Statistical Measurement Space for Document Clustering Based on Correlation Preserving Indexing

Uppe Nanaji, Majji Nagaraju
dept of CSE
Avanthi Saint theressa Institute of Engineering & Technology , JNTU Kakinada

Abstract--- **This paper presents a spectral clustering method named as correlation preserving indexing (CPI), which is performed in the correlation similarity measure space for sample data in live areas . In this framework, the documents are projected into a low-dimensional semantic space in which the correlations between the documents in the local patches are maximized while the correlations between the documents outside these patches are minimized simultaneously. Finally it Gives the statistical values for the existing with present system. Since the intrinsic geometrical structure of the document space is often embedded in the similarities between the documents, correlation as a similarity measure is more suitable for detecting the intrinsic geometrical structure of the document space than Euclidean distance. Consequently, the proposed CPI method can effectively discover the intrinsic structures embedded in high-dimensional document space. The effectiveness of the new method is demonstrated by extensive experiments conducted on various data sets and by comparison with existing document clustering techniques.**

***Keywords*-----Document clustering, correlation measure, correlation latent semantic indexing dimensionality reduction.**

I.INTRODUCTION

Document clustering aims to automatically group related documents into clusters. It is on of the most important tasks in machine learning and artificial intelligence and has received much attention in recent years [1]-[3]. Based on various distance measures, and a number of methods have been proposed to handle document clustering [4]-[10]. A typical and widely used distance measure is the Euclidean distance. The k-means method [4] 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 dimensionality, it is preferable to find a low-dimensional representation of the documents to reduce the 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) [7] 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).

However, because of the high dimensionality of the document space, a certain representation of documents usually reside on a nonlinear manifold embedded in the between the documents. Thus, it is not able to effectively capture the nonlinear manifold structure embedded in the similarities between them [12]. 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 [8]. 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.

In recent years ,some studies suggest that correlation as a similarity measure can capture the intrinsic structure embedded in high-dimensionality data, especially when the input data is sparse. In probability theory statistics , correlation indicates the strength and direction of liner relationship between two random variable which revels the nature of data 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 this paper, we propose a new document clustering methods based on correlation preserving indexing(CPI), which explicitly consider the manifold structure embedded in the similarities between the documents. It aims to find an optimal semantic subspace by simultaneously maximizing the correlation between the documents in the local patches and minimizing the correlations between the documents outside these patches. which is implemented by Reuters sample data and presented the correlation measurements and comparisons between the CPI and K-means.

In recent years, some studies [13]suggest that correlation as a similarity measure space can capture the intrinsic structure embedded in high-dimensional data, especially when the input data is sparse[19]-[20]. 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. The usage of correlation as a similarity measure can be found in the canonical correlation analysis (CCA)method [21]. 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 = $\{x_1, x_2, ..., x_n\}$ and Y = $\{y_1, y_2, ..., y_n\}$ 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

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

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 [20]-[21]. 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. In this paper, we propose a document metric that is correlation preserving indexing (CPI).

II.DOCUMENT 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 documents 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-dimensionality document space.

Mathematically, the correlation between vector U and V is defined as

$$corr(u,v) = \frac{u^T v}{\sqrt{u^T u}\sqrt{v^T v}} = \left\langle \frac{u}{\|u\|}, \frac{v}{\|v\|} \right\rangle \quad (2)$$

Note that the correlation corresponds to an angle θ such that $\cos\theta = corr(u,v)$. The larger the value of $corr(u,v)$ is the stronger the association between the two vectors u and v .

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

1)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 dimensionality, it is preferable to find a low dimensional representation of the documents to reduce computation complexity.

2)*Correlation Based Clustering Criteria:* 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 the above assumptions A1) and A2) can be expressed as

$$\max \sum_i \sum_{x_j \in N(x_i)} corr(y_i, y_j) \quad (3)$$

$$\min \sum_i \sum_{x_j \notin N(x_i)} corr(y_i, y_j) \quad (4)$$

respectively, where $N(x_i)$ denotes the set of nearest neighbours 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 neighbours of each other.

