Volume 4, Issue 2, February 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Document Clustering Based on Correlation Preserving Indexing

N.Vijayakumar, M.Sindhu, S.Venkatesh

Asst.Prof, Department of CSE,
Pollachi Institute of
Engineering and Technology,
Pollachi, Tamil Nadu, INDIA

Abstract - ***Document clustering is used to group the documents into clusters. Document clustering is a more specific technique for unsupervised document organization, automatic topic extraction, fast information retrieval and filtering. Two documents close to each other tend to be grouped into the same cluster and two documents far away from each other tend to be grouped into different clusters. The spectral clustering method is used to cluster the documents. The correlation preserve indexing algorithm proposed to cluster the documents. The correlation preserve indexing algorithm projects the documents in low dimensional semantic space and uses similarity measure to find correlation between the documents. The correlations between the documents in local patches are maximized and outside the patches are minimized. The proposed correlation preserve indexing algorithm can effectively discover the intrinsic structures embedded in high dimensional document space.***

Keywords— ***Correlation measure, document clustering, dimensionality reduction, spectral clustering***

I. INTRODUCTION

Document clustering [1] is a fundamental operation used in unsupervised document organization and information retrieval. Document clustering is to group automatically related documents into clusters. It is one of the most significant tasks in machine learning and artificial intelligence and has received much attention in recent years [2, 3, and 4]. Based on a variety of distance measures, a number of methods have been proposed to handle document clustering. A distinctive and widely used distance measure is the euclidean distance. The k-means method is one of the ways that use the euclidean distance, it minimizes the sum of the squared euclidean distance between the data points and their corresponding cluster centers. The document space is of high dimensionality forever, it is preferable to find a low-dimensional representation of the documents to reduce computation complexity. Low computation cost is attained 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)[8] is one of the effective spectral clustering methods, intended at finding the best subspace approximation to the original document space by minimizing the global reconstruction error (euclidean distance). It 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. It is not bright 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 speculation. The lpi method applies a weighted function to each pair wise distance attempting to focus on capturing the similarity structure, instead of dissimilarity structure, of the documents. It does not overcome the essential limitation of Euclidean distance. Moreover, 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 the input data is sparse. It is a scale-invariant association measure usually used to calculate the similarity between two vectors. In a lot of cases, correlation can effectively represent the distributional structure of the input data to conventional euclidean distance cannot explain. The usage of correlation as a similarity measure can be found in the canonical correlation analysis (cca) method. The cca method is to find projections for paired data sets such that the correlations between their low-dimensional representatives in the projected spaces are mutually maximized. Specifically, given a paired data set consisting of matrices in the equation (1).

$$x = \{ x_1, x_2, \ldots, x_n \} \text{ and } y = \{ y_1, y_2, \ldots, y_n \} \qquad (1)$$

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 shown in the equation (2) anywhere $\langle . , . \rangle$ and || . || denote the operators of inner product and set, in the order.

$$\max_{w_x w_y} \frac{\langle X_{w_x}, Y_{w_y} \rangle}{\|X_{w_x}\| \cdot \|Y_{w_y}\|} \qquad (2)$$

As a great statistical technique, the cca method has been applied in the field of pattern recognition and machine 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. Learning. To propose a new document clustering method based on correlation preserving indexing (cpi), it clearly considers the manifold structure embedded in the similarities between the documents. It aims to locate an optimal semantic subspace by simultaneously maximizing the correlations between the documents in the local patches and minimizing the correlations between the documents outer these patches. This is different from lsi and lpi, that are based on a dissimilarity measure (euclidean distance), and are focused n detect 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. As 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.

In multivariate statistics and the clustering of data, spectral clustering techniques make use of the spectrum (eigen values) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. The similarity matrix is provided as an input and consists of a quantitative assessment of the relative similarity of each pair of points in the dataset. An efficiency improvement of spectral clustering is the spectral neighborhood (span) algorithm, it performs spectral clustering without explicitly computing the similarity matrix, and therefore dramatically improves the scalability of the standard spectral clustering algorithm.

II. RELATED WORKS

In Document Clustering [1] in Correlation Similarity Measure Space proposed by Taiping Zhang[15] 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 that the documents are first projected into a low dimensional semantic space and then a traditional clustering algorithm is applied to finding document clusters. Because of the high dimensionality of the document space, a certain representation of documents usually reside on a nonlinear manifold embedded in the similarities between the data points .Unfortunately, the Euclidean distance is a dissimilarity measure and describes the dissimilarities rather than similarities between the documents. Thus, it is not able to efficiently capture the nonlinear manifold structure embedded in the similarities between them. So a new document clustering method based on correlation preserving indexing (CPI), it clearly considers the manifold structure embedded in the similarities between the documents. It aims to locate 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. The traditional vector space information retrieval model in document clustering use words as measure to find similarity between documents [14]. In reality the concepts, semantics, and topics are used to describe the documents. VSM ignores semantic relations in the middle of 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 our explore. The semantic relationships between documents are not explored in the most of the clustering methods.

