Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 8, August 2013, pg.237 – 242

RESEARCH ARTICLE

Hierarchical Document Clustering Using Correlation Preserving Indexing

L. Prabhakar[1], Srikanth. B[2]

[1]M.Tech 2nd year, Dept. of CSE, ASCET, Gudur, India

[2]Associate Professor, Dept. of CSE, ASCET, Gudur, India

[1] lakkuprabhakar09@gmail.com; [2] BSrikanth@yahoo.com

Abstract— This paper presents a spectral clustering method called as correlation preserving indexing (CPI). This method is performed in the correlation similarity measure space. Correlation preserving indexing explicitly considers the manifold structure embedded in the similarities between the documents. The aim of CPI method is to find an optimal semantic subspace by maximizing the correlation between the documents in the local patches and simultaneously correlation in the patches outside are minimized. Correlation is a similarity measure can capture the intrinsic structure in high dimensional data. In an effort to reduce the computational cost of CPI method, we propose to apply the bi-iterative least square method to reduce the dimensions. On comparison of the effectiveness of the CPI method with other clustering methods, using the Bi-iterative least square method there has been a considerable reduction in the time computation.

Key Terms: - Document clustering; Correlation preserving indexing; Singular value decomposition; Dimensionality reduction; Correlation measure; QR decomposition

I. INTRODUCTION

Document clustering is one of the most important tasks in machine learning and artificial intelligence. Document clustering is the act of collecting similar documents into clusters, where similarity is some function on a document. To handle the document clustering various distance measures have been proposed based on various distance measures [1], [2], [3]. A typical and widely used distance measure is Euclidean distance. In this, the ordinary distance between two points or objects or items is called Euclidean distance. The unsupervised learning can be transformed into semi-supervised learning using these two assumptions that are, if two documents are close to each other in the original document space they can be grouped into same cluster. The other assumption is, if two documents are far away from each other in the original document space then they can be grouped into different clusters. Based on these assumptions we can propose a new spectral clustering in the correlation similarity measure space through the nearest neighbors graph learning.

Many clustering Methods have been proposed. The k-means method [4] is one of the methods that use the Euclidean distance between the data points and their corresponding centers. One of the key limitation is this method is that you must specify the number of clusters as input to the algorithm. In spectral clustering methods it lowers the dimensional subspace and lowers the computation cost. Latent semantic indexing (LSI) [5] is one of the effective spectral clustering methods, aimed at finding best subspace approximation to the original document space by minimizing the Euclidean distance. Euclidean distance is a dissimilarity measure which describes the dissimilarities rather than similarities between the documents. Locality preserving indexing (LPI) method is a different spectral clustering method based on graph partitioning theory. LPI is an optimal

unsupervised approximation to the Linear Discriminant Analysis algorithm which is supervised. LPI can have more discriminant power than LSI. Correlation as a similarity measure can capture the intrinsic structure embedded in high dimensional data, especially when input data is sparse [6], [7]. Correlation indicates the strength and direction of a linear relationship between two random variables. It reveals the nature of data represented by the classical geometric concept of an "angle".

Correlation as a similarity measure can be found in the canonical correlation analysis (CCA) method [8]. The CCA method is used to find projections for paired data sets such that the correlations between their low dimensional representatives in the projected spaces are maximized. Here the CCA method cannot be directly used for clustering [9]. Mathematically, the correlation between two vectors u and v is defined as [10].

$$\text{Corr}(u,v)= \frac{u^T v}{\sqrt{u^T u}\sqrt{v^T v}}$$

II. RELATED WORK

A. *Document Preprocessing*

Document Preprocessing is the process to identification of unique words, removal of stop words and stemming. In document clustering with similarity measure is necessary to identification of unique words in the document. The removal of stop words is the most common term filtering technique used. There are standard stop word lists available but in most of the applications. These are modified depending on the quality of the dataset.

Stemming is the process of reducing words to their base, stem or root form. For example 'receive', 'receiving', 'received' are all forms of the same word used in the different constraints but measuring similarity these should be considered as same. Term weighting is useful to provide an information retrieval and text categorization. Conceptually related documents are grouped in document clustering.

B. *Document representation*

In this document clustering each document is represented as a term frequency vector. That can be computed as follows.

1. After word stemming operation transform the documents to a list of terms.
2. Remove the stop words. Stop words are the common words, which contain no semantic content.
3. Compute the term frequency vector using the TF / IDF weighting scheme. We now combine the definitions of term frequency and inverse document frequency to produce a composite weight for each term in each document. The TF / IDF weighting scheme assigns to term t and weight in document d given by

$$tf - id\hat{f}_{t,d}= tf_{td} \times idf_t$$

Using n documents we construct an m×n term document matrix X. This process can be done by using the text to matrix generator (TMG) [11]. Text to Term Matrix Generator is used to perform the preprocessing and filtering steps that are typically performed in the context of Information Retrieval applications.

C. *Clustering algorithm*

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

1. First construct the local neighbor patch, and compute the matrices M_I 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$\sum V^T$. Here all zero singular values in $\sum$ have been removed.

