International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 01 | Jan -2021 ISSN: 2582-3930

Document Clustering Based On Correlation-A Comparative Study

Aswathy V Shaji[1], Arjun V S[2]

[1]Department of Computer Science, SN College Cherthala
[2]Department of Computer Science, SN College Cherthala

Abstract—**Document clustering is a useful technique that automatically organises large quantities of documents into a small number of coherent groups called clusters. In this paper, aSpectral clustering method called Correlation Preserving Indexing (CPI) has been proposed for clustering the related documents together based on the correlation similarity measure.In this work, comparative study of the existing document clustering method, Locality Preserving Indexing (LPI) and the proposed method CPI is done over different similarity measures, Correlation and Jaccard coefficient. Experimental evaluation shows that Correlation Preserving Indexing (CPI) method is more efficient than Locality Preserving Indexing (LPI) method as CPI is based on similarity measure. In CPI method the similarity measure Corelation and Jaccard coefficient are taken and it is clear that Jaccard coefficient is a better metric for text clustering.**

Keywords —**Document clustering, correlation measure, correlation latent semantic indexing, locality preserving indexing, dimensionality reduction.**

1. INTRODUCTION

Document clustering is one of the most important techniques to organize the documents in an unsupervised manner.It is a significant area of interest in the field of machine learning. To handle document clustering different distance measures are used. A widely used distance measure is the Euclidean distance. Spectral clustering methods are employed to improve the computational cost, in which the documents are arranged into a low dimensional semantic space and traditional clustering algorithm is applied for finding document clusters.

The Euclidean distance could be used as a dissimilarity measure. It specifies dissimilarities rather than similarities between the documents. Hence, it is not able to effectively capture the nonlinear manifold structure available in the similarities between the documents. An effective document clustering method must be able to find a low dimensional representation of the documents. Locality preserving indexing (LPI) [1] method is a different spectral clustering method which is based on graph partitioning theory. The LPI method applies a weighted function for the purpose of capturing the similarity structure, rather than the dissimilarity structure of the documents. But the problem with LPI is that the selection of weighted function is a tedious task and also it does not overcome the difficulties with the Euclidean method. In correlation preserving indexing (CPI), the similarity between the documents is considered rather than the dissimilarity. 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. In short the methods LSI and LPI are based on dissimilarity measure but in CPI method is based on similarity measure. Hence CPI can detect the intrinsic geometric space of the document, by minimizing the intra-cluster distances and maximizing the inter-cluster distance.

The remaining part of the paper is organized as follows: Section 2 presents the related work on this topic. Section 3 presents an overview of clustering based on correlation preserving indexing, Section 4 presents the algorithm for Locality Preserving Indexing, Section 5 presents Pre- processing of documents, Section 6 present the different distance measures, Section 7 deals with the algorithm implementation, Section 8 deals with experimental analysis and Section 9 presents the conclusion and future work.

2. RELATED WORK

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 needs some form of dimensionality reduction. One of the basic assumptions behind data clustering is that, if two data points are close toeach other in the high dimensional space, they tend to be grouped into the same cluster.

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. In K-means computational complexity is high.

Latent semantic indexing (LSI) is one of the effective spectral clustering methods, aimed at finding the best subspace approximation to the original document space by minimizing the global reconstruction error (Euclidean distance). In LSI 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. Thus, it is not able to effectively capture the nonlinear manifold structure embedded in the similarities between them.LSI is optimal in the sense of reconstruction. It respects the global Euclidean structure while fails to discover the intrinsic geometrical structure especially when the document space is non-linear.

Locality preserving indexing (LPI) [1] 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. LPI method does not overcome the essential limitation of Euclidean distance. Furthermore, the selection of the weighted functions is often a difficult task.