In Document Clustering Using Locality Preserving Indexing [11] the document space is of high dimensionality, in broad ranging from several thousands to tens of thousands. Learning in a high-dimensional space is extremely difficult due to the curse of dimensionality. Consequently, document clustering necessitates some form of dimensionality reduction. One of the basic assumptions at the back data clustering is that, the two data points are close to each other in the high dimensional space, they be inclined to be grouped into the same cluster. The optimal document indexing method should be able to discover the local geometrical structure of the document space. To this last part, the LPI algorithm is of particular attention. LSI is optimal in the sense of renovation. It respects the global Euclidean structure while failing to discover the intrinsic geometrical structure, especially the document space is nonlinear. Another consideration is due to the discerning power. One can expect that the documents should be projected into the subspace in which the documents with different semantics can be well alienated, while the documents with common semantics can be clustered. As indicated LPI is an optimal unsupervised approximation to the Linear Discriminate Analysis algorithm is supervised. so, LPI can have additional discriminant power than LSI. There are a quantity of other linear subspace learning algorithms, such as informed projection and Linear Dependent Dimensionality 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, Yihong Gong Clusters each of which 0corresponds to a coherent topic. 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, 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. Because it is more normal 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. Furthermore, it is also quite common that the topics comprising a document corpus are not completely independent of each other, and there are some overlaps among them. In such a case, the axes of the semantic space that capture each of the topics are not necessarily orthogonal.

III. PROPOSED WORK

In high-dimensional document space, the semantic structure is typically implicit. It is pleasing to find a low dimensional semantic subspace in which the semantic structure can become clear. consequently, discovering the intrinsic structure of the document space is often a primary concern of document clustering. Since the manifold structure is frequently embedded in the similarities between the documents, correlation as a similarity measure is appropriate for capturing the manifold structure embedded in the high dimensional document space. Mathematically, the correlation between two vectors (column vectors) u and v is defined in the equation (3).

$$\mathrm{Corr}(u,v) = \frac{u^T v}{\sqrt{u^T u}\sqrt{v^T v}} = \langle \frac{u}{\|u\|}, \frac{v}{\|v\|} \rangle \qquad (3)$$

The correlation corresponds to an angle θ such that
$\cos\theta = \mathrm{Corr}(u,v)$.

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

Online document clustering aims to group documents into clusters, it belongs unsupervised learning. though, it can be transformed into semi-supervised learning by using the following side information:
1) If two documents are close to each other in the original document space, then they tend to be grouped into the same cluster [8].
2) If two documents are far away from each other in the original document space, they tend to be grouped into different clusters.

a.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. It need to eliminate those stop words for finding such similarity between documents.

Stemming is the processes for reducing derived words to their stem, base or origin form 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 the variant forms of a word are reduced to a frequent form, the following all the 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 are shown in the fig. (1)

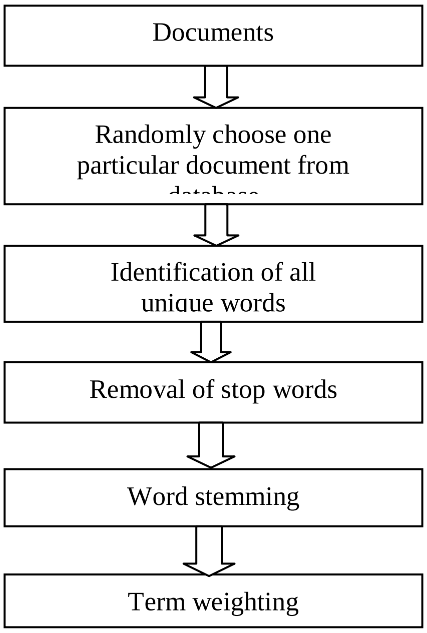
Documents
Randomly choose one particular document from database
Identification of all unique words
Removal of stop words
Word stemming
Term weighting

Fig. 1 Document preprocessing

b.Term Frequency and Inverse Document Frequency

Every 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 following 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 in the equation (4)

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

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

Here is the term frequency of the term t_i in document d_j, and t_i is the number of occurrences of the considered term as mentioned in the equation (5)

$$d_j.idf_i = \log\left(\frac{|D|}{|d:t_j \epsilon d|}\right) \qquad (5)$$

The inverse document frequency is a measure of the general importance of the term t_i, and $|D|$ is the total number of documents in the corpus and $|\{d : t_i \epsilon d\}|$ is the number of documents in the term t_i appears. Let $V = \{t_1, t_2 \dots t_m\}$ 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 in the equation (6)

$$X_j = [x_{1j}, x_{2j}, \dots, x_{mj}],$$
$$x_{ij} = (tf/idf)_{i,j}. \qquad (6)$$

