2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology

# Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering

Ammar Ismael Kadhim[1,2], Yu-N Cheah[1], and Nurul Hashimah Ahamed[1]

[1]*School of Computer Sciences,* Universiti Sains Malaysia, Penang, Malaysia
[2]*Department of Computer Science,* College of Medicine , Baghdad, Iraq
ammarusm70@gmail.com, yncheah@cs.usm.my, and nurulhashimah@cs.usm.my

***Abstract* - Text mining defines generally the process of extracting interesting features (non-trivial) and knowledge from unstructured text documents. Text mining is an interdisciplinary field which depends on information retrieval, data mining, machine learning, parameter statistics and computational linguistics. Standard text mining and retrieval information techniques of text document usually rely on similar categories. An alternative method of retrieving information is clustering documents to preprocess text. The preprocessing steps have a huge effect on the success to extract knowledge. This study implements TF-IDF and singular value decomposition (SVD) dimensionality reduction techniques. The proposed system presents an effective preprocessing and dimensionality reduction techniques which help the document clustering by using k-means algorithm. Finally, the experimental results show that the proposed method enhances the performance of English text document clustering. Simulation results on BBC news and BBC sport datasets show the superiority of the proposed algorithm.**

***Keywords - Text mining; retrieval information; clustering; singular value decomposition; dimension reduction; clustering; k-means.***

## I. INTRODUCTION

All before going to implement any operation on the text data, the data have to preprocess. Due to the data often involves some special formats or non-informative features like date formats, number formats, and the most widely words that non-informative to help text mining like prepositions, articles, and pro-nouns can be removed. Dimension reduction is one of important process in text mining and enhances the performance of clustering techniques via reducing dimensions so that text mining procedures process text documents with a reduced number of features. SVD is a way used to reduce the dimension of a vector that is used in linear transformation approach. Clustering is to make the retrieval information easy. K-means algorithm is an efficient clustering technique which is performed for clustering text documents [1].

The focus of the paper is on the problems of weighting terms using TF-IDF, dimension reduction using singular value decomposition (SVD) and documents clustering using k-means algorithm. These steps are implemented on the same platform dataset (BBC news and BBC sport) but with varying sizes of the document.

The use of SVD in dimension reduction leads to an effective learning algorithm which can enhance the generalization ability of documents clustering.

## II. RELATED WORK

In machine learning, clustering [1] is an example of unsupervised learning. In the same context of machine learning, clustering deals with unsupervised learning (or classification). Classification allocates data objects in a set to target classes. The major issue of classification is to precisely expect the target class for each occurrence in the data. So classification algorithm needs to train data. Unlike classification, clustering need to train data. Clustering does not allocate any per-defined label to each and every set. Clustering sets a set of objects and discovers whether there is some correlation between the objects.

As stated by [2], [3], Data clustering algorithms can be broadly divided as follows:

- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Model-based methods
- Frequent pattern-based clustering
- Constraint-based clustering

With partitional clustering the algorithm generates a group of data non-overlapping subsets (clusters) such that each data object is in correctly one subset. These advances need choosing a value for the preferred number of clusters to be produced. A few public heuristic clustering techniques are k-means and a variant of k-means-bisecting k-means, k-medoids, PAM (Kaufman and Rousseeuw, 1987), CLARA (Kaufmann and Rousseeuw, 1990), CLARANS (Ng and Han, 1994) etc.

With hierarchical clustering the algorithm generates a nested group of clusters that are arranged as a tree. Such hierarchical algorithms can be agglomerative. Agglomerative algorithms, also described the bottom-up algorithms, initially treat each object as a split cluster and sequentially combine the pair of clusters that are close to one another to produce new clusters until all of the clusters are combined into one. The public hierarchical techniques are ROCK [4], Chamelon [5], BIRCH [6] and UPGMA.

