An Efficient Productive Feature Selection and Document Clustering (PFS-DocC) Model for Document Clustering

Document Clustering using PFS-DocC Model

Perumal Pitchandi
Department of Computer Science and Engineering
Sri Ramakrishna Engineering College, Coimbatore, India

***Abstract*—In mining, document clustering pretends to diminish the document size by constructing the clustering model which is extremely essential in various web-based applications. Over the past few decades, various mining approaches are analysed and evaluated to enhance the process of document clustering to attain better results; however, in most cases, the documents are messed up and degrade the performance by reducing the level of accuracy. The data instances need to be organized and a productive summary have to be generated for all clusters. The summary or the description of the document should demonstrate the information to the users' devoid of any further analysis and helps in easier scanning of associated clusters. It is performed by identifying the relevant and most influencing features to generate the cluster. This work provides a novel approach known as Productive Feature Selection and Document Clustering (PFS-DocC) model. Initially, the productive features are selected from the input dataset DUC2004 which is a benchmark dataset. Next, the document clustering model is attempted for single and multiple clusters where the generated output has to be more extractive, generic, and clustering model. This model provides more appropriate and suitable summaries which is well-suited for web-based applications. The experimentation is carried out in online available benchmark dataset and the evaluation shows that the proposed PFS-DocC model gives superior outcomes with higher ROUGE score.**

***Keywords*—Benchmark standards; document clustering; productive feature selection; multiple clustering; web applications**

I. INTRODUCTION

With the vast expansion towards the web and Internet applications along with the growth of mobile phones leads to the growth of enormous textual information [1]. This drastic explosion of data generation not only produces mess over document clustering and summarization. This complexity is not only encountered by the humans' but also by the machines which lag in processing the massive data generated from various sources like (applications, technologies, and organizations) [2]. The evaluations towards the huge amount of data are generally non-structural and quite a challenging task. The drastic eruption of documents over the web necessitates the path for document clustering and summarization process [3]. It attempts to give a shorter version of documents by maintaining the necessary information. The extensive insight of data makes the researchers to take appropriate decision by document clustering [4]. Thus, document clustering turns to be an essential approach in the growing world.

The document clustering or summarization helps in attaining a wider insight towards the data and offer decision making process [5]. For example, various social media applications like Facebook, Twitter and so on are used for personal causes for political and marketing purposes [6]. Recently, most of the political campaigns are made over these social media sources all over the world to reach the supporters in various regions. Therefore, the process of extracting the textual data is essential for successful political and marketing strategies [7]. Various real-time applications of document clustering are not constraint with these political and marketing strategies. For example, it is also employed for compressing the content for searching the outcomes over search engines along with the keyword for direct subscription towards the application [8]. Moreover, a proficient document clustering process over social media resources can preserve the user's trust relies on navigation among various contents [9].

The document summarization process includes huge challenge and the preliminary attempt is performed in 1950's when it uses features like phrase and word frequency to extracting essential sentences [10]. It is also considered as a huge demand in the field of research owing to its applicability. The finest way of summarization has to preserve the preliminary factors while assisting the users to have better insight towards the enormous volume of data in a faster manner [11]. The preliminary idea behind document clustering is to gather the more essential information in a clustered or with a compressed manner for certain tasks/users [12]. The clustering is also depicted as the gathering of data instances or the shortest document version which is gathered from the machine to attain most essential information is specific manner without human interventions. Moreover, the foremost definition is provided by [13], as 'text is gathered from one or more documents that provides essential information based on the source content and provides the shorter version of it". Based on this definition, there are three different factors that have to be concentrated: 1) clustering can be done with one or more documents; 2) clustering should preserve the essential parts of original content and 3) clustering have to gather the

original source content without any reduction or alteration with original content [14].

There is various classification of document clustering process. Moreover, the process of document summarization is partitioned as: abstractive or extractive manner. The former model is to understand the textual content of the document profoundly and expresses the text in shorter manner. Subsequently, the target is to extract the document content to choose the most essential information [14]. It is extremely harder for the machine to generate the clustering of multiple documents which is smoother and understandable by the humans. In common practise, extractive approaches are generally used. From the various categories of document clustering process, recently, learning approaches are used for various documents clustering process [15]. The extractive process can be either supervised or unsupervised. In the former model, the problem is based on binary classification where the classes are defined with the summary; similarly, in the latter model, the ultimate target is to attain representative sentences. This research proposes a novel Productive Feature Selection and Document Clustering (PFS-DocC) model which is beneficial to handle the supervised and unsupervised challenges in an interpretable way. The anticipated model possesses the following characteristics:

1) Here, the challenges identified in clustering are considered as a single-objective problem. The clustering process attempts to identify the underlying data structure and provides the information for further classification purpose. Therefore, it enhances the performance of clustering algorithm.

2) The features are extracted with dynamical process via selective manner for all clusters. The clustering process should include the weight of the document by label discrimination to cluster the document.

3) The sentences are chosen in a way that it produces the clustering process in a non-redundant and coherent manner. The complex documents are placed at the top while remaining sentences are selected to gather the essential information with the redundancies.