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 neighbours, β_i, are minimized and the global angles between the points outside the local neighbours, α_i , 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 hyper-sphere can be expressed as

$$d_G(Z,Z^{'}) = \arccos(Z^T Z^{'}) = \arccos(corr(Z,Z^{'})) \qquad (5)$$

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[14].Based on this conclusion, CPI can effectively capture the intrinsic structures embedded in the high-dimensional document space.

The Proposed system shows 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 [12]. Lebanon in [21] proposed a distance metric for text documents, which was defined as:

$$d_{F_\lambda^* \Im(x,y)} = \arccos\left(\sum_{i=1}^{n+1} \lambda_i \frac{\left(\sqrt{(x_i y_i)}\right)}{\langle x,\lambda\rangle\langle y,\lambda\rangle} \right) \qquad (6)$$

This distance is very similar to the distance defined by (5). Since the distance is local(thus it captures local variations within the space) and is defined on the entire embedding space [21], 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 can be obtained based on the training data and it can be used for classification rather than clustering.

If we use the notations $\bar{x} = \left(\sqrt{\left(x_1, \sqrt{x_2},....\right)}\right), \bar{y} = \left(\sqrt{\left(y_1, \sqrt{\left(y_2,....\right)}\right)}\right)$ and set $\lambda_1 = \lambda_2 = ... = \lambda_n$, then the distance metric $d_{F_\lambda^* \Im(x,y)}$ reduced to

$$d_{F_\lambda^* \Im(x,y)} = \arccos\left(\sum_{i=1}^{n+1} \frac{x_i y_i}{\langle \overline{x}\overline{x}\rangle\langle \overline{y}\overline{y}\rangle} \right) = \arccos\left(Corr\left(\bar{x}, \bar{y}\right)\right). \qquad (7)$$

This distance is very similar to the distance defined by (5).

Since the distance $d_{F_\lambda^* \Im(x,y)}$ is local (thus it captures local variations within the space) and is defined on the entire embedding space [21], 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 $d_{F_\lambda^* \Im(x,y)}$ can be obtained based on the training data and it can be used for classification rather than clustering.

III.RELATED WORK:

1)*Clustering Algorithm Based on CPI:*

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

a)Construct the local neighbour patch, and compute the matrices M^S and M^T .

b)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 $X = U^T X$.

c) Compute CPI Projection. Based on the multipliers $\lambda_0, \lambda_1, ...\lambda_n$ obtained from 19 and 20 , one can computer the matrix $M = \lambda_0^* M_T + \lambda_1^* x_1 x_1^T + ... + \lambda_n^* x_n x_n^T$. 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 $Y = W_{CPI}^T \tilde{x} = W^T X$ where $W = UW_{CPI}$ is the transaction matrix.

d)Cluster the documents in the CPI semantic subspace. Since the documents were projected on the unit hyper-sphere, the inner product is a natural measure of similarity. we seek a partitioning $\{\pi_j\}_{j=1}^k = \sum_{j=1}^k \sum_{x \in \pi_j} x^T c_j$ 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.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 a, we need to compute the pair wise distance which needs O(n^2d) operations. Secondly, we need to find the k nearest neighbours for each data point which needs O(kn^2) operations. Thirdly, 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^2d + kn^2 + n(n − k)d). In step 'b' 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 'b' costs O(d^3+n^{2}d).Then, transforming the documents into m-dimensions semantic subspace requires O(mn^2) operations In step 'c', 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, the step 'b' will dominate the computation. To reduce the computation cost of step 'b', one can apply the iterative SVD algorithm [18] rather than matrix decomposition algorithm or feature selection method to first reduce the dimension.