Accordingly, the vectors in U and V that correspond to the zero singular values have been removed, and the document vectors in the SVD subspace can be obtained by using $\bar{X}=U^T X$.

3. Compute CPI projection based on multipliers $\lambda_0, \lambda_1, \ldots\ldots, \lambda_n$ obtained. One can compute the matrix M=$\lambda_0^* M_T+\lambda_1^* x_1 x_1^T+\ldots\ldots\ldots$ Let W_{CPI} be the solution of the generalized Eigen value problem $M_I W = \lambda MW$.

And the low dimensional representation of the document can be computed as

$$Y=W_{CPI}^T \bar{X}=W^T X.$$

Where W= U W_{CPI} is the transformation matrix.

4. Cluster the documents in the CPI semantic subspace. In this, the documents were projected on the unit hyper sphere, but 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 objective function is [12].

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

In this CPI clustering algorithm step 2 will dominate the computation in document clustering applications. To reduce the computation cost of step 2, we are used Bi-LS method based algorithms.

D. Bi-SVD Method

Singular Value Decomposition (SVD) plays an important role in signal processing because they can be used to split a signal into a set of desired and a set of unwanted components. Eigen based methods have been extensively researched in adaptive signal processing, and these methods were obtained in array processing, source localization, and high resolution frequency estimation. Eigen based methods were originally discussed in a block processing context. In this SVD method to reduce the computational burden, in this we can use only one iteration for each new data vector. The basic Bi-SVD subspace tracking algorithm has the computational complexity is O (NLr).

Bi-iterative SVD method for SVD computation:

Initialization: $Q_B(0) = \begin{bmatrix} I_r \\ 0 \end{bmatrix}$

For k=1, 2,…… until convergence Do

Step1:

A (K) = X Q_B(k-1)

A (K) = $Q_A(k) R_A(K)$ Skinny QR decomposition

Step2:

B (K) = $X^H Q_A(K)$

B (K) = $Q_B(k) R_B(K)$ Skinny QR decomposition

III. BI-LS METHOD

The subspace of a vector sequence is well known for its importance in a wide range of signal processing applications, such as channel estimation, frequency estimation, target localization, multiuser detection, image feature extraction etc. let x(1),x(2),…….,x(L) be a sequence of L vectors, each of which is N-dimensional. And the span of this vector sequence can be divided into a principal subspace and a minor subspace, and these two subspaces are orthogonal complement of each other. The minor subspace can be computed uniquely from the corresponding principal subspace and vice versa. The computation of the principal subspace can be done by computing the singular value decomposition (SVD) [13] of the matrix that consists of the L vectors as columns or rows. But computation cost is high. The Bi-LS method is easier to simplify than the Bi-SVD method.

The Bi-LS method is different from the Bi-SVD method and the latter of which has served as the fundamental basis of several other subspace tracking algorithms [14]. The Bi-iterative Least-Square method provides a more convenient framework than the Bi-SVD method. In linear complexity subspace tracking most of the high accuracy linear complexity algorithms belongs to a family of power – based algorithms or simply called the power family [15]. A key feature of the power family is that the primary new information in the updated subspace comes from multiplying the old subspace matrix by the underlying new data.

Optimal low rank matrix approximation using Bi-LS method:

Initialization: $Q_B(0) = \begin{bmatrix} I_r \\ 0 \end{bmatrix}$

For k=1, 2, …… until convergence Do

Step1:

A (K) = X Q_B(k-1)

A (K) = $Q_A(k) R_A(K)$ QR decomposition

Step2:

B (K) = $X^H Q_A(K) R_A^{-H}(K)$

B (K) = $Q_B(k) R_B(K)$ QR decomposition

Subspace tracking using Bi-LS algorithm:

Initialization: $Q_B(0) = \begin{bmatrix} I_r \\ 0 \end{bmatrix}$

For t=1, 2, ……, Do:

Step1:

A (t) = X (t) Q_B (t-1)

A (t) = Q_A(t) $R_A(t)$

Step2:

B (t) = $X^H Q_A(t) R_A^{-H}(t)$

B (t) = $Q_B(t) R_B(t)$

Based on the Bi-LS method introduced different algorithms, and these algorithms having the different time complexities. In Bi-LS-1 algorithm principal computational complexity is 6Nr + 9Lr + $6r^2$ + O(r), for each iteration. In Bi-LS-4 algorithm principal computation complexity is 3Nr + $2.5r^2$ + O(r).

A. Clustering Performance

The testing data used for evaluating the proposed document clustering method are formed by mixing documents from multiple clusters randomly selected from the document corpus. The result is evaluated by comparing the cluster label of each document with its label provided by the document corpus. The accuracy (AC) and the normalized mutual information ($\overline{MI}$) metrics are used to measure the document clustering performance.