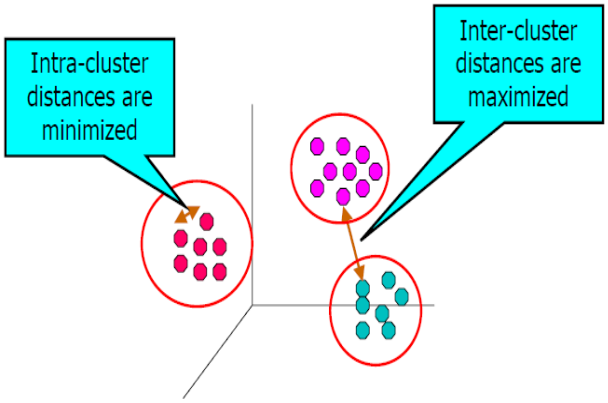

Figure 1: Good clustering scheme

3. CLUSTERING BASED ON CORRELATION PRESERVING INDEXING

When the document space is high dimensional, the semantic structure is not at all clear. Hence it is desirable to find a low dimensional semantic subspace in which the semantic structure can become clear. Discovering the intrinsic structure of the document space is often a major concern involved in document clustering. Since the manifold structure is often embedded in the similarities between the documents, correlation is suitable for capturing the manifold structure embedded in the high-dimensional document space. Mathematically, the correlation between two vectors *u* and *v* is defined as,

$$Corr(u,v) = \frac{u^T v}{\sqrt{u^T u}\sqrt{v^T v}} = \frac{u}{||u||}, \frac{v}{||v||} \quad (1)$$

where*u* and *v* are document vectors, which is obtained after preprocessing and term weighting procedures.

The main result of a correlation is called the correlation coefficient or *r*. It ranges from -1.0 to +1.0. The results are summarized as follows:

1. The closer *r* is to +1 or -1, the more closely the two variables are related.
2. If *r* is close to 0, it means there is no relationship between the variables.
3. If *r* is positive, it means that as one variable gets larger, the other gets larger.
4. If r is negative it means that as one variable gets larger, the other gets smaller.

3.1Clustering criterion based on correlation

Suppose $y_i \in Y$ is the low-dimensional representation of the i^{th}document $x_i \in X$ in the semantic subspace, where i = 1,2, . .. n. Then we can summarize the equation for the correlation as

$$max \sum_i \sum_{x_j \in N(x_i)} Corr(y_i, y_j) \quad (2)$$

and

$$min \sum_i \sum_{x_j \notin N(x_i)} Corr(y_i, y_j) \quad (3)$$

respectively, where N (x_i) denotes the set of nearest neighbors of x_i. The maximization problem (2) is an attempt to ensure that if x_i and x_j are close, then y_i and y_j are close as well. Similarly, the minimization problem (3) is an attempt to ensure that if x_i and x_j are far away, y_i and y_j are also far away. The optimization of (2) and (3) is equivalent to the following metric learning:

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

where d(x, y) denotes the similarity between the documents x and y, α corresponds to whether x and y are the nearest neighbors of each other. Since the following equality is always true.

$$\sum_i \sum_{y_{j \in N(y_x)}} Corr(y_i, y_j) + \sum_i \sum_{y_{j \notin N(y_x)}} Corr(y_i, y_j) = \sum_i \sum_j Corr(y_i, y_j) \quad (4)$$

The simultaneous optimization of (2) and (3) can be achieved by maximizing the following objective function.

$$\frac{\sum_i \sum_{x_{j \in N(x_i)}} Corr(y_i, y_j)}{\sum_i \sum_j Corr(y_i, y_j)} \quad (5)$$

Without loss of generality, we denote the mapping between the original document space and the low-dimensional semantic subspace by *W*, that is, $W^T x_i = y_i$. Where tr(.) is the trace operator.

$$\frac{\sum_i \sum_{x_{j \in N(x_i)}} Corr(y_i, y_j)}{\sum_i \sum_j Corr(y_i, y_j)} = \frac{\sum_i \sum_{x_{j \in N(x_i)}} \frac{y_i^T y_j}{\sqrt{y_i^T y_i y_j^T y_j}}}{\sum_i \sum_j \frac{y_i^T y_j}{\sqrt{y_i^T y_i y_j^T y_j}}}$$

$$= \frac{\sum_i \sum_{x_{j \in N(x_i)}} \frac{tr(W^T x_i x_j^T W)}{\sqrt{tr(W^T x_i x_i^T W) tr(W^T x_j x_j^T W)}}}{\sum_i \sum_j \frac{tr(W^T x_i x_j^T W)}{\sqrt{tr(W^T x_i x_i^T W) tr(W^T x_j x_j^T W)}}} \quad (6)$$

Physically, this model may be interpreted as follows: all documents are projected onto the unit hyper-sphere (circle for 2D). The global angles between the points in the local neighbors, β_i are minimized and the global angles between the points outside the local neighbors, α_j, are maximized simultaneously.

