Semantic Analysis Based Correlation Preserving Index for Document Clustering

Mrs.Sasikala.G[#1], Mr.S.Nandagopal M.E., (Ph.D).,[#2]
[#]*M.E- CSE, Nandha College of Technology, Erode, India.*
[#]*Associate Professor, Nandha College of Technology, Erode, India.*
[#]sasi91kala@gmail.com
[#]asnandu@yahoo.com

Abstract

Clustering is one of the most important techniques in machine learning and data mining tasks. Similar data grouping is performed using clustering techniques. Hierarchical clustering model produces tree structured results. Partitioned clustering produces results in grid format. The documents are projected into a low-dimensional semantic space and then a traditional clustering algorithm is applied to finding document clusters. The euclidean distance is a dissimilarity measure describes the dissimilarities between the documents. Correlation indicates the strength and direction of a linear relationship between two random variables. A scale-invariant association measure is used to calculate the similarity between two vectors. Correlation preserving index (CPI) based clustering is used for document clustering process. The similarity-measure-based CPI method is used for detecting the intrinsic structure between nearby documents. In CPI method the documents are projected into a low-dimensional semantic space. Correlations between the documents in the local patches are maximized. Correlations between the documents outside these patches are minimized simultaneously.

The spectral clustering is applied on the correlation similarity model with nearest neighbor learning process. The Ontology repository is used to manage the term concept relations. Local patch extraction is carried out with Ontology support. Term frequency based weight is replaced with concept weight based model. The document preprocess operations are carried out to extract term information. Stopword elimination and stemming process are applied on the term collection. Porter stemming algorithm is used for suffix analysis. Ontology is used to extract term relationships.

I. INTRODUCTION

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 [3]. Based on various distance measures, a number of methods have been proposed to handle document clustering [10]. 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 centers. Since the document space is always of high dimensionality, it is preferable to find a low-dimensional representation of the documents to reduce computation complexity.

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) 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.

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. Unfortunately, the euclidean distance is a dissimilarity measure which describes the dissimilarities rather than similarities between the documents [1]. 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) method is a different spectral clustering method based on graph partitioning theory. The LPI method applies a weighted function to each pairwise 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.

In recent years, some studies [9], [4] suggest that correlation as a similarity measure can capture the intrinsic structure embedded in high-dimensional data, especially when the input data is sparse. In probability theory and statistics, correlation indicates the strength and direction of a linear relationship between two random variables which reveals the nature of data represented by the classical geometric concept of an "angle." 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.

II. CORRELATION SIMILARITY MEASURE

The usage of correlation as a similarity measure can be found in the canonical correlation analysis (CCA) method [6]. 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 $X = \{x_1, x_2 \ldots, x_n\}$ and $Y = \{y_1, y_2 \ldots, y_n\}$, we would like to find directions w_x for X and w_y for Y that maximize the correlation between the projections of X on w_x and the projections of Y on w_y. This can be expressed as

$$\max_{w_x,\ w_y} \frac{\langle Xw_x,\ Yw_y \rangle}{\|Xw_x\| . \|Yw_y\|} \qquad (1)$$

Where $\langle . , . \rangle$ 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 [5]. 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, the cluster label (Y) is not. So the CCA method cannot be directly used for clustering.

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. This is different from LSI and LPI, which are based on a dissimilarity measure 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. At this point, it is similar to Latent Dirichlet Allocation (LDA) which attempts to capture significant intradocument statistical structure via the mixture distribution model.

III. 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. Mathematically, the correlation between two vectors *u* and *v* is defined as

$$\text{Corr}(u,v) = \frac{u^{\mathrm{T}} v}{\sqrt{u^{\mathrm{T}} u}\sqrt{v^{\mathrm{T}} v}} = \left\langle \frac{u}{\|u\|} , \frac{v}{\|v\|} \right\rangle \qquad (2)$$

Note that the correlation corresponds to an angle θ such that $\cos\theta = \text{Corr}(\mathcal{U},\mathcal{V})$. The larger the value of $\text{Corr}(\mathcal{U},\mathcal{V})$, the stronger the association between the two vectors $\mathcal{U}$ and $\mathcal{V}$.

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.

Based on these assumptions, we can propose a spectral clustering in the correlation similarity measure space through the nearest neighbors graph learning.

