International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 2, April 2019, pp. 1313~1320
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i2.pp1313-1320

Feature selection, optimization and clustering strategies of text documents

A. Kousar Nikhath[1], K. Subrahmanyam[2]
[1]Department of Computer Science and Engineering, Koneru Lakshamaiah Education Foundation, India
[2]Department of Computer Science and Engineering, Koneru Lakshamaiah Education Foundation, India

Article Info

Article history:

Received Dec 28, 2017
Revised Sep 16, 2018
Accepted Oct 1, 2018

Keywords:

Feature extraction
Feature selection
Semi-supervised learning
Unsupervised learning

ABSTRACT

Clustering is one of the most researched areas of data mining applications in the contemporary literature. The need for efficient clustering is observed across wide sectors including consumer segmentation, categorization, shared filtering, document management, and indexing. The research of clustering task is to be performed prior to its adaptation in the text environment. Conventional approaches typically emphasized on the quantitative information where the selected features are numbers. Efforts also have been put forward for achieving efficient clustering in the context of categorical information where the selected features can assume nominal values. This manuscript presents an in-depth analysis of challenges of clustering in the text environment. Further, this paper also details prominent models proposed for clustering along with the pros and cons of each model. In addition, it also focuses on various latest developments in the clustering task in the social network and associated environments.

Corresponding Author:

A. Kousar Nikhath,
Department of Computer Science and Engineering,
Koneru Lakshamaiah Education Foundation,
Guntur-522502, AP, India.
Email: kousarnikhath@vnrvjiet.in

1. INTRODUCTION

Clustering of documents is an essential process and efficient algorithms need to be employed to ensure effective document clustering. The process of clustering includes categorization of given documents into individual groups. These clusters should be meaningful and provide right description of the documents. However, for an efficient clustering, researchers often face the challenge of complexity in terms of large number of words. When the clustering is done in the form of matrices, each and every document is considered as an instance and all the terms associated will be features. In general, the volume of features is almost equal to a dictionary, posing strong challenges for algorithm developers. The clustering algorithm efficiency largely fluctuates with increasing number of words. Accordingly, researchers search for non-context related, redundant words and stop words and attempt to ignore or remove such words to boost efficiency of algorithm.

Document clustering contains particular methods and algorithms built on unsupervised document management [1]. In clustering the assets, memberships, and number of the classes not recognized in advance. Documents can group together built on an exact type, such as legal, economic, and medical. Machine learning algorithms have become prevalent in numerous domains, impacting a wide diversity of applications. In the past periods, the machine-learning community has elaborated to decrease the labeling work done by the human for supervised machine learning procedures or to develop unsupervised learning with only smallest supervision. Nevertheless, there are still several cases that neither semi-supervised learning nor transfer learning can help. Providentially, with the propagation of general-determination knowledge bases

(or knowledge graphs), e.g., WikiTaxonomy Wikipedia, Freebase, Probase, TextRunner, DBpedia, NELL and Knowledge Vault, we have profusion of available world knowledge. We call these knowledge bases world knowledge. The concept of representing the given document by the group of words included in the document is underlying many text mining studies. Often referred to as bag-of-words, the concept requires accurate description of the word position in the given document. Accordingly, researchers opt for vector representation of the word position and assign an 'importance' value to each word. The vector space model is versatile because vector representation can use as a feature vector for a large number of clustering algorithms. The vector-based document models do not have the information about the order by which the words occur in a document. In previous articles, researchers developed a much-advanced document model termed 'STD model'. The approach is based on storing complete word sequence data. Overlapping between strings in the combined suffix tree is used to represent the document similarity. A novel model relying on linear convex mix of documents is studied by researchers in. To enable feature basis as this mixture, convex-NMF approach is proposed. The model also attained similar factorization as attained by CF factorization approach.

2. TAXONOMY

2.1. Sub section 1 feature extraction

Feature Extraction (FE) process is categorized into three types including Syntactical, Semantic and Morphological Analysis. Of these, MA is primarily engaged in dealing with each and every word (individual words) of the given text document. Predominantly, it comprises tokenization, stop word elimination and stemming [2]. In tokenization process, the text document is often considered as word strings which are word sequences and divides them by eliminating punctuations [3]. The researchers in [4] attempted to understand the exact logic represented by a particular sentence. That is, a sentence should have proper grammatical connectives. SA caters understanding of the grammatical arrangement of a certain language, often referred to as "syntax". Further, POS Tagging process allows adding of contextual grammar knowledge for a specific word in the given sentence. By identifying the open word class, linguistic analysis can be performed easily [5]. Numerous approaches were proposed in scientific literature aiming to implement POS Tagging process depending on the dictionaries [6].

