2nd International Conference on Science, Technology and Management
University of Delhi (DU), Conference Centre, New Delhi (India)
27 September 2015, www.conferenceworld.in, www.icstmdu.com
978-81-931039-6-8

USING DIMENSIONALITY REDUCTION METHODS IN TEXT CLUSTERING

D.Madhavi[1], Dr. A. Mary Sowjanya[2]

[1] M.Tech. [2] Assistant Professor, Dept. of Computer Science and Systems Engineering,
A.U. College of Engineering,Andhra University, Visakhapatnam, Andhra Pradesh, (India)

ABSTRACT

High dimensionality of the feature space is one of the major concerns owing to computational complexity and accuracy consideration in the text clustering. Therefore, various dimension reduction methods have been introduced in the literature to select an informative subset (or sub list) of features. As each dimension reduction method uses a different strategy (aspect) to select a subset of features, it results in different feature sub lists for the same dataset. Hence, a hybrid approach, which encompasses different aspects of feature relevance, altogether for feature subset selection, receives considerable attention. Traditionally, union or intersection is used to merge feature sub lists selected with different methods. The union approach selects all features and the intersection approach selects only common features from considered features sub lists, which leads to increase the total number of features and loses some important features, respectively. Therefore, to take the advantage of one method and lessen the drawbacks of other, a novel integration approach namely modified union is proposed. This approach applies union on selected top ranked features and applies intersection on remaining features sub lists. Hence, it ensures selection of top ranked as well as common features without increasing dimensions in the feature space much. In this study, feature selection methods term variance (TV) and document frequency (DF) are used for features' relevance score computation. Next, a feature extraction method principal component analysis (PCA) is applied to further reduce dimensions in the feature space without losing much information. Finally clustering is performed.

***Keywords*: Document Frequency, Feature Selection, Feature Extraction, Principal Component Analysis, Text Clustering, Term Variance.**

I. INTRODUCTION

In recent years, a tremendous growth in the volume of text documents available on the Internet, digital libraries, news sources, and company-wide intranets has been witnessed. This has led to an increased interest in developing methods that can help users to effectively navigate, summarize, and organize this information with the ultimate goal of helping them to find what they are looking for. Fast and high-quality text clustering algorithms play an important role towards this goal as they have been shown to provide both an intuitive navigation or browsing mechanism by organizing large amounts of information into a small number of meaningful clusters as well as to greatly improve the retrieval performance either via cluster-driven dimensionality reduction, term-weighting, or query expansion [1].

Text clustering is an automatic way of grouping the digital documents in a form of clusters based on their intrinsic characteristics. Due to automatic and proficient processing of the digital documents, text clustering is applied to several application domains such as organization of the results returned by a search engine in response to a user's query [2], browsing large document collections, topic detection, and generating a hierarchy of web documents, Various clustering methods are k-means [3],expectation–maximization clustering, density based clustering, have been proposed in the past several years to achieve these tasks.

In text clustering, documents are traditionally represented as bag-of-words (Salton & Yang, 1975), where each distinct term present in a document collection is considered as a separate dimension (feature). Hence, a document is represented by a multi-dimensional feature vector where each dimension corresponds to a weighted value of the term within the document collection. This weighted value is computed using term frequency inverse document frequency (tfidf). As features originate from distinct terms, a corpus of even moderate-sized documents results in hundreds of thousands of dimensions. One of the most important issue in the text clustering is therefore to deal with hybrid dimensionality of feature space.

Dimensionality reduction or dimension reduction is the process of reducing the number of random variables under consideration and can be divided into feature selection and feature extraction.

The feature extraction methods transform a high dimensional feature space into a distinct low dimensional feature space through a combination or Transformation of the original feature space. Many components like Principal component analysis [4], latent semantic indexing[5],independent component analysis[6], multi dimensional scaling [7] ,and partial least square [8] are few examples of feature extraction methods. In this study, we use PCA to reduce dimensions in the feature space.