978-1-4799-7910-3/14 $31.00 © 2014 IEEE
DOI 10.1109/ICAIET.2014.21



IEEE computer society

Density-based clustering techniques group the data objects with arbitrary shapes. Clustering is done during to a density (number of objects). The public density-based methods are DBSCAN and its extension, OPTICS and DENCLUE [6].
Grid-based clustering methods utilize multi-resolution grid structure to collect the data objects. The advantage of this method is its velocity in processing time. Model-based methods utilize a model for each group and conclude the fit of the data to the given model. It is also utilized to automatically conclude the number of clusters.
Frequent pattern-based clustering utilizes models which are extracted from subsets of dimensions, to collect the data objects.

Constraint-based clustering techniques achieve clustering depend on the user-specified or application-specific constraints. It requires user's constraints on clustering such as user's requirement or explains properties of the involved clustering results.

## III. Proposed System

The proposed approach can be divided into four main stages such as: preprocessing stage, term weighting stage, dimensional reduction and dimension reduction as shown in Figure 1.

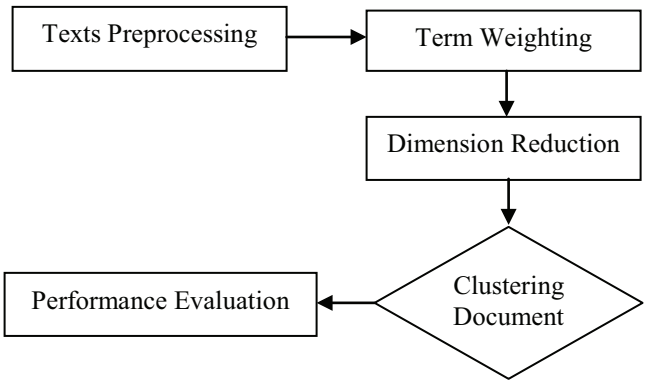


Figure 1. The stages of proposed system.

### *A. Texts Preprocessing*

Texts preprocessing step after reading the input text documents, it divides the text document to features which are called (tokenization, words, terms or attributes), it represents that text document in a data representation as a vector space whose components are that features and their weights which are obtained by the frequency of each feature in that text document, after that it removes the non-informative features such as (stop words, numbers and special characters). The remaining features are next standardized by reducing them to their root using the stemming process. In spite of the non-informative features removal and the stemming process, the dimensionality of the feature space may still be too high. Therefore; the study applies specific thresholds to reduce the size of the feature space for each input text document based on the frequency of each feature in that text document. The objective of this step is that it improves the quality of features and at the same time reduces the difficulty of mining process.

Tokenization

Tokenization in our sense does not only separate strings into basic processing units, but also interprets and groups isolated tokens to create higher level tokens. Raw texts are preprocessed and segmented into textual units. The data must be processed in the three operations: the first operation is to convert document to word counts which is equal to bag of word (BOW). The second operation is removing empty sequence i.e. this step comprises cleansing and filtering (e.g., whitespace collapsing, stripping extraneous control characters). Finally, each input text document is segmented into a list of features which are also called (tokens, words, terms or attributes).

Stop Words Elimination

A stop words list is a list of commonly repeated features which emerge in every text document. The common features such as conjunctions such as or, and, but and pronouns he, she, it etc. need to be removed due to it does not have effect and these words add a very little or no value on the categorization process (i.e., each feature should be removed when it matches any feature in the stop words list). For the same reason, if the feature is a special character or a number then that feature should be removed. In order to find the stop words, we can arrange our list of terms by frequency and pick the high frequent ones according to their lack of semantics value. They should be removed from words, can also remove very rare words, e.g., words that only occur in m or fewer document, for example m=6.

Stemming

Stemming is the process of removing affixes (prefixes and suffixes) from features i.e. the process derived for reducing inflected (or sometimes derived) words to their stem. The stem need not to be identified to the original morphological root of the word and it is usually sufficiently related through words map to the similar stem. This process is used to reduce the number of features in the feature space and improve the performance of the clustering when the different forms of features are stemmed into a single feature. For example: (connect, connects, connected, and connecting) from the mentioned above example, the set of features is conflated into a single feature by removal of the different suffixes -s, -ed, -ing to get the single feature connect. This study applied standard Porter Stemming Algorithm for find the root words in the document.