The proposed Productive Feature Selection and Document Clustering (PFS-DocC) model obviates the requirement of feature engineering in a document clustering Even though, the most crucial phase over learning process in feature selection and extraction, various work concentrates in sentence clustering process. In recent time, various attempts to make to predict the optimal feature set for clustering process. This process considers the feature relevance as binary issues, that is, whether the features are attained from feature patterns. The overview of the Productive Feature Selection and Document Clustering (PFS-DocC) model is shown in Section 3. The samples of the document are chosen based on the feature vectors. The final outcomes need to similar group of samples with the features of similar group. The weighted features show similar features with clustering. In document clustering process, these clusters specify whether the document is efficient. The preliminary contributions of this clustering process are given below:

1) This work introduces the theoretical model based on productive manner. Here, a novel concept is the process of document clustering. This model facilitates the process of clustering the documents which helps in choosing the document. More specifically, the process of designing the clustering model is to measure the document sentences by labelling ‘0’ and ‘1’, respectively.

2) The proposed Productive Feature Selection and Document Clustering (PFS-DocC) model have the ability to measure the significance of the features by class discrimination which is clustered with various dataset over the reported dataset.

3) Here, evaluation is done with online available dataset to compute the clustering process in an efficient manner. It validates that the clustering process is less redundant and possess more information in a competitive manner.

4) Also, based on the comparison with prevailing approaches, Productive Feature Selection and Document Clustering (PFS-DocC) model gives added advantages which are less interpretable. It clearly states that the process tracks the cluster of document which is essential to explain the decision performed by the end-users.

II. Related Work

Document clustering is considered as an unsupervised approach for semantic clustering with the similar documents. The embedded documents are determined as a vector space and predict the neighbours over the space along with the clustering model based on word extraction which is extensively utilized. There are various investigations that enhance the performance with cluster initialization and automatic parameterization. Moreover, these approaches consider that all the provided documents are autonomous and do not determine the relationship strength among them. The document clustering model helps to get rid of various limitations that determine the relationship along with the document significance which is extensively investigated.

Network-based document clustering [16] is determined based on the interconnection among the documents and carry out document clustering based on network characteristics. It is depicted as the graphs that comprises of vertices related to the edges. Based on this analysis, generally it is considered as the vertices pair related with the edges to project semantic relationships. Then the assumption is based on the link strength and authority over the provided documents measured and the documents are clustered based on provided parameters [17]. This document clustering model performs hyper-linked web document classification based on academic papers and society, and citations. These approaches are utilized to demonstrate the semantic relevance among the news [18]. Therefore, it employs various kinds of meta-data and applied to various document ranges. This model is utilized to link a document that relies on content. It inter-connects various shared words with preliminary text documents. Therefore, documents are clustered relies on dependent association among the prevailing network; even in case of meta-data with absolute completeness.

The process is formally defined to carry out network-based document clustering. Kusner et al., [19] depicts network based document clustering formally with probabilistic generative model and utilized to cluster the documents. Therefore, the modelling of probabilistic generative model is anticipated and not applicable for multi-label clustering where the document is provided for multiple clusters. Moreover, the model is not suitable for various domains which encounter highly complex documents like certain documents and mobile applications which are allocated to multiple clusters [20]. Based on various analyses, network based document clustering offers multi-labelling process. Here, neighbourhood graph-based weighted matrix is used to evaluate the relationship strength among the documents with clustering process, concept factorization, and matrix factorization [21].

Moreover, diverse ranking models like search rank over search engines, paper classification and hubness value process which are utilized to compute the link strengths and document significance between the documents [22]. Hubness values are extensively utilized for evaluating the document significance. For instance, HITS and PageRank approaches are used for analysing the flow of web pages to search the documents and allocate higher value authorization to possess enormous number of inter-links [23]. Thus, these approaches are adopted over various documents and assigns higher authority values with huge amount of inter-links. Moreover, these approaches are considered to be the favourable older documents and assigns low authority values for all the newer documents. Thus, meaning-based search engine is adopted to handle these issues and projects the meaning-based information with document significance and un-important factors [24]. Thus, it enhances the processing speed.

Thus, search engine based significant ranking documents are based on semantic relevance and concentrates on internal meaning information [25]. The limitations over these methods are extremely prone for abusing which specifies internal inclusion of essential irrelevant words in context to actual documents. Based on various approaches, the proposed model makes use of document significance with indices that are autonomous independently with the document content like number of downloads over the mobile apps [26]. The given model preliminarily reduces the abuse by handling these issues over the network-based document significance examination.

The embedding documents are considered as the conversion of documents which includes word set with latent vectors [27]. It is utilized to evaluate the distance among the provided documents and consequently clusters the similar documents during document clustering model [28]. This model is extensively utilized for embedding document techniques which is composed of inverse document frequency and term frequency, topic modelling approaches termed as Latent Dirichlet allocation [29]. Various investigators consider document clustering by adopting topic modelling document embedment with k-means algorithm. The functionality of topic modelling is enhanced using the measure of documents with network modelling. This enhanced model uses document clustering.

