Computer Systems Science & Engineering
DOI:10.32604/csse.2022.020459
Article

Tech Science Press

Document Clustering Using Graph Based Fuzzy Association Rule Generation

P. Perumal*

Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Coimbatore, India
*Corresponding Author: P. Perumal. Email: perumalp@srec.ac.in
Received: 25 May 2021; Accepted: 25 October 2021

Abstract: With the wider growth of web-based documents, the necessity of automatic document clustering and text summarization is increased. Here, document summarization that is extracting the essential task with appropriate information, removal of unnecessary data and providing the data in a cohesive and coherent manner is determined to be a most confronting task. In this research, a novel intelligent model for document clustering is designed with graph model and Fuzzy based association rule generation (gFAR). Initially, the graph model is used to map the relationship among the data (multi-source) followed by the establishment of document clustering with the generation of association rule using the fuzzy concept. This method shows benefit in redundancy elimination by mapping the relevant document using graph model and reduces the time consumption and improves the accuracy using the association rule generation with fuzzy. This framework is provided in an interpretable way for document clustering. It iteratively reduces the error rate during relationship mapping among the data (clusters) with the assistance of weighted document content. Also, this model represents the significance of data features with class discrimination. It is also helpful in measuring the significance of the features during the data clustering process. The simulation is done with MATLAB 2016b environment and evaluated with the empirical standards like Relative Risk Patterns (RRP), ROUGE score, and Discrimination Information Measure (DMI) respectively. Here, DailyMail and DUC 2004 dataset is used to extract the empirical results. The proposed gFAR model gives better trade-off while compared with various prevailing approaches.

Keywords: Document clustering; text summarization; fuzzy model; association rule generation; graph model; relevance mapping; feature patterns

1 Introduction

The strength and volume of data after the increasing use of web is an unceasing repository pool where the resources are products of the internet given for the sake of humans [1]. Various document clustering and text summarization approaches are available in the past few decades with the analysis of unlabeled data attained from the unsupervised data analysis [1]. The data (document) clustering process is also an unsupervised learning approach which uses unstructured information and provides the relationship among the data [2]. The document clustering process is highly acknowledged as a most appropriate element in

computer vision, data mining, pattern recognition, machine learning, computational biology, and clinical diagnostic approaches [3]. Information is considered as the personal source of repositories over the public environment which is generally accessible in a textual form [4]. The term big data is attained from the evolution of enormous unstructured, structural, and semi-structured data [5]. These sort of data are constantly evolves over the Internet with various forms like news, social networking, discussions, articles, e-mails, educational materials and so on [6]. Generally, text data comes under the unlabeled data structure format. It is determined as an efficient candidate model to examine the clustering techniques.

The data clustering approaches are deployed for efficient computation in past few years. Various transformation in clustering model are provided to coin the single and multi-objective constraint where the model works prominently under complex environment [7]; however, it does not entails better outcomes with better prediction accuracy. These approaches possess various natural biases to generate superior outcomes on datasets [7]. Therefore, the degradation over the generalized data clustering model promotes the result blending from various approaches to acquire superior clustering results. There are diverse methods where the prevailing approaches are merged or hybridized to acquire better outcomes. Multi-view, consensus, accumulated, and ensemble clustering are some common terms provided for document clustering. The concept of the consensus clustering mechanism is performed with diverse strategies like utilization of diverse clustering model with same parameters, adoption of single objective strategy with multiple clustering factors or the hybridization of both the models.

Similarly, other concepts like training data partitioning and the utilization of every data partition with diverse clustering model is provided with an efficient outcomes. The clustering outcomes from diverse baseline models are provided to establish the common relationship among these strategies [8]. The relationship is provided in a simpler manner with cluster-to-cluster mapping and individual cluster relationship. The successive stages of consensus mechanism are to attain expatiated clustering relationship where the clustering concept is determined from the similar way. It can allocate various newer labels with efficient voting formula and the variations with the baseline segmentation process (data partitioning) during clustering process [9]. The voting process adopts feature selection from base cluster representation, energy minimization with graphical representation, pair-wise similarity establishment among the cluster assignments and data points respectively. Certain consensus methods follow the below-given strategies: iterative voting, hierarchical agglomeration, clustering consensus, expectation-maximization, furthest consensus, and fragment-based clustering model.

There are an enormous amount of unstructured text documents that are used by the humans every day in their life. It is not so easier to process the text without the accessibility of any automatic approach. The automatic corpus processing model relies on dynamical information grouping in the process of text retrieval [10]. The text document grouping efficiency reduces with the larger corpus size. Moreover, text document grouping is done by partitioning the multiple text documents to associated subsets which is considered as the problematic task of text processing. Moreover, the effectual information retrieval process reduces with the absence of certain corpus grouping. Ultimately, less efficient grouping model enhances the effort and cost which involves in the process of information retrieval [11]. The experts in text mining explain processing time and cost based on the information diversity and information size of the underlying corpus. The efficient text processing is enhanced by lesser time consumption with less relative corpus [12]. Hence, the optimal time utilization and cost evaluation for processing corpus is more complex during information retrieval process. The appropriate text extraction and meta-data analysis is used to handle these issues effectually by playing the significant role during the process of cost, time and effort respectively.

This research concentrates on proposing an efficient document clustering and text summarization with proper rule generation by data mapping and connecting the relevance of data based on weighted graph model.

This model intends to reduce the error during data clustering as the rules are formed in an efficient manner. The text clustering model uses fuzzy classifier for training the data labels produced form the preliminary clusters. The clustering process is utilized to enhance the outcomes of document classification in the earlier stage; however the concept behind classification is to enhance the outcomes of document clustering as an added contribution towards the result.

