# Usage of Cosine Similarity and term Frequency count for Textual document Clustering

**B Sindhuja, Mrs. VeenaTrivedi**

*Abstract*— **This paper presents textual document clustering using two approaches namely cosine similarity and frequency and inverse document frequency. With the combination of these approaches a similarity measure values are generated between keywords in the documents and between the documents. Using this approach, the best related document can be identified on the basis of clustering method called correlation preserving index in which related documents are stored in an index format.**

*Index Terms*— **Document Clustering, Cosine similarity, Tf-idf, Correlation preserving index.**

## I. INTRODUCTION

Textual document clustering [1] is a fundamental operation used in unsupervised document organization and information retrieval. Document clustering aims to automatically group related documents into clusters. It is one of the most important tasks in machine learning and artificial intelligence and has received much attention in recent years [2] [3] [4]. 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 square Euclidean distance between the data points and there is always of high dimensionality, it is preferable to find a low-dimensional representation of the documents to reduce computation complexity corresponding cluster centers. Since the document space. 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. Latent semantic indexing (LSI)[6][7][10] is one of the effective spectral clustering methods, aimed to finding the best subspace approximation term to the original document space by minimizing the global reconstruction error (Euclidean distance)[10]. However, because of the high dimensionality of the document space, a certain representation of documents usually resides on a nonlinear manifold embedded in the similarities between the data points. 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. An effective document clustering method must be able to find a low-dimensional representation of the documents that can best preserve the similarities between the data points. Locality preserving indexing (LPI) [6] [10] method is a different spectral clustering method based on graph partitioning theory. The LPI method applies a weighted function to each pair wise distance attempting to focus on capturing the similarity structure, rather than the dissimilarity structure, of the documents. However, it does not overcome the essential limitation of Euclidean distance. Furthermore, the selection of the weighted functions is often a difficult task. Correlation as a similarity measure can capture the intrinsic structure embedded in high-dimensional data, especially when the input data is sparse [7] [10]. It is a scale-invariant association measure usually used to calculate the similarity between two vectors. In many cases, correlation can effectively represent the distributional structure of the input data which conventional Euclidean distance cannot explain.

Cosine similarity together with term frequency and inverse document frequency approach is used to find out semantic structure between the documents and generated similarity measure values between the keywords to the documents and between the documents.

## II. RELATED WORK

In Document Clustering [1][2][3][10] in Correlation Similarity Measure Space an effective document clustering method must be able to find a low-dimensional representation of the documents that can best preserve the similarities between the data points. Low computation cost is achieved in spectral clustering methods, in which the semantic space and then a traditional clustering algorithm are applied to finding document clusters [1] [10]. Because of the high dimensionality 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 [10]. Thus,

**Manuscript received August 18, 2014**.
**B. Sindhuja**, Information Technology, Gokaraju Rangarju Institute of Engineering and Technology, Hyderabad, India, 9032663923
**Mrs. VeenaTrivedi**, Information Technology, Gokaraju Rangarju Institute of Engineering and Technology, Hyderabad, India , 9052988260.,

it is not able to effectively capture the nonlinear manifold structure embedded in the similarities between them. So a new document clustering method based on correlation preserving indexing (CPI) [10], 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. Christopher *et al.,* (2008) study shows that the traditional vector space information retrieval model in document clustering use words as measure to find similarity between documents. In reality the concepts, semantics, features, and topics are used to describe the documents. VSM (Vector Space Model) [10] ignores semantic relations among terms. For instance, having "automobile" in one document and "car" in another document does not contribute to the similarity measure among these two documents. Several factors contribute to this problem and motivate the research. The semantic relationships between documents are not explored in the most of the clustering methods. In Document Clustering Using Locality Preserving Indexing [6] [10] Deng Cai, Xiaofei He, and Jiawei Han Generally, the document space is of high dimensionality, typically ranging from several thousands to tens of thousands. Learning in such a high-dimensional space is extremely difficult due to the curse of dimensionality. Thus, document clustering necessitates some form of dimensionality reduction. One of the basic assumptions behind data clustering is that, if two data points are close to each other in the high dimensional space, they tend to be grouped into the same cluster. Therefore, the optimal document indexing method should be able to discover the local geometrical structure of the document space. To this end, the LPI [6] algorithm is of particular interest. LSI [5] [6] is optimal in the sense of reconstruction. It respects the global Euclidean structure while failing to discover the intrinsic geometrical structure, especially when the document space is nonlinear. Another consideration is due to the discriminating power. One can expect that the documents should be projected into the subspace in which the documents with different semantics can be well separated, while the documents with common semantics can be clustered. As indicated LPI is an optimal unsupervised approximation to the Linear Discriminant Analysis algorithm which is supervised. Therefore, LPI [6] [7] [10] can have more discriminant power than LSI. There are some other linear subspace learning algorithms, such as informed projection and Linear Dependent Dimensionality [6] [10] Reduction. However, none of them has shown discriminating power. Finally, it would be interesting to note that LPI is fundamentally based on manifold theory. LPI tries to find a linear approximation to the Eigen functions of the Laplace Beltrami operator on the compact Riemannian manifold. Therefore, LPI is capable of discovering the nonlinear structure of the document space to some extent. Document Clustering Based on Non-negative Matrix Factorization Wei Xu, Xin Liu, and Yihong Gong Clusters each of which corresponds to a coherent topic [10]. Each document in the corpus either completely belongs to a particular topic, or is more or less related to several topics. To accurately cluster the given document corpus [10], it is ideal to project the document corpus into a *k*-dimensional semantic space in which each axis corresponds to a particular topic. In such a semantic space, each document can be represented as a linear combination of the *k* topics [10]. Because it is more natural to consider each document as an additive rather than subtractive mixture of the underlying topics, the linear combination coefficients should all take non-negative values. Tf-Idf short for term frequency–inverse document frequency [9] [10], is a numerical statistic that is intended to reflect 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[9][10]. Variations of the Tf–idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. Tf–idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