Feature selection is a process that chooses a subset from the original feature set according to some criterions. Subset selection evaluates a subset of features as a group for suitability. Subset selection algorithms can be broken up into Wrappers, Filters and Embedded. Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. Wrappers can be computationally expensive and have a risk of over fitting to the model. Filters are similar to Wrappers in the search approach, but instead of evaluating against a model, a simpler filter is evaluated. Embedded techniques are embedded in and specific to a model.

Hence, filter methods are widely used to reduce dimensions, especially when dimensions in the feature space are huge. DF[8], TV[9], term strength (TS), information gain (IG), and chi-square (CHI), odds Ratio (OR), mutual Information (MI),gini index (GI), improved Gini index (GINI) distinguishing feature selector (DFS), genetic algorithm (GA), ant colony optimization (ACO), trace oriented feature analysis (TOFA), are few examples of the feature selection methods.

II. RELATED WORK

All single dimension reduction methods consider only one aspect of the features for the feature subset selection. Consideration of wider (different) aspects altogether is not possible with a single dimension reduction method. Therefore, recently hybrid methods have received considerable attention for dimension reduction. They integrate different dimension reduction methods considering different aspects of the features into one.

Dimension reduction methods are classified as feature extraction and feature selection methods. The feature extraction methods also known as feature construction methods transform a high dimensional feature space into a distinct low dimensional feature space through a combination or transformation of the original feature space. In this paper [2], the minimum classification error (MCE) training algorithm (which was originally proposed for optimizing classifiers) is investigated for feature extraction. A generalized MCE (GMCE) training algorithm is proposed to mend the shortcomings of the MCE training algorithm. LDA, PCA, and MCE and GMCE algorithms extract features through linear transformation. Support vector machine (SVM) is a recently developed pattern classification algorithm, which uses non-linear kernel functions to achieve non-linear decision boundaries in the parametric space. In this paper[2], SVM is also investigated and compared to linear feature extraction algorithms. Principal component analysis (Pearson, 1901), latent semantic indexing, independent component analysis (Comon, 1994), multi dimensional scaling, and partial least square are few examples of feature extraction methods. In this study, we use PCA to reduce dimensions in the feature space.

The dimension reduction methods presented above either use union approach (Tsai & Hsiao, 2010) or intersection approach (Tsai & Hsiao, Zhang et al., 2008) for feature subset selection. The union approach merges all the features present in the considered feature subsist into one, which leads to increase the total number of features. On the hand, intersection approach selects only common features. It reduces the total number of features, however, it losses those feature which attains highest relevance score with respect to only one feature selection method. Therefore, to take advantage of one method and lessen the drawback of other, we present a mid-approach namely modified union for dimension reduction. This approach integrates union and intersection into one and selects informative features without increasing dimensions in the feature space much. Term Variance (TV) and Document frequency (DF) feature selection methods are used to assign relevance scores to each feature.

TV [3] assigns a relevance score to a feature based on its deviation from the mean value. Mean value depicts on an average distribution of the feature among given set of documents. A large deviation of a feature from the mean value shows that it is non-uniformly distributed among the given set of documents and vice versa.

DF [3] is one of the simplest and effective methods to assess relevance of the features. It assigns a relevance score to a feature based on the number of documents covered by it. A fundamental premise of this method is that a frequently occurring feature is more important than non-frequently occurring features.