In our algorithm, we use SVD projection in our data preprocessing step to remove those components corresponding to the zero singular value. If the rank of original term-document matrix X equals to the number of documents, the X will be a full rank square matrix after SVD projection. In document clustering, the number of terms is often larger than the number of documents, thus if all the document vector x are linearly independent, the X will be a full rank square matrix after SVD projection

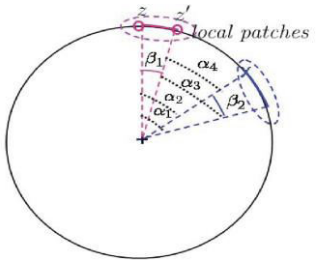

Figure 2:2D Projections on CPI

.The main aim of this work is to develop document clustering method based on overall architecture given in Figure 3.

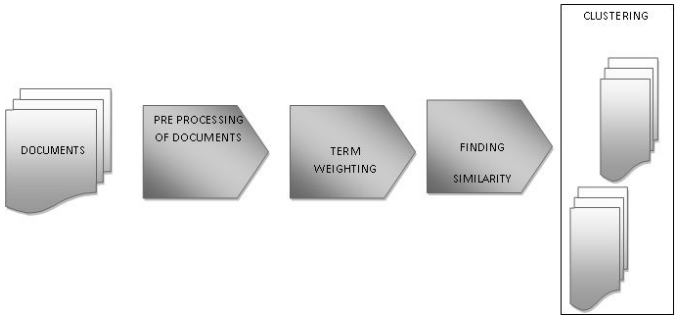

Figure 3: Overall system architecture

4. LOCALITY PRESERVING INDEXING (LPI) ALGORITHM FOR CLUSTERING

A set of documents $x_1, x_2, \ldots, x_n \in R^n$ is given. Let X denotes the document matrix. The algorithm for document clustering based on LPI can be summarized as follows:

1. Constructing the adjacency graph: Let G denote a graph with n nodes. The i-th node corresponds to the document x_i. We put an edge between nodes iand j if x_i and x_j are "close".
2. Choosing the weights: If nodes iand j are connected, put $S_{ij} = x_i^T x_j$ Otherwise, put $S_{ij} = 0$. The weight matrix S of graph G models the local structure of the document space. We define D as a diagonal matrix whose entries are column (or row, since S is symmetric) sums of S, i.e., $D_{ii} = \sum_j S_{ij}$. We also define L = D − S, which is called the Laplacian matrix.
3. 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 $X = U \sum V^T$. Here all zero singular values in $\sum$ 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 SVD subspace can be obtained by $\tilde{X} = U^T X$.
4. LPI Projection: Compute the eigenvectors and eigen values for the generalized eigen problem
$$\tilde{X} L \tilde{X}^T a = \lambda \tilde{X} D \tilde{X}^T$$
5. Cluster the documents in LPI semantic subspace.

5. PRE PROCESSING OF DOCUMENTS

We choose a set of n document which has to be group together for detecting intrinsic structure of the document space using correlation. Pre processing steps of documents includes stop words removal and stemming.

1. *Stop word removal:* A stop word is defined as a term which is not thought to convey any meaning as a dimension in the vector space. Stop words are the most common words (e.g., "and", "or" , "in") in a language, but they do not convey any significant information so they are stripped from the document set.
2. *Stemming:* This is known as the process of reducing words to their base form, or stem. For example, the words "connected", "connection", "connections" are all reduced to the stem "connect". Porter's algorithm [10] is the de facto standard stemming algorithm. Smaller number of distinct terms results in a saving of memory space and processing time.