Cosine Similarity [8] [9] [10] is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. Two vectors with the same orientation have a Cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0, 1].Note that these bounds apply for any number of dimensions, and Cosine similarity is most commonly used in high-dimensional positive spaces. For example, in Information Retrieval and text mining, each term is notionally assigned a different dimension and a document is characterized by a vector where the value of each dimension corresponds to the number of times that term appears in the document. Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter [8]

## III. SIMILARITY MEASURES

### A) Cosine Similarity Method

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 [10]. 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. Mathematically, the correlation between two vectors (column vectors) a and b is defined as

$$Corr(a,b) = a^T b \Big/ \sqrt{a^T a}\sqrt{b^T b} = \left(\frac{a}{\|a\|}.\frac{b}{\|b\|}\right)$$

The correlation corresponds to an angle Θ such that

$$\cos\theta = Corr(a,b)$$

The larger the value of Corr (a, b) the stronger the association between the two vectors a and b.

Online document clustering aims to group documents into clusters, which belongs unsupervised learning. However, it can be transformed into semi-supervised learning by using the following side information:

A1. If two documents are close to each other in the original document space, then they tend to be grouped into the same cluster [8].

A2. If two documents are far away from each other in the original document space, they tend to be grouped into different clusters.

**B) Term Frequency And Inverse Document Frequency**

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 assigned to the term ($t_i$) in document ($d_j$) is given by

$$\left(tf/_{idf}\right)_{i,j} = tf_{i,j} \times idf_i$$

$$tf_{i,j} = n_{i,j} \Big/ \sum_k n_{k,j}$$

Here ($tf_{i,j}$) is the term frequency of the term $t_i$ in document $d_j$, where $n_{k,j}$ is the number of occurrences of the considered term $t_i$ in document

$$d_j . idf_i = \log\left(\frac{|D|}{|d: t_i \in d|}\right)$$

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 $|d: t_i \in d|$ is the number of documents in which the term $t_i$ appears. Let V = $\{t_1, t_2 \ldots t_m\}$ be the list of terms after the stop words removal and words stemming operations. The term frequency vector $X_j$ of document $d_j$ is defined as

$$X_j = [x_{1j}, x_{2j}, \ldots, x_{mj}]$$

$$x_{ij} = \left(tf/_{idf}\right)_{i,j}$$

Using n documents from the corpus, we construct a $m \times n$ term-document matrix X.

## IV. ALGORITHM

Clustering Algorithm Based on CPI

1. Construct the local neighbor patch, and compute the matrices MS and MT.
2. Project the document vectors into the SVD subspace by throwing away the zero singular values. The singular value decomposition of X can be written as X= $x = a \sum h^T$. Here all zero singular values in have been removed. Accordingly, the vectors in a and b that correspond to these zero singular values have been removed as well.
3. CPI Projection is computed. The matrix value is computed.
4. Documents are clustered in the CPI semantic subspace.

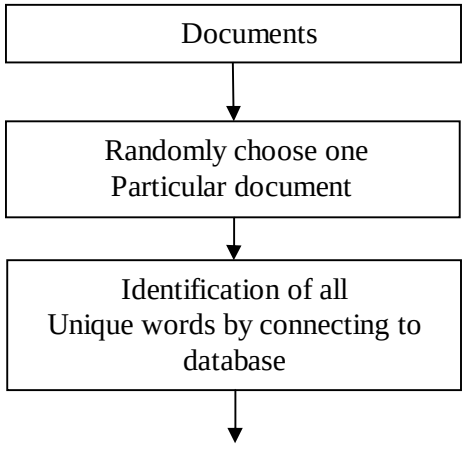
I) If two documents are close to each other in the original document space, then they tend to be grouped into the same cluster

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

## V. EXPERIMENT AND RESULTS

**Evaluation**

It is very difficult to conduct a systematic study comparing the impact of similarity measures on cluster quality, because objectively evaluating cluster quality is difficult in itself. In practice, manually assigned category labels are usually used as baseline criteria for evaluating clusters. As a result, the clusters, which are generated in an unsupervised way, are compared to the pre-defined category structure, which is normally created by human experts. This kind of evaluation assumes that the objective of clustering is to replicate human thinking, so a clustering solution is good if the clusters are consistent with the manually created categories.