The automatic data clustering and summarization process over the unstructured text is considered to be a most challenging task [12]. The preliminary cause behind this process is lack of completeness and robustness to deal with the actual text with lesser amount of words or lines. It is complex to ensure the shorter form of extracted data with the underlying textual information. Based on the literature analysis, the investigators need to examine the problem related to specification of text in two different forms known as abstraction and extraction [13]. The meta-data processing or abstraction processing technique provides keyword-based information regarding the text as abstractive summarization. Similarly, collection of meta-data document provides relevant set of documents indeed of document processing with individual meta-data. The keyword-based clustering model produces meta-data group in the dictionary format [13]. This is considered as the most challenging factors, i.e., semantic dependency over restrictive languages. Moreover, it has extensive application with the provided documents which comprises of relative unstructured text format. It is extremely appropriate for tasks like document title, headlines, keywords, or websites, sentence fusion, sentence compression and so on. The appropriate keyword selection process holds various limitations and concerns based on the analysis. It cannot fulfill the comprehensive text analysis process. The significant contributions of the research are:

1. To select the more appropriate online accessible dataset for data clustering and summarization. Here, DailyMail and DUC 2004 dataset is used to extract the empirical results.
2. A novel intelligent model for document clustering is designed with graph model and Fuzzy based association rule generation (gFAR).
3. The empirical analysis is done using MATLAB simulation environment to evaluate metrics like Relative Risk Patterns (RRP), ROUGE score, and Discrimination Information Measure (DMI) respectively.

This work is arranged as follows: Section 2 includes extensive analysis with the background studies related to data clustering; Section 3 explains the novel intelligent model for document clustering using graph model and Fuzzy based association rule generation (gFAR). Section 4 is numerical results and discussion. Section 5 is conclusion with future research directions.

2 Related Works

There are various extensive analyses carried out for unsupervised and other clustering approaches termed as topic modelling; however, it provides a lower-dimensional embedding process for exploring the textual datasets. Certain appropriate approaches are concentrated on website search-based outcomes. Some highlighted models are recommended with user interpretation for extracting the outcomes in the descriptive keywords, titles, or phrases that summarize the semantic content of topics and clusters for certain users [14]. When the dataset holds well-known document clustering and ground truth topic categories which are computed with certain measures; however comparisons of certain descriptive labels are generally relay on human computation [15]. Specifically, the descriptive measures are more challenging for certain datasets, paradigm modelling or clustering and descriptive forms changes in an extensive manner [16]. Even though, there is a certain user computation process that is focused on labels with a user's preference where the evaluation is focused on whether the description assists the users in analyzing the most effective clustering process [17]. Based on the user's knowledge, no prior models are

provided for descriptive clustering as an identification issue with quantification objective based on classification performance. Some unique features of this model include principle model to choose the number of descriptive features and automatic methods are chosen based on the number of clusters which is autonomous towards clustering model [18]; however, it is dependent on binary feature representation and cluster assignments.

The objective behind the descriptive clustering model towards text datasets is utilized as an information retrieval model [19]. The users can predominantly sense the relevance descriptions *vs.* the cluster determination of the appropriate cluster model by manual evaluation of document samples [20]. An iterative model known as the scatter-gathering model utilizes successive descriptive clustering stages to assist the users to predict the appropriate relevant documents. The initial clustering process is provided with the certain description or projecting the chosen clusters are merged and clustered in a successive manner [21]. This process is continued till the user sharpens the appropriate document set. The automatic description quality is more crucial to predict the relevant clusters. Some exploratory model is contrasted to some traditional query-based information retrieval system [22]. With query-based systems predominant exploratory analysis, web searchers are resourceful when users consider this model as a non-relevant topic with the certain corpus (the variations are identified among a set of complete text documents towards the set of shorted summaries of every outcome retrieved from the search engines) or it is more appropriate for formulating the query model to attain the appropriate instances [23]. The descriptive clustering models are carried out with the initial clustering model and then it predicts the set of feature (characteristics) related to every cluster.

It facilitates various applicable clustering models to be utilized. The chosen features gain superior users; information over the cluster content process (cause of this concept) which is more challenging. The preliminary concept behind the process is determined based on the clusters which are more likely towards the cluster words, over the cluster center, titles (available) or phrases with certain content alike of words [24]. Moreover, these data features are not more optimal for establishing discrimination among the various clusters. Some scoring criteria like information gain, mutual information indeed of point-wise mutual information which is utilized for choosing to discriminate these features (phrases or keywords) for certain clusters. Some other approaches are simultaneously grouped into a set of features which influence the cluster [25]. The grouping features are mitigated with some issues with noise and feature sparsity and the utilization of these clustered features is provided to enhance the clustering performance. Some novel techniques are grouped directly to gain knowledge regarding the joint vector space specification where both the instances and the features are given descriptively [26]. The vectors specifying the instances and features are optimized towards the nearer vectors that are semantically similar to one another. With these provided joint vector space, the cluster samples are labelled with certain features that reside over the boundaries [27].

With the above-mentioned model, it is not obvious to project how these objective measures are chosen based on the feature labels and lists that function as a descriptive model. The hypothetical condition of these descriptions is essential when it facilitates the user with an appropriate prediction of cluster content. In some cases, predicting the feature set characteristics is provided to determine the feature selection process [28]. For example, the decision tree is trained for every cluster to categorize the instances directly over the absence or presence of these features. The corresponding Boolean expression for the decision tree provides a descriptive set. The clustering and decision tree is generated concurrently from the hierarchical clustering model with the occurrence or absence of certain individual phrases or features with suffix-tree-based clustering [29]. Some methods can link explicitly with organization description of documents and the outcomes are provided with an arbitrary clustering model.

Various techniques for text analysis are done with word embeddings that are produced with a neural network model. The word-based embedding model is represented in a distributive manner of single words [30]. In contrary to the one-hot vector utilized for bag-of-words, the lower dimensionality and distributed vectors are more dense and continuous. With the word embedding model, the words possess the same meaning for certain specifications and thus word embedding is more appropriate for assisting in learning towards latent document semantics. The semantic word information is captured from word embeddings which is improved by various methods [31]. For instance, word mover distance is used for evaluating the semantic association among the documents with the use of earth movers' distance-based word embeddings. When compared to the conventional text classification process based on text similarity, the above-mentioned model is proved to be more efficient than the conventional LDA and LSI respectively [32]. Moreover, it is conventionally expensive for huge documents. However, certain semantic level document information cannot be attained with word-based embedding indeed of word-level semantic data owing to the fact that word2 vec based feature specifications are specified as input. It cannot capture the document-level semantic information which is completely efficient [32]. Generally, it is complex for the word2 vectors for interpreting the learned data with appropriate higher-level semantics (for instance: document level semantics). Specifically, these models are considered to be more sensitive during the training process of data distribution. Therefore, it is poorly attained by the distributed data due to wrong generalization [33]. These issues are handled efficiently with the proposed graph model which is used to map the relationship among the data (multi-source) followed by the establishment of document clustering with the generation of association rule using the fuzzy concept. This method shows benefit in redundancy elimination by mapping the relevant document using graph model and reduces the time consumption and improves the accuracy using the association rule generation with fuzzy. This framework is provided in an interpretable way for document clustering.