2.1.1. Feature selection

A feature refers to an individual measurable property of a process, which is being observed. Through the use of a set of features, any machine learning algorithm is capable of performing classification. Over the past years in the applications of pattern recognition or machine learning, the domain of features has generally extended from tens to hundreds of features or variables which are employed in those applications. Numerous techniques have been invented so as to effectively address the problem of reducing irrelevant, as well as redundant variables that are a burden on challenging tasks [7]. It is imperative that Feature Selection (variable elimination) is highly beneficial in understanding data, minimizing computation requirement, minimizing the effect of curse of dimensionality besides enhancing the predictor performance.

2.1.2. Filter methods

Filter techniques use variable ranking approaches as the main standards for variable selection through ordering. Ranking techniques are employed because of their simplicity. At the same time, good success is often reported for practical applications. A highly appropriate ranking principle is employed in scoring the variables. Again, a threshold is often employed for the removal of variables below the threshold. Ranking techniques are filter methods because they are used prior to classification for filtering out the variables, which are less relevant. A simple property of a unique feature is to have highly beneficial information regarding the diverse classes in the given data.

2.1.3. Wrapper methods

Wrapper techniques generally employ the predictor as a black box and the predictor presentation as objective function for the evaluation of the variable subset. Because the evaluation of 2N subsets has become an NP-hard problem, suboptimal subsets can be got through the use of search algorithms, which find a subset heuristically. Numerous search algorithms may be adopted for finding a subset of variables, which maximizes the objective function that is the classification presentation [8]. We generally categorize the Wrapper techniques into Sequential Selection Algorithms, as well as Heuristic Search Algorithms. Sequential selection algorithms commence with an empty set (full set). It thereafter adds features (remove features) up to the point of achievement of maximum objective function.

2.1.4. Embedded methods

The embedded methods aim to perform feature selection throughout the training procedure and are essential and distinct to the various machine learning algorithms implemented. Embedded techniques [9] want to minimize the computation time which is taken up in the reclassification of diverse subsets that is done in wrapper techniques. The major approach entails the incorporation of the feature selection as an element of the process of training.

2.1.5. Hybrid approaches

The approach combines filter, as well as the wrapper-based techniques. Filter approach selects a cluster of candidate features from high dimensional and efficient original feature set. Then, by utilizing a wrapper technique, this candidate feature set will be refined. It generally exploits the various kinds of advantages which are brought about by the use of the two methods. Feature selection [5] generally plays huge role in the detection of the anomalies of networks. In the anomaly based detection systems, by monitoring the performance of the regular data thoroughly in contrast with the ones which are irregular, inconsistency will be identified within the network. Thus, this kind of detection system will play a vital role in recognizing various intrusions depending on the distinct characteristics of network traffic.

2.2. Similarity measures

Prior to clustering, there is the need for the determination of a similarity or a distance measure. Generally, the measure reflects the proximity of the targeted objects or the degree of vector separation. It should relate different characteristics used to separate the clusters. In several circumstances, these characteristics vary in accordance with data and can also depend on the problem context. However, as each clustering problem differs from other, no such measure is existing to satisfy every kind of clustering problem. Further, selecting an appropriate similarity measure will be a key driver in Cluster Analysis, predominantly for specified clustering models [10]. Thus, realizing the significance and efficiency of various measures will support the selection of the most suitable one. This value in-turn relies on two distinct factors such as the properties of both objects and on the measurement metrics. The five measures have been discussed below. The different measure brings about different final partition. At the same time, it also imposes diverse requirements for similar clustering algorithm.

2.2.1. Euclidean distance

Euclidean distance refers to a standard metric used for geometrical problems. At the same time, it can be defined as the ordinary distance between two points. Measuring it can easily be done through the use of a ruler in two- or in three-dimensional space. In addition, it is also observed that Euclidean distance will also be selected in clustering problems, which comprises clustering text.