III. HYBRID DIMENSION REDUCTION METHODOLOGY

The processing of the proposed methodology starts with a collection of documents. Here, we use text dataset Reuters-21,578 for experimental analysis. For representing the documents in a form of VSM, we need to pre-process the documents with standard pre-processing steps, e.g., stop words removal, stemming, tokenization, and term weighting. After processing the documents with these pre-processing steps, we use feature ranking methods TV and DF to assign relevance score to each feature based on the considered criteria. Next, feature sublists merging approach is used to select discriminative subset of features. Traditionally, union or intersection approach is used to perform this task. In this study, we introduce a novel merging approach namely modified union approach to achieve this task. Next, feature extraction method PCA is applied to further refine the selected feature space. Finally, the clusters of documents is created with reduced feature space in order to assess

effectiveness of the dimension reduction methods. The pictorial representation of the proposed methodology is shown in Fig. 1

3.1 Pre-Processing

3.1.1 Stop Word Removal

Words such as a conjunction, pronoun in a text document which does not concern the concept are called as stop-words. This process involves removing the most frequent word that exists in a text document such as 'a', 'an', 'the' etc... Removing these words will save spaces and increase classification performance because stop-words exist nearly in all of the text documents. In the study, stop words were removed in accordance with the existing stop word list (www.unine.ch/Info/clef/) with 571 words.

3.1.2 Stemming

Stemming converts inflectional/derivationally related forms of a word to their root form. For example, introduction, introduce, and introducing all have the common root introduc'. Various stemming methods have been proposed in the literature to achieve this task (Hull, 1996). The most commonly used stemming method is Porter Stemmer. In this study, we also use porter stemmer for stemming purpose.

3.1.3 Tokenization

In tokenization, a document gets split into independent terms called tokens. The length of token varies from a single term (unigram) to a consecutive sequence of n-terms (n-grams). In this study, we use single terms for documents representation.

3.1.4 Removing Numbers and Punctuations

In Removing Numbers and Punctuations, documents have unnecessary values are present then we remove those numbers. To remove noise data in text by using these pre-processing steps.

3.1.5 Strip White Spaces

In Strip White spaces, to remove the unwanted white spaces in the documents. It will remove all types of white spaces.

3.1.6 Term Weighting

To cluster documents, a vector representation model is used to map textual documents into a compact vector representation. In the vector representation, each document o_p {p = 1, 2, 3,. .n} and term t_f {f= 1, 2, 3, s} present in the collection is considered for representation. Various terms weighting schemes have been proposed in the literature to map textual content into a numeric format. The most widely used term weighting scheme is term frequency inverse document frequency (tfidf). Mathematically, it is formulated as follows:

$$\text{Tfidf}_{pf} = \begin{cases} (|Y_{pf}|)1/2\ln(n/df_f), \text{ if } Y_{pf} \geq 1 ; \\ 0 \quad , \text{otherwise.} \end{cases}$$

where, tfidf_{pf} is tfidf of the f^{th} term in the p^{th} document, Y_{pf} is frequency of the f th term in the p^{th} document, n is total number of documents in the corpus and df_f is document frequency of the f th term, i.e., the number of documents in the corpus that include the f th term.

3.2 Feature Relevance Score Computation

Our next step is relevance score computation. To assign relevance score to each feature, we use two different feature scoring methods TV and DF. A detailed description of these methods is presented below:

3.2.1 Term Variance (TV)

TV assigns a relevance score to each on the basis of deviation of the feature from its mean value. A fundamental premise of this method is that the features which are non-uniformly distributed over all documents are comparatively more descriptive than the uniformly distributed features. Mathematically TV is defined as follows:

$$TV_i = 1/n \sum (X_{ij} - \bar{X}_j)^2$$

3.2.2 Document Frequency (DV)

DF is one of the simplest and effective methods to assess relevance of the features. It assigns a relevance score to a feature based on the number of documents covered by it. A fundamental premise of this method is that a frequently occurring feature is more important than non-frequently occurring features.

3.3 Merge Feature Sub Lists

Usually, union or intersection approach is used for feature sub lists merging. A mathematical formulation of these approaches is as follows.