3 Methodology

Here, the novel intelligent model for document clustering is designed with a graph model and Fuzzy based association rule generation (gFAR). Fig. 1 illustrates the flow diagram of the gFAR model. Initially, the graph model is used to map the relationship among the data (multi-source) followed by the establishment of document clustering with the generation of association rule using the fuzzy concept. This method shows benefit in redundancy elimination by mapping the relevant document using graph model and reduces the time consumption and improves the accuracy using the association rule generation with fuzzy.

3.1 Graph-Based Document Relationship Establishment

The proposed model introduces the graph-based document clustering and summarization approach that facilitates multi-labelling and determines the number of clusters automatically based on a set of documents. The most essential task during document clustering is to evaluate the relationship among the document. The relationship is learned based on document vectors to measure the document quality. Here, a framework employs a probabilistic model to consider the graph-based model for extracting the data resources. The data clustering model includes the below-given components.

a) **Cluster initialization**: 'W' is the words that are extracted from the provided from the document set during cluster initialization (words), i.e., {C_k^i}_(k = 1)^k where C_k^i shows the initialized clusters. The summarized data is extracted from the provided document title where the association tiles are utilized to extract the related words. Some unnecessary (stop words) are eliminated from the extracted nouns. Moreover, the weights of the documents are allocated towards the remaining words where the nouns (higher weight) are extracted from the salient features.

b) **Cluster analysis:** weight-based probabilistic generative p(x| c_k^i) model is learned where the documents are allocated with clusters C_k^i along with the posterior probability. Subsequently, documents are provided C_k^e = {x_i} where p(C_k^i| x)>threshold and x_i εV(G) leads to {C_k^e}_(k = 1)^k construction where C_k^e specifies the extended cluster.

c) **Cluster summarization:** The similarity among the pairs, i.e., C_k^e and C_k^e that belong to C^e. The clusters with better similarity are integrated to design the {C_k^m}_(k = 1)^(k′) where K ≠ K^′ is specified with the integrated cluster module.

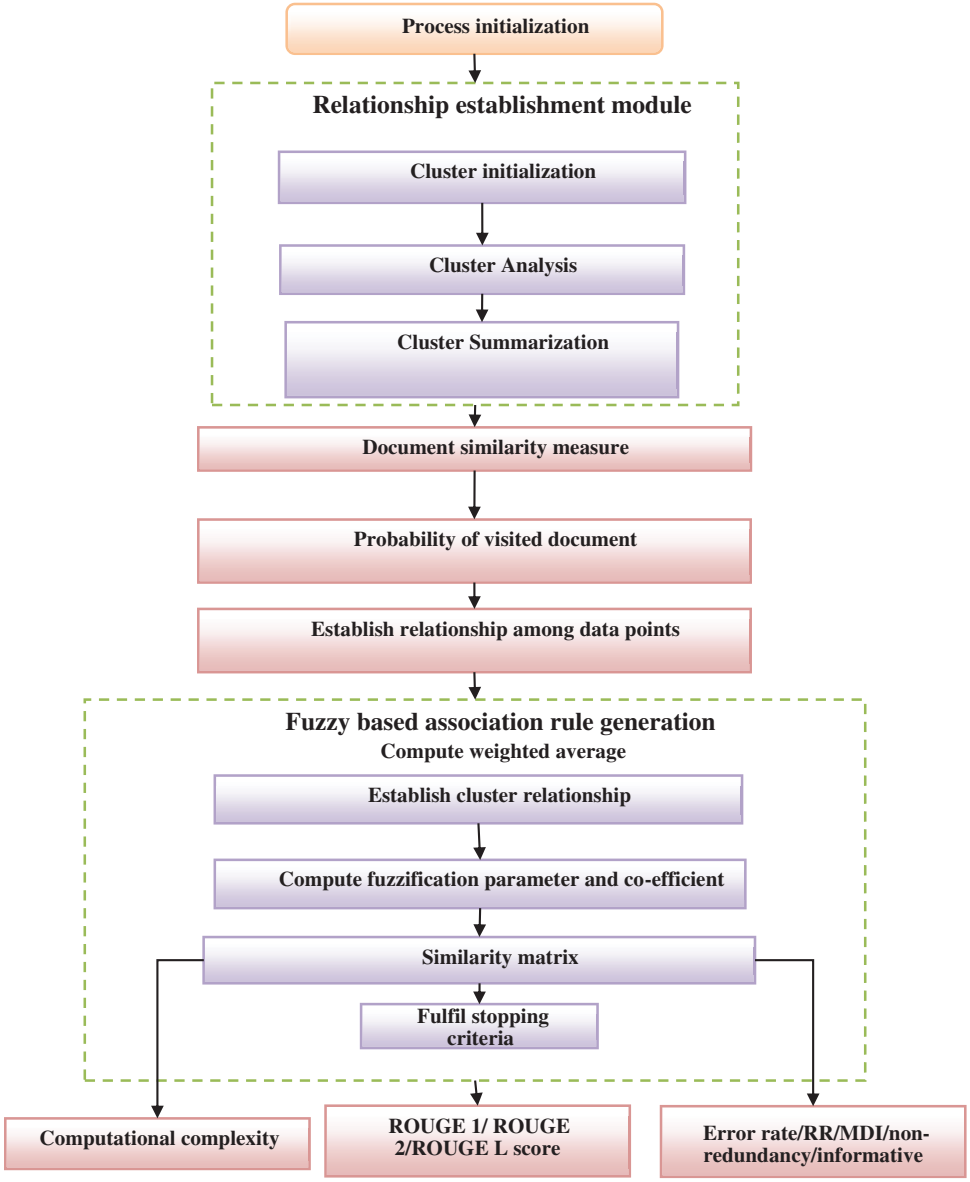

Figure 1: Flow diagram of gFAR model

The initial clusters are composed of documents (words) with similar words. The cluster is expanded during the document allocation where p(C_k^i |x) is superior than cluster threshold C_k^i. The extracted