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

Clustering Algorithm Based on CPI

1. Input text document from dataset.
2. Transform the documents to a list of terms after words stemming operations.
3. Construct the local neighbor patch, and compute the matrices M_S and M_T.
4. Compute CPI projection based on the multipliers
   tf *idf.
5. Compute CPI Projection. Based on the multipliers λ_0, λ_1 . . . λ_n Obtained one can compute the matrix M=$\lambda_0^* M_T + \lambda_1^* M x_1 x_1^T + \dots + \lambda_n^* x_n x_n^T$. Let w_{CPI} be the solution of the generalized eigen value problem M_SW=λMW. Then, the low dimensional representation of the document can be computed by Y=W_{CPI}^T $X^{\sim} = W^T X$ is the transformation matrix.
6. Finding correlation between the numbers of documents.
7. Finding distance between the numbers of documents.

Cluster the documents in the CPI semantic subspace. The Documents are projected on the unit hyper sphere, the inner product is a natural measure of similarity.

To seek a partitioning $\{\pi_j\}_{j=1}^k$ of the document using the maximization of the following objection function.

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

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

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 steps a, need to compute the pair wise distance which needs $O(n^2 d)$ operations. Secondly, we need to find the k nearest neighbors for each data point that needs $O(kn^2)$ operations. Thirdly, computing the matrices M_s and M_T requires $O(n^2 d)$ operations and $O(n(n-k)d)$ operations, respectively. Thus, the computation cost in step 1 is $O(2n^2 d + k n^2 + n(n-k)d)$.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, 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.

IV. DATASET DESCRIPTION

The 20 newsgroups corpus3 consists of roughly 20,000 documents that come from 20 specific Usenet newsgroups. We repeated the experiment [5] to illustrate the performance of the proposed CPI algorithm and other competing algorithms. The first set of experiments involved binary clustering. In each experiment, we randomly chose 50 documents from the two selected newsgroups and 100 runs were conducted for each algorithm to obtain statistically reliable clustering result.

The means and standard deviations of the test results were recorded. We also tested other competing methods under same experimental setting, including K-means, pK-means, p-QR and Spectral . It can be seen from Table that CPI achieves the best clustering accuracy on all six data sets. LPI performs the second best, pK-means and p-QR outperform K-means, and K-means performs the worst. Under Normalized mutual information metric, CPI also performs the best. K-means performs better than p-K-means and p-QR, and pK-means or p-QR performs the worst. It can see that the CPI method outperforms with statistical significance other competing methods in most of the data sets.

Reuter

The Reuters corpus4 contains 21,578 documents in 135 topics. Many documents have multiple category labels. A subset of Reuters contained the total 8,067 documents in 30 categories with unique category labels is used in this experiment. The proposed method was compared with five methods, including .Kmeans on original data used with cosine similarity measure . Kmeans with cosine similarity measure after LSI . Kmeans with cosine similarity measure after LPI . von Mises-Fisher model (vMF) nonnegative matrix factorization (NMF) . The experiments were performed with the number of clusters ranging from 2 to 8. For each given c that is the number of clusters, 50 document sets with different clusters were randomly selected from the corpus. Since all the tested algorithms depend heavily on the initial partition, we performed 100 runs for each set of documents.

RESULTS

The proposed technique is done using correlation preserving indexing, this technique improves the document clustering accuracy based on similarity between the documents. Experiments were performed on NG20, 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.