Given a document d_i, let l_i be the cluster label and α_i be the label provided by the document corpus. The accuracy (AC) is defined as follows.

$$AC = \frac{\sum_{i=1}^{n} \delta(\alpha_i, map(l_i))}{n}$$

Here n denotes the total number of documents in the test, $\delta(x,y)$ is the delta function, $\delta(x,y)$=1 if (x=y) and $\delta(x,y)$ =0 if (x≠y) and map(l_i) is the mapping function that maps each cluster label l_i to the equivalent label from the document corpus. The best mapping can be achieved by using the Kuhn- munkres algorithm. The normalized mutual information ($\overline{MI}$) is defined as

$$MI(C,C') = \frac{MI(C,C')}{\max(H(C),H(C'))}$$

B. Generalization Capability

The LSI, LPI, and CPI methods are trying to find a low-dimensional semantic subspace by preserving the relational structure among the documents, but where the mapping between the original document space and the low dimensional semantic subspace is explicit. We may use the part of documents to learn such mapping in practical applications, and then transform the documents into the low dimensional semantic subspace which can reduce the computing time. So it is very important for a clustering method to have the capability of predicting the new data by using knowledge formerly acquired from training data without learning once again. The performance on the new samples reflects the generalization capability of the methods.

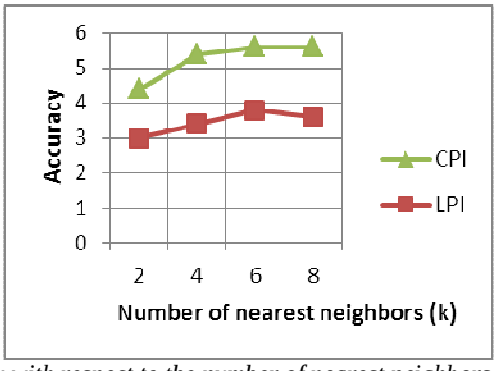

Fig.1. The accuracy with respect to the number of nearest neighbors on Reuter's corpus.

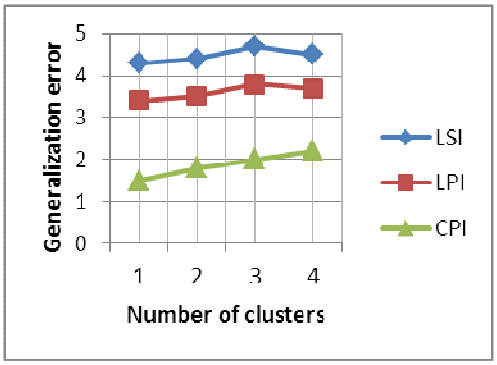

Fig.2. The generalization capability of the LPI, LSI and CPI methods on the dataset Reuters corpus.

In Fig.2 Clearly, the CPI method has smaller generalization error than the LSI and LPI methods, and the CPI method has better generalization capability. CPI can find a low-dimensional semantic subspace in which the documents related to the same semantic are close to each other. So correlation is an appropriate metric for measuring similarity between the documents.

IV. CONCLUSION

In this paper, based on correlation preserving indexing we present a document clustering method. Correlation preserving indexing (CPI) based clustering algorithm is used for the clustering process. It simultaneously maximizes the correlation between the documents in the local patches and minimizes the correlation between the documents outside these patches. And consequently, a low dimensional semantic subspace is derived where the documents corresponding to the same semantics are close to each other. The Bi-LS method is designed to construct the optimal low rank approximation of a matrix. In subspace tracking, more efficient algorithms derived from the Bi-LS method than from the Bi-SVD method, and both methods have the same accuracy of subspace tracking. Furthermore, the CPI method has good generalization capability 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," Int'l Conf, 1994.
[2] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, 1999.
[3] S. Kotsiants and P. Pintelas, "Recent Advances in Clustering: A Brief Review," WSEAS Trans. Information Science and Applications, 2004.
[4] J.B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. Fifth Berkely Symp. Math. Statistics and Probability, vol. 1, pp. 281-297, 1967.
[5] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman, "Indexing by Latent Semantic Analysis," Information Science, 1990.
[6] Y. Fu, S. Yan, ant 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.
[7] 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.
[8] 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, 2004.
[9] R.D. Juday, B.V.K. Kumar, and A. Mahalanobis, Correlation Pattern Recognition. Cambridge University, Press, 2005.
[10] T. Zhang, Y.Y. Tang, and B. Fang, "Document Clustering in Correlation Similarity Measure Space," IEEE Trans. Vol. 4, 2012.
[11] I.S Dhillon and D.M. Modha, "Concept Decompositions for Large Sparse Text Data Using Clustering," Machine Learning, vol. 42, no. 10, pp. 143-175, 2001.
[12] D. Zeimpekis and E. Gallopoulos, "Design of a Matlab Toolbox for Term-Document Matrix Generation," Fifth SIAM Int'l Conf, Data Mining (SDM' 05), pp. 38-48, 2005.
[13] P. Strobach, "Bi-Iteration SVD Subspace Tracking Algorithms," IEEE Trans. Signal Processing, vol. 45, pp. 1222-1240, May 1997.

[14] M. Clint, A. Jennings, "A Simultaneous Iteration Method for the Unsymmetric Eigen Value Problem," J. Inst. Math. Appl, vol. 8, pp. 111-121, 1971.
[15] Y. Hua, Y. Xiang, T. Chen, K. Abed-Meraim, and Y. Miao, "A New Look at the Power Method for Fast Subspace Tracking," Digital Signal Processing, vol. 9, Oct. 1999.