V. DOCUMENT REPRESENTATION:

In all experiments, 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: $\left({}^{tf}\!/_{idf}\right)_{i,j} = tf_{i,j} \times idf_i$, Here $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ 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 $d_j.idf_i = \log\left(\frac{|d|}{d : t_i \in d}\right)$ 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 $v = \{t_1, t_2 ..., t_m\}$ 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:

$$X_j = \left[x_{1j}, x_{2j} .....x_{mj}\right],$$
$$X_{ij} = (tf / idf)_{i,j}$$

Using n documents from the corpus, we construct an m× n term-document matrix X. The above process can be completed by using the text to matrix generator (TMG) code.

VI.CLUSTERING RESULTS:

1)*Correlation Measurement Values Generated:* Experiments were performed on Reuters, and OHSUMED data sets. . We compared the proposed algorithm with other competing algorithms under same experimental setting. In all experiments, our algorithm performs better than or competitively with other algorithms.

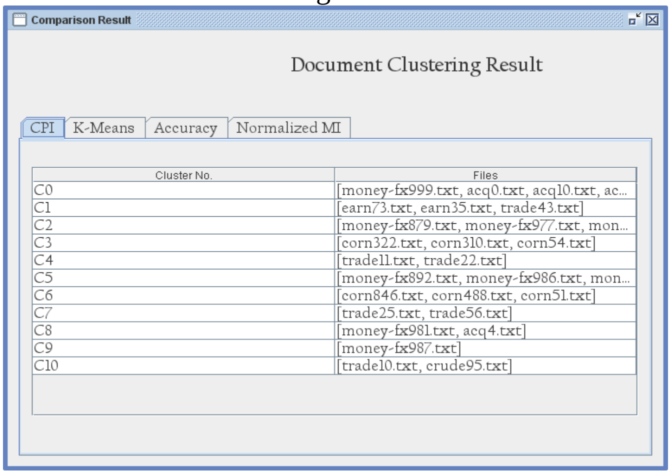
cor 0 : 1 :0.04395113068664207 : 0.04395113068664207

:

:

cor 26 : 27 : 0.051122792602862815 : 0.051122792602862815

TABLE 1
Document clustering Results Based on CPI

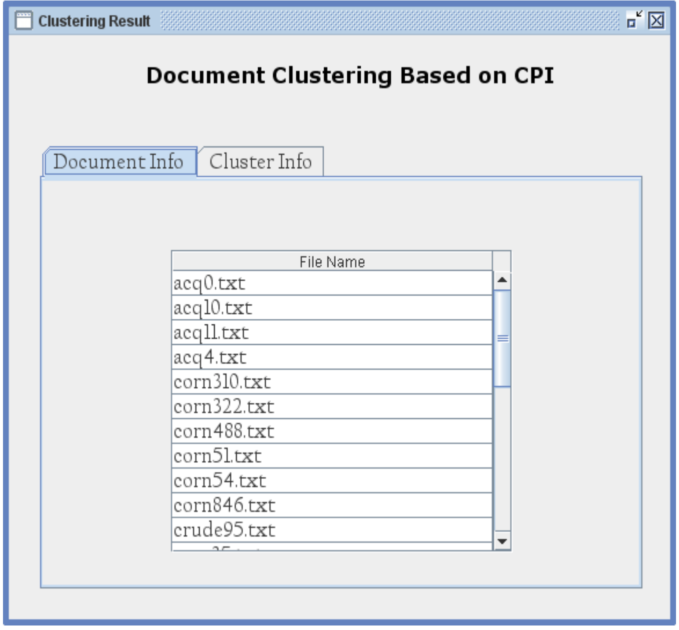


Comparison Result

Document Clustering Result

Cluster No.	Files
C0	[money-fx999.txt, acq0.txt, acq10.txt, ac...
C1	[earn73.txt, earn35.txt, trade43.txt]
C2	[money-fx879.txt, money-fx977.txt, mon...
C3	[corn322.txt, corn310.txt, corn54.txt]
C4	[trade11.txt, trade22.txt]
C5	[money-fx892.txt, money-fx986.txt, mon...
C6	[corn846.txt, corn488.txt, corn51.txt]
C7	[trade25.txt, trade56.txt]
C8	[money-fx981.txt, acq4.txt]
C9	[money-fx987.txt]
C10	[trade10.txt, crude95.txt]

TABLE 2
Reauters Documents sets from different Data Sets i.e sample data of Reauter Data.