Additionally, various researches are underway with word/document embedment with neural network approaches. The representative NN model is composed of word2vec which identifies the similar words form the input words. Similarly, Doc2vec predicts the word that offers the input document. In recent times, Doc2Vec, LDA, and TF-IDF are adopted to include the documentation [29]. The performance of document-clustering process is improved with the adoption of semi-supervised approaches that include the construction of initial-clusters which relies on words and enhances the similarity among the documents over the provided clusters via learning process. Word2Vec-based documents are used to predict, classify, and visualize social network neighbourhood [30]. Subsequently, embedding algorithm is alike of word2vec with certain exception and identifies the neighbourhood indeed of context words. In this research, a novel Productive Feature Selection and Document Clustering (PFS-DocC) model is proposed to reflect the document significance based on feature selection and document clustering. This model provides better performance based on consistent document information, document meta-data, and information clustered with input document. It is explained in the section given below.

III. Methodology

This research model includes three different processes: pre-processing, feature selection, and summarization. The evaluation is done with MATLAB environment using DUC 2004 dataset. The comparison is done with various metrics like accuracy, precision, F1-score, recall, ROUGE 1 and ROUGE 2 score. Also, the evaluation is done with DUC 2003 and DUC 2004 benchmark dataset. An extensive analysis is done with a proposed Productive Feature Selection and Document Clustering (PFS-DocC) model. Fig. 1 depicts the block diagram of proposed PFS-DocC model.

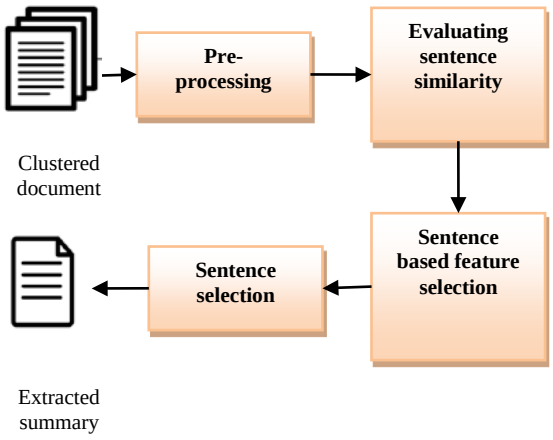

Fig. 1. Block Diagram of Productive Feature Selection and Document Clustering (PFS-DocC) Model.

A. Dataset

The DUC 2004 uses paper documents, newswire from TDT and TREC collections. The data is used for training, summarization exploration from the produced by machine translation. The task involves summarization by question and represents various tasks. The official ROUGE measures of DUC 2004 were 1-gram, 2-gram, 3-gram, and 4-gram and longest sub-string scores. The manual summarization is used for running ROUGE was provided to available participants. Thus, the truncated summaries longer than the targeted length before evaluation and generates summaries lesser than target length. The maximal target length was depicted based on bytes (punctuation, whitespace, and alphanumeric) included. The maximal target length for short summaries was 75 bytes. The shorter summaries are 665 bytes.

B. Pre-processing

The data (document) pre-processing is composed of linguistics, tokenization which provides a mathematical mode. It transforms the document content into sequence of terms which avoids punctuation and carries out removal of stop word ('a', 'an', 'in', 'etc.') are removed. There are enormous numbers of stop words.

C. Productive Clustering

The target of adopting productive clustering is used as a process of information retrieval. The user needs to scan the provided descriptors for relevancy measure and demonstrate that the clusters are relevant by manual processing of various document instances. The iterative process uses multiple stages of productive clustering to assist user for predicting the appropriate documents. The initial clustering is provided with description or clusters to the users who selects clusters of own interest. The text instances over the chosen clusters are merged and clustered. This process is continued with appropriate set of documents. The automatic description of quality is crucial for facilitating users to predict which clusters the relevant text.

The productive clustering is performed by initially clustering and predicts the set of features related with cluster. It facilitates appropriate clustering algorithm (see Algorithm 2) to be adopted. The chosen features provide best information to the users based on the users' content (cluster). The preliminary process is to demonstrate the clusters with likely words over the cluster. But, the features are not optimal for establishing discrimination among various clusters. The scoring criterion includes information gain (mutual information).

D. Feature Selection

For the provided clusters, the prediction of instances from the input clusters handle the conventional classification problem and selection of appropriate feature subset is more essential. The selection of smaller subset with maximal feature prediction is a complex task. In smaller feature subset, step-wise similarity measure is carried out to enhance the classification performance. The model should fit with the features which cannot scale the features that are encountered with the textual data. The proposed model should trace number amount of features. The feature selection process has to ensure the process by positive correlation with target class, that is, feature occurrence rate of provided class which is higher than average rate.