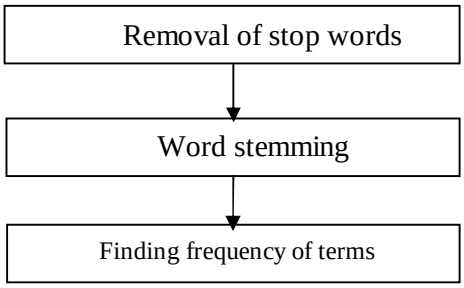



Fig 1: Document preprocessing

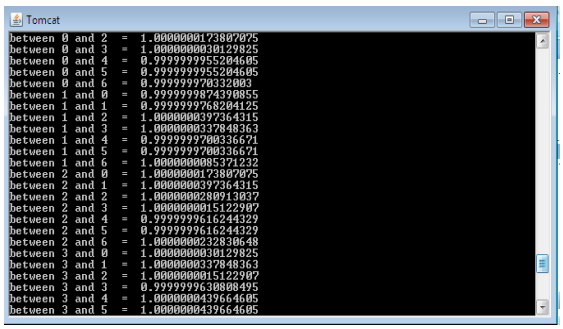
Preprocessing is the phase to remove stop words, stemming and identification of unique words. Identification of unique words in the document is necessary for clustering of document with similarity measure. And after that we remove the stop words that is the non informative word for example the, end, 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
- Groping 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.

**Results**


Fig 2: Similarity measure values between keyword and documents



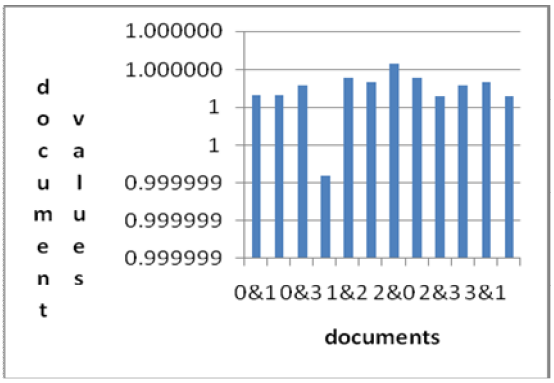




Fig 3: Similarity measure values between documents

## VI. CONCLUSION AND FUTURE ENHANCEMENT

Currently limited documents are used for calculation purpose. More documents can be used for calculation and hence better results can be obtained. The research can be further enhanced using N-gram similarity approximation method.

## VII. REFERENCES

[1]R.T. Ng and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB), pp. 144-155, 1994.
[2] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.
[3] S. Kotsiantis and P. Pintelas, "Recent Advances in Clustering: A Brief Survey," WSEAS Trans. Information Science and Applications, vol. 1, no. 1, pp. 73-81, 2004.
[4] R. Mihalcea and C. Corley "Measuring the Semantic Similarity of Texts," Proc. ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, 2005, page. 13-18.
[5] X. Liu, Y. Gong, W. Xu, and S. Zhu, "Document Clustering with Cluster Refinement and Model Selection Capabilities," Proc. 25th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '02), page. 191-198, 2002.
[6] D. Cai, X. He, and J. Han, "Document Clustering Using Locality Preserving Indexing," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 12, pp. 1624-1637, Dec. 2005.
[7]S.C. Deerwester, S.T. Dumais, "Indexing by Latent Semantic Analysis," J. Am.Soc. Information Science, vol. 41, no. 6, pp. 391-407, 1990.
[8] K.P.N.V.Satya Sree1, Dr.J V R Murthy2
"Clustering Based On Cosine Similarity Measure" International Journal Of Engineering Science & Advanced Technology Volume-2, Issue-3,2012
[9] An improved TF-IDF approach for textclassification ZHANG Yun-tao, GONG Ling 2004
[10]Taiping Zhang, Yuan Yan Tang, Bin Fang and Yong Xiang "Document Clustering in Correlation Similarity Measure Space" Ieee Transactions On Knowledge And Data Engineering, Vol. 24, No. 6, June 2012.

**B. Sindhuja** Masters of Technology, Student, Gokaraju Rangaraju Institute of Engineering and Technology, Bachupally, Hyderabad. Area of Interest: Data Mining.

Assoc Prof Mrs. Veena Trivedi has Bachelor's and Master's Degree in the field of Computer Science and Engineering. She has Keen interest in the area of Data mining and Computer Networks and has published several papers in national and International conferences and journals. She has fifteen years of experience in teaching undergraduate as well as post graduate students. She has attended several workshops and faculty development programs.