# Implementation on Document Clustering using Correlation Preserving Indexing

B.Nageswara Rao, P.Keerthi, V.T. Sree Ramya, S.Santhosh Kumar, T.Monish

*Computer Science & Engineering,*
*Lendi Institute of Engineering & Technology, Vizianagaram, India.*

*Abstract* -- **Document clustering is a technique for unsupervised document organization, automatic topic extraction and fast information recovery. In correlation preserving indexing, the documents are first assign into a low-dimensional semantic space. The documents with in the cluster are highly related to each other while the documents outside the cluster are dissimilar. The document space is always of high dimensional and it is preferable to find a low dimensional representation of the documents to reduce computation complexity. The intrinsic geometrical structure of the document space is often set in the similarities between the documents. Consider Correlation as a similarity measure for detecting the intrinsic geometrical structure of the document space than Euclidean distance.**

Keywords**: Document Clustering, Correlation Preserving Indexing**, **unsupervised learning, intrinsic semantic structure, manifold structure.**

## I. INTRODUCTION

Document Clustering is an automatic grouping of text documents into clusters. Cluster is a subset of objects which are “similar”. A subset of objects such that the distance between any two objects in the cluster is less than the distance between any object in the cluster and any object not located inside it. A connected region of a multidimensional space containing a relatively high density of objects. Based on various distance measures, a number of methods have been proposed to handle document clustering. A typical and widely used distance measure is the Euclidean distance. The k-means method is one of the methods that use the Euclidean distance, which minimizes the sum of the squared Euclidean distance between the data points and their corresponding cluster center. Since the document space is always of high dimensional, it is preferable to find a low-dimensional representation of the documents to reduce computation complexity than k-means. Low computation cost is achieved in spectral clustering methods, in which the documents are first projected into a low-dimensional semantic space and then a traditional clustering algorithm is applied to finding document clusters. However, because of the high dimensional of the document space, a certain representation of documents usually resides on a nonlinear manifold embedded in the similarities between the data points. Unfortunately, the Euclidean distance is a dissimilarity measure which describes the dissimilarities rather than similarities between the documents. Thus, it is not able to effectively capture the nonlinear manifold structure embedded in the similarities between them.

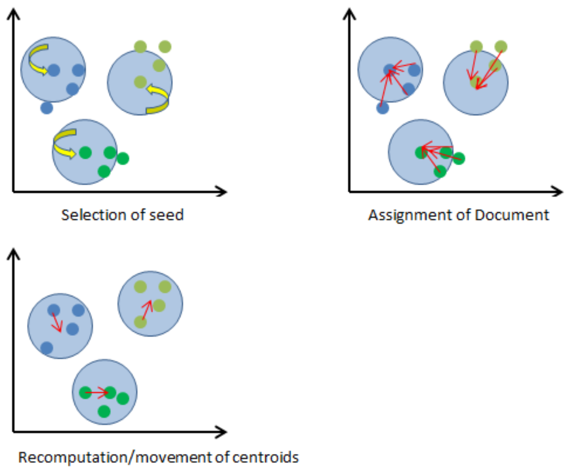
Document clustering involves the use of descriptors and their extraction. Descriptor is the sets of words that describe the content in the document within the cluster. It is generally considered to be a centralized process. Examples includes web document clustering for search users. Document clustering can be categorized to two types, offline and online. Online applications are usually constrained by efficiency problems when compared offline applications.

## II. K-MEANS ALGORITHM

The k-means clustering algorithm is known to be efficient in clustering large data sets. This clustering algorithm was developed by Macqueen, and is one of the simplest and the best known unsupervised learning algorithms that solve the well-known clustering problem. The K-Means algorithm aims to partition a set of objects, based on their attributes/features, into *k* clusters, where *k* is a predefined or user-defined constant. The main idea is to define *k* centroids, one for each cluster. The centroids of a cluster is formed in such a way that it is closely related (in terms of similarity function, similarity can be measured by using different methods such as cosine similarity, Euclidean distance) to all objects in that cluster.