words are clustered (words and synonyms). Then, the cluster produces higher similarity that needs to be integrated. For cluster integration, the similarities among the cluster vectors are provided in an efficient manner. The vectors are provided based on average document vectors of similar clusters. To evaluate the similarity, the documents are vectors from document and multi-source of the network. The clusters are initialized based on time complexity with O(n) as words are extracted with association rules O(n) time. It is provided as O(Kn) time where 'K' is initial document cluster. The cluster integration with O(K^′2) is computed as the cluster pair. The value of K^′ is smaller than 'n' which is scaled with larger documents.

3.2 Probabilistic Graph for Data Clustering

The objects with higher network links are considered for further connection establishment. Here, $p(x|G)$ shows the probability of document visited over the graph 'G' which shows significance of other models. The document clustering model is considered for evaluating the probabilistic model for clustering $p(x|G)$. The relationship among the data is mapped with object reflected with the edge weight. The probability of visiting object is proportional to sum of neighbours. It is expressed as in Eq. (1):

$$p(x|G)\alpha \sum_{x_j \varepsilon N_G} p(x_j|G)^{w}_{x,x_j} \tag{1}$$

Here, $N_G(x)$ specifies object neighbour 'x' in graph 'G' and w_{x,x_j} specifies weighted edge and x_j neighbourhood. The posterior probability of the document x_i comes under the graph model G_k which is expressed as in Eq. (2):

$$p(G_k|x_i) = p(x_i|G_k) * \ p(G_k)/p(x_i) \tag{2}$$

From the above Eq. (2), it is observed that the $p(x_i)$ values are identical. Similarly, w_{x_i,x_j} is provided to establish the association and strength of the data. The neighbourhood document with higher relationship shows better significance over the posterior probability. When x_j significance is higher; then the probability of visiting x_j over G_k. Thus, the function provides the relationship establishment among the data which is expressed as in Eq. (3):

$$p(x_j|G_k)^{w_{x_j},x_j} = f_d(x_i, \ x_j) * p(x_j|G_k) \tag{3}$$

The probability of visiting the document (f_d) is computed by evaluating the relationship among x_j and G_k. The average association is established using Eq. (4):

$$p(x_j|G_k)^{w_{x_i},x_j} = f_d \ (x_i, \ x_j) * f_d(x_j, \ \ avg(V(G_k)) \tag{4}$$

The score function is directly proportional to posterior probability where the object belongs to cluster and expressed as in Eq. (5):

$$score(C_k, \ \ x_i) = \frac{|V(G_k)|}{|V(G)|} \sum_{x_j \varepsilon N_G k(x)} f_d(x_i, \ \ x_j) * f_d(x_j, \ \ avg(V(G_k)) \tag{5}$$

The process is determined based on the number of objects belongs to the graph by total amount of objects. It is expressed based on the data relationship towards the document as the score is directly proportional to the document relationship during the clustering process.

3.3 Fuzzy Based Document Association Rule Generation

Here, fuzzy-based clustering approach is utilized for clustering the documents based on the association rule generation among the clusters. The data is partitioned into various clusters where x_l, ($l = 1, 2, 3, \ldots, k$) are related to cluster centric C_l. The cluster relationship and the association among the mapped data point are

considered as fuzzy. The membership function $u_{i,j}\,\varepsilon\,[0, 1]$ which specifies the data points and cluster centre. The data point set $S = \{X_i\}$ is provided with minimal distortion where $u_{i,j}$ is membership and it specifies the cluster as shown in Eq. (6):

$$J = \sum_{j=1}^{k}\sum_{i=1}^{N} u_{i,j}^{q} d_{i,j} \tag{6}$$

Here, 'N' specifies number of data points, 'q' fuzzifier parameters, number of clusters are specified as $d_{i,j}$ and 'k' which is squared with Euclidean distance among the clusters (X_i *and* C_j). The fuzzification parameters are provided with appropriate clusters. The mappings of association among the vectors are improved with data point partitioning. The process is initiated with initial cluster centres and forwarded until the stopping criteria are fulfilled. When no two documents (clusters) are similar, then $d_{i,j} < n$ and $u_{i,j} = 1$ and $u_{i,j} = 0$ for all $l \neq j$ with 'n' positive numbers. The clusters are provided for computation and update the membership function with Eq. (7). The membership degree of all documents with clusters is provided with the association among the documents. It is known as fuzzification coefficient and expressed as in Eq. (7):

$$u_{i,j} = ((d_{i,j})^{\frac{1}{m}-1} \sum_{l=1}^{k}\left(\frac{1}{d_{il}}\right)^{\frac{1}{m}-1} \tag{7}$$

Here, $d_{i,j} < n$ and values of 'n' is extremely smaller and $u_{i,j} = 1$. It is adopted to predict the topics and words of the matrix. The clusters are evaluated to attain better clusters which are specified as $C_p + 1$. The generation of newer clusters are provided as in Eq. (8):

$$C_j(p) = \frac{\sum_{j=1}^{N} u_{i,j}^{m} X_i}{\sum_{j=1}^{N} u_{i,j}^{m}} \tag{8}$$

The process is stopped when $(\|C_j(p) - C_j(p-1)\|) < \in$ and $j = 1$ *to* k. Stop when $\varepsilon > 0$ is positive number. Else, set $p + 1 \rightarrow p$ and move to fuzzy association algorithm. The total number of computation and the complexity of this model is provided as $O(N_{kt})$ and 't' specifies various iterations. The algorithm for the proposed model is given below:

Algorithm 1: Graph based weighted document measurement

Initialize: set of documents

Output: Document based ROUGE score

1. Compute average weighted matrix [n][n];
2. Compute array score [n];
3. for i ←1 to n do
4. for j ←1 to n do
5. Evaluate the similarity based on document clusters (C[i],C[j]);
6. Compute the probability of visited document using Eq. (1);
7. Compute posterior probability using Eq. (2):
8. Evaluate the data relationship by mapping the data points;
9. Compute the score function using Eq. (5);