Vector Space Model

In present document clustering after performing tokenization keywords, removing stop words and performing stemming keywords in the documents are remained. For example suppose if the term and the document matrix. Rows represent words in each document and columns represent documents. Each cell represents the word count in the particular document. If the number of documents is increasing the size of the matrix will also be increased. Thereby this leads to high-dimensionality. So for

efficient clustering algorithms is used to reduce the high dimensionality. Data mining is a method to find useful features from database. Clustering algorithms are widely used to group these features from a large dataset. The algorithms must be prepared to handle data of limited length. For this High-dimensionality of data must be reduced. To reduce the high-dimensionality of data, the study uses SVD technique. Processing high dimensional data very expensive in terms of execution time and storing the data. In this case the clustering is not done in an effective method. High dimensionality is challenging to achieve efficient clusters of the input documents; by using one of high dimensionality reduction techniques to reduce the size of the vector space model. The dimensionality reduction techniques are SVD, Independent Component Analysis (ICA) and Principle component Analysis (PCA).

The major objective of document indexing is to increase the efficiency by extracting from the resulting document a selected set of terms to be used for indexing the document. Document indexing involves choosing the suitable set of keywords based on the whole corpus of documents, and assigning weights to those keywords for each particular document, thus transforming each document into a vector of keyword weights. The weight normally is related to the frequency of occurrence of the term in the document and the number of documents that use that term.

### *B. Term Weighting*

Term weighting can be as simple as binary representation or as detailed as a mix of term and dataset existence probabilities stemmed from complex information theoretic underlying concepts. TF-IDF is the most widely known and used weighting method, and it is remain even comparable with novel methods. In TF-IDF terms weighting, the text documents are modeled as transactions. Selecting the keyword for the feature selection process is the main preprocessing step necessary for the indexing of documents. This study utilized two different TF-IDF methods such as Global and normal to weight the terms in term-document matrices of our evaluation datasets. A common practice to reduce the possible impacts resulting from it, is the normalization of the TF-IDF scores for each document in the collection by using the Euclidean norm is calculated by using Equations (1&2) respectively, given by [7] as follows:

$$\text{TF-IDF}=\ln(\text{TF}) \times \ln(\text{IDF}) \quad (1)$$

$$\text{TF-IDF}=\text{TF} \times \text{IDF} \quad (2)$$

### *C. Dimension Reduction*

Dimension reduction is a one of the parameters is used in different applications such as retrieval information, text classification and data mining. The main goal of dimension reduction is to introduce high-dimensional data in a lower-dimensional subspace, while essential features of the original data are kept as much as possible. In this study, in spite of the non-informative features removal such as tokenization and stop words and the stemming process, a number of features in the feature space may still be too large. Thereby, some features may be negative to the clustering task and sometimes decrease accuracy. Such features will remove without affecting the clustering performance. The large number of features may affect the cluster learning due to the most machine learning algorithms which are implemented on clustering cannot deal with this huge number of features.

Singular value decomposition:

The major purpose of SVD is to reduce a dataset involving essentially values fewer values. In this study, the SVD of the TF-IDF matrix will be employed which is used in linear transformation approach. This matrix factorization leads to decompose a given $n \times m$ matrix H into a correlated set of singular values and two orthogonal bases of singular vectors, as below:

$$H = USV^T \quad (3)$$

where U and V are unitary matrices (i.e. $U^T = U^{-1}$ and $V^T = V^{-1}$) of dimensions $n \times n$ and $m \times m$, S is an $n \times m$ diagonal matrix (i.e. $s_{ij} = 0$ if $i \neq j$), and T defines transposition. The diagonal elements of S are known as the singular values of H, and the columns of U and V are known as the ''left'' and ''right'' singular vectors of H, respectively. Observe that SVD does not represent a space reduction method per se. Indeed, the n columns of U represents an orthogonal basis for the vector space spanned by the rows of H. Be reminded that the rows of a TF-IDF matrix correspond to vocabulary terms, so if H is a TF-IDF matrix, the columns of U would represent an orthogonal basis for the correlated document space. Similarly, the m columns of V represent an orthogonal basis for the vector space spanned by the columns of H. In this way, as the columns of a TF-IDF matrix correspond to documents, V would provide an orthogonal basis for the associated word space. These orthogonal bases are optimal values in the intelligence that they focus on the variability of the data in as few dimensions as possible. On the other hand, the first singular vector is supported with the direction in which the data shows its maximal variability, the second singular vector is aligned with the orthogonal direction (with respect to the first singular vector) in which the data shows maximal variability, the third singular vector is supported with the orthogonal direction (with respect to both the first and second singular vectors) in which the data shows maximal variability, and so on [8].