### *A. Basic K-Means Algorithm*

1. Choose k number of clusters to be determined
2. Choose k objects randomly as the initial cluster center
3. Repeat
   3.1. Assign each object to their closest cluster
   3.2. Compute new clusters, i.e. Calculate mean points.
4. Until
   4.1. No changes on cluster centers
   4.2. No object changes its cluster



Selection of seed

Assignment of Document

Recomputation/movement of centroids

Fig (1) Process of Document clustering

## III PROBLEM STATEMENT

- The way to initialize the means was not specified. One popular way to start is to randomly choose k of the samples.
- The results produced depend on the initial values for the means, and it frequently happens that suboptimal partitions are found. The standard solution is to try a number of different starting points.
- It can happen that the set of samples closest to $\mathbf{m}_i$ is empty, so that $\mathbf{m}_i$ cannot be updated. This is an annoyance that must be handled in an implementation, but that we shall ignore.
- The results depend on the metric used to measure $\| \mathbf{x} - \mathbf{m}_i \|$. A popular solution is to normalize each variable by its standard deviation, though this is not always desirable.
- The results depend on the value of k.

## IV. METHODOLOGY

### A Correlation-Based Clustering With TF-IDF

The low-dimensional representation of the 'i'th document xi ϵ X in the semantic subspace, where i=1, 2, 3 ….n.

- D1 = If two documents are close to each other in the original document space, then they tend to be grouped into the same cluster.
- D2 = If two documents are far away from each other in the original document space, they tend to be grouped into different clusters.

$$\max \ \Sigma_i \Sigma_{xj\epsilon N(xi)} \operatorname{corr}(y_i,yj) \qquad (1)$$

And $$\min \Sigma_i \Sigma_{xj\epsilon N(xi)} \operatorname{corr}(y_i,yj) \qquad (2)$$

Where N (xi) denotes the set of nearest neighbours of xi. The equivalent metric learning

$$d(x,y)=\alpha*\cos(x,y) \qquad (3)$$

Where d(x,y) denotes the similarity between the documents x and y, $\alpha$ corresponds to whether x and y are the nearest neighbours of each other.0

### B.Document Representation

Each document is represented as a term frequency vector. The term frequency vector can be computed as follows:

1. Transform the documents to a list of terms after words stemming operations.
2. Remove stop words. Stop words are common words that contain no semantic content.
3. Compute the term frequency vector using the TF/IDF weighting scheme. The TF/IDF weighting scheme assign to

$$(tf/idf)_{i,j} = tf_{i,j}*idf_i \qquad (4)$$

$$tf_{i,j} = \frac{n_{i,j}}{\Sigma_k n_{k,j}} \qquad (5)$$

is the term frequency of the term $t_i$ in document $d_j$, where $n_{i,j}$ is the number of occurrences of the considered term $t_i$ in document $d_j$.

$$idf_i=\log\left(\frac{|D|}{|djti\epsilon d|}\right) \qquad (6)$$

is the inverse document frequency which is a measure of the general importance of the term $t_i$, where $|D|$ is the total number of documents in the corpus and $|\{d_j t_i \epsilon d\}|$ is the number of documents in which the term $t_i$ appears. Let V = {t1, t2. . . tm } be the list of terms after the stop words removal and words stemming operations. The term frequency vector $x_j$ of document dj is defined as

$$X_j=[x1j,x2j,........,xm_j] \qquad (7)$$

$$xij= (tf/idf)_{i,j} \qquad (8)$$

### C. Module Description:

#### 1) Pre-processing:

A new document clustering method based on correlation preserving indexing (CPI), which explicitly considers the manifold structure embedded in the similarities between the documents. It aims to find an optimal semantic subspace by simultaneously maximizing the correlations between the documents in the local patches and minimizing the correlations between the documents outside these patches This is different from K-means, which are based on a dissimilarity measure (Euclidean distance), and are focused on detecting the intrinsic structure between widely separated documents rather than on detecting the intrinsic structure between nearby documents. The similarity-measure-based CPI method focuses on detecting the intrinsic structure between nearby documents rather than on detecting the intrinsic structure between widely separated documents. Since the intrinsic semantic structure of the document space is often embedded in the similarities between the documents, CPI can effectively detect the intrinsic semantic structure of the high-dimensional document space.

