PRELATION HOMOGENEITY MEASURE: AN ADEQUATE APPROACH FOR DOCUMENT CLUSTERING

Kota Dinesh 1*, N Anjaneyulu 2*

1. *II.M.Tech , Dept of CSE, AM Reddy Memorial College of Engineering & Technology, Petlurivaripalem.*
2. *Asst Prof, Dept. of CSE, AM Reddy Memorial College of Engineering & Technology, Petlurivaripalem.*

Abstract : *This paper presents the latest spectral clustering procedure called correlation preserving indexing (CPI), which is completed in the correlation similarity measure living space. In this kind of framework, the paperwork are usually estimated right low-dimensional semantic space that the correlations between the document in the spots are usually maximized while correlations between the document external these types of spots are usually reduced concurrently. Since the intrinsic geometrical framework of the file can often be stuck in the commonalities between the documents, correlation being a similarity evaluate will be more desirable intended for detecting the intrinsic geometrical framework of the cluster in comparison with euclidean length. Therefore, the planned CPI procedure can successfully discover the intrinsic high-dimensional file living space. The effectiveness of the brand new procedure will be shown simply by extensive findings carried out upon several data models as well as by comparison with existing document clustering approaches.*

Keywords : Document Clustering, Correlation Measure

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. Based on various distance measures, a number of methods have been proposed to handle document clustering. 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 (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 similarities between the data points. Unfortunately, the euclidean distance is a dissimilarity measure which describes the dissimilarities rather than similarities between the documents. Thus, it is not able to effectively capture the nonlinear manifold structure embedded in the similarities between them. 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.

II Previous Work

In many traditional approaches to machine learning, a target function is estimated using labeled data, which can be thought of as examples given by a "teacher" to a "student." Labeled examples are often, however, very time consuming and expensive to obtain, as they require the efforts of human annotators, who must often be quite skilled. For instance, obtaining a single labeled example for protein shape classification, which is one of the grand challenges of biological and computational science, requires months of expensive analysis by expert crystallographers. The problem of effectively combining *unlabeled* data with labeled data is therefore of central importance in machine learning. The semi-supervised learning problem has attracted an increasing amount of interest recently, and several novel approaches have been proposed. Among these methods is a promising family of techniques that exploit the "manifold structure" of the data; such methods are generally based upon an assumption that similar unlabeled examples should be given the same classification. In this paper we introduce a new approach to semi-supervised learning that is based on a random field model defined on a weighted graph over the unlabeled and labeled data, where the weights are given in terms of a similarity function between instances.

Unlike other recent work based on energy minimization and random fields in machine learning and image processing, we adopt *Gaussian* fields over a continuous state space rather than random fields over the discrete label set. This "relaxation" to a continuous rather than discrete sample space results in many attractive properties. In particular, the most probable configuration of the field is unique, is characterized in terms of harmonic functions, and has a closed form solution that can be computed using matrix methods or loopy belief propagation. In contrast, for multi-label discrete random fields, computing the lowest energy configuration is typically NP-hard, and approximation algorithms or other heuristics must be used. The resulting classification algorithms for Gaussian fields can be viewed as a form of nearest neighbour approach, where the nearest labelled examples are computed in terms of a random walk on the graph. The learning methods introduced here have intimate connections with random walks, electric networks, and spectral graph theory, in particular heat kernels and normalized cuts. [1]

In recent years, spectral clustering based on graph partitioning theories has emerged as one of the most effective document clustering tools. These methods model the given document set using a undirected graph in which each node represents a document, and each edge (*i, j*) is assigned a weight *wij* to reflect the similarity between documents *i* and *j*. The document clustering task is accomplished by finding the best cuts of the graph that optimize certain predefined criterion functions. The optimization of the criterion functions usually leads to the computation of singular vectors or eigenvectors of certain graph affinity matrices, and the clustering result can be derived from the obtained eigenvector space. Many criterion functions, such as the Average Cut, Average Association, Normalized Cut, Min-Max Cut, etc, have been proposed along with the efficient algorithms for finding their optimal solutions. It can be proven that under certain conditions, the eigenvector spaces computed by these methods are equivalent to the latent semantic space derived by the LSI method. As spectral clustering methods do not make naive assumptions on data distributions, and the optimization accomplished by solving certain generalized eigenvalue systems theoretically guarantees globally optimal solutions, these methods are generally far more superior than traditional document clustering approaches.