5.1 Document Representation

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

The TF/IDF weighting scheme assigned to the term t_i in document $d_j$is given by

$$(tf\ /idf)_{i,j} = tf_{i,j} * idf_i \quad (7)$$

Here,

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (8)$$

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.$idf_i = \log(\frac{|D|}{|d:t_i \in d|})$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 test set and $|\{d: t_i \in d\}|$is the number of documents in which term t_i occurs. Let v={t_1,t_2,..t_m} 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

$$X_J = [x_{1j}, x_{2j}, \dots \dots x_{mj}],$$

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

Using n documents from the corpus, construct an $m \times n$ Term-document matrix X.

6. SIMILARITY MEASURES

To cluster similar documents together, clustering algorithms require a similarity measure between two documents dl and d2.

Euclidean distance: It is a standard metric for geometrical problems. It is the ordinary distance between two points and can be easily measured with a ruler in two- or three-dimensional space. Euclidean distance is widely used in clustering problems, including clustering text. Measuring distance between text documents, given two documents d_a and d_b represented by their own term vectors t_a and t_b respectively, the Euclidean distance of two documents is defined as

$$D_E(t_a, t_b) = (\sum_{t=1}^{m} |w_{t,a} - w_{t,b}\ |)^{1/2} \quad (9)$$

where the term set is T={t_1,t_2,…..t_m}.As mentioned previously, we use tf/idf value as term weights, that is $w_{t,a}$=tf/idf(d_a,t)

Correlation Similarity: It can be calculated as Corr(u, v) = (u/|u|, v/|v|),where u and v are vectors over the term set. An important property of the cosine similarity is its independenceof document length. For example, combining two identical copies of a document d to get a new pseudo document d^1, the correlation similarity between d and d^1 is 1, which means that these two documents are regarded to be identical. Meanwhile, given another document l, d and d^1 will have the same similarity value to l, that is, $sim(t_d, t_l) = sim(t_d^1, t_l)$.In other words, documents with the same composition but different totals will be treated identically.

Jaccard Coefficient: Jaccard coefficient measures similarity as the intersection divided by the union of the objects. For text document, the Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms.

$$sim_j(t_a, t_b) = \frac{t_a.t_b}{|t_a|^2 + |t_b|^2 - t_a.t_b} \quad (10)$$

The Jaccard coefficient is a similarity measure and ranges between 0 and 1. It is 1 when the t_a =t_b and 0 when t_a and t_b are disjoint, where 1 means the two objects are the same and 0 means they are completely different.

7. CPI ALGORITHM FOR CLUSTERING[2]

A set of documents x_1, x_2, . . . ,$x_n \in R^n$ is given. 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_S and M_T where $M_T = \sum_i \sum_j (x_i x_j^T)$and $M_S = \sum_i \sum_{x_j \in N(x_i)} (x_i x_j^T)$
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 $X = U \sum V^T$. Here all zero singular values in $\sum$ 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 SVD subspace can be obtained by $\tilde{X} = U^T X$
3. Compute CPI Projection. Based on the multipliers λ_0,λ_1,...$\lambda_n$one can compute the matrix $M=\lambda_i * M + \lambda 1 * x_1 x_1^T + \dots + \lambda n * x_n x_n^T$. Let WCPI be the solution of the generalized eigen value problem $M_S W = \lambda M W$. Then, the low dimensional representation of the document can be computed by $Y = W_{CPI}^T \tilde{X} = W^T X$ Where $W=UW_{CPI}$ is the transformation matrix.
4. Cluster the documents in CPI semantic subspace.

7.1 Complexity Analysis

The time complexity of the CPI clustering algorithm can be analyzed as follows: consider n documents in the d-dimensional space (d >> n). In Step 1, we first need to compute the pair wise distance which needs O (n^2d) operations. Second, we need to find the k nearest neighbors for each data point which needs O (kn^2) operations. Third, computing the matrices M_S and M_T requires O (n^2d) operations and O (n (n – k) d) operations, respectively. Thus, the computation cost in Step 1 is O($2n^2$d +kn^2 + n(n –k)d) In Step 2, the SVD decomposition of the matrix X needs O(d^3) operations and projecting the documents into the n dimensional SVD subspace takes O(mn^2) operations. As a result, Step 2 costs O (d^3 + n^2d). In Step 3, we need to solve the generalized eigen value problem Msw=λMw in order to find the m generalized eigenvectors associated with the m-largest eigen values which needs O (n^3) operations. Then, transforming the documents into m-dimensional semantic subspace requires O (mn^2) operations. Consequently, the computation cost of Step 3 is O (n^3 + mn^2). In Step 4, it takes O(lcmn) operations to find the final document clusters, where l is the number of iterations and c is the number of clusters.