Algorithm 1: Evaluating sentence score for similarity measure

Input: Array of sentences
Output: Similarity scores
1. Average weighted matrix [n][n];
2. Scores [n];
3. for $i \rightarrow 1\ to\ n$ do;
4. for $j \rightarrow 1\ to\ n$ do;
5. predict = $identity - similarity\ (s[i], s[j])$;
6. Average similarity matrix [i][j] = average value (id);
7. end
8. end
9. score = id (average similarity matrix);
10. return scores;

The productive clustering model is composed of two preliminary tasks: identifying the original occurrence of features based on cluster allocation and identifying the instances of certain cluster with smaller feature dimensionality set that functions as the cluster descriptions. This task offers an objective to automatically choose from clustering with various numbers of clusters. The cluster is related with various feature distributions with lesser frequency instances over the clusters. When the instances are allocated with similar cluster have same feature distribution where the cluster allocation is productive of feature occurrence. The successive task is the prediction of clustering membership with dependency over the selected clusters. The total information attained by clustering increases with clusters; however the complexity of finding the cluster membership increases with the fine-grained clusters. Also, the added numbers of clusters are more inherent and provide better trade-off among the number of features and prediction performance. The traceability process needs to be performed with number of available clusters and features for multi-document clustering process. The feature selection model is evaluated and chosen for candidate set.

In cluster creation process, the set of probable clusters are generated that varies from the total number of clusters which arises from various clustering process or various data specification. For all clusters, the model is trained to find the feature occurrence from allocated clusters. The association among the clusters are more productive. The feature subsets are chosen based on the selection mechanism. Specifically, feature subsets are predicted with positive constraints by changing regularization process. Every stage is allocated with standard modelling like clustering and model selection.

The productive framework is provided by allocating clusters that gives flat clustering. The clustering process is effectually executed with sparse data when the similarity among the data is utilized and changes the number of clusters to generate set of clustering process $\phi = \{\phi_1, \dots, \phi_K\}$. Here, various clustering process are chosen from the feature vectors that are attained from cluster assignments. The probability of the feature set occurrence is provided with cluster assignment $\boldsymbol{y} = [\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(C)}]$ which is expressed as in Eq. (1):

$$prob\ (X = 1|y) = \frac{1}{1+e^{(-X_0)}} \quad (1)$$

The above equation is expressed with binary features where X_0 is bias compactness with coefficient of parameter vector $u' = [x_0, x]$ and constant features are added with cluster allocation. The cluster is a supervised learning problem which includes both feature selection and training process. The cluster $c \in \{1, \dots, C\}$ with feature subset prediction is expressed as in Eq. (2):

$$\arg\min E[Loss\ (x^{(c)}, \hat{y}^{(c)} + \alpha\ \Omega\ (w^{(c)}) \quad (2)$$

Here, $w^{(c)}$ is weighted co-efficient for providing the feature ranking with original cluster vectors. The feature constraints are related with the clusters positively. The feature subset reduces the computational cost of the provided model. For the provided cluster, the probability instances are allocated with the cluster as conditional random variables. It is expressed as in Eq. (3):

$$prob\ (Y = 1|x) = f_{w'}(x) \quad (3)$$

It is provided as the bias compactness with the integration of co-efficient and constant which is included at the feature vectors. The minimization problem with appropriate solution in expressed as in Eq. (4):

$$\arg \min_{w'} -\ln L(w') + \frac{\tau}{2} ||w||_2^2 \quad (4)$$

Generally, the features are related with various non-zero coefficients. The equivalent constraints provide solution when the coefficients are zero. It is shown in Eq. (5):

$$\arg \min_{w'} -\ln L(w') + \gamma\ ||w||' \quad (5)$$

Here, γ influences the number of features with non-zero co-efficient where the larger value of γ which yields better solution with non-zero co-efficient. These non-zero co-efficient are provided with chosen features. The suitable feature subsets are determined by sweeping the γ values. The feature subset and the weighted co-efficient are used to choose appropriate feature subset. Consider a feature subset $S_1, \dots, S_j$ for certain cluster and optimal feature subset is chosen with Eq. (6):

$$\hat{j} = \arg min - \ln L(w_j) + |S_j| \ln \sqrt{n} \quad (6)$$

Here, S_j is set of features that do not include feature subset. The coefficients are generated from the provided subset. The bias value sometimes influences the chosen subsets. The numbers of interpretable features are stable over the sample size variations. In practical condition, the feature subsets are restricted based on the size. The user needs to deal with enormous features to demonstrate the clustering process. Sometimes, the limit may reduce the productive performance; also it reduces the computational complexity with number of feature subsets during evaluation process. The analysis is done with publicly available dataset. The anticipated model is based on set of predictive features. The numbers of features are restricted with positive correlation among the clusters and classes. The positive constraints are provided based on classification performance.

For the computation of cluster predictions, here f1-score is used for individual clusters or classes where the summarization is resulted with the average of computed F1-score. The data instances are allocated with multiple clusters and not allocated with various available clusters. The instance possesses equivalent weight among distributed among the assigned values. The un-allocated instances are determined based on the valid group of added clusters. The mutual information is extracted from the discrete variables partitioning. The computation is done with automatic selection of total clusters whether the numbers of clusters correlate the maximal information content. The major drawback associated with existing approaches is the evaluation of multi-modal distributions of all features with higher computational complexity $O\ (N^2)$. The redundancy elimination is done with candidate features by setting the divergence among the multi-modal distributions. The scalability is done with the features of higher score over the targeted clusters.

