8]. Finally the system introduced hybrid game involving both cooperative game and non cooperative game into our system in which the players within a communication session are cooperative and among sessions, they are non-cooperative. However it only considered Channel Allocation.

Bora Karaoglu et al. introduced a Cooperative Load balancing and Dynamic Channel Allocation scheme in Cluster-Based Mobile Ad Hoc Networks. The Cooperative Load balancing and Dynamic channel allocation (DCA) algorithm executed based on spectrum sensing. In cooperative load balancing algorithm, the nodes select their channel access providers based on the availability of the resources. Dynamic Channel Allocation scheme is introduced for increases the importance of bandwidth efficiency while maintaining tight requirements on energy consumption delay and jitter. The proposed scheme is integrated into TRACE protocol. The CDCA protocol achieves higher bandwidth efficiency [9]. However it does not avoid duplicates of same packets.

II. PROPOSED SYSTEM

1. Creation of network

An undirected graph G (V, E) where the set of vertices V represent the mobile nodes in the network and E represents set of edges in the graph which represents the physical or logical links between the sensor nodes. Nodes are placed at a same level. Two nodes that can communicate directly with each other are connected by an edge in the graph. Let N denote a network of m mobile nodes, \( N_{1}, N_{2}, ..., N_{m} \) and let D denote a collection of n data items \( d_{1}, d_{2}, ..., d_{n} \) distributed in the network. For each pair of mobile nodes \( N_{i} \) and \( N_{j} \), let \( t_{ij} \) denote the delay of transmitting a data item of unit-size between these two nodes.
2. MH-TRACE protocol

In MH-TRACE, particular nodes consider as channel coordinators, which is called cluster-heads. All CHs broadcast the periodic Beacon packets for announce their attendance to the nodes in their neighborhood. If the node does not receive Beacon packet from any CH for a predefined amount of time, it assumes the role of a CH. This method guarantees the existence of at least one CH around every node in the mobile adhoc network. In this protocol time is divided into super frames of equal length. It further divided into frames. By utilizing any one of the frames in super frame, every CH operates. The CH offers channel access to nodes within its communication range. Each super frame divided into sub frames. They are control sub-frame and data sub-frame.

The control sub-frame is used for signaling among nodes and the CH, and the data sub-frame is used to transmit the data payload. In Beacon slot, the Cluster head broadcast their survival and the number of available data slots in the present frame. The CHs working in the same frame means, the interferes is occurred. The CA slot is used to determine the interferences. The mobile nodes are broadcast their channel access requests to cluster head by using Contention slots. After listening to the medium, it makes the schedule to nodes. Then CH sends a header message to nodes which containing transmission schedule. The nodes send short packets summarizing the information in IS slot. By listening to the IS packets, receiver nodes become aware of the data that are going to be sent and may choose to sleep during the corresponding data slots.

3. Dynamic Channel Allocation for TRACE

DCA-TRACE consists of two extra mechanisms. They are i) method to maintain track of the interference level from the other CHs in each frame; and ii) a method to sense the interference level from the transmitting nodes in each data slot in each frame. The interference level of the Beacon and CA slots are updated with the measured values in that frame using

$$I_{kt} = \begin{cases} M_{kt} & \text{if } I_{k,t-1} < M \\ I_{k,t-1} & \text{if } I_{k,t-1} \geq M \end{cases}$$

Where,

$$I_{k,t}$$ and $$I_{k,t-1}$$ represent the interference levels of the kth slot in the present and the previous super frame respectively.

$$M_{kt}$$ - measured interference level of the kth slot in the current super frame

$$a$$ - smoothing factor

In DCA-TRACE, CHs mark a frame as unavailable if there is another cluster that uses the frame and resides closer than a certain threshold, Thintf, measured through the high interference value of that frame. Even under high local demand, CHs refrain from accessing these frames that have high interference measurements, in order to protect the stability of the clustering structure and the existing data transmissions. At the end of each superframe, CHs determine the number of frames that they need to access, m, based on the reservations in the previous frame. Depending on the interference level of each frame, they choose the least noisy m frames that have an interference value also below a common threshold, Thintf. If the number of available frames is
less than m, the CHs operate only in the available frames. ThinIF prevents excessive interference in between co-frame clusters that can potentially destabilize the clustering structure.

4. Collaborative Load Balancing for TRACE

In DCA-TRACE protocol, once CH1 allocates all of its available slots, an additional frame is selected by using algorithm. If other frames interference level is large, additional frame selection is not possible. And also, accessing additional frames increases interference on the IS and data slots of the new frame and decreases the potential extent these packets can reach. To overcome this problem, the system introduced CMH-TRACE and CDCA-TRACE. These are add cooperative CH monitoring and reselection on top of MH-TRACE and DCA-TRACE, respectively. In this CMH-TRACE and CDCA-TRACE, the nodes always monitor the presented data slots at the cluster heads by using Beacon messages. The cooperative load balancing algorithm is triggered once all the obtainable data slots for a CH are allocated with a probability p. The cooperative load balancing algorithm is triggered by active nodes in the network. When the cooperative load balancing mechanism is triggered, the node that is presently using a data slot from the heavily loaded CH contends for data slots from other nearby CHs while maintaining and using its reserved data slot until it secures a new data slot from another CH.