It is satisfying all the four main conditions which have been given above and as a result, it is a true metric. At the same time, it is the default distance measure that is used with k-means algorithm. Resolving the distance measure between text documents d_x and d_y will be denoted by their respective term vectors called $\vec{t_x}$ and $\vec{t_y}$. Hence, the Euclidean metric of these two documents could be defined as:

$$D_E\left(\vec{t_x},\vec{t_y}\right)=\left(\sum_{t=1}^{n}\left|w_{t,x}-w_{t,y}\right|^2\right)^{1/2}, \quad (1)$$

In which the term set is $T=\{t_1,....t_n\}$. As discussed in above section, *tfidf* value can be considered as term weights, i.e., $w_{t,x}=tfidf(d_x,t)$.

2.2.2. Cosine similarity

As pointed above, the text documents are indicated as term vectors. In this scenario, the similarity measure between 2 text documents implies the association in between the selected vectors. In general, this is evaluated as the Cosine functions between given term vectors and is called Cosine Similarity. It is worth pointing out that cosine similarity forms part of the most popular measure of similarity that is used in order to text documents.

The Cosine Similarity (CS) measure for $\vec{t_x}$ document and $\vec{t_y}$ document is depicted:

$$SIM_Z\left(\vec{t_x},\vec{t_y}\right)=\frac{\vec{t_x}.\vec{t_y}}{\left|\vec{t_x}\right|\times\left|\vec{t_y}\right|}, \quad (2)$$

Where $\vec{t_x}$ and $\vec{t_y}$ are called multidimensional vectors of the Vector-term set $T = \{t_1, ...t_n\}$. Each dimension contains its own weight and corresponds to a term set. The value of these dimensions is always more than zero. Hence, the CS holds positive values and will always be bound between [0, 1].

A noteworthy property of this kind of similarity is that it is independent of document length. For instance, by merging two copies of a particular text document d to generate a pseudo-document d', the CS value computed between d and d' will be equal to 1. This refers that, matching should be carried out among two documents. When fed with another document m, d and d' would likely to result in same similarity to m and is $sim\left(\vec{t_d}, \vec{t_m}\right) = sim\left(\vec{t_{d'}}, \vec{t_m}\right)$. On the other hand, it can also be expressed as, for text documents with similar content or words, diverse totals will be managed identically. However, this is unable to satisfy the metric's second condition because with the consolidation of two similar copies, a completely dissimilar object will be obtained from original text document. In addition, it is essential to note that, if the vectors are normalized to a fixed unit length, this case reflects similar notations for both d and d0.

2.2.3. Jaccard coefficient

Jaccard Coefficient or Tanimoto Coefficient is also proposed to calculate similarity. According to this computation, similarity is measured as the "intersection to combined specified objects ratio". For the given text document, this coefficient evaluates the total weight of the mutual terms existing in both documents with the total weight of all terms existing in at least one of the two documents but unique terms. Based on this computation, matching among the documents will be carried out. The general computation formula has been depicted:

$$SIM_H\left(\vec{t_x}, \vec{t_y}\right) = \frac{\vec{t_x}.\vec{t_y}}{\left|\vec{t_x}\right|^2 + \left|\vec{t_y}\right|^2 - \vec{t_x}.\vec{t_y}} \quad (3)$$

Jaccard coefficient is a similarity measure and it bounds between 0 and 1. The measure will be 1 if both the documents are similar and 0 when they are dissimilar. In general, coefficient value of 1 represents that both given objects are same, whereas, coefficient value of 0 denotes that the specified objects are extremely different. In addition, dissimilarity should also be observed in this similarity measure- the Jaccard distance measure [11]. The dissimilarity among the given objects will be computed using distance metrics and is $D_H = 1 - SIM_H$. D_H can also be used as an alternative in following experiments.

2.3. All about clustering

Data mining refers to the process which mainly entails the extraction of implicit, previously unknown as well as potentially beneficial information from data. It is imperative that document clustering, which is a subgroup of data clustering, refers to a data mining approach that includes various concepts from information retrieval, natural language processing, as well as machine learning fields [12]. The high-quality and efficient document clustering methods play a vital role in supporting the clients in terms of effective navigation, summarizing and organizing diversified set of information effectively. A specified document will always have a probability to occur in multiple clusters [13] in the overlapping partition. Further, in disjoint partition, the text document will appear in only one cluster.