Let D be the set of documents and after pre-processing of the documents T features are selected. Let T { t_1, t_2, t_3, . . . , t_f} be the original feature sets. $FS_1 = \{t_{11}, t_{12}, t_{13}, \ldots ; t_q\}$ is sublist of features selected with TV (M_1), where $t_q$indicates that q number of features are selected with M_1 and f. $FS_2 = \{t_{21}, t_{22}, t_{23}, \ldots , t_l\}$ is sublist of features selected with DF (M_2), where t_l indicates that total number of features are selected with M_2 and l << f

Definition 1. Union FS_1 is the set of features selected with model M_1. It contains q number of features. FS_2 is the another feature sublist selected with model M_2, which contains l number of features. To create a feature sub list FS_3 with union approach, we simply merge all features present in the feature sub lists FS_1 and FS_2.

$$FS_3 = FS_1 \cup FS_2$$

Created feature sub list FS_3 contains f^1 number of features, where $f_1 \geq \{q, l\}$.

Definition 2. Intersection FS_1 is the set of features selected with model M_1, which contains q number of features and FS_2 is another feature sub list selected with model M_2, which retains l number of features. To create a feature sub list FS_4 with intersection approach, we only include those features that are present in both the feature sub lists, i.e., FS_1 and FS_2.

$$FS_4 = FS_1 \cap FS_2$$

Created feature sub list FS_4 contains f'' number of features, where $F' \leq \{q, l\}$.

Traditionally, these two approaches are used to merge feature sub lists selected by two different methods. Union approach selects all feature present in the considered feature sub lists, however, it increases the total number of features. On the other hand, intersection approach selects only common features. It reduces the total number of features; however, it loses those features which are proficient in one aspect. Hence, to select top ranked features as well as common features, we propose modified union approach. It first assesses relevance of each feature and then sorts them based on their relevance score from highest to lowest score. The highest scored feature achieves first place in the list, the second place is occupied by second highest scored feature and so on. The lowest ranked

feature gets last position in the list. Thus, to select high scored features that are present in both feature lists, we apply union approach over top ranked features (C_1) and apply an intersection approach over remaining feature sub lists (C_2). Here, the union approach ensures not to lose highly ranked features and intersection approach ensures the selection of features which are present in both features sub lists. Mathematical formulation of this approach is given below.

Definition 3. Modified Union To create feature sub list with modified union methodology, we apply union as well as intersection approach over given feature sub lists. Here, we apply union approach on top C_1% of the features and then apply intersection approach on remaining C_2% of the features.

$$FS_5=\{C_1\%\{FS_1\}UC_1\%\{FS_2\}\}U\{C_2\%\{FS_2\}\cap\{C_2\ \%\{\ FS_2\}\}\}$$

This step creates feature sub list containing f^{111} number of features, where $f^{111} \geq \{q, l\}$. Pictorial representation of the all approaches, i.e., union, intersection, modified union is given in Fig. 1.

After merging the feature sub lists with modified union approach, next, we apply PCA to further reduce dimensions in the feature space without losing much information.

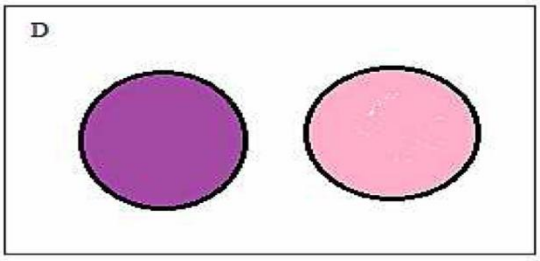
3.4 Feature Extraction

PCA also known as Karhunen–Loeve or K–L method is proposed by Karl Pearson in 1901. PCA uses orthogonal transformation to transform high dimensional feature space into low dimensional feature subspace. Dimension of the reduced feature space may be less than or equal to original feature space.