#### 2) Documentation clustering based on Correlation Preserving Indexing:

In high-dimensional document space, the semantic structure is usually implicit. It is desirable to find a low dimensional semantic subspace in which the semantic structure can become clear. Hence, discovering the intrinsic structure of the document space is often a primary concern of document clustering. Since the manifold structure is often embedded in the similarities between the documents, correlation as a similarity measure is suitable for capturing the manifold structure embedded in the high-dimensional document space.

#### 3) K-means on Document sets:

The *k*-means method is one of the methods that use the Euclidean distance, which minimizes the sum of the squared Euclidean distance between the data points and their corresponding cluster centres. Since the document space is always of high dimensional, it is preferable to find a low dimensional representation of the documents to reduce computation complexity.

#### 4) Classification of Documents into clusters:

Document clustering aims to group documents into clusters, which belongs unsupervised learning. However, it can be transformed into semi-supervised learning as:

- If two documents are close to each other in the original document space, then they tend to be grouped into the same cluster.

- If two documents are far away from each other in the original document space, they tend to be grouped into different clusters.

Pre-processing is the phase to remove stop words, stemming and identification of unique words. Identification of unique words in the document is necessary for clustering of document with similarity measure. And after that we remove the stop words that is the non informative word for example the, end, have, more etc. The stop words which should be removed are given directly. We need to eliminate those stop words for finding such similarity between documents.

Stemming is the process for reducing derived words to their stem, base or root forms generally a written word form. The stem need not be identical to the root of the word it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. A stemming algorithm is a process in which the variant forms of a word are reduced to a common form.

For example, Removal of suffix to generate word stem grouping words Increase the relevance .Finally term weighting is to provide the information retrieval and text categorization. In document clustering groups together conceptually related documents. Thus enabling identification of duplicate words.

Instead, a typically smaller list of "rules" is stored which provides a path for the algorithm, given an input word form, to find its root form. Some examples of the rules include if the word ends in 'ed', remove the 'ed' ; if the word ends in 'ing', remove the 'ing' ; if the word ends in 'ly', remove the 'ly' Suffix stripping approaches enjoy the benefit of being simpler to maintain than brute force algorithms, assuming the maintainer is sufficiently knowledgeable in the challenges of linguistics and morphology and encoding suffix stripping rules. Finally term weighting is to provide the information retrieval and text categorization. In document clustering groups together conceptually related documents. It also provides metadata characterization the content of given document cluster. Tf–idf, term frequency–inverse document frequency, is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others.