The load balancing algorithm triggered for probabilistically reduces the load. An introduced Cooperative load balancing method is does not alter the clustering structure. The load cannot transfer to nearby CH while the source nodes may not be in the surrounding area of another CH. To solve this problem DCA algorithm is required.

The proposed system includes additional frame selection algorithm of DCATRACE with some delay. A fully loaded CH resets a counter, $N_{DCA} = 0$, and starts incrementing it at the beginning of each super frame while it remains fully loaded. The CH attempts to (subject to the interference levels in the frames) access an additional frame when $NDCA >= TDCA$. It offers time for the active member nodes to trigger the cooperative load balancing algorithm and transfer their load to nearby CHs. A small p value leads to a slower response time for the cooperative load balancing algorithm.

5. Energy efficient (EE - TRACE) Protocol

In EE - TRACE, we include the source ID and the packet sequence number in the IS packet, so that nodes that already received a particular data packet avoid receiving a duplicate of the same packet, which saves a considerable amount of energy. The nodes can selectively choose what data to receive based on information from the IS packets, enabling the nodes to avoid receiving multiple receptions of the same packet. In MH-TRACE, certain nodes assume the roles of channel coordinators, here called cluster-heads. All CHs send out periodic Beacon packets to announce their presence to the nodes in their neighborhood. We have to select cluster head based on residual energy and node stability which is having high value and assign them as cluster head.
Residual energy

The residual energy RE can be calculated by using the following formula

\[ RE = E_i - E_c(t) \]

Where

\( E_i \) - The initial energy of a node

\( E_c(t) \) - energy consumed by a node after time t Stability factor computation:

Nodes that are relatively more stable as compared to the others in the localities are given more favorite. The node's stability is computed as the ratio of number of link failures \((f_j)\) and new connection established \((C_j)\) per unit time to the total number of nodes surrounding that node \((n_j)\).

Therefore, stability of a node \( j \) is \( \frac{c_j f_j}{n_j} \). As the values of \( c_j \) and \( f_j \) increase, the stability of the node decreases.

III. EXPERIMENTAL RESULT

The existing CDCA-TRACE protocol and proposed EE-TRACE protocols are evaluated in terms of end to end delay, packet delivery ratio and Throughput.

Performance evaluation

1. End to end delay

End-to-end delay refers to the time taken for a packet to be transmitted across a network from source to destination.

From the above figure 3.1 can be proved that the proposed methodology provides better result than the existing approach. In this figure x axis plots the number of nodes and y axis plots the time. The time taken for a packet to be transmitted across a network from source to destination is efficient compare to existing one. The proposed algorithm EE-TRACE achieves low end to end delay compared with existing DCA.

2. Energy consumption

Power consumption is the consumption of energy or power.

Figure 3.2: Energy consumption
The Energy consumption of the network is shown in this graph. In the X-axis number of nodes is taken. Y-axis energy of the network is taken. This graph clearly shows that if the number of node is increases the energy of the network is decreased using existing DCA. The proposed algorithm high energy efficiency compared with existing DCA.

3. Throughput

It is defined as number of packets successfully delivered to the destination. From figure 3.3, x axis plots the number of nodes and y axis plots the Throughput value. This graph clearly shows that if the number of nodes is increases the throughput of the network is decreased and achieve high energy efficiency. The coordinator nodes are selected based on remaining battery and node stability. Finally an introduced CDCA-TRACE protocol combines dynamic channel allocation and cooperative load balancing algorithms to achieve efficient channel allocation and load balancing. The experimental results show that the proposed EE-TRACE protocol provides high performance compared to the existing methods in terms of throughput, energy consumption and end to end delay.

REFERENCES


ENERGY EFFICIENT COOPERATIVE LOAD BALANCING AND DYNAMIC CHANNEL ALLOCATION MECHANISM IN MOBILE AD HOC NETWORKS


AUTHOR'S BIOGRAPHY

**R. Nagaraj** is working as Asst. Professor in Department of Computer Science, Kaamadhenu Arts and Science College, Sathyamangalam, Erode, Tamilnadu, India. He has 10 years of experience in teaching. He has published many research articles in the National / International journals. He is currently pursuing doctor of programme in Computer Science at Karpagam Academy of Higher Education, Coimbatore, Tamilnadu, India. His current area of research interests Datamining.

**Dr. X.Agnise Kala Rani** is working as Professor in the Department of Computer Applications at Karpagam University, Coimbatore. She received her Ph.D. in 2011 from Mother Teresa University. She holds the Master of Engineering degree from VMKV University and Master of Computer Applications degree from Madras University. She has several publications including scientific journals and top-tier networking conferences to her credit.