IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue, 4, Feb 2019 E-ISSN: 2347-2693

Document Object Mapping and Clustering Using Semantic Indexing Process

V. Geetha[1*], C. Vivekeswari[2]

[1]Department of Computer science, STET Women's College, Mannargudi, India
[2]M.Sc., Computer science, STET Women's College, Mannargudi, India

Corresponding Author: kkmannaig@gmail.com

Available online at: www.ijcseonline.org

***Abstract*—** 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 During this framework, the documents are projected into a low-dimensional semantic area during which the correlations between the documents within the native patches are maximized whereas the correlations between the documents outside these patches are minimized simultaneously. Since the intrinsic geometrical structure of the document area is usually embedded within the similarities between the documents, correlation as a similarity live is additional appropriate for detecting the intrinsic geometrical structure of the document area than Euclidean distance. Consequently, the proposed CPI technique will effectively discover the intrinsic structures embedded in high-dimensional document area. The effectiveness of the new technique is demonstrated by in depth experiments conducted on varied information sets and by comparison with existing document clustering strategies.

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

I. INTRODUCTION

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 clus-tering 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 subspaceapproximation 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 docu-ment 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 cap-turing the similarity structure, rather than the dissimi-larity 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 asimilarity measure 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

$X = \{ x_1, x_2, \ldots\ldots\ldots x_n\}$ and $Y = \{y_1, y_2, \ldots\ldots\ldots\ldots y_n\}$

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

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

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 [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 new document clustering method based on correlation preserving indexing (CPI).

II. DOCUMENT CLUSTERING BASED ON CORRELATION PRESERVING INDEXING

In high-dimensional document house, the Semantic structure is sometimes implicit. it's fascinating to find an occasional dimensional semantic subspace in which the semantic structure will become clear. Hence, discovering the intrinsic structure of the documents house is usually a primary concern of document clustering since the manifold structure is usually embedded within the similarities between the documents, correlation as a similarity live is appropriate for capturing the manifold structure embedded within the high-dimensionality document house.

Mathematically, the correlation between vector U and V defined as

$$corr(u, v) \quad \frac{u^T v}{u^T u \; v^T v} \qquad \frac{u}{\|u\|}, \frac{v}{\|v\|} \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Note that the correlation corresponds to an angle θ such that cosθ = 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 belong 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.

A. 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.

B. Correlation Based Clustering Criteria

Suppose $y_i \epsilon Y$ is the low-dimensional representation of the i^{th} document $x_i \epsilon X$ in the semantic subspace, where i = 1, 2, · · · , n. Then the above assumptions A1) and A2) can be expressed as

$$\max \sum_{i} \sum_{j \in N(x_i)} corr(y_i, y_j) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

$$\min \sum_{i} \sum_{j \in N(x_i)} corr(y_i, y_j) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

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

$$\alpha \quad d(x, y) = \quad * \cos(x, y)$$

Where d(x, y) denotes the similarity between the documents x and y, corresponds to whether x and y are the nearest neighbours of each other.

The maximization problem (3) is an attempt to ensure that if $y_i$and$y_j$are close as well. Similarly, the minimization problem (4) is an attempt to ensure that if the maximization problem (3) 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 thefollowing equality is always true

$$corr(y_i, y_j) + \quad corr(y_i, y_j) = \quad corr(y_i \quad, y_j) \quad ..(5)$$

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

$$\frac{corr(y_i, y_j)}{corr(y_i, y_j)} \quad ..... \quad (6)$$

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

$$\frac{corr(y_i, y_j)}{corr(y_i, y_j)} = \frac{\frac{y_i^T y_i}{y_i^T y_i y_i^T y_j \quad y_j}}{\frac{y_i \quad y_j}{y_i^T y_i y_j^T \quad y_j}} = \frac{\frac{tr(y_i y_j^T)}{tr(y_i y_i^T) tr(y_j y_j^T)}}{\frac{tr(y_i \quad y_j^T)}{tr(y_i y_i^T) tr(y_j y_j^T)}} = \frac{\frac{tr(W^T x \quad x^T W)}{tr(W^T x \quad x^T W) tr(W^T x \quad x^T W)}}{\frac{tr(W^T x \quad x^T W)}{tr(W^T x \quad x^T W) tr(W^T x \quad x^T W)}} \quad ..(7)$$

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

$$\arg\max_{w} \frac{tr(W^T x \quad x^T W)}{tr(W^T x \quad x^T W)} = \arg\max_{w} \frac{tr(W^T(\quad (x \quad x^T))W)}{tr(W^T(\quad (x_i \quad x^T))W)} \quad (8)$$

Consider a mapping $W \quad R^{m*d}$ 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

$$\arg\max \frac{\frac{d}{w \quad M \quad w}}{\frac{d}{w^T M \quad w}} \quad ..... \quad (10)$$

Subject to $w x \quad w \quad =1$, j=1,2,.,n.

Here the matrices M_T and M_S^1 are defined as

$$M_T \quad (x x^t) \qquad M_S \quad (x x^t)$$

correlations between the document points among the nearest neighbours are preserved, we call this criterion "correlation preserving indexing" or CPI.

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

$$d_G(Z, Z') \quad \arccos(Z^t Z') \quad \arccos(corr(Z, Z') \quad (12)$$

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

It is worth nothing that semi-supervised learning using the nearest neighbours graphs approach in the Euclidean distance space was originally proposed in the literatures [15] and [16],

correlations between the document points among the nearest neighbours are preserved, we call this criterion "correlation preserving indexing" or CPI.