As points out, document clustering can be grouped into two main subcategories, which includes: Soft (overlapping) and Hard Clustering. Overlapping Clustering is clustered into Hierarchical clustering, Partitioning and Itemset-based Clustering.

a. Disjoint (Hard): It will compute disjoint assignments of a specified text document towards a cluster. That is, as mentioned above, hard clustering will always assign a document to single cluster, which then caters a set of different clusters.
b. Overlapping (Soft Clustering): This type of clustering process soft assignments will be carried out. That is, every text document is can be presented in distinct clusters. Hence, soft clustering produces multiple overlapping clusters.
c. Partitioning: It is primarily engaged in assigning documents into a specific volume of Non-Empty Clusters. In particular, k-means along with its alternatives are highly repudiated partitioning techniques as per [1].
d. Hierarchical: It involves developing dendrograms, where clusters are organized in hierarchical tree patterns. In the tree, the Leaf node represents the sub-set of given document collection. Both HAC clustering and UPGMA clustering are grouped in the hierarchical structure [14].

2.3.1. Document clustering

Document Clustering plays a vital role in clustering the given documents into numerous topics without having any information of the structure of the category available in a given document collection. Each and every Semantic Information is obtained from within the given documents and is Un-supervised. On the other side, document classification is concerned with assigning the text documents to pre-defined categories, where labeled instances for learning from the clustering for classification is called supervised learning in which a given classifier is learned from the labeled examples. It is then used for predicting classes of unseen documents. Document clustering is employed in numerous diverse contexts, like exploring the structure in a given document collection for the discovery of knowledge [8], dimensionality contraction for all other tasks such as classification [15], grouping search outcomes to ranked list [9] for executing an alternative presentation and also employed for pseudo-relevance feedback.

2.4. Cluster evaluation measures

Evaluation of document clustering is a difficult task. Built-in quality measures like distortion or log possibilities imply how a certain algorithm optimizes a given representation. Meanwhile, internal measures could not be compared among different representations. In addition, it's a noteworthy point that external views of truth are human-made. They continue to suffer from the major shift for humans to understand different document topics in a distinct manner. Predominantly, whether the certain document belongs to that particular topic or not might be subjective. However, as clustering of a document has feasibility to execute in a number of ways, above mentioned scenario could even complicate the conditions.

The major advantage of this measure in compared to evaluation through text classification is that there is no need of such conditions which are depicted above. This measure does not include either a test bed platform (comprises labeled documents) or consistency factor amid clusters and targeted categories. On the other hand, it approximately evaluates the outcome of text clustering [13], only when the labeled documents are utilized as test bed. Text classification parameters like accuracy, re-call, F1 and precision measures were used for estimating the presentation of text clustering in [14], [16]. Based on properly classified text documents and each and every document present in the test bed, the rate of accuracy will be computed. Further, the measure is the simplest measuring parameter in associated classification problems. This measure is directly applicable to the Multi-Classification Problems.

However, significant measures like precision, re-call and F1 can be directly applied to the binary classification tasks. Hence, to evaluate the classification performance by making the use of those measures, the respective problem is to be split into binary classification problems. Each and every class corresponds to a specific binary classification task in Multi-Classification task. Of the classes, positive ones represent "Belonging to the class" and the Non-Positive ones represent "Not-Belonging to the class". The evolution measure majorly concentrates on the positive class.

In the text categorization, re-call measure will be obtained by the ratio of the specific true positive document to all documents that are true. Precision measure referred as the rate of classified true positive documents to every classified positive document includes both true positives and false positives. Whereas, F1 is used to determine a value using both Re-call R and Precision P measures by using (4).

$$F1-measure=\frac{2\times R\times P}{R+P} \quad (4)$$

Various metrics like F1, accuracy, detection costs are employed in text categorization. These are primarily employed to calculate the performance metric in text clustering. When these measures are used there always exist two conditions. Each and every given document inside the specified test bed should contain target categories and must be labeled. It is somewhat critical in real-time in terms of getting labeled document when compared to the document which is unlabeled. Meanwhile, the process which is engaged in labeling documents follows in practice with clustering documents. In addition, it is also significant to note that, vast time will be consumed by the process which is engaged in evaluating the approaches to text clustering when preparation of labeled documents is ongoing. Secondly, the cluster number must be constant with target categories number. For example, when a sequence of documents having same target category will be partitioned into two clusters, then the evolution measures of text characterization will not be applicable in such case.