Algorithm 2: Document clustering

Input: Array of sentences
Output: sentence score
1. Similarity matrix [n][n];
2. Array scores [n];
3. for $i \rightarrow 1\ to\ n$ do;
4. for $j \rightarrow 1\ to\ n$ do;
5. DocC [i][j] = measure similarity (S[i], S[j]);
6. end
7. end
8. DocC = Similarity matrix;
9. Hyper-linked similarity matrix;
10. for $i \rightarrow 1\ to\ n$ do;
11. score [i] = average summarization;
12. end
13. return scores;

This work concentrates in computing the appropriate selection of number of clusters with the Productive Feature Selection and Document Clustering (PFS-DocC) model. This model enhances and maximizes the information attained by the clustering algorithm. The experimentation is done to compute the information among the original clusters and the chosen clusters are varied based on proportional cluster number. The productive document clustering facilitates both the number of features and clusters which is utilized to determine the cluster. The user needs to select appropriate range data clusters with computational feasibility. The user needs to enhance the range of more optimal clustering process. The anticipated model is utilized to any data with weighted features. The productive r with productive features and cluster assignment is prediction with feature subset. The outcomes are attained based on the every cluster with minimal amount of feature subset which is essential to identify the instances that belongs to certain clusters. The productive clustering model is used to predict the cluster membership of given document. The relevance of the proposed PFS-DocC model is efficient to give higher amount of information with reduced data redundancy. The section below discusses the numerical outcomes attained with the analysis of proposed PFS-DocC model.

IV. Numerical Results and Discussion

The performance of the proposed Productive Feature Selection and Document Clustering (PFS-DocC) model based on clustering, information extraction, and non-redundancy and overall processing is evaluated. Some metrics like accuracy (%), recall (%), F1-score (%), and precision (%) are measured. For this evaluation, online available DUC 2004 dataset is a generic model for document clustering. It includes 50 clusters of new documents. These clusters include the summaries of various human references which are considered by the researchers for extracting the outcomes. It is essential to set the length of document clusters. The clusters over DUC 2004 organize 665 bytes where the pre-processing step is extremely needed for accuracy evaluation. Here, some essential pre-processing steps are performed with text documents. Generally, the documents are processed to predict the document source information from textual components. The initial process needs to eliminate the information tags such as <TEXT>, <DOC>, and so on for processing the documents.

The experimentation performance is measured with evaluation toolkit known as ROUGE which is a recall based evaluation metrics. It computes the efficiency of document clustering for evaluating the summaries generated by the humans. The ROUGE score evaluates the number of successive terms. After the completion of pre-processing steps, the similarity measures among the sentences are evaluated using the proposed Productive Feature Selection and Document Clustering (PFS-DocC) model. The probability occurrences of the words from the input clusters are used to identify the productive words. The clusters are summarized with the clusters over the dataset. The outcome of the discriminant analysis is measured with metrics like True Negative (TN), True Positive (TP), False Positive (FP), and False Negative (FN) are known as correct predictions with negative samples, correct prediction with positive instances, incorrect prediction with positive samples, and incorrect predictions with negative instances, respectively. It is expressed as in Eq. (7) - Eq. (10):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$F1-score = 2 * \frac{Precision*Recall}{(Precision+recall)} \quad (8)$$

$$Recall = \frac{TP}{TP+FP} \quad (9)$$

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

The simulation is carried out in MATLAB environment. Here, six different methods along with the Productive Feature Selection and Document Clustering (PFS-DocC) model are compared. The six methods are FLSA (ProbIDF), FLSA (Normal), FLSA (IDF), FLSA (Entropy), LDA, and LSA respectively. Similarly, metrics like Accuracy (%), F1-score (%), Recall (%), and precision (%) is evaluated. The accuracy of proposed Productive Feature Selection and Document Clustering (PFS-DocC) model is 98.9% which is 1.9%, 7.9%, 3.9%, 1.9%, and 38.9% higher than the prevailing methods. The F1-score of PFS-DocC is 99% which is 29.7%, 27.6%, 1.3%, 3.5%, 7.8%, and 1.3%, respectively. The recall of PFS-DocC is 99% which is 27%, 26%, 4%, 6%, 10%, and 4% higher than the other models. Similarly, precision of PFS-DocC is 99% which is 33%, 30%, 4%, 6%, 10%, and 4% higher than other models. All these process includes 50 topics. It is shown in Table I. Fig. 2 depicts the performance metrics evaluation. Fig. 3 depicts the F1-score computation.

TABLE I. Comparison of Performance Metrics

Methods	Accuracy (%)	F1-score (%)	Recall (%)	Precision (%)	Topics
FLSA (ProbIDF)	97	69.3	72	66	50
FLSA (Normal)	91	71.4	73	69	50
FLSA (IDF)	95	97.7	95	95	50
FLSA (Entropy)	97	95.5	93	93	50
LDA	60	91.2	89	89	50
LSA	57	97.7	95	95	50
PFS-DocC	98.9	99	99	99	50

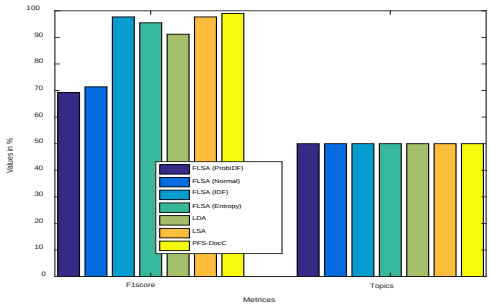

Fig. 2. Performance Metrics Evaluation.