Since k << n, l << n, and m; n << d in document clustering applications.

8. EXPERIMENTAL RESULT

This section presents the experimental evaluation of our method and compares its result with existing method, Locality Preserving Indexing (LPI). A set of text documents are selected and cluster analysis is performed.

8.1Dataset

All datasets used for evaluation are text documents. These include abstract documents from IEEE Xplore and ACM. It contains document abstracts in the field of cloud computing, data mining, mobile computing and computer networks.

Newspaper articles are also taken for evaluation purpose.We only use the articles that are uniquely assigned to exactly one topic.Datasets are heterogeneous in terms of document size, cluster size and document distribution.

Initially the documents are pre-processed and the frequency measure steps are completed. In first step we implement the CPI algorithm. This algorithm is based on similarity measure. In this work Correlation and Jaccard coefficient are taken. For experimental analysis both the methods LPI and CPI are used.

8.2 Performance Evolution

Out of 200 documents about 50 documents were taken for training purpose. The rest of the documents were used for testing. It is very important for a clustering method to have the capability of predicting the new data by using the information acquired during training.

Finally the cluster label of the testing documents was predicted by using the knowledge formerly acquired from training.

$$precision(k_i, c_j) = \frac{n_{ij}}{|c_j|} \quad (11)$$

wheren_{ij} is the number of members of class K, in cluster Cj .

PRECISION			
domain	Euclidean (LPI)	Correlation (CPI)	Jaccard (CPI)
Abstract documents	0.287	0.724	0.773
News paper articles	0.32	0.687	0.77

Table 1: Precision Results

As shown in Table 1, Euclidean distance performs worst while the performances of the other two measures are quite similar. On average, the Jaccard is slightly a good metric for text clustering.

Recall means text search on a set of documents in which the number of correct results divided by the number of results that should have been returned. The increase value of recall also indicates a good clustering.

$$recall(k_i, c_j) = \frac{n_{ij}}{|k_i|} \quad (12)$$

RECALL			
domain	Euclidean (LPI)	Correlation (CPI)	Jaccard (CPI)
Abstract documents	0.267	0.684	0.792
News paper articles	0.289	0.597	0.725

Table 2: Recall Results

Results from Table 2, shows that Jaccard coefficient has better Recall capability compared to the other two distance metric, Euclidean as well as the Correlation.

For evaluating the quality of clustering frequently used metric is the F-Measure. It is a measure that combines the precision and recall ideas from information retrieval. F measure can be calculated as follows.

$$F(k_i, c_j) = 2 * \frac{recall(k_i,c_j)*precision(k_i,c_j)}{recall(k_i,c_j)+precision(k_i,c_j)} \quad (13)$$

F MEASURE			
domain	Euclidean (LPI)	Correlation (CPI)	Jaccard (CPI)
Abstract documents	0.298	0.7034	0.7823
News paper articles	0.303	0.6388	0.7468

Table 3: F-measure Results

High value of F measure indicates that both precision and recall is high. In short we can conclude that clustering accuracy is high. From Table 3, it is clear that Jaccard has high value for the F measure.

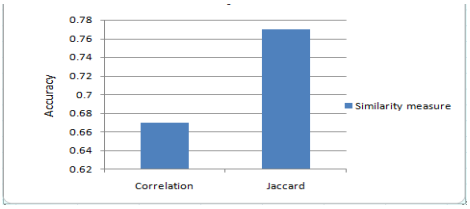

Figure 4:Accuracy graph based on CPI

The accuracy result of the similarity measure comparison is shown in Figure 4.The graph in the Figure 4 represents the values for two distance metric.The results from graph shows that jaccard coefficient is also a good metric for text clustering.