(continued)

Algorithm 1: (continued)

10. Compute weighted average with average values.
11. end
12. end
13. ROUGE score evaluation with Eq. (9);
14. return score;

Algorithm 2: Establishing Fuzzy based association rule among the documents

Initialize: set of documents

Output: Associate among the documents (clusters)

1. Establish cluster based relationship by mapping the data points;
2. Measure the data points length using Euclidean distance measure;
3. Compute the fuzzification parameters using Eq. (7);
4. Derive fuzzy co-efficient;
5. Generate similarity matrix using topics and words as (S[i],S[j]);
6. Generate newer clusters using Eq. (8);
7. Repeat //until stopping criteria is fulfilled
8. Measure the computational complexity // O(N_kt)
9. Categorize the document clusters as 1 and 2 as in Eqs. (10) and (11);
10. Extract discrimination information using Eqs. (12) & (13);
11. Compute the error based on weighted average;
12. end process;
13. Return error rate.

The weighted edges of the documents are measured with the association among the data points. The ROUGE scores are attained and maintained over the outcomes. The process can be merged and the scores are evaluated. The probability of word occurrence over the input clusters is attained from the background knowledge to predict the most appropriate words. For all input clusters, the background corpus is similar other than the chosen cluster. By summarizing the first cluster of the DUC 2002 dataset, 29 clusters use background corpus. Three different ROUGE scores are measured for providing the significance of the model.

4 Numerical Results and Discussion

Here, a network traffic dataset termed as NSL-KDD dataset is introduced in association with the anticipated model. To analyze the functionality of this method, diverse experimental comparisons are performed. This dataset includes both testing and training sets. The features chosen determine the dataset description with preliminary statistical and contents information towards network connection. The feature size is given as 41. The dataset label includes five diverse network events like a probe, normal, denial of service (DoS), user to root, and remote to local (R2L). Various investigators consider the NSL-KDD

dataset as an authoritative benchmark standard in intrusion detection. Thus, the NSL-KDD dataset is considered in this work for valuating the semi-supervised approach. It comprises of various attack patterns that are more appropriate for validating generalization capability. Here, random samples are chosen and remaining samples are utilized as unlabeled data. Here, intrusion detection is considered a multi-class problem. The experimentation is performed in PC. The system configurations are given as Intel i5 processor, Windows 7 OS, 8 GB RAM @3.00 GHz.

There are two diverse features known as numerical and symbolic. The anticipated model deals with symbolic features and values are not distributed randomly. It also triggers negative effect over learning process. To get rid of this problem, data normalization and one-hot encoding approach are used before learning process. The feature values are sequence encoded with 0 and 1. Dimensionality change based on distinctive values of symbolic features. Features like 'protocol type', 'service', and 'flag' are encoded when values are higher than 2. Symbolic features are treated as Boolean type with 1 or 0.

The above Tab. 1 portrays the statistics of different data sets. Here, three different metrics are used for evaluation purposes, i.e., Relative Risk Patterns (RRP), ROUGE score, and Discrimination Information Measure (DMI) respectively. ROUGE is expanded as Recall oriented understudy for gisty evaluation is adopted for automatic evaluation of provided summaries. It relies on the comparison of n-grams among the number of hand-written references and summary for evaluation. It is expressed in Eq. (9):

$$ROUGE = \frac{\begin{matrix} S \in [reference\ summaries] \\ \sum\sum_{gam\_(n \in s)} count\ match\ (gram_n) \end{matrix}}{\begin{matrix} S\epsilon\ (reference\ summaries) \\ \sum \sum^{g} ram\ n\varepsilon s\ (gram_n) \end{matrix}} \tag{9}$$

ROUGE toolkit is utilized to compute the performance of proposed gFAR model which includes ROUGE-1, ROUGE-2, and ROUGE-L respectively.

Table 1: Computational statistics with different data sets

Datasets	Documents	Classes	Instance size		Data length	
	Total	Total	Max	Average	Min	Average
DailyMail	8094	4	4203	2774	2033	39
Re0	1504	13	608	116	11	69
DUC 2004	9649	165	4725	131	1	42
WebKB	4199	4	1641	1050	504	124

Tab. 2, and Fig. 2 depict the results attained from the ROUGE-1/ROUGE-2/ROUGE-L score which shows that the proposed gFAR outperforms the best value of all the ROUGE metrics that are used in the experimentation with the DUC 2002 dataset respectively. It projects that the hierarchical model provides better document specification and the finest sentences. It shows the abstract features which give better performance of task summarization. Similarly, in case of Tab. 3, and Fig. 3 depicts the results attained from the Daily Mail dataset. The outcomes show that the proposed gFAR model outperforms the existing model with baseline ROUGE metrics.

Table 2: Performance evaluation with DUC 2002

Model	DUC 2002		
	ROUGE-1	ROUGE-2	ROUGE-L
URank	0.490	0.220	-
Tgraph	0.485	0.230	-
Lead-3	0.440	0.220	0.405
Cheng'16	0.475	0.240	0.440
SummaRunner	0.475	0.230	0.425
HSSAS	0.530	0.230	0.490
gFAR	0.621	0.255	0.500