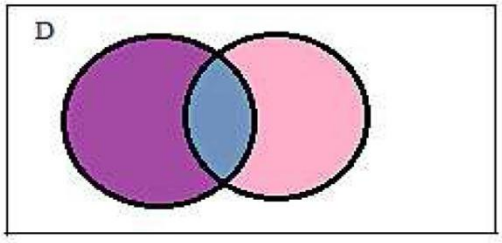
Let p is the number of features in the original feature space and P1 be the number features obtained after transformation of the original feature space, then $p1 \leq p$. The transformation of the feature space is carried out in such a way that the highest variance lies in the first component, next highest variance lies in the second component and so on. The interested reader may refer for a detailed description of the PCA. The identification of the principal components, i.e., the number of the components p1 up to which original feature space has to be transformed, is one of the important steps in the PCA method. In our study, we use cumulative variance (CV) criteria for the component (dimension) selection as used. The reasonable range of CV varies in the range of 70–90. In this study, we set it at 70%.

3.5 Clustering

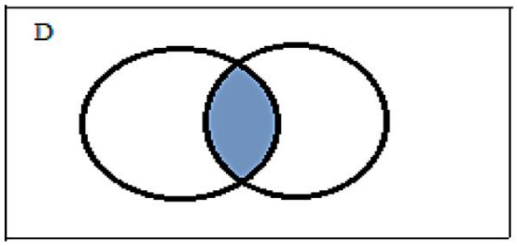
Finally clusters of the documents are created using k-means with feature subspace created with different dimension reduction methods. A pictorial summary of the proposed model is presented in Fig. 2.



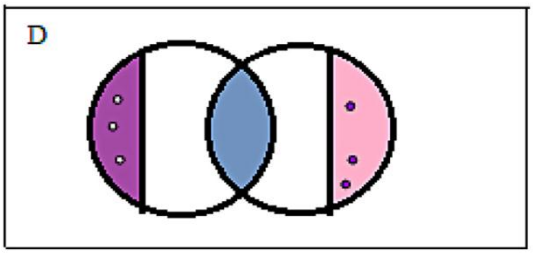
(a)Two different measures



(b)Union Approach



(c)Intersection Approach



(d) Modified Union

IV. RESULTS AND OBSERVATION

Here, we illustrate stepwise refinement in performance of the underlying clustering method (k-means) from the traditional dimension reduction method to the hybrid dimension reduction method.

Step1:

Figure explains about preprocessing step of stemming, in this step we remove stemmed words in xml document.

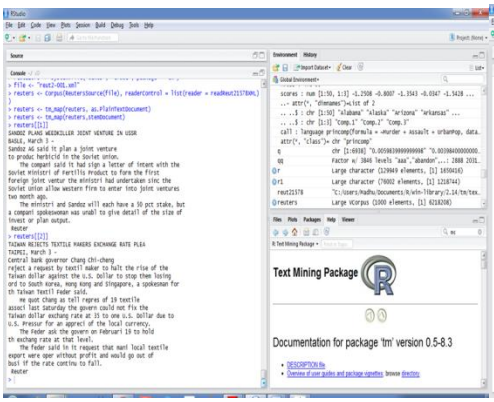


Fig3: In Preprocessing Stemming

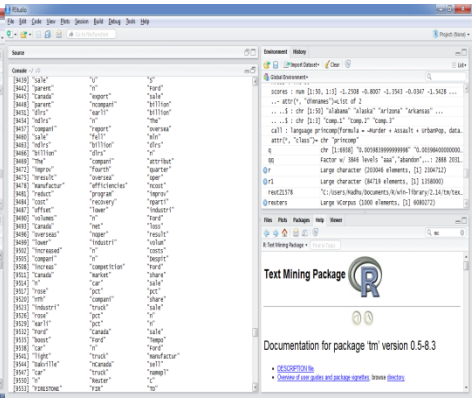


Fig4: In Preprocessing Data

Step2:

In this step we perform all preprocessing steps like stop word removal, removing numbers and punctuations, then perform tokenization in fig 4.

Step3:

In this step we have to convert tokens into weighted valuesI.e. term weighting .they calculates term frequency and inverse document frequency of each term.

Step4:

In this step we have to perform relevance score computation i.e. term variance and document frequency.