3. REVIEW OF LITERATURE

3.1. Feature extraction strategies

These techniques are introduced based on keywords. These keywords are employed to depict various emotions which exist inside the text [17]. In contrast, the main disadvantage of this method is that it relies on presence of various affective words in the text. To overcome such drawbacks and to achieve accurate extractions and outcomes, the authors proposed a novel model called Semantic Networks in [15]. These networks represent events, relationships and various concepts among them. Unlike feature extraction, these semantic networks are independent on keywords to depict the human emotions in the text. Hence, [18] made valuable conclusions about the process of achieving enhanced performance in detecting the human emotions through semantic networks. In these networks, human emotions will be identified through contextual information. In particular, and presented a range of explanations of this approach. However, they failed to explain the respective outcomes of the experiments. Moreover, there is a necessity of huge databases like SentiWordNet and WordNet-Affect to improve the accuracy of results.

3.2. Feature selection strategies

Multiple feature identification programs, are implemented for classification. However, all projected algorithms have common goal, i.e., searching for efficient features set which caters results in terms of best classifications. In general, various algorithms involved in feature selection employ distinct evaluation metrics like information gain and correlation. In addition, they often use population-based heuristics such as ant colony optimization, simulated annealing, particle-swarm optimization, and genetic algorithms. According to by using feature similarity, Un-Supervised Feature Sub-Set Selection Technique were proposed [19]. This approach is used to avoid the duplications among the selected features. This approach uses new metrics called MIC Index for calculating the similarity measure between two different variables for selecting a feature. In Fuzzy rough set theory is employed for the selection of feature by considering the natural properties of both fuzzy logic t-norms and t-conorms. Additionally, in MIFS-U algorithm is introduced to handle restrictions linked with MIFS. The primary objective of this approach is to get improved similar information among input characteristics and output classes of the MIFS. Similarly, [12] also proposed feature selection technique called Max-Relevance and Min-Redundancy (MRMR) based on mutual information concept. In general, this technique minimizes the redundancy between the features as well as maximizes the dependency between a class label and sub-set of features.

3.3. Clustering techniques

Clustering Algorithms are characterized based on two major properties. The first property primarily deals with whether certain membership of cluster is distinct. The hard or disjoint clustering algorithms allocate each and every document to justify a single cluster. The other side, the soft or overlapping clustering algorithms allocate dissimilar documents to single or multiple clusters in discrete membership degree. On the other hand, the second property controls the clusters structure. In general, the structure may be observed in either flat or hierarchical. On flat clustering technique front, it generates rigid clusters, without any correlation between them. On the contrary, the hierarchical algorithms are engaged in generating clusters in a tree structure. It follows bottom-up approach, as it involves executing the procedure from its bottom most cluster (at the root) of the tree structure.

3.3.1. Partitioning and hierarchical document clustering

The majority of traditional clustering algorithms are categorized into two main groups including partitioning algorithms and hierarchical algorithms [18]. The hierarchical clustering algorithms are primarily involved in decomposing a specified dataset hierarchically. Hence, it forms a dendrogram tree where given dataset is split repeatedly into small sub-sets. Thus, the documents will be represented in Multi-Level structure as depicted. These algorithms are often grouped into either divisive algorithms or agglomerative algorithms, On the other hand, in agglomerative procedure, each document is allocated to a separate cluster. Later, the procedure involves merging similar clusters repeatedly until termination criterion is observed. While on the divisive algorithms front, it increases the number of clusters at each iterative stage by splitting the whole document into a specified quantity of clusters. In addition, another clustering algorithm based on Partitioning is one of the most studied categories [17]. It upholds extreme realistic techniques to cluster big datasets as represented, unlike dendrogram tree structure, these techniques cluster data in a single level. In general, these approaches are engaged in dividing a given document collection into distinct clusters, which in-turn increases the pre-defined objective value. By containing efficient clustering characteristics in terms of quality and accuracy, these hierarchical clustering algorithms do not offer re-allocation of documents. This is a major drawback of this approach and hence there can be chances of poor performance measures in the early

stages of clustering. Accordingly, in various data objects, the time taken to execute a hierarchical algorithm is Quadratic.