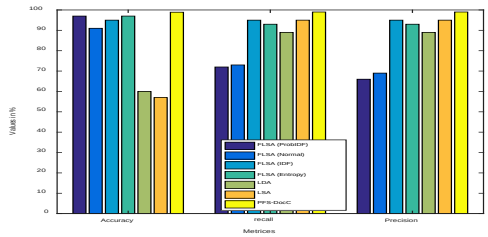

Fig. 3. F1-Score Computation for 50 Topics.

Table II depicts comparison of ROUGE 1 score and ROUGE 2 score with the evaluation toolkit. The comparison is done for ExDoS, Banditsum, HSSAS, summaRunner, NN-SE, LEAD-3, and PFS-DocC respectively. Rouge 1 score is 45 which is 3%, 4%, 3%, 6%, 10%, and 6% higher than other models. Rouge 2 score of PFS-DocC is 2%, 1.9%, 3%, 4%, 7%, and 4.7%, respectively. Finally, Rouge L score of PFS-DocC is 39 which are 4%, 7%, 4%, 2%, 1.5%, and 1% higher than other models (see Fig. 4). Table III shows the amount of information extracted, non-redundant, overall percentage achieved. PFS-DocC based information extraction is 30%; however for other approaches it is 27%, 23.5%, 20.5%,

17.6%, 13.5%, and 13% respectively (see Fig. 5). The avoidance of non-redundant data from PFS-DocC is 25% where the other data is 22.5%, 22.6%, 16.5%, 19.5%, 21%, and 23% respectively (see Fig. 6). The overall performance of PFS-DocC w.r.t information extraction and non-redundancy avoidance is 27%; whereas for other models it is 25%, 18.5%, 21.6%, 16.8%, 20.8%, and 22%, respectively.

TABLE II. ROUGE SCORE EVALUATION

Methods	Rouge 1 score	Rouge 2 score	Rouge L score
ExDoS	42	18.5	35
Banditsum	41	18.6	32
HSSAS	42	17.5	35
SummaRunner	39	16.5	37
NN-SE	35	13.5	37.5
LEAD-3	39	15.8	38
PFS-DocC	45	20.5	39

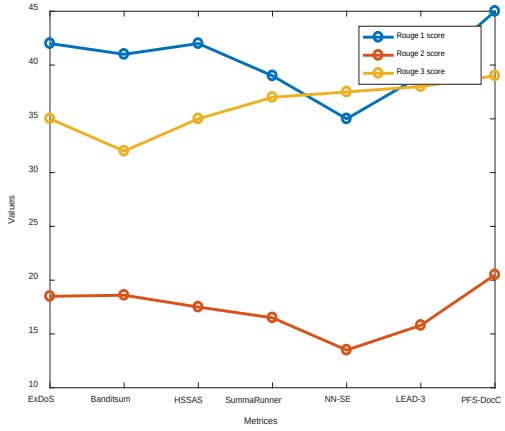

Fig. 4. ROUGE Score Computation.

TABLE III. INFORMATION EXTRACTION AND NON-REDUNDANCY PERCENTAGE

Methods	Information extraction	Non-redundancy	overall
ExDoS	27%	22.5%	25%
Banditsum	23.5%	22.6%	18.5%
HSSAS	20.5%	16.5%	21.6%
SummaRunner	17.6%	19.5%	16.8%
NN-SE	13.5%	21%	20.8%
LEAD-3	13%	23%	22%
PFS-DocC	30%	25%	27%

Table IV depicts the comparison of PFS-DocC without feature extraction is done with benchmark datasets, like DUC2002-ROUGE 1, DUC2002-ROUGE 2, Main-ROUGE 1, Main-ROUGE 2, DUC2004-ROUGE1, and DUC2004-ROUGE 2. The values of PFS-DocC (without feature extraction) are 53, 26.7, 42.5, 19, 55, and 57 respectively. Similarly, the values of PFS-DocC are 46, 22.5, 39.7, 15, 50, and 53 respectively (see Fig. 7 and Fig. 8).

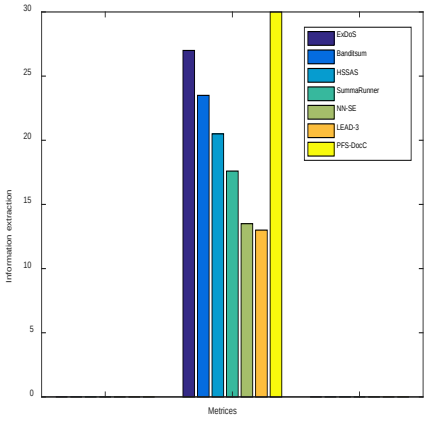

Fig. 5. Information Extraction.