Step5:

In this step we perform hybrid dimensionality reduction,i.e Merge feature sub list (modified union). Performing union and intersection, they give most frequent terms and some relevant terms based on the calculations. After we perform modified union fig 4.

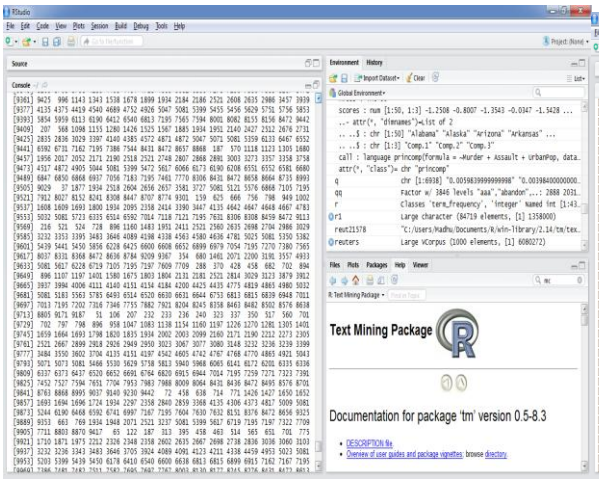


Fig5: Term Weiggting

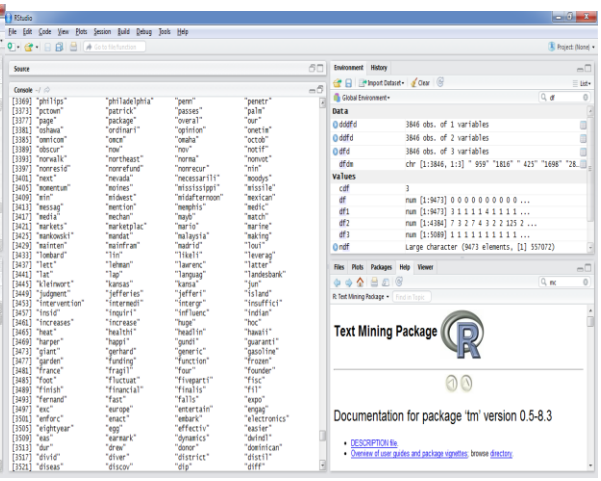
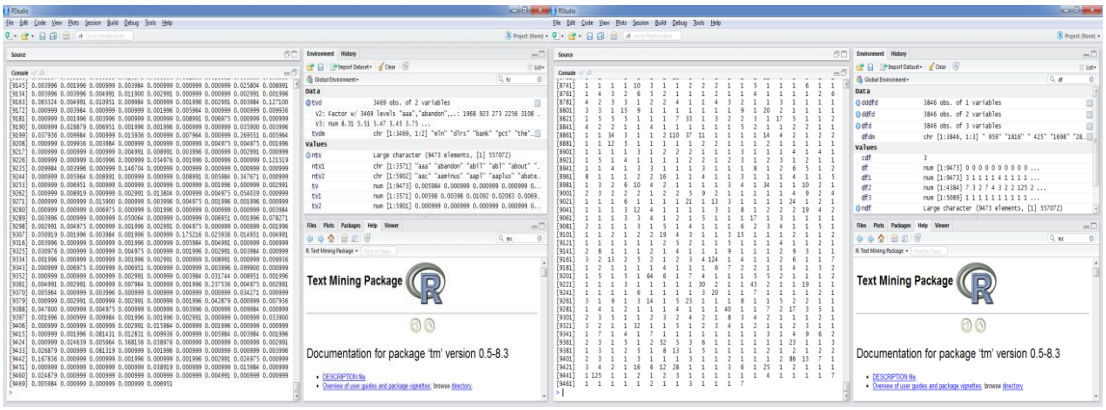