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, pi, are minimized and the global angles between the points outside the local neighbours, qi, 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

dG(Z', Z ') arccos(ZTZ') arccos(corr(Z , Z') (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 [14].Based on this conclusion, CPI can effectively capture the intrinsic structures embedded in the high-dimensional document space.

and LPI is also based on this idea. Differently, CPI is a semi-supervised learning using nearest neighbours 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 better metric for text clustering is the cosine similarity[12]-[13].Lebanon in [17] proposed a distance metric for text documents, which was defined as:

It is easy to validate that the matrix M_T is semi-positive definite. Since the documents are projected in the low-dimensional semantic sub subspace in which the

$$d_{F^*}(x, y) \quad \arccos \sum_{i=1}^{n} \frac{x_i \, y_i}{x, \quad y,}$$

This distance is very similar to the distance defined by

(12). Since the distance d_{F^*} (x, y) is local and is defined on the entire embedding space [17], 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^*} (x, y) can be obtained based on the training data and it can be used for classification rather than clustering.

III. RELATED WORK

A. Clustering algorithm based on CPI

Given a set of documents $x_1, x_2 \ldots\ldots x_n \in R^n$ 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 $X=U\Sigma V^T$ 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 $X=U^TX$.

IV. COMPLEXITY ANALYSIS

The time complexity of the CPI clustering algorithm can be analysed as follows: Consider n documents in the d-dimensional space ($d \gg n$). 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 $O(d^3+n^2d)$.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 $k \ll n$, $l << n$ and $m, n \ll d$ 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 dj is given by:

$$\left(tf/idf\right)_{i,j} = tf_{i,j} \times idf_i \quad \text{Here} \quad tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

is the term frequency of the term t_i in document dj 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_i of document d_i is defined as:

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

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

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.

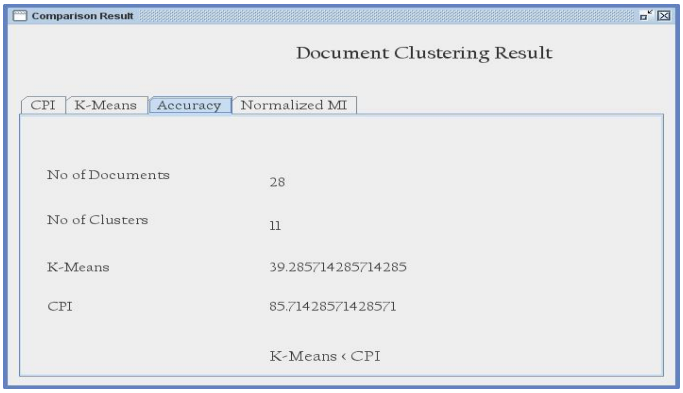
Cor 0: 1:0.04395113068664207:

0.04395113068664207

:

:

Cor 26: 27: 0.051122792602862815:
0.051122792602862815



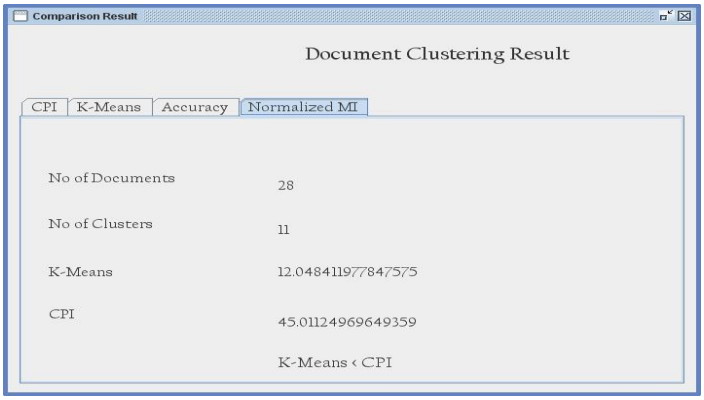


Fig: 5.1 Performance comparisons of different clustering methods using Reuter's dataset

VII. CONCLUSIONS

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.

REFERENCES

[1] P. Mell, T. Grance, The NIST definition of cloud computing, 2011.
[2] Y. Cui, X. Ma, H. Wang, I. Stojmenovic, J. Liu, A survey of energy efficient wireless transmission and modeling in mobilecloud computing, Mobile Networks and Applications 18 (1) (2013) 148–155.
[3] M. Satyanarayanan, P. Bahl, R. Caceres, N. Davies, The case for VM-based cloudlets in mobile computing, Pervasive Computing, IEEE 2009;8(4):14–23.
[4] B.-G. Chun, S. Ihm, P. Maniatis, M. Naik, A. Patti, CloneCloud: elastic execution between mobile device and cloud, Proceedings of the Sixth Conference on Computer Systems. ACM; 2011:301–314.
[5] S. Kosta, A. Aucinas, P. Hui, R. Mortier, X. Zhang, ThinkAir: dynamic resource allocation and parallel execution in the cloud for mobile code offloading, 2012 Proceedings IEEE INFOCOM. 2012:945–953.
[6] A.R. Khan, M. Othman, S.A. Madani, S.U. Khan, A survey of mobile cloud computing application models, Communications Surveys & Tutorials, IEEE 2014;16(1):393–413.