In the recent past, it was demonstrated that the partition techniques are best suitable for applications which comprise big datasets due to its Minimized Computational Complexity, Time complexity is comparatively less as compared to hierarchical techniques and is linear. Thus, partitioning techniques are highly adaptable for big scale clustering. In addition, to lessen the limitations brought about by the conventional partition clustering techniques discussed earlier, multiple models are introduced in the recent few years. These methods focused on implementing optimization techniques over a pre-determined clustering through objective function.

3.3.2. Machine learning based document clustering

Numerous knowledge bases like Cyc project, Freebase, KnowItAll, Wikipedia, TextRunner, WikiTaxonomy, Probase, DBpedia, YAGO, NELL [8] as well as Knowledge Vault generally play a highly vital role in the process of document clustering with regards to context, concept and semantic relations. So, as to notice all of these relations between the documents, a prior knowledge is vital. This will point out the need of highly sophisticated learning techniques to notify the relations. On the contrary, the aforesaid knowledge bases have the ability of training the learning approaches so as to cluster the given documents based on a single or additional context, concept as well as semantic relations. The argument illustrates the huge role of machine learning when it comes to Document Clustering.Usage of multiple existing knowledge bases is primarily aimed at enhancing document's features of multi-set of words representation. For example, using WordNet, a linguistic knowledge base, resolves synonyms while introducing various WordNet concepts. Utilization of such innovative knowledge base concepts improves the quality of text document as depicted in [4]. By mapping the given content to the semantic space which is offered through Wikipedia pages, it has been proved as an efficient knowledge base and is best suitable for Short Text Classification and Document Clustering [9], [20]. In addition, in [16], other two knowledge bases including Probase and Taxonomy are introduced. These knowledge bases are majorly involved in enhancing the ad keywords features in order to build a novel taxonomy of keywords which are domain dependent. Thus, it might be significant to consider the knowledge as "Supervision" to direct the other Machine Learning Techniques and distinct tasks. Distant Supervision learning scheme employs information entities and respective relations from Freebase knowledge bases as supervision to execute entity and relation extraction [15], [13] and [14]. In addition, it also employs knowledge supervision for extracting more entities and relationships from the novel content or also used for generating an efficient installation of both entities and relations. Thus, exploitation of direct supervision is restricted to knowledge entities and relations among them.

4. CONCLUSION

This research paper discusses a detailed survey of different clustering approaches for data mining in the text environment. An efficient text clustering approach must choose optimal attributes along with the right algorithm for execution. Of various types of algorithms found in literature, distance-based approaches are observed to be both efficient and widely implemented across different domains. Over the past few years, researchers working on text clustering focused on two types of applications.

a. Dynamic: Huge voluminous information generated in dynamic environments including social networking platforms or online chat resulted in a strong requirement for streaming information. These applications should be adaptable to scenarios where the text is often not clear like the social networking platforms.
b. Heterogeneous: In these applications, the text is often present as links and different multimedia formats. For instance, in platforms like Flickr, text clustering should be adapted. Accordingly, it is important to efficiently implement text mining approaches in this environment.

This manuscript observes that the area of clustering in text mining applications is wide and is challenging to completely present in one paper. Certain approaches like committee-driven clustering could not be clearly categorized into any groups as they incorporate multiple clustering techniques to generate the final outcome. The intention of this work is to put forward a complete brief of prominent approaches proposed for text mining, to serve as an initial step for other proposed research contributions.

REFERENCES

[1] Jain, Anil K., "Data Clustering: 50 Years beyond K-Means," *Pattern recognition letters*, vol. 31(8), pp. 651-666 2010.
[2] Vijay Sonawane, D. Rajeshwara Rao, "An Optimistic Approach for Clustering Multi-version XML Documents Using Compressed Delta," *International Journal of Elecctrical and Computer Engineering*, vol.5 (6), pp.1472-1479, Dec 2015.