A. Correlation-Based Clustering Criterion

Suppose $\mathcal{Y}_i \in Y$ is the low-dimensional representation of the ith document $\mathcal{X}_i$ 2 X in the semantic subspace, where i = 1, 2, . . . , n. Then the above assumptions (A1) and (A2) can be expressed as

$$\max \sum_i \sum_{x_j \in N(x_i)} \mathrm{Corr}(\mathcal{Y}_i, Y_j), \quad (3)$$

$$\text{and } \min \sum_i \sum_{x_j \notin N(x_i)} \mathrm{Corr}(\mathcal{Y}_i, Y_j), \quad (4)$$

respectively, where $N(x_i)$ denotes the set of nearest neighbors of x_i. The optimization of (3) and (4) 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.

The maximization problem (3) 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 (4) is an attempt to ensure that if x_i and x_j are far away, y_i and y_j are also far away. Since the following equality is always true

$$\sum_i \sum_{y_j \in N(y_x)} \mathrm{Corr}(\mathcal{Y}_i, Y_j) + \sum_i \sum_{y_j \notin N(y_x)} \mathrm{Corr}(\mathcal{Y}_i, Y_j) = \sum_i \sum_j \mathrm{Corr}(\mathcal{Y}_i, Y_j), \quad (5)$$

The simultaneous optimization of (3) and (4) can be achieved by maximizing the following objective function

$$\frac{\sum_i \sum_{x_j \in N(x_i)} \mathrm{Corr}(\mathcal{Y}_i, Y_j)}{\sum_i \sum_j \mathrm{Corr}(\mathcal{Y}_i, Y_j)} \quad (6)$$

Without loss of generality, we denote the mapping between the original document space and the low-dimensional semantic subspace by W, that is, $W^T x_i = y_i$. Following some algebraic manipulations, we have

$$\frac{\sum_i \sum_{x_j \in N(x_i)} \mathrm{Corr}(\mathcal{Y}_i, Y_j)}{\sum_i \sum_j \mathrm{Corr}(\mathcal{Y}_i, Y_j)} = \frac{\sum_i \sum_{x_j \in N(x_i)} \frac{y_i^T y_j}{\sqrt{y_i^T y_i y_j^T y_j}}}{\sum_i \sum_j \frac{y_i^T y_j}{\sqrt{y_i^T y_i y_j^T y_j}}}$$

$$= \frac{\sum_i \sum_{x_j \in N(x_i)} \frac{\mathrm{tr}(y_i y_j^T)}{\sqrt{\mathrm{tr}(y_i y_i^T)\mathrm{tr}(y_j y_j^T)}}}{\sum_i \sum_j \frac{\mathrm{tr}(y_i y_j^T)}{\sqrt{\mathrm{tr}(y_i y_i^T)\mathrm{tr}(y_j y_j^T)}}} \quad (7)$$

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

where tr(.)is the trace operator. Based on optimization theory, the maximization of (7) can be written as

$$\arg\max_w \frac{\sum_i \sum_{x_j \in N(x_i)} \mathrm{tr}(W^T x_i x_j^T W)}{\sum_i \sum_j \mathrm{tr}(W^T x_i x_j^T W)} = \arg\max_w \frac{\mathrm{tr}(W^T(\sum_i \sum_{x_j \in N(x_i)}(x_i x_j^T))W)}{\mathrm{tr}(W^T(\sum_i \sum_j (x_i x_j^T))W)}, \quad (8)$$

With the constraints

$\mathrm{Tr}(W^T x_i x_i W) = 1$ for all i = 1,2,……,n. (9)

Consider a mapping $W \in IR^{m*d}$, where m and d are the dimensions of the original document space and the semantic subspace, respectively. We need to solve the following constrained optimization:

$$\mathrm{Argmax} \frac{\sum_{i=1}^d w_i^T M_s w_i}{\sum_{i=1}^d w_i^T M_T w_i} \quad (10)$$

subject to $\sum_{i=1}^d w_i^T x_j x_j^T w_i = 1$, j = 1,2,….,n.(11)

Here, the matrices M_T and M_S[1] are defined as

$$M_T = \sum_i \sum_j (x_i x_j^T)$$

$$M_S = \sum_i \sum_{x_j \in N(x_i)} (x_i x_j^T).$$

It is easy to validate that the matrix MT is semi positive definite. Since the documents are projected in the low dimensional semantic subspace in which the correlations between the document points among the nearest neighbors are preserved, we call this criterion "correlation preserving indexing."