However, because of the use of singular vector or eigenvector spaces, all the methods in this category have the same problem as LSI, i.e., the eigenvectors computed from the graph affinity matrices usually do not correspond directly to individual clusters, and consequently, traditional data clustering methods such as K-means have to be applied in the eigenvector spaces to find the final document clusters.[3]

Latent Semantic Indexing (LSI) is one of the most popular linear document indexing methods which produces low dimensional representations.

LSI aims to find the best subspace approximation to the original document space in the sense of minimizing the global reconstruction error. In other words, LSI seeks to uncover the most representative features rather the most discriminative features for document representation. Therefore, LSI might not be optimal in discriminating documents with different semantics which is the ultimate goal of clustering. Recently, Xu et al. applied the Nonnegative Matrix Factorization (NMF) algorithm for document clustering. They model each cluster as a linear combination of the data points, and each data point as a linear combination of the clusters. They also compute the linear coefficients by minimizing the global reconstruction error of the data points using Nonnegative Matrix Factorization. Thus, the NMF method still focuses on the global geometrical structure of document space. Moreover, the iterative update method for solving NMF problem is computationally expensive. [4]

Lower dimensional representations are useful for visualizing high-dimensional data. However, these methods assume strict conditions that are often violated in real world, high-dimensional data. The obtained sub manifold is tuned to the training data and new data points will likely lie outside the sub manifold due to noise. It is necessary to specify some way of projecting the off-manifold points into the manifold. There is no notion of non-Euclidean geometry outside the sub manifold and if the estimated sub manifold does not fit current and future data perfectly, Euclidean projections are usually used.

Another source of difficulty is estimating the dimension of the sub manifold. The dimension of the sub manifold is notoriously hard to estimate for high-dimensional sparse data sets. Moreover, the data may have different lower dimensions in different locations or may lie on several disconnected sub manifolds, thus violating the assumptions underlying the sub manifold approach. [6]

II. PROPOSED SYSTEM

A. *Corellation Vectors*

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 (column vectors) u and v is defined as Note that the correlation corresponds to an angle _ such that cosine. The larger the value of Corrpus; the stronger the association between the two vectors u and 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. 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.

B. *Max – Min Criteria*

Suppose yi 2 Y is the low-dimensional representation of the I th document xi 2 X in the semantic subspace, where i ¼ 1; 2; . . . ; n. Then the above assumptions (A1) and (A2) can be expressed as Max X Correlation and min X respectively, where N denotes the set of nearest neighbours of xi. The optimization of (3) and (4) is equivalent to the following metric learning denotes the similarity between the documents x and y, corresponds to whether x and y are the nearest neighbours of each other.

C. *Maximization Criteria*

The maximization problem (3) is an attempt to ensure that if xi and xj are close, then yi and yj

are close as well. Similarly, the minimization problem (4) is an attempt to ensure that if xi and xj are far away, yi and yj are also far away. Since the following equality is always true the simultaneous optimization of (3) and (4) can be achieved by maximizing the following objective function Without loss of generality, we denote the mapping between the original document space and the low-dimensional semantic subspace by W. Following some algebraic manipulations, we have 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 (circle for 2D). The global angles between the points in the local neighbors, i, are minimized and the global angles between the points outside the local neighbors, j, are maximized simultaneously. On the unit hypersphere, a global angle can be measured by spherical arc, that is, the geodesic distance..

D. *CPI Clustering*

Given a set of documents x1; x2; . . . ; xn 2 IRn. 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 MS and MT . 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 UV. Here all zero singular values 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 UTX. Compute CPI Projection. Based on the multipliers and obtained, one can compute the matrix M. Let WCPI be the solution of the generalized eigen value problem MSW. Then, the low dimensional representation of the document can be computed by Y CPI where UWCPI is the transformation matrix. 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 of the document using the maximization of the objection function; with cj, mj where mj is the mean of the document vectors contained in the cluster j..

III. RESULTS

The concept of this paper is implemented and different results are shown below, The proposed paper is implemented in Java technology on a Pentium-IV PC with 20 GB hard-disk and 256 MB RAM with apache web server. The propose paper's concepts shows efficient results and has been efficiently tested on different Datasets.