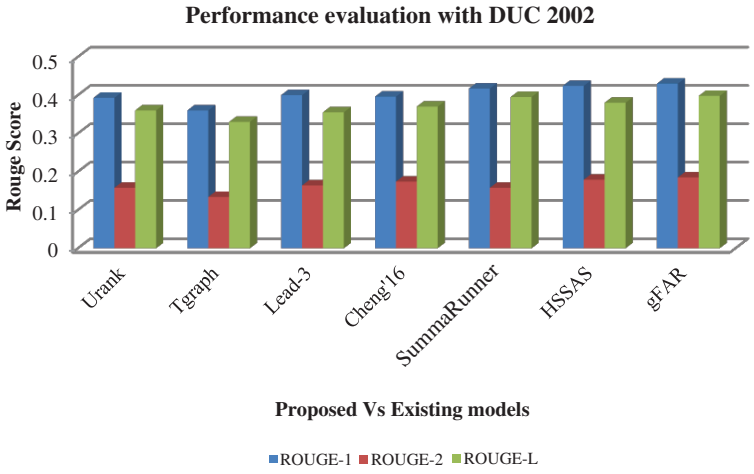

Figure 2: Performance evaluation with DUC 2002

Table 3: Performance evaluation with Daily Mail

	DUC 2002		
Model	ROUGE-1	ROUGE-2	ROUGE-L
URank	0.393	0.158	0.360
Tgraph	0.360	0.134	0.330
Lead-3	0.400	0.164	0.355
Cheng'16	0.396	0.174	0.370
SummaRunner	0.417	0.158	0.395
HSSAS	0.424	0.179	0.380
gFAR	0.430	0.185	0.398

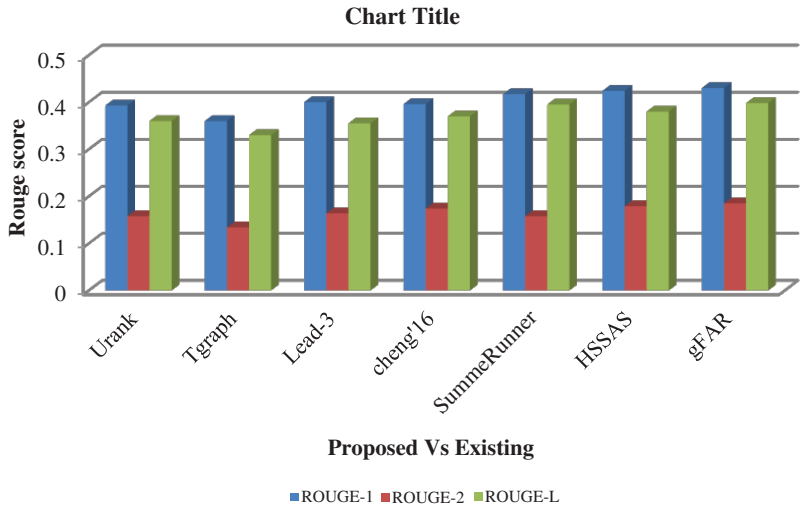

Figure 3: Performance evaluation with DUC 2002

With the news articles, it is extremely essential for the information that is initiated from the article. It provides better ROUGE metrics that beat the performance of URank, Tgraph, Lead-3, Cheng'16, SummeRunner, and HSSAS respectively. When dealing with the abstractive model, the ROUGE measures overlap with one another with minor variations which are not so essential during readable form. The preliminary ideal behind this ROUGE discrete metrics does not fulfill the increase in readability and quality of produced summary. It is provided to justify the ROUGE scores over the abstractive baselines utilized in this work. Then, the problem related to ROUGE metrics is increased with the numRR/MDIber of referral summaries of the given document. The ROUGE score inflexibility produces reference summary for all documents is much lesser than others when compared to multiple reference summaries. Finally, the proposed gFAR model attains better outcomes when compared to the prevailing models.

It is proven that the outcomes of gFAR have attained better quality outcomes. The outcomes of RR and MDI are compared between two models, i.e., clustering-based discrimination information maximization (CDIM) and gFAR respectively. The significant decision is attained with the choice established among the discrimination values (measurement of discrimination information (MDI) and relative risk (RR)) respectively. The performance of CDIM is poor when compared to gFAR model. The difference between the RR/CDIM and MDI/CDIM is not satisfied when compared to the performance of RR/gFAR and MDI/gFAR respectively. Some simple patterns are sensed from these outcomes. The results of RR are stronger with small 'K' values; however, MDI is stronger with higher 'K' values respectively. Tab. 4 depicts the comparison of RR and MDI of CDIM and gFAR respectively. Fig. 4 depicts the graphical representation of the metrics.

MDI is utilized to evaluate the term discrimination information for measuring the semantic relationship among the terms. It measures are provided as $ifd_{I1\varepsilon}$ and $ifd_{I2\varepsilon}$ respectively. It is the quantified discrimination among the distribution divergence among the combined data/category 1 and combined data/category 2 respectively. During the data clustering process, the category 1 and 2 are related with the provided data clusters C_k and $\bar{C}_k$ respectively. It is expressed as in Eqs. (10) & (11):

$$ifd_{I1\varepsilon}\ (x_j) = p(x_j|C_k)\log\frac{p(x_j|C_k)}{\lambda_1 p(x_j|C_k) + \lambda_2 p(x_j|\bar{C}_k)} \tag{10}$$

Table 4: RR and MDI evaluation

Name	RR/CDIM	MDI/CDIM	RR/gFAR	MDI/gFAR
Pu	0.763	0.608	0.820	0.620
Movie	0.605	0.575	0.650	0.580
Citeseer	0.440	0.405	0.450	0.410
Hitech	0.430	0.485	0.445	0.490
Tr31	0.630	0.598	0.675	0.610
Cora	0.365	0.340	0.385	0.350
Re0	0.412	0.440	0.425	0.450
wap	0.475	0.512	0.480	0.530

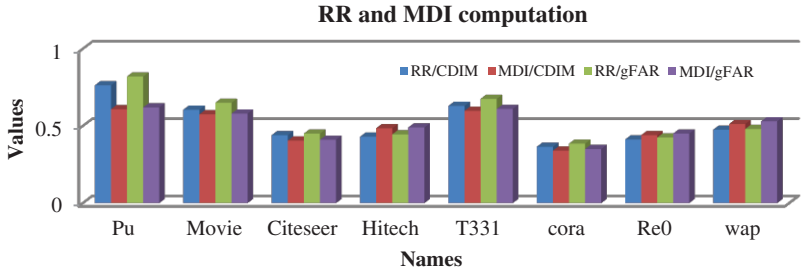

Figure 4: RR and MDI computation

$$ifd_{I2\varepsilon}(x_j) = p(x_j|C_k)\log\frac{p(x_j|\bar{C}_k)}{\lambda_1 p(x_j|C_k) + \lambda_2 p(x_j|\bar{C}_k)} \tag{11}$$

Here, λ_1 and λ_2 are prior probabilities of C_k and $\bar{C}_k$ respectively.