Physically, this model may be interpreted as follows: all documents are projected onto the unit hypersphere. 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, $\propto_j$, are maximized simultaneously, as illustrated in Fig. 1. On the unit hyper sphere, a global angle can be measured by spherical arc, that is, the geodesic distance. The geodesic distance between z and z' on the unit hypersphere can be expressed as

$$dG(z; z') = \arccos(z^T z') = \arccos(Corr(z; z')). \quad (12)$$

Since a strong correlation between z and z' means a small geodesic distance between z and z', then CPI is equivalent to simultaneously minimizing the geodesic distances between the points in the local patches and maximizing the geodesic distances between the points outside these patches. The geodesic distance is superior to traditional euclidean distance in capturing the latent manifold. Based on this conclusion, CPI can effectively capture the intrinsic structures embedded in the high-dimensional document space.

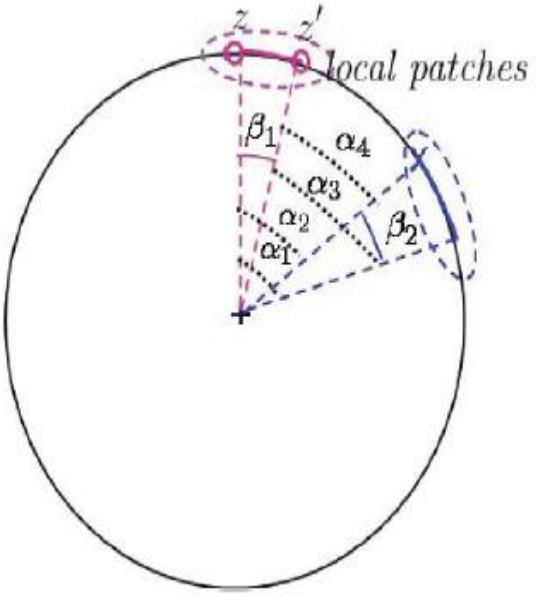

Fig. 1. 2D projections of CPI.

It is worth noting that semi-supervised learning using the nearest neighbors graph approach in the euclidean distance space was originally proposed in the literature [2] and LPI is also based on this idea. Differently, CPI is a semi-supervised learning using nearest neighbors graph approach in the correlation measure space. Zhong and Ghosh showed that euclidean distance is not appropriate for clustering high dimensional normalized data such as text and a better metric for text clustering is the cosine similarity. Lebanon proposed a distance metric for text documents, which was defined as

$$d_{F^*_{\lambda}\mathcal{J}(x,y)} = \arccos = \left(\sum_{i=1}^{n+1} \lambda_i \frac{\sqrt{x_i y_i}}{\langle x, \lambda\rangle\langle y, \lambda\rangle}\right)$$

If we use the notations $\bar{x} = (\sqrt{x_1}, \sqrt{x_2}, \ldots..))$, $\bar{y} = (\sqrt{y_1}, \sqrt{y_2}, \ldots..))$, and set $\lambda_1 = \lambda_2 = \ldots..,= \lambda_n$, then the distance metric $d_{F^*_{\lambda}\mathcal{J}(x,y)}$ reduces to

$$d_{F^*_{\lambda}\mathcal{J}(x,y)} = \arccos = \left(\sum_{i=1}^{n+1} \lambda_i \frac{\sqrt{x_i y_i}}{\langle x, \lambda\rangle\langle y, \lambda\rangle}\right) = \arccos(\bar{x}, \bar{y}).$$

This distance is very similar to the distance defined by (12). Since the distance $d_{F^*_{\lambda}\mathcal{J}(x,y)}$ is local and is defined on the entire embedding space [7], correlation might be a suitable distance measure for capturing the intrinsic structure embedded in document space. That is why the proposed CPI method is expected to outperform the LPI method. Note that the distance $d_{F^*_{\lambda}\mathcal{J}(x,y)}$ can be obtained based on the training data and it can be used for classification rather than clustering.