fig 6: Term Variance



A. Term Variance **B. Document Frequency**

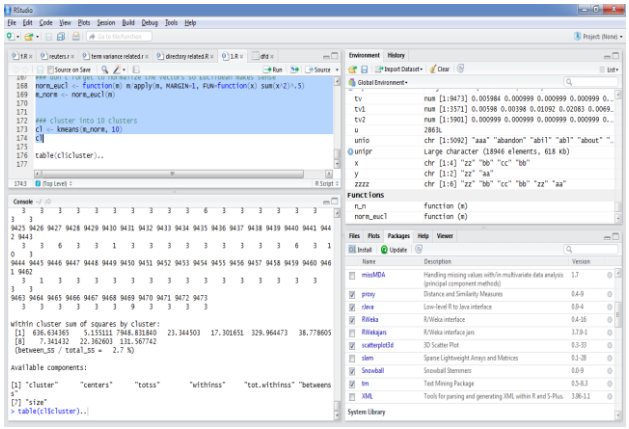


Fig5: K-Means Clustering

Step6:

In this we perform feature extraction and clustering. In this paper we perform k-means clustering fig5.

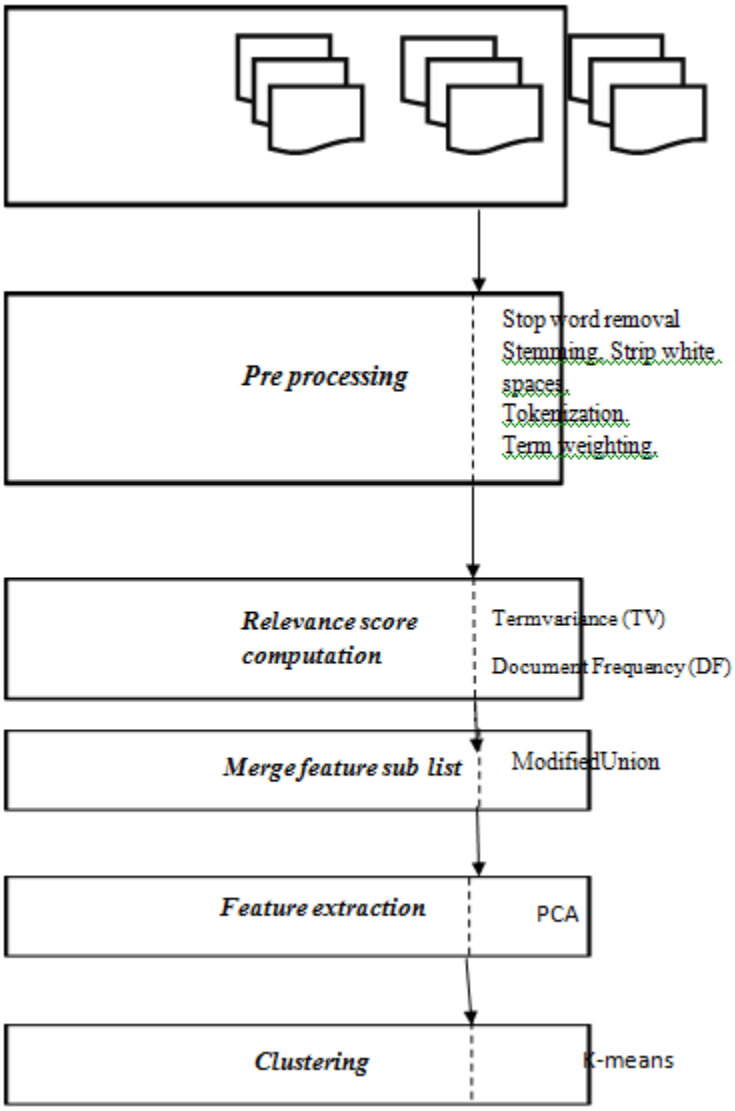


Fig2 The Flowchart of the Proposed Methodology

III. CONCLUSION

Selection of an informative subset of features is one of the major concerns in the text clustering due to computational complexity and accuracy consideration. The simple dimension reduction methods select subset of features by considering only one aspect. Consideration of different aspects for feature subset selection is not possible with a single dimension reduction method. Moreover, each feature selection method has its own advantages and disadvantages.