[3] Srividya Sivasankar, Sruthi Nair, M.V. Judy, "Feature Reduction in Clinical Data Classification using augmented-Genetic Algorithm," *International Journal of Electrical and Computer Engineering*, vol. 5(6), pp.1516-1524, Dec 2015.
[4] Mugunthadevi K., *et al.*, "Survey on Feature Selection in Document Clustering," *International Journal on Computer Science and Engineering*, vol.3 (3), pp. 1240-1241, 2011.
[5] Forman, George, and Evan Kirshenbaum, "Extremely Fast Text Feature Extraction for Classification and Indexing," *Proceedings of the 17th ACM conference on Information and knowledge management. ACM*, 2008.
[6] Li, Yanjun, Soon M. Chung, and John D. Holt, "Text Document Clustering Based on Frequent Word Meaning Sequences," *Data & Knowledge Engineering*, vol. 64 (1), pp. 381-404, 2008.
[7] Chandrashekar, Girish, and FeratSahin, "A Survey on Feature Selection Methods," *Computers & Electrical Engineering*. Vol. 40 (1), pp. 16-28, 2008.
[8] Pedram Vahdani Amoli, Omid Sojoodi Sh, "Scientific Documents Clustering Based on Text Summerization," *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 5 (4), pp. 782-787, Aug 2015.
[9] Law, Martin HC., Mario AT. Figueiredo, and Anil K. Jain, "Simultaneous Feature Selection and Clustering using Mixture Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26(9), 1154-1166, 2004.
[10] Gabrilovich, Evgeniy, and Shaul Markovitch. "Feature Generation for Text Categorization Using World Knowledge," *IJCAI*, vol. 5, 2005.
[11] A. Kousar Nikhath, K. Subrahmanyam, "Incremental Evolutionary Genetic Algorithm Based Optimal Document Clustering," *Journal of Theoritical and Applied Information Technology*, vol. 87(3), May 2018.
[12] A. Kousar Nikhath, K. Subrahmanaym, "Conceptual Relevance Based Document Clustering Using Concept Utility Scale," *Asian Journal of Scientific Research*, vol. 11(1), pp. 22-31, 2018.
[13] Whissell, John S., and Charles LA. Clarke. "Improving Document Clustering using Okapi BM25 Feature Weighting," *Information retrieval*, vol. 14 (5), pp. 466-487, 2011.
[14] Kwak, Nojun, and Chong-Ho Choi, "Input Feature Selection for Classification Problems," *IEEE Transactions on Neural Networks*, vol. 13 (1), pp. 143-159, 2002.
[15] Lee, Kyung Soon, W. Bruce Croft, and James Allan. "A Cluster-based Resampling Method for Pseudo-relevance Feedback," *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM*, 2008.
[16] Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-relevance, and Min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27 (8), pp. 1226-1238, 2005.
[17] Sun Y., C. F. Babbs, and E. J. Delp. "A Comparison of Feature Selection Methods for the Detection of Breast Cancers in Mammograms: Adaptive Sequential Floating Search vs. Genetic Algorithm," *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the. IEEE*, 2006.
[18] Lu, Yijuan, *et al.* "Feature Selection using Principal Feature Analysis," *Proceedings of the 15th ACM international conference on Multimedia. ACM*, 2007.
[19] Wolf, Lior, and AmnonShashua. "Feature Selection for Unsupervised and Supervised Inference: The emergence of Sparsity in a Weight-based Approach," *Journal of Machine Learning Research*, vol. 6, pp. 1855-1887, Nov 2005.
[20] Sun, Zhanquan, *et al.* "A Parallel Clustering Method Combined Information Bottleneck Theory and Centroid-Based Clustering," *The Journal of Supercomputing*, vol. 69 (1), pp. 452-467, 2014.

BIOGRAPHIES OF AUTHORS

A. Kousar Nikhath is currently working as Asst. Professor in Computer Science & Engineering Department at VNRVJIET, Hyderabad. She is into teaching profession for the past 13 years. She is currently pursuing Ph.D. at Koneru Lakshamaiah Education Foundation, Guntur.She has published nearly about 10 papers in various Journals/ International conferences. Her research area interest included Text mining, Data mining, Document Clustering, Artificial Intelligence and Neural Network.

Dr. K Subrahmanyam, a Gold Medalist from Andhra University (1992-93) is currently working as a Professor in Computer Science & Engineering Department of Koneru Lakshamaiah Education Foundation, Guntur. He is in teaching profession for the past 25 years and prior to joining Koneru Lakshamaiah Education Foundation he worked as Programme Leader in the School of Engineering, Science & Technology at KDU University, Malaysia for about 10 years. He has published more than 40 papers in both national and international journals and conferences and attended various workshops in Malaysia, Singapore, USA & India. His research interests include Knowledge & Software Engineering, Data Mining, Soft Systems Methodologies. He has guided 100 over students towards their Master's and Bachelor Dissertations, and currently guiding 8 towards their PhD.