The relative risk of the data cluster C_k over other clusters $\bar{C}_k$ is considered as the discrimination information for all clusters $\bar{C}_k$. The discrimination information x_j for all clusters C_k and $\bar{C}_k$ is expressed as in Eqs. (12) & (13):

$$w_{jk} = \begin{cases} \dfrac{p(x_j|C_k)}{p(x_j|\bar{C}_k)} & p(x_j|C_k) - p(x_j|\bar{C}_k) > 0 \\ 0 & else \end{cases} \tag{12}$$

$$\bar{w}_{jk} = \begin{cases} \dfrac{p(x_j|\bar{C}_k)}{p(x_j|C_k)} & p(x_j|\bar{C}_k) - p(x_j|C_k) > 0 \\ 0 & else \end{cases} \tag{13}$$

Here, $p(x_j|C_k)$ is the conditional probability of x_j over the cluster C_k. The range of discrimination information relies on ($0 \rightarrow$ *no discrimination informaiton* or greater than 1) where the larger value specifies higher discriminative power.

Tab. 5 and Fig. 5 depict the human evaluation based on participant's percentage for all approaches. gFAR shows better results when compared to the prevailing models like URank, Tgraph, lead-3, Cheng'16, SummaRunner, and HSSAS respectively. It is more interesting that the outcomes of gFAR gives better data clustering efficiency without influencing the constraint optimization process. The feature significance

of gFAR is capable of learning the feature relevance based on the classes. The weights are average weight over the feature set. DUC 2002 gives a better feature-based selection of document summaries. The similarity shows a major impact on class discrimination owing to the summarization model. The error reduction rate is computed based on weighted factors as shown in Tab. 6 respectively. The error reduction rate is higher for gFAR when compared to other models.

Table 5: Human evaluation

Model	Informative	Non-redundancy	Overall (%)
URank	23%	21%	20%
Tgraph	13%	19%	16%
Lead-3	15%	16%	21%
Cheng'16	19%	22%	18%
SummaRunner	17%	22%	25%
HSSAS	20%	25%	27%
gFAR	28%	27%	28%

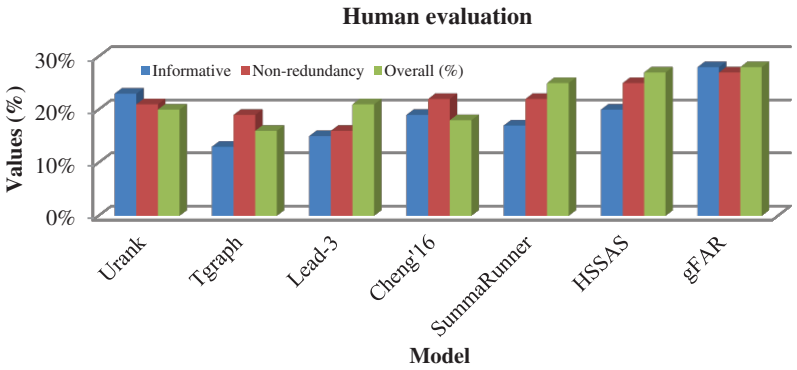

Figure 5: Human evaluation

Table 6: Error reduction rate using weighted graph

	DUC2002-ROUGE1	DUC2002-ROUGE2	Mail/ROUGE1	Mail/ROUGE2
gFAR+ weight	53.7	25.7	42.1	19.5
gFAR-weight	44.3	21.1	39.7	15.3

5 Summary

This work provides a novel intelligent model for document clustering is designed with a graph model and Fuzzy based association rule generation (gFAR). The unique characteristics of the dataset are maintained (DailyMail and DUC 2002) respectively. This framework is provided in an interpretable way for document clustering. It iteratively reduces the error rate during relationship mapping among the data (clusters) with the assistance of weighted document content. The simulation is done with MATLAB 2016b environment and empirical standards like Relative Risk Patterns (RRP), ROUGE score, and Discrimination Information Measure (DMI) respectively are measured. The error reduction rate of gFAR

+weighted value is 53.7 and 25.7 (DUC2002ROUGE 1 and ROUGE 2) and 42.1 and 19.5 (DailyMail ROUGE 1 and ROUGE 2) respectively. The information attained with the document clustering is 28%, 27% non-redundancy, and 28% overall performance. The ROUGE score (1/2/L) of gFAR with DUC 2002 is 0.621, 0.255, and 0.500 respectively. Similarly, the ROUGE score (1/2/L) of gFAR with DailyMail is 0.430, 0.185, 0.398 respectively. The proposed gFAR shows a better trade-off in contrast to prevailing approaches.

In the future, this research is extended by considering how the embedded words are merged with multi-source textual data to enhance the multi-source model and attain better multi-document text clustering to a certain semantic extent.

Funding Statement: The author received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] K. Salomatin, Y. Yang and A. Lad, "Multi-field correlated topic modelling," in *Proc. SIAM Int. Conf. on Data Mining*, Philadelphia, PA, USA: SIAM, pp. 628–637, 2009.

[2] L. Hong, B. Dom, S. Gurumurthy and K. Tsioutsiouliklis, "A time-dependent topic model for multiple text streams," in *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery Data Mining*, San Diego, USA, pp. 832–840, 2011.

[3] K. Sheng Tai, R. Socher and D. C. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proc. of the 53rd Annual Meeting of Association for Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Processing*, Beijing, China, pp. 1556–1566, 2015.

[4] Q. Jipeng, L. Yun, Y. Yunhao and W. Xindong, "Short text clustering based on pitman-yor process mixture model," *Applied Intelligence*, vol. 48, no. 7, pp. 1802–1812, 2018.

[5] W. Shao, L. He and P. S. Yu, "Clustering on multi-source incomplete data via tensor modeling and factorization," in *Proc. of Pacipic-Aisa Conf. on Knowledge Discovery and Data Mining (PAKDD)*, Springer, Ho Chi Minh City, Vietnam, pp. 485–497, 2015.