### *D. Clustering Documents*

Clustering is the process of automatically grouping the objects in a given collection according to the similarities of their prominent features. While elements within each group or cluster are predicted to show large similarities among them, elements across different groups or clusters are

predicted to show large differences among them. The process is defined to as unsupervised as far as no information about the categories coexisting in the collection is previously know or used during the process. Unsupervised clustering is a very powerful procedure for detection the hidden structure, if any, of a given data collection. In this study we will focus our attention into a specific method of cluster analysis denominated k-means clustering.

K-Means Algorithm:
The k-means clustering algorithm is a refined algorithm in which a collection of n objects is partitioned into k clusters, which are updated recursively until they concentrate into a regular partition. Each algorithm's repetition is achieved in two steps: assignment and updating. In the assignment step, each element in the collection is allocated to the closest cluster by means of a distance metric. The distance between a cluster and a given element is calculated in a vector space representation by measuring the distance between the given element representation and the centroid of the cluster. In the updating step, all cluster centroids are updated by getting into consideration the new partition produced during the previous step. In k-means clustering, centroid vectors are denoted by the mean vector of all elements belonging to the matching clusters. The algorithm begins from a predefined set of centroids (which can be produced either randomly or by means of any other criterion) and achieves sequential repetitions of the assignment and updating steps until no more modifies to the partition are noticed over sequential repetitions. The procedure of the clustering implemented as following steps.
Step1: Initialize the parameters for k-means clustering.
Step2: Define the number of clusters to be considered.
Step3: Define the initial set of centroids and apply the k-means clustering algorithm to the test set.
Step4: To identify correspondence among cluster and category tags:
- Compute cosine distances among cluster and category centroids.
- Select the best cluster-to-category assignment (with minimum overall distance among centroids).
Step5: Consider all possible assignments.
Step6: Compute overall distances for all cases.
Step7: Get the best assignment
Step8: Go back to step 3, stop when no more new assignment.

## IV. Exprimental Results

### A. Dataset

There are two sets:

- The first set collects BBC English Dataset is an English dataset collected from BBC news website corresponding to stories in five topical areas from 2004-2005. The dataset consists of 2225 text documents and 9636 words unique. The dataset is classified into five natural classes such as business, entertainment, politics, sport and tech.
- The second set collects BBC sport dataset is collected from BBC sport website corresponding to sports news articles in five topical areas from 2004-2005. The dataset consists of 737 text documents and 4613 words unique. The dataset is classified into five natural classes such as athletics, cricket, football, rugby and tennis.

### B. Performance Evaluation

- The comparison has to be implemented on the same algorithm and under the same environment. Hence, the algorithm must be implemented using the same technique and on the same platform. Thus, we used the same platform dataset but with varying sizes of the document. In addition, we used the same algorithm for the dimension reduction using SVD and for the clustering using K-means. Figures 2 and 4 show the document size (No. of Words) for each dataset as mentioned earlier. Figures 3 and 5 show the clustering of the document as well as the resulting three clusters.

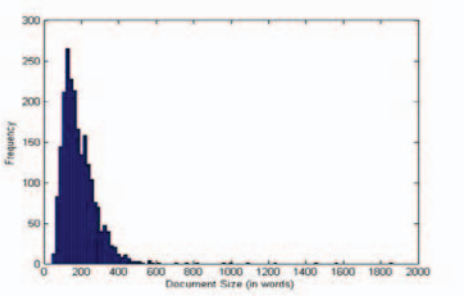


Figure 2. Nunmber of words with frequency for BBC news dataset.