B. Algorithm Derivation

The optimization problem (10) with the constraints (11) can be solved by maximizing the objective function $\sum_{i=1}^{d} w_i^T M_S w_i$ under the constraints

$$\sum_{i=1}^{d} w_i^T M_T w_i = 1, \quad (13)$$

and

$$\sum_{i=1}^{d} w_i^T x_j x_j^T w_i = 1, \; j = 1,2 \ldots..n. \quad (14)$$

To do this, we introduce an additional Lagrangian function with multipliers λ_i as follows:

$$J_L(W, \Lambda) = \sum_{i=1}^{d} \frac{1}{\lambda} w_i^T M_S w_i - \lambda_i \left(\sum_{i=1}^{d} \frac{1}{\lambda} w_i^T M_S w_i - 1\right) - \lambda_1 \left(\sum_{i=1}^{d} w_i^T x_1 x_i^T w_i - 1\right), \ldots. - \lambda_n \left(\sum_{i=1}^{d} w_i^T x_n x_n^T w_i - 1\right), \quad (15)$$

where $W = [w_1, \ldots, w_d]$ and $\Lambda = [\lambda, \lambda_o \lambda_1, \ldots\ldots, \lambda_n]$. The additional Lagrange multiplier $1/\lambda \neq 0$) is introduced as a multiplicative factor for $\sum_{i=1}^{d} w_i^T M_T w_i$, which does not affect the

solution. The additional Lagrangian function $J_L(W, \Lambda)$ will be maximized by setting the partial derivatives of $J_L(W, \Lambda)$ with respect to w_i to be zero, i.e., $\frac{\partial J_L(W,\Lambda)}{\partial w_i} = 0$. This yields

$$\frac{1}{\lambda} M_S w_i - \lambda_0 M_T w_i - \lambda_1 x_1 x_1^T w_i - \lambda_n x_n x_n^T w_i = 0, \quad (16)$$

or equivalently

$$M_S w_i - \lambda (\lambda_0 M_T + \lambda_1 x_1 x_1^T + \cdots + w_i - \lambda_n x_n x_n^T w_i = 0, \quad (17)$$

Equation (17) means that the w_i, i = 1, 2, . . . , d are the set of generalized eigenvectors of the matrix M_S and the matrix

$$M = \lambda_0 M_T + \lambda_1 x_1 x_1^T + \cdots + \lambda_n x_n x_n^T, \quad (18)$$

Corresponding to the d largest generalized eigenvalues.

In order to find the optimal solution, we first need to fix the value of the Lagrange multipliers λ_i. In (15), we suppose that the function

$$F(W, \Lambda) = \sum_{i=1}^{d} \frac{1}{\lambda} w_i^T M_S w_i,$$

Obtains a relative extremum: $F^* = \sum_{i=1}^{d} w_i^{*T} M_S w_i^*$, together with the optimal values λ_i^* and w_i^* under the constraints

$$\sum_{i=1}^{d} w_i^{*T} M_T w_i^* = b_0$$

and

$$\sum_{i=1}^{d} w_i^{*T} x_j x_j^T w_i^* = b_j.$$

Based on the interpretation of the Lagrange multipliers the values of F*, w_i^*, and λ_i^* depend on the values of b_i on the right-hand sides of the constraint equations. Suppose that λ_i^* and w_i^* are continuously differentiable functions of b_i in some ε-neighborhood of b_i. Then F* is also continuously differentiable with respect to b_i. The partial derivatives of F* with respect to bi are equal to the corresponding Lagrange multipliers λ_i^*, i.e.,

$$\lambda_i^* = \frac{\partial F^*}{\partial b_i}, \quad i = 0, 1, 2, \ldots, n.$$

Let $\Omega = W^* W^{*T}$.

Then, the values of λ_i^* can be computed by

$$\lambda_0^* = \frac{\partial F^*}{\partial b_0} = \frac{\frac{\partial F^*}{\partial \Omega}}{\frac{\partial b_0}{\partial \Omega}} = \frac{tr(M_s)}{tr(M_T)},$$

(19)

$$\lambda_j^* = \frac{\partial F^*}{\partial b_j} = \frac{\frac{\partial F^*}{\partial \Omega}}{\frac{\partial b_j}{\partial \Omega}} = \frac{tr(M_s)}{tr(x_j x_j^T)}, \quad j = 1, 2, \ldots, n. \quad (20)$$

Note that in document clustering, the matrix M in (18) is often singular as the dimension of the documents is generally larger than the number of documents. To circumvent the requirement of M being nonsingular, we may first project the document vectors into the SVD subspace by throwing away the zero singular values.