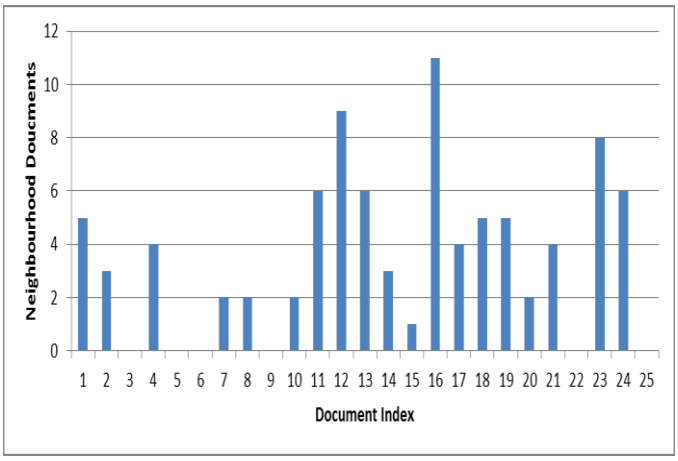

Fig. 1 Number of Neighbourhoods Set

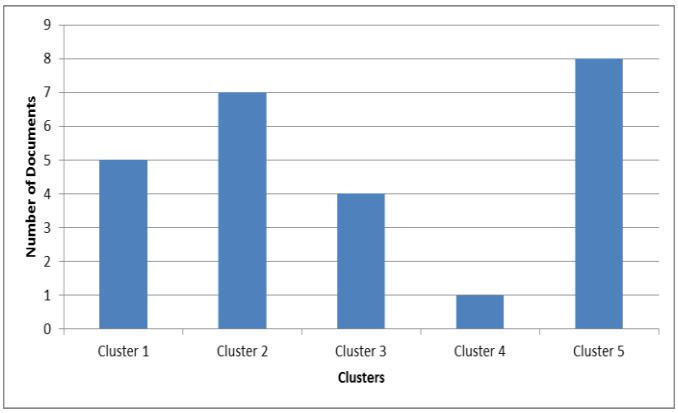

Fig. 1 Document Clusters using CPI Clustering

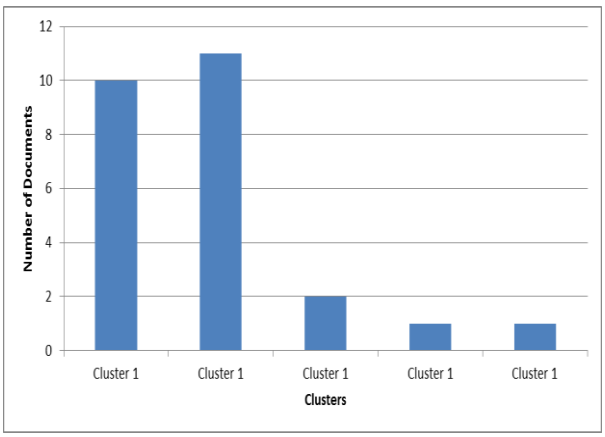

Fig. 3 Document Clusters using CPI Pattern Clustering

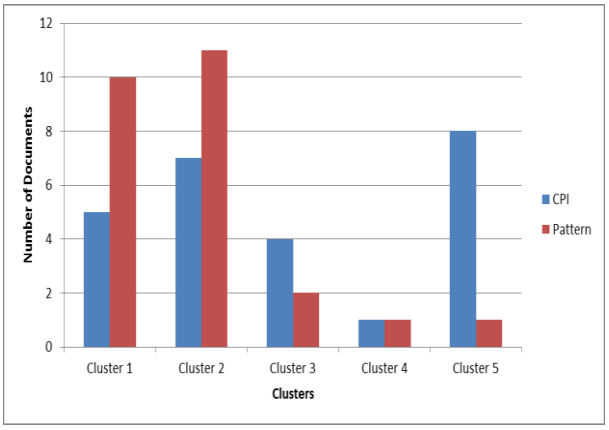

Fig. 4 Comparative Graphs of clustering

IV. CONCLUSIONS

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, and OHSUMED corpora show that the proposed CPI method outperforms other classical clustering methods. Furthermore, the CPI method has good generalization capability and thus it can effectively deal with data with very large size.

V. REFERENCES

[1] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions," Proc. 20th Int'l Conf. Machine Learning (ICML '03), 2003.

[2] 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.

[3] W. Xu, X. Liu, and Y. Gong, "Document Clustering Based on Non-Negative Matrix Factorization," Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Informaion Retrieval (SIGIR '03), pp. 267-273, 2003.

[4] 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.

[5] 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,

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

[7] 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.

[8] S. Zhong and J. Ghosh, "Scalable, Balanced Model-Based Clustering," Proc. Third SIAM Int'l Conf. Data Mining, pp. 71-82, 2003.

[9] 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.

[10] D.R. Hardoon, S.R. Szedmak, and J.R. Shawe-taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods," J. Neural Computation, vol. 16, no. 12, pp. 2639-2664, 2004.

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