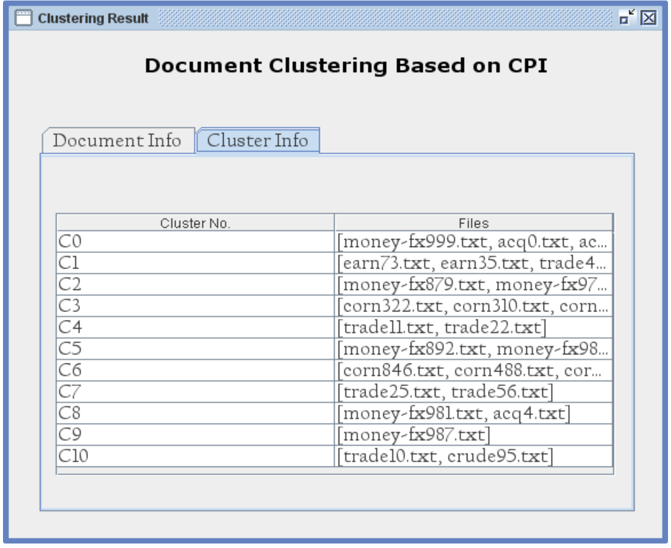
Clustering Result

Document Clustering Based on CPI

File Name
acq0.txt
acq10.txt
acq11.txt
acq4.txt
corn310.txt
corn322.txt
corn488.txt
corn51.txt
corn54.txt
corn846.txt
crude95.txt

TABLE 3

Final Document Clusters based on CPI each cluster has some similar Group of Datasets

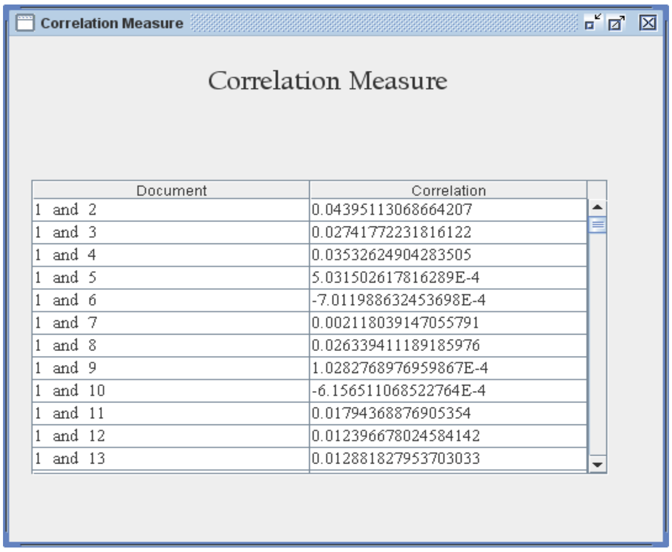


Document Clustering Based on CPI

Cluster No.	Files
C0	[money-fx999.txt, acq0.txt, ac...
C1	[earn73.txt, earn35.txt, trade4...
C2	[money-fx879.txt, money-fx97...
C3	[corn322.txt, corn310.txt, corn...
C4	[trade11.txt, trade22.txt]
C5	[money-fx892.txt, money-fx98...
C6	[corn846.txt, corn488.txt, cor...
C7	[trade25.txt, trade56.txt]
C8	[money-fx981.txt, acq4.txt]
C9	[money-fx987.txt]
C10	[trade10.txt, crude95.txt]

TABLE 4

Statistical Measurements of correlation measuring values between the documents which is belongs to Reauters Data set.



