Dimensionality Reduction Techniques for Document Clustering - A Survey

G. Kalaivani
Thiyagarajar College of Engineering
Department of Computer Science and Engineering
kalaitce22@gmail.com

K. Aravi
Thiyagarajar College of Engineering, Department of Computer Science and Engineering
vanigk4@gmail.com

Abstract — Dimensionality reduction technique is applied to get rid of the inessential terms like redundant and noisy terms in documents. In this paper a systematic study is conducted for seven dimensionality reduction methods such as Latent Semantic Indexing (LSI), Random Projection (RP), Principle Component Analysis (PCA) and CUR decomposition, Latent Dirichlet Allocation (LDA), Singular value decomposition (SVD), Linear Discriminant Analysis (LDA).


1 INTRODUCTION
A text document is stored in a large database, if a user needs to search or retrieve a specific document; it is tough to try and do this task. So clustering the text documents makes easy for searching and retrieving the data. But high dimensional datasets are not efficient for clustering and information retrieval. High dimensionality denotes that quite 10 thousand terms in a document. It slows down computing the distance between documents and it makes the clustering process also slow. The illustration of every and each term within the documents is projected in a vector space. The preprocessing steps are done before dimensionality reduction such as stop word removal and stemming process. The redundant and noisy data to be reduced for efficient document clustering. It improves the performance of clustering.

2 DIMENSIONALITY REDUCTION TECHNIQUES
The curse of dimensionality refers that when the dimensions are increasing and the size of the space also increases, it leads to sparseness. It increases running time. Dimensionality reduction is the transformation of high dimensional data into a low dimensional space. The main idea for this technique is to reduce the number of dimensions without much loss of data.

2.1 Principle Component Analysis
PCA aims to seek out the correlation between the documents. The correlation value ranges between -1 to 1. It reduces the sparseness in documents. The objective is to maximize the variance of the projected data.

Step 1: The original matrix is the documents are projected in vector space. Taking Xi,..., Xn as column vectors, each of which has M rows. Place the column vectors into a single matrix X of dimensions M x N.

Step 2: Calculate mean for each documents

\[ M = \frac{1}{n} \sum_{k=1}^{n} x_k \]  

Step 3: Calculate Covariance Matrix

\[ \Sigma = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \]

\[ \bar{x} = \text{Mean of 1st data set} \]
\[ x_i = 1st \text{ dataset raw score} \]
\[ \bar{y} = \text{Mean of 2nd data set} \]
\[ y_i = 2nd \text{ dataset raw score} \]

Step 4: Calculate Eigen vectors and Eigen values for covariance matrix

\[ \Sigma = E_k \Sigma E_k^T \]

The number of the terms of the original data matrix X is reduced by multiplying with a d x k matrix Ek which has k eigenvectors corresponding to the k largest Eigen values. The result matrix is X PCA.

PROS and CONS
It works well for sparse data.
It can be applied only for linear datasets.

2.2 Random Projection
Random projection is a technique which projects the documents to a randomly chosen low-dimensional space. Its equation is as follows:

\[ X^{PCA} = E_k^T X \]

The number of the terms of the original data matrix X is reduced by multiplying with a d x k matrix Ek which has k eigenvectors corresponding to the k largest Eigen values. The result matrix is X PCA.
\[ X_{k \times d}^{EP} = R_{k \times d} X_{d \times N} \]

N = Total no of Documents
k = desired dimension.
d = Original dimension of documents.
X = Matrix of those documents projected in vector space. We reduce the number of the features by multiplying the matrix R to the original document matrix X.

2.3 Singular Value Decomposition

SVD is a technique for matrices dimensionality reduction and based on a theorem of linear algebra which says that a rectangular matrix M can be divided into the product of three matrices – U is an orthogonal