## V. Related Work

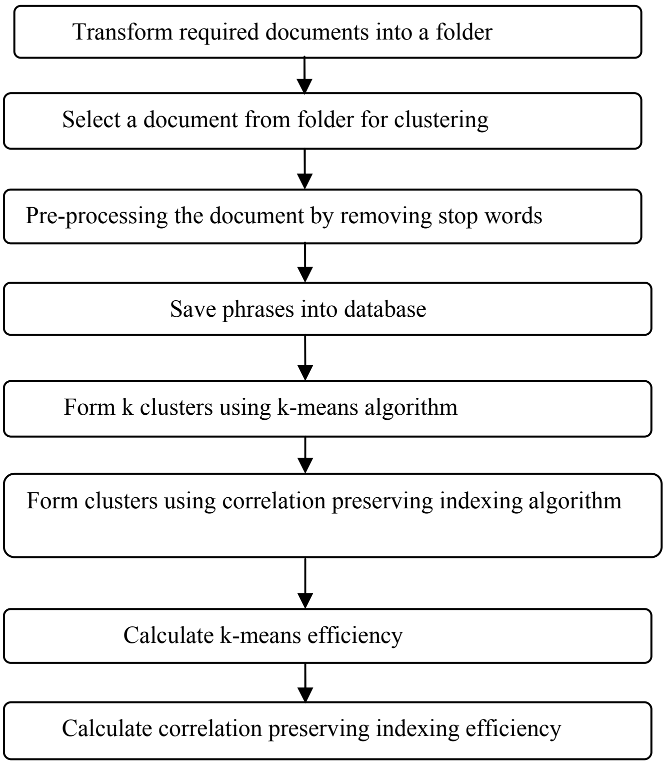
### A *Correlation Preserving Indexing*

The usage of correlation as a similarity measure can be found in the canonical correlation analysis (CCA) method. The CCA method is to find projections for paired data sets such that the correlations between their low-dimensional representatives in the projected spaces are mutually maximized. Specifically, given a paired data set consisting of matrices X= {x1; x2; . . .; xn ;} and Y {y1; y2; . . . ; yn;} we would like to find directions wx for X and wy for Y that maximize the correlation between the projections of X on wx and the projections of Y on wy. This can be expressed as

$$\text{Max}_{wx,wy} \frac{<Xwx,Ywy>}{||Xwx||.||Ywy||}$$

Where <·,·> and ||.|| denote the operators of inner product and norm, respectively. As a powerful statistical technique, the CCA method has been applied in the field of pattern recognition and machine learning]. Rather than finding a projection of one set of data, CCA finds projections for two sets of corresponding data X and Y into a single latent space that projects the corresponding points in the two data sets to be as nearby as possible. In the application of document clustering, while the document matrix X is available. So the CCA method cannot be directly used for clustering.

In this paper, we propose a new document clustering method based on correlation preserving indexing (CPI), which explicitly considers the manifold structure embedded in the similarities between the documents. It aims to find an optimal semantic subspace by simultaneously maximizing the correlations between the documents in the local patches and minimizing the correlations between the documents outside these patches.


Transform required documents into a folder
Select a document from folder for clustering
Pre-processing the document by removing stop words
Save phrases into database
Form k clusters using k-means algorithm
Form clusters using correlation preserving indexing algorithm
Calculate k-means efficiency
Calculate correlation preserving indexing efficiency

**Figure(2)**

### B *Clustering Algorithm Based on CPI*

Given set of documents x1,x2,x3...xn є IRn. Let X denotes the document matrix. The algorithm for document clustering based on CPI can be summarized as follows:

1. Construct the local neighbour patch, and compute the matrices $M_sW$ and λMW.

2. Project the document vectors into the subspace by throwing away the zero singular values. The singular value decomposition of X can be written as $X=U\Sigma V^T$.

.3. Here all zero singular values in $\Sigma$ have been removed. Accordingly, the vectors in U and V that correspond to these zero singular values have been removed as well. Thus the document vectors in the subspace can be obtained by $\tilde{X}=U^TX$.

4. Compute CPI Projection. Based on the multipliers $\lambda_1$ , $\lambda_2$ ,.... $\lambda_n$ one can compute the matrix

$$M = \lambda_0{}^*MT + \lambda_1{}^* x_1x_1{}^T+....+\lambda_n{}^*x_nx_n{}^T.$$

5. Let $W_{CPI}$ be the solution of the generalized Eigenvalue problem $M_SW = \lambda MW$. Then, the low dimensional representation of the document can be computed by

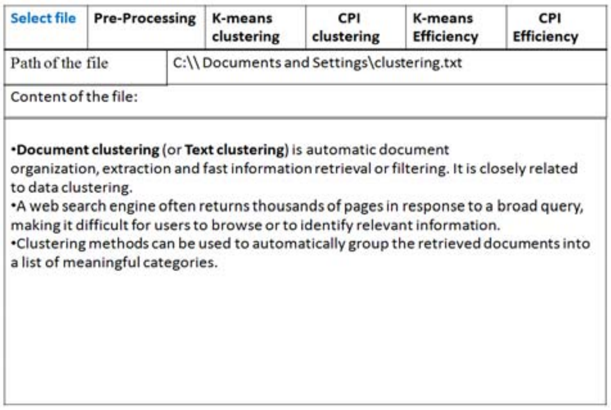
$$Y=W^T{}_{CPI}\tilde{X}{}^=W^TX \qquad (10)$$