Correlation Measure

Document	Correlation
1 and 2	0.04395113068664207
1 and 3	0.02741772231816122
1 and 4	0.03532624904283505
1 and 5	5.031502617816289E-4
1 and 6	-7.011988632453698E-4
1 and 7	0.002118039147055791
1 and 8	0.026339411189185976
1 and 9	1.0282768976959867E-4
1 and 10	-6.156511068522764E-4
1 and 11	0.01794368876905354
1 and 12	0.012396678024584142
1 and 13	0.012881827953703033

VII.CONCLUSTION

We present a new document clustering method based on correlation preserving indexing. It simultaneously maximizes the correlation between the documents in the local patches and minimizes the correlation between the documents outside these patches. Consequently, a low-dimensional semantic subspace is derived where the documents corresponding to the same semantics are close to each other. Extensive experiments on NG20, Reuters and OHSUMED corpora show that the proposed CPI method outperforms other classical clustering methods. Furthermore, the CPI method has good capturing and capability of assigning clustering formation and thus it can effectively deal with data with very large size.

REFERENCES:

[1] R. T. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," in Proceedings of the 20th VLDB Conference, Santiago. New York, NY, USA: ACM, 1994, pp. 144-155

[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 Transactions on Information Science and Applications, vol. 1, no. 1, pp. 73-81, 2004.

[4]J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, 1967,pp 281-297.

[5]L. D. Baker and A. K. McCallum, "Distributional clustering ofwords for text classification," in Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, Melbourne, AU:. New York, NY, USA: ACM, 1998, pp. 96-103.

[6]X. Liu, Y. Gong, W. Xu and S. Zhu, "Document clustering withcluster refinement and model selection capabilities," in SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 2002, pp. 191-198.

[7]S. C. Deerwester, S. T. Dumais. T. K. Landauer, G. W. Furnas, andR. A. Harshman, "Indexing by latent semantic analysis," Journal of the American Society of Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[8]D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," IEEE Trans. on Knowledge and Data Engi-neering, vol. 17, no. 12, pp. 1624-1637, 2005.

[9]W. Xu, X. Liu, and Y. Gong, "Document clustering based onnon-negative matrix factorization," in SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. New York, NY, USA: ACM, 2003, pp. 267-273.

[10]S. Zhong and J. Ghosh, "Generative model-based document clustering: a comparative study," Knowl. Inf. Syst., vol. 8, no. 3, pp.374-384, 2005.

[11]D. K. Agrafiotis and H. Xu, "A self-organizing principle for learning nonlinear manifolds," Proceedings of the National Academyof Sciences of the United States of America, vol. 99, no. 25, pp. 15 869-15872, 2002.

[12]S. Zhong and J. Ghosh, "Scalable, balanced model-based clustering," in Proc. 3rd SIAM Int. Conf. Data Mining. New York, NY,USA: ACM, 2003, pp. 71-82.

[13]Y. Fu, S. Yan, and T. S. Huang, "Correlation metric for generalized feature extraction," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 12, pp. 2229-2235, Dec. 2008.

[14]J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction." Science, vol. 290, no. 5500, pp. 2319-2323, December 2000.

[15]X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in ICML-03, 20th International Conference on Machine Learning, 2003.

[16]X. Zhu, "Semi-supervised learning literature survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep., 2005.

[17]G. Lebanon, "Metric learning for text documents," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 28, no. 4, pp. 497-507, 2006.

[18]P. Strobach, "Bi-iteration svd subspace tracking algorithms," IEEE Transactions on Signal Processing, vol. 45, no. 5, pp. 1222 -1240, may 1997.

[19]Y. Ma, S. Lao, E. Takikawa, and M. Kawade, "Discriminant analysis in correlation similarity measure space," in ICML '07: Proceedings of the 24th international conference on Machine learning, 2007, pp. 577-584.

[20]R. D. Juday, B. V. K. Kumar, A. Mahalanobis, Correlation Pattern Recognition. Cambridge University Press, 2005.

[21]D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-taylor, "Canonicalcorrelation analysis: An overview with application to learning methods," Neural Comput., vol. 16, no. 12, pp. 2639-2664, 2004.