Hence, a hybrid feature selection method, which integrates advantage of one method and lessens drawback of the other method, receives considerable attention. Traditionally, union or integration approach is used for feature sub lists merging. The union approach combines all features, which are present in the considered feature lists.

Hence, it increases the total number of features. On the other hand, intersection approach selects only those features, which are common in all the considered feature lists. This approach selects comparatively less number of features; however, it sometimes loses highly ranked features. Therefore, we present a mid-approach namely modified union, which selects all highly ranked features as well as common features from the considered feature lists without increasing the dimensions much in the feature space much is designed and implemented. To further refine the selected feature subspace, PCA is applied. This step reduces dimensions in the feature space without losing much information.

Despite good results, a major weakness of this method is its dependency on the parameters C1 and C2. The performance of the method varies with these parameters values. Hence, in future, a method to automatically determine the values of parameters C1 and C2 can be developed. In this study, the interaction between terms has not been considered. Hence, in future this information can also be incorporate this information also.

REFERENCES

[1] Y. Zhao & G. Karypis. "Criterion Functions for Document Clustering: Experiments and Analysis," Technical Report #01-40, Department of Computer Science, University of Minnesota, November 2001.

[2] Zamir, O., Etzioni, O., Madani, O., & Karp, R. M. (1997). Fast and intuitive clustering of web documents. In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (Vol. 97, pp. 287–290)..

[3] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley symposium on mathematical statistics and probability (Vol. 1, pp. 281–297). California, USA.

[4] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,2, 559–572.

[5] Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M. A., & Meira, W. Jr., (2011).Word co-occurrence features for text classification. Information Systems, 36,843–858.

[6] Hsu, H., Hsieh, C., & Lu, M. (2011). Hybrid feature selection by combining filters andwrappers. Expert Systems with Applications, 38, 8144–8150.

[7] Huang, S., Peng, X., Niu, Z., & Wang, K. (2011). News topic detection based onhierarchical clustering and named entity. In 2011 7th international conference onnatural language processing and knowledge engineering (NLP-KE) (pp. 280–284).

[8] IEEE.Hull, D. A. (1996). Stemming algorithms: A case study for detailed evaluation.Journal of the American Society for Information Science, 47, 70–84.

[9] Janaki Meena, M., Chandran, K., Karthik, A., & Vijay Samuel, A. (2012). An enhancedACO algorithm to select features for text categorization and its parallelization. Expert Systems with Applications, 39, 5861–5871.

[10] Kabir, M. M., Shahjahan, M., & Murase, K. (2012). A new hybrid ant colonyoptimization algorithm for feature selection. Expert Systems with Applications, 39, 3747–3763.

[11] "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.Wu, Y.-L., Tang, C.-Y., Hor, M.-K., & Wu, P.-F. (2011). Feature selection using geneticalgorithm and cluster validation. Expert Systems with Applications, 38,2727–2732.

[12] Yang, Y. (1995). Noise reduction in a statistical approach to text categorization. InProceedings of the 18th annual international ACM SIGIR conference on research anddevelopment in information retrieval (pp. 256–263). ACM.

[13] Yan, J., Liu, N., Yan, S., Yang, Q., Fan, W., Wei, W., et al. (2011). Trace-oriented featureanalysis for large-scale text data dimension reduction. IEEE Transactions onKnowledge and Data Engineering, 23, 1103–1117.

[14] Zamir, O., Etzioni, O., Madani, O., & Karp, R. M. (1997). Fast and intuitive clusteringof web documents. In Proceedings of the 3rd International Conference onKnowledge Discovery and Data Mining (Vol. 97, pp. 287–290).

[15] Zhang, Y., Ding, C., & Li, T. (2008). Gene selection algorithm by combining Relief and MRMR. In IEEE 7th international conference on bioinformatics andbioengineering (pp. 127–132).