## VI. Results And Discussions

Performance comparison of k-means and CPI:

| *Data set* | *k-means* | *CPI* |
|---|---|---|
| DOC1 | 74.14±16.10 | 97.08±4.41 |
| DOC2 | 63.92±11.49 | 83.30±11.40 |
| DOC3 | 67.78±14.18 | 86.18±12.07 |
| DOC4 | 64.50±10.87 | 76.38±12.36 |
| DOC5 | 73.56±13.51 | 95.18±7.07 |

Initially, select a file from set of documents in a folder by providing path of the file. After pre-processing the content, it removes stop words and form phrases. Later these phrases are separated as keywords which have semantic meaning. Based on these keywords clusters are formed using k-means algorithm and correlation preserving indexing algorithms. Calculate and compare efficiencies of both k-means and correlation preserving indexing algorithms. As a result correlation preserving indexing performs better results than k-means. Correlation preserving indexing gives results with low computation cost.



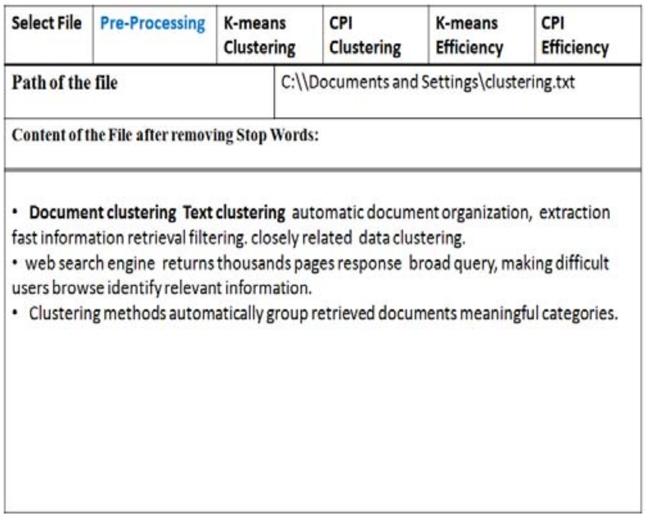
| Select file | Pre-Processing | K-means clustering | CPI clustering | K-means Efficiency | CPI Efficiency |
|---|---|---|---|---|---|

Path of the file C:\\ Documents and Settings\clustering.txt

Content of the file:

•**Document clustering** (or **Text clustering**) is automatic document organization, extraction and fast information retrieval or filtering. It is closely related to data clustering.
•A web search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information.
•Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories.

Fig 3(a) Selecting file from folder



| Select File | Pre-Processing | K-means Clustering | CPI Clustering | K-means Efficiency | CPI Efficiency |
|---|---|---|---|---|---|

Path of the file C:\\Documents and Settings\clustering.txt

Content of the File after removing Stop Words:

- **Document clustering Text clustering** automatic document organization, extraction fast information retrieval filtering. closely related data clustering.
- web search engine returns thousands pages response broad query, making difficult users browse identify relevant information.
- Clustering methods automatically group retrieved documents meaningful categories.

Fig 3(b) after pre-processing document

## Conclusions

In this paper, we present a new document clustering method based on correlation preserving indexing. It simultaneously maximizes the correlation between the documents inside the clusters and minimizes the correlation between the documents outside the clusters. Consequently, a low dimensional semantic subspace is derived where the documents corresponding to the same semantics are close to each other. It reduces the computational cost.

## References

- http://www.comp.hkbu.edu.hk/~yytang/stopwords.txt
- http://www.Articles.com/439890/Text-Documents-Clustering-using-K-Means-Algorithm
- http://en.wikipedia.org/wiki/Document_clustering
- http://www.milanmirkovic.com/wp-content/uploads/2012/10/pg049_Similarity_Measures_for_Text_Document_Clustering.pdf
- http://www.cscjournals.org/csc/manuscript/Journals/IJDE/volume2/Issue4/IJDE-63.pdf