C. Clustering Algorithm Based on CPI

Given a set of documents $x_1, x_2, \ldots, x_n \in IR^n$. Let X denote the document matrix. The algorithm for document clustering based on CPI can be summarized as follows:

1. Construct the local neighbor patch, and compute the matrices M_S and M_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 $X = U\Sigma V^T$. Here all zero singular values in Σ 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 $X = U^T X$.
3. Compute CPI Projection. Based on the multipliers $\lambda_0, \lambda_1, \ldots\ldots, \lambda_n$ obtained from (19) and (20), one can compute the matrix $M = \lambda_i^* M_T + \lambda_1^* x_1 x_1^T + \ldots\ldots + \lambda_n^* x_n x_n^T$. Let W_{CPI} be the solution of the generalized eigenvalue problem $M_S W = \lambda M W$. Then, the low dimensional representation of the document can be computed by

$$Y = W_{CPI}^T X = W^T X,$$

where $W = UW_{CPI}$ is the transformation matrix.

4. Cluster the documents in the CPI semantic subspace. Since the documents were projected on the unit hypersphere, the inner product is a natural measure of similarity. We seek a partitioning $\{\pi_j\}_{j=1}^{k}$ of the document using the maximization of the following objection function:

$$Q(\{\pi_j\}_{j=1}^{k}) = \sum_{j=1}^{k} \sum_{x \in \pi_j} x^T c_j,$$

with $c_j = \frac{m_j}{\|m_j\|}$, where m_j is the mean of the document vectors contained in the cluster π_j.

IV. DOCUMENT CLUSTERING WITH SEMANTIC CPI

The system is designed to group up relevant documents. Subspace based document clustering mechanism is used in the system. Correlation Preserving Index (CPI) is used in the document clustering process. The system is divided into six major modules. They are Document preprocess, Term analysis, Semantic analysis, Dimensionality reduction process, CPI estimation and Clustering process.

Document parsing, stopword elimination and stemming process are carried out under document preprocess module. Term weight estimated under term analysis module. Semantic analysis is performed to identify concept relationships. Dimensionality reduction process module is designed to reduce document vector size. CPI estimation module is designed to perform correlation similarity and index estimation process. Clustering process module is designed to partition the documents.

A. Document Preprocess

Document preprocess is performed to parse the text documents into words. Document cleaning is applied to remove stop words. Stemming process is applied to detect the base term. Terms are updated with their frequency values.

B. Term Analysis

The term analysis is performed to estimate the term weight values. Statistical method is used for the term weight estimation process. Term frequency (TF) and Inverse Document Frequency (IDF) are used for the term weight estimation process. Terms and associated weight values are updated into the database.

C. Semantic Analysis

The semantic analysis is performed to identify the concept relationships. Ontology is constructed for the selected domains. Terms and associated concept relationships are identified using the Ontology. Semantic weights are assigned with reference to the concept relationship type.

D. Dimensionality Reduction Process

The document vector is constructed with term and weight values. The term weight based document vector is build with high dimensionality. Infrequent term elimination is performed with threshold values. The document vector is updated with reduced term collections.

E. CPI Estimation

The terms and its structure information are used to identify the subspaces. The correlation similarity is estimated with the subspace information. The correlation similarity is estimated between the documents. Correlation Preserving Index (CPI) is prepared with similarity values.

F. Clustering Process

The clustering process is applied to group up the documents using the similarity values. The CPI intervals are analyzed with threshold values. The cluster count is collected from the user. The clustering process is performed with term and semantic weight models.

V. CONCLUSION

The text documents are high dimensional data elements. Term and geometric patch information are used in the similarity analysis. Correlation Preserving Index (CPI) based clustering algorithm is used for the clustering process. Concept relationships are extracted and weight values are used in the clustering process. The dimensionality is reduced with concept relationships. Label based patch extraction model. Patches are represented with patch weight and term weight values. The system produces the clustering results in a hierarchical manner.

REFERENCES

[1] 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.

[2] X. Zhu, "Semi-Supervised Learning Literature Survey," technical report, Computer Sciences, Univ. of Wisconsin-Madison, 2005.

[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] 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.

[5] R.D. Juday, B.V.K. Kumar, and A. Mahalanobis, Correlation Pattern Recognition. Cambridge Univ. Press, 2005.

[6] D.R. Hardoon, 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.

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

[8] 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.

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

[10] 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.