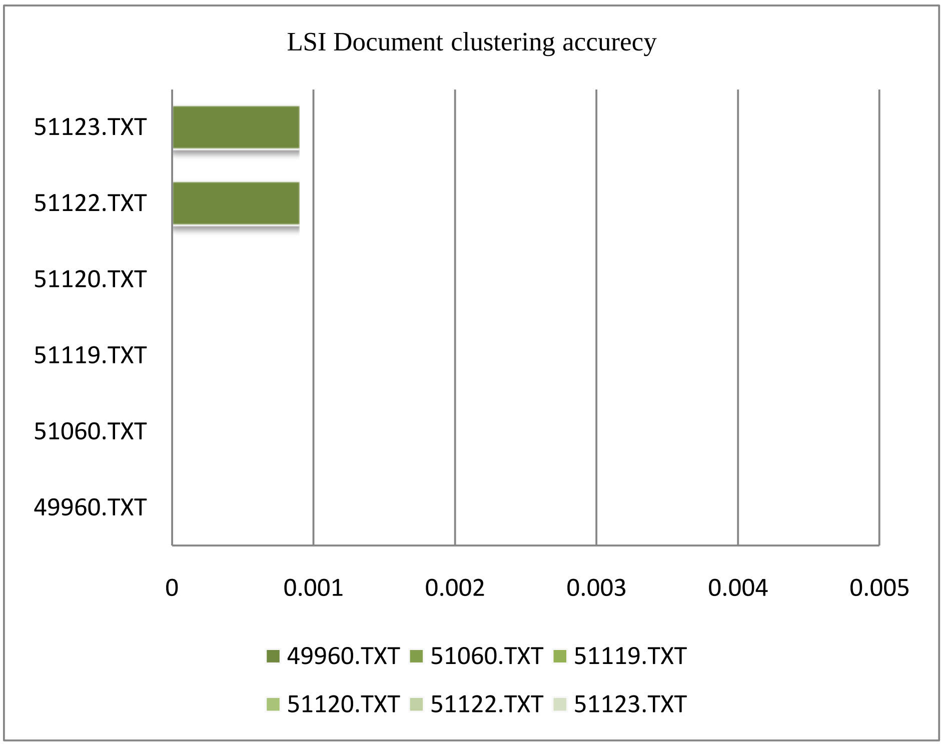

Fig 2. Lenten semantic indexing

The Figure 2 shows the Latent semantic analysis algorithm production. It process the number of text documents to cluster documents and produced the document clustering accuracy with minimum distance.

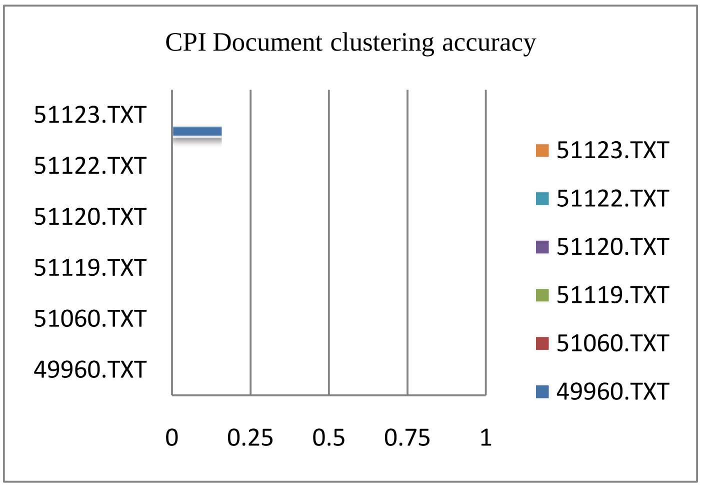

Fig 3. Correlation preserving indexing

The Figure 3 shows the Correlation preserving indexing algorithm production. It process the number of text documents to cluster documents and produced the document clustering accuracy with minimum distance. While comparing with previous algorithm the LPI produce more than accuracy.

CONCLUSION

In this paper, 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, show that the proposed CPI method outperforms other classical clustering methods.

REFERENCES

[1] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman, “Indexing by Latent Semantic Analysis” J. Am. Soc. Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[2] 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.

[3] 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.

[4] L.D. Baker and A.K. McCallum, “Distributional Clustering of Words for Text Classification”, Proc. 21st Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 96-103, 1998.

[5] 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.

[6] M. Belkin and P. Niyogi, “Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering”, Advances in Neural Information Processing Systems 14, pp. 585-591, Cambridge, Mass.:MIT Press, 2001.

[7] I.S. Dhillon and D.M. Modha, “Concept Decompositions for Large Sparse Text Data Using Clustering”, Machine Learning, vol. 42, no. 1,pp. 143-175, 2001.

[8] 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), pp. 191-198, 2002.

[9] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet Allocation”,J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[10] 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.

[11] D. Cai, X. He, and J. Han, “Document Clustering Using Locality Preserving Indexing”, IEEE Trans. Knowledge and Data Eng., vol. 17, no. 12, pp. 1624-163 7, Dec. 2005.

[12] D. Cheng, R. Kannan, S. Vempala, and G. Wang, “A Divide-and- Merge Methodology for Clustering” ACM Trans. Database Systems, vol. 31, no. 4, pp. 1499-1525, 2006.

[13] H. Chim and X. Deng, “Efficient phrase-based document similarity for clustering”, IEEE Trans. On Knowl. And Data Eng., Vol. 20,no. 9, pp. 1217-1229, 2008.

[14] P. Lakkaraju, S. Gauch, and M. Speretta, “Document similarity based on concept tree distance”, in proc. Of the 19th ACM conf. on Hypertext and hypermedia, 2008, pp. 127-132.

[15] 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.