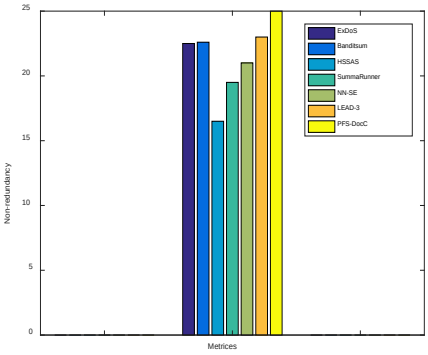

Fig. 6. Non-Redundant Data Extraction.

TABLE IV. PFS-DOCC COMPARISON (WITH / WITHOUT FEATURE SELECTION)

Methods	DUC2002-ROUGE 1	DUC2002-ROUGE 2	Mail-ROUGE1	Mail-ROUGE2	DUC2004-ROUGE 1	DUC 2004 – ROUGE 2
PFS-DocC	53	26.7	42.5	19	55	57
PFS-DocC + feature extraction	46	22.5	39.7	15	50	53

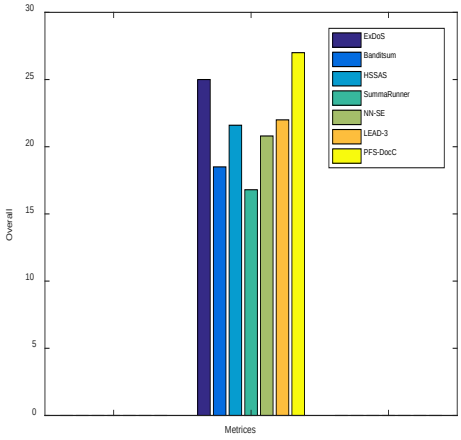

Fig. 7. Overall Performance Measure of PFS-DocC.

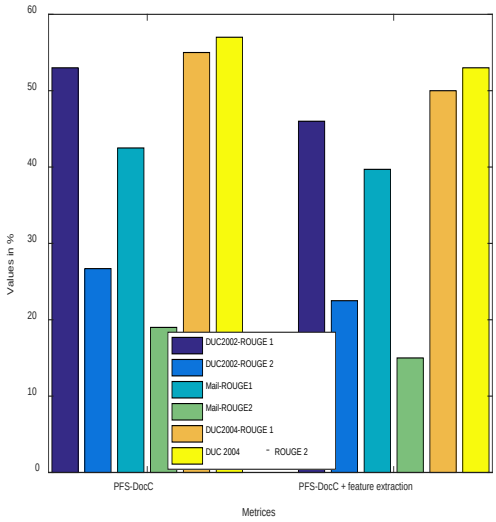

Fig. 8. PFS-DocC Performance (with / without Feature Extraction).

The similarity score is examined based on the sentence that is extracted from the central score. Initially, the document clusters are converted to connected sentence using various similarity scores. Based on the experimentation, the proposed PFS-DocC model enhances the summarization process attained from document clustering. After the extraction process, the sentences are provided with high score and include the summary length. It is essential to improve the extracted sentences which do not possess any redundant information. Therefore, to diminish the redundancy over any sentences with the similarity measure of extracted summary the proposed PFS-DocC is used.

V. Conclusion

This research concentrates on proposing a novel Productive Feature Selection and Document Clustering (PFS-DocC) model with three essential steps that includes background knowledge, pre-processing, feature selection, and summarization (clustered document). It is to enhance the performance of the proposed PFS-DocC model. Here, DUC 2004 online available dataset is used for evaluation. The input from the dataset is given for pre-processing and further process is carried out. The similarity and the correlation among the clustered document are examined and summarized to extract the essential features for provided document. Therefore, the proposed PFS-DocC model enhances the performance of the clustering algorithm. The simulation is done with MATLAB environment.

The performance of the PFS_DocC model is evaluated with the adoption of DUC 2004 benchmark dataset. The performance is measured using the ROUGE score toolkit. Various metrics like accuracy, F1-score, recall, and precision are measured for PFS-DocC model with 98.5% accuracy and 99% F1-score, recall, and precision. The outcome of the proposed PFS-DocC model is higher when compared to other approaches like FLSA (ProbIDF), Prob (Normal), FLSA (IDF), FLSA (Entropy), LDA, and LSA respectively. Similarly, the comparison is done with two benchmark dataset known as DUC 2003 and DUC 2004 for evaluating the performance of PFS + DocC with and without feature selection process. Also, the information extracted and the non-redundant data evaluation is also done for the PFS + DocC model. The performance show better trade-off in contrast to prevailing approaches. However, there is a constraint, as the proposed PFS + DocC model does not provided for classification. It will be concentrated in future along with the optimization process.

Acknowledgment

I thank the Management, Principal and Head of the department to support and provide the resources to carry out this research work.

References

[1] Vinaitheerthan Renganathan, Text mining in biomedical domain with emphasis on document clustering,' Healthcare Inform. Res., 23 (2017), 141_146, http:// 10.4258/hir.2017.23.3.141.

[2] Cheng and M. Lapata, Neural summarization by extracting sentences and words, 2016, arXiv:1603.07252. [Online]. Available: https://arxiv.org/abs/1603.07252.