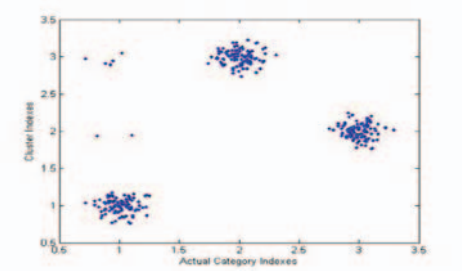


Figure 3. Doucuments clustered into three different topics for BBC news dataset.

The accuracy for the clustering of the BBC news (Figure 3) and BBC sport (Figure 5) dataset is 95 and 94.6667 respectively. The figures show that when the dimension is reduced from 9,636 and 4,613 (from BBC news and BBC sport respectively) to 100 (as in Figure 4), the accuracy of clustering is approximately the same.

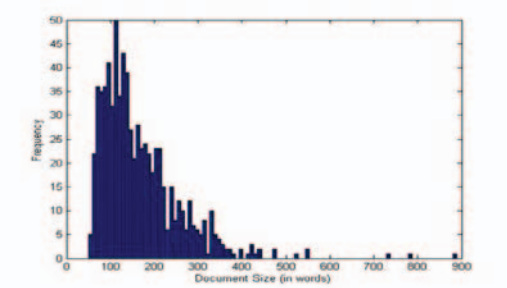


Figure 4. Nunmber of words with frequency for BBC news dataset.

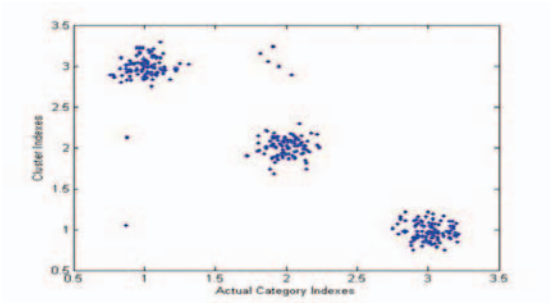


Figure 5. Doucuments clustered into three different topics for BBS Sport dataset.

## V. CONCLUSION

The paper introduces a new algorithm combing TF-IDF and SVD to reduce the high dimensionality. It's clearly that the hybrid algorithm and clustering can be achieved to cluster topics in the similar category. K-means algorithm based on clustering is very important techniques which is used in unsupervised learning machine. The present paper shows that the dimension is reduced from 9,636 and 4,613 (from BBC news and BBC sport respectively) to 100, the accuracy of clustering is approximately the same. The results show that the reduced dimensions can enhance the recognition degree between documents and the system performance for clustering documents.

## REFERNECES

[1] MK. Wong, SSR. Abidi, and D. Jonsen. "A multi-phase correlation search framework for mining non-taxonomic relations from unstructured text.", Knowledge and information systems, vol. 38.3 2014, pp. 641-667.

[2] J. Han, and M. Kamber. "Data Mining, Southeast Asia Edition: Concepts and Techniques" . Morgan kaufmann, 2006.

[3] DT Larose "Discovering knowledge in data: an introduction to data mining". John Wiley & Sons, 2014.

[4] R. Ahmad, and A. Khanum. "Document topic generation in text mining by using cluster analysis with EROCK.", International Journal of Computer Science & Security (IJCSS) vol. 4.2, 2008, pp. 176.

[5] G. Karypis, EH. Han, and V. Kumar. "Chameleon: Hierarchical clustering using dynamic modeling.". Computer vol. 32.8, 1999, pp. 68-75.

[6] T. Zhang, R. Ramakrishnan, and M. Livny. "BIRCH: an efficient data clustering method for very large databases." ACM SIGMOD Record. vol. 25. 2, 1996.

[7] S. Ramasundaram, and SP. Victor. "Text Categorization by Backpropagation Network". International Journal of Computer Applications, vol. 8, 2010, pp. 1-5.

[8] R. E. Banchs. "Text Mining with MATLAB" . Springe, 2012.