[6] Y. Yan, R. Huang, C. Ma, L. Xu, Z. Ding *et al.,* "Improving document clustering for short texts by long documents via a dirichlet multinomial allocation model," in *Proc. of the 1st Int. Conf. on Web and Big Data*, Beijing, China, pp. 626–641, 2017.

[7] J. Xu, B. Xu, P. Wang, S. Zheng, G. Tian *et al.,* "Self-taught convolutional neural networks for short text clustering," *Neural Networks*, vol. 88, pp. 22–31, 2017.

[8] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," in *Proc. of 20th ACM SIGKDD Int. Conf. on Knowledge Discovery Data Mining (KDD)*, New York, USA, pp. 233–242, 2014.

[9] A. Mohebi, S. Aghabozorgi, T. Y. Wah, T. Herawan and R. Yahyapour, "Iterative big data clustering algorithms: A review," *Software: Practice and Expertise*, vol. 46, no. 1, pp. 107–129, 2016.

[10] C. H. Chung and B. R. Dai, "A fragment-based iterative consensus clustering algorithm with a robust similarity," *Knowledge and Information System*, vol. 41, no. 3, pp. 591–609, 2014.

[11] F. Schwenker and E. Trentin, "Pattern classification and clustering: A review of partially supervised learning approaches," *Pattern Recognition. Letters*, vol. 37, pp. 4–14, 2014.

[12] T. Boongoen and N. Iam-On, "Cluster ensembles: A survey of approaches with recent extensions and applications," *Computer Science Review*, vol. 28, pp. 1–25, 2018.

[13] L. M. Abualigah, A. T. Khader, M. A. Al-Betar and O. A. Alomari, "Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering," *Expert System with Applications*, vol. 84, pp. 24–36, 2017.

[14] J. Zhao, X. Xie, X. Xu and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[15] F. Nie, G. Cai and X. Li, "Multi-view clustering and semi-supervised classification with adaptive neighbours," in *Proc. of Thirty-First AAAI Conf. on Artificial Intelligence*, San Francisco, California, USA, pp. 2408–2414, 2017.

[16] M. Lu, X. J. Zhao, L. Zhang and F. Z. Li, "Semi-supervised concept factorization for document clustering," *Information Science*, vol. 331, pp. 86–98, 2016.

[17] Y. Wang and Y. Pan, "Semi-supervised consensus clustering for gene expression data analysis," *BioData Mining*, vol. 7, no. 1, pp. 1–7, 2014.

[18] S. Wang and R. Koopman, "Clustering articles based on semantic similarity," *Scientometrics*, vol. 111, no. 2, pp. 1017–1031, 2017.

[19] H. Gan, N. Sang, R. Huang, X. Tong and Z. Dan, "Using clustering analysis to improve semi-supervised classi_cation," *Neurocomputing*, vol. 101, no. 3, pp. 290–298, 2013.

[20] G. N. Corrêa, R. M. Marcacini, E. R. Hruschka and S. O. Rezende, "Interactive textual feature selection for consensus clustering," *Pattern Recognition. Letters*, vol. 52, pp. 25–31, 2015.

[21] M. T. Hassan, A. Karim, J. B. Kim and M. Jeon, "CDIM: Document clustering by discrimination information maximization," *Information Science*, vol. 316, pp. 87–106, 2015.

[22] K. N. Junejo and A. Karim, "Robust personalizable spam filtering via local and global discrimination modeling," *Knowledge AndInformation Systems*, vol. 34, no. 2, pp. 299–334, 2013.

[23] E. Baralis, L. Cagliero, N. Mahoto and A. Fiori, "Graphsum: Discovering correlations among multiple terms for graph-based summarization," *Information Science*, vol. 249, pp. 96–109, 2013.

[24] D. Parveen, H. Ramsl and M. Strube, "Topical coherence for graph-based extractive summarization," in *Proc. of Conf. on Empirical Methods Natural Language Processing*, Lisbon, Portugal, pp. 1949–1954, 2015.

[25] M. Mendoza, S. Bonilla, C. Noguera, C. Cobos and E. León, "Extractive single-document summarization based on genetic operators and guided local search," *Expert System with Applications*, vol. 41, no. 9, pp. 4158–4169, 2014.

[26] R. Ferreira, L. S. Cabral, F. Freitas, R. D. Lins, G. F. Silva *et al.,* "A multi-document summarization system based on statistics and linguistic treatment," *Expert System with Applications*, vol. 41, no. 13, pp. 5780–5787, 2014.

[27] D. Parveen and M. Strube, "Integrating importance, non-redundancy and coherence in graph-based extractive summarization," in *Proc. of 24th Int. Joint Conf. on Artificial Intelligence*, Buenos Aries, Argentina, pp. 1298–1304, 2015.

[28] L. Yang, X. Cai, Y. Zhang and P. Shi, "Enhancing sentence-level clustering with ranking-based clustering framework for theme-based summarization," *Information Science*, vol. 260, pp. 37–50, 2014.

[29] C. Li, X. Qian and Y. Liu, "Using supervised bigram-based ILP for extractive summarization," in *Proc. of 51st Annual Meeting Association for Computational Linguistics*, Sofia, Bulgaria, vol. 1, pp. 1004–1013, 2013.

[30] P. Li, Z. Wang, W. Lam, Z. Ren and L. Bing, "Salience estimation via variational auto-encoders for multi-document summarization," in *Proc. of 31st AAAI Conf. on Artificial Intelligence*, California, USA, pp. 3497–3503, 2017.

[31] P. Ren, Z. Chen, Z. Ren, F. Wei, L. Nie *et al.,* "Sentence relations for extractive summarization with deep neural networks," *ACM Transactions on Information Systems*, vol. 36, no. 4, pp.1–32, 2018.

[32] P. Ren, F. Wei, Z. Chen, J. Ma, M. Zhou *et al.,* "A redundancy-aware sentence regression framework for extractive summarization," in *Proc. of 26th Int. Conf. on Computational Linguistics Technical Papers*, Osaka, Japan, pp. 33–43, 2016.

[33] Z. Cao, W. Li, S. Li and F. Wei, "Retrieve, rerank and rewrite: soft template based neural summarization," in *Proc. of 56th Annual Meeting Association for Computational Linguistics*, Melbourne, Australia, pp. 152–161, 2018.