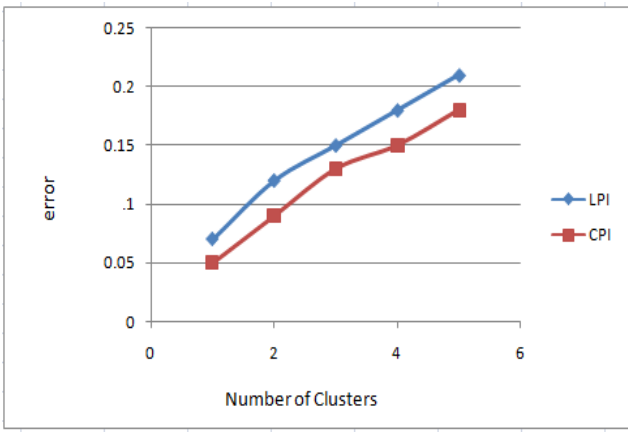

Figure 5:Generalisation capability

Inorder to test the generalisation capabilities of LPI and CPI clutering abilities is tested.From Figure 5 it is clear that CPI has better grneralisation capability.

Next we have to compare the generalisation capability of Correlation and Jaccard coefficient. From Figure 5, CPI with Jaccard coefficientperforms much better than CPI with Correlation similarity measure

Figure 5:Generalisation capability of CPI

9. CONCLUSION AND FUTURE WORK

In this paper we have discussed the problem of clustering documents in Correlation similarity measure space.CPI can detect the intrinsic geometric space of the document, by minimizing the intra-cluster distances and maximizing the inter-cluster distance. In this we have projected the document in low- dimensional semantic subspace where the manifold structure of the original document space is retained. For the purpose of finding a distance metric for clustering, correlation and Jaccard coefficient are taken. From this we can conclude that Jaccard is also an efficient means for doing text clustering. This work can be extended to include other distance measures. While distance measures are often capable of properly describing similarity between objects, in some application areas there is still potential to fine-tune these measures with additional information provided in the data sets. In future we can combine such traditional distance measures for document analysis with link information between documents to improve clustering results.

REFERENCES

[1] D.Cai, X.He, and J.Han, “Document Clustering Using Locality Preserving Indexing”, IEEE Trans.Knowledge and Data Engg. vol.17, no.12, 1624-1637, Dec-2005.

[2] Taiping Zhang,,Yuan Yan Tang,,, Bin Fang,, and G.Wang, “Document Clustering In Correlation Similarity Measure Space,” IEEE Trans.Knowledge and Data Engg, Vol 24, No. 6, June 2012

[3] Y.Fu, S.Yan, and T.S.Huang, “Correlation Metric for Generalized Feature Extraction Pattern Analysis and Machine Intelligence,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol.30, no.12, pp.2229-2235, Dec-2008.

[4] D.R.Hardoon, S.R.Szedmak, and J.R.Shawe-taylor, “Canonical correlation analysis; An overview with application to learning methods,” J.Neural Computation, vol.16, no.12, pp.2639-2664, 2004.

[5] G.Lebanon, “Metric Learning for Text Documents,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol.28, no.4, pp.497-507, Apr-2006.

[6] Y.Ma, S.Lao, E.Takikawa, and M.Kawade, “Discriminant Analysis in Correlation Similarity Measure Space,” Proc.24th Int’l Conf. Machine Learning (ICML ’07) pp.577-584, 2007.

[7] D.Zeimpekis and E.Gallopoulos, “Design of a Mat lab toolbox for term-document matrix generation,”proc. Workshop Clustering HighDimensional Data and its Applications at the FifthSIAM Int’l Conf.data Mining (SDM ’05) pp 38-48, 2005.

[8] S.Zhong and J.Ghosh, “Generative Model-based Document Clustering: A Comparative Study,” Knowledge of Information System, vol .8, no.3, pp.374-384, 2005.

[9] X.Zhu “Semi Supervised learning using Gaussian Fields and Harmonic Functions,” Technical report Computer Sciences Univ. of Wisconsin-Madison, 2005.

[10] M. F. Porter, ″An algorithm for suffix stripping″, Program automated library and information systems, vol. 14(3), pp. 130137,1980.