[3] Yang, X. Cai,Y. Zhang, and P. Shi, Enhancing sentence-level clustering with ranking-based clustering framework for theme-based summarization, Inf. Sci. 260 (2014), 37-50.

[4] Hong, M. Marcus, and A. Nenkova, System combination for multidocument summarization, in Proc. Conf. Empirical Methods Natural Lang. Process. 2015, 107-117.

[5] Wang, W. Lam, Z. Ren, and L. Bing, ``Salience estimation via variational auto-encoders for multi-document summarization," in Proc. 31st AAAI Conf. Artif. Intell. 2017, 1-9.

[6] Cao, F. Wei, L. Dong, S. Li, and M. Zhou, ``Ranking with recursive neural networks and its application to multi-document summarization," in Proc. 29th AAAI Conf. Artif. Intell. 2015, 1-9.

[7] Ren, Z. Chen, Z. Ren, F.Wei, L. Nie, J. Ma, and M. De Rijke, Sentence relations for extractive summarization with deep neural networks, ACM Trans. Inf. Syst. 36 (2018), 1-32.

[8] Ren, F. Wei, Z. Chen, J. Ma, and M. Zhou, A redundancy-aware sentence regression framework for extractive summarization, in Proc. 26th Int. Conf. Comput. Linguistics Tech. Papers. 2016, 33-43.

[9] Cao, F.Wei, S. Li,W. Li, M. Zhou, and W. A. N. G. Houfeng, Learning summary prior representation for extractive summarization,' in Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics. 2 (2015), 829-833.

[10] Hong and A. Nenkova, Improving the estimation of word importance for news multi-document summarization, in Proc. 14th Conf. Eur. Cha ter Assoc. Comput. Linguistics. (2014), 712-721.

[11] Cao, W. Li, S. Li, and F. Wei, Retrieve, Rerank and rewrite: Soft template based neural summarization, in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics. (2018), 152-161.

[12] Tang, L. Yan, Z. Yang, and Q. H. Wu, Improved document ranking in ontology-based document search engine using evidential reasoning, IET Software. 8 (2014), 33–41.

[13] Huang and X. X. Zhou, Knowledge model for electric power big data based on ontology and semantic web, CSEE Journal of Power and Energy Systems. 1 (2015), 19–27.

[14] MacQueen, Some methods for classification and analysis of multivariate observations, in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1 (1967), 281–297.

[15] Vega-Pons and J. Ruiz-Shulcloper, A survey of clustering ensemble algorithms, International Journal of Pattern Recognition and Artificial Intelligence. 25 (2011), 337–372.

[16] Vega-Pons, J. Correa-Morris, and J. Ruiz-Shulcloper, Weighted partition consensus via kernels, Pattern Recognition. 43 (2010), 2712–2724.

[17] Zhang, Y.-F. Pu, S.-Q. Yang, J.-L. Zhou, and J.-K. Gao, An ontological Chinese legal consultation system,' IEEE Access. 5 (2017), 18250-18261.

[18] Kim, H. Jang, H. J. Kim, and D. Kim, A document query search using an extended centrality with the Word2vec,' in Proc. ICEC, Suwon, South Korea, 2016, Art. no. 14.

[19] Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, from Word Embeddings to Document Distances,' in Proc. ICML, Lille, France, 2015, 1-10.

[20] Koniaris, G. Papastefanatos, and Y. Vassiliou, Towards automatic structuring and semantic indexing of legal documents,' in Proc. PCI, Patras, Greece, 2016.

[21] Zhang, Y.-F. Pu, and P. Wang, An ontology-based approach for Chineselegal information retrieval,' in Proc. CENet, Shanghai, China, 2015, 1-7.

[22] Zhang, Y. Xu, and W. Zhang, Clustering scientific document based on an extended citation model, IEEE Access. 7 (2019), 57037–57046.

[23] Yoon, J. Lee, S.-Y. Park, and C. Lee, Fine-grained mobile application clustering model using retrofitted document embedding, ETRI J. 39 (2017), 443−454.

[24] Nayak, Fine-grained document clustering via ranking and its application to social media analytics, Soc. Netw. Anal. Min. 29 (2018).

[25] Kim, D. Seo, S. Cho, and P. Kang, Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec, Inf. Sci. 477 (2019), 15–29.

[26] Chen, F. S. C. Tseng, and T. Liang, An integration of WordNet and fuzzy association rule mining for multi-label document clustering, Data Knowl. Eng. 69 (2010). 1208−1226.

[27] Duan, Y. Li, R. Li, R. Zhang, X. Gu, and K. Wen, LIMTopic: A framework of incorporating link based importance into topic modeling, IEEE Trans. Knowl. Data Eng, 26 (2014), 2493–2506.

[28] Chali and S. A. Hasan. Query-focused multi-document summarization: Automatic data annotations and supervised learning approaches, Natural Language Engineering. 18 (2012), 109-145.

[29] Liu, J. Flanigan, et al. Toward Abstractive Summarization Using Semantic Representations. In HLT-NAACL (2015).

[30] Li, D. Cheng, L. He, et al. Joint Event Extraction Based on Hierarchical Event Schemas from FrameNet[J]. IEEE Access, 7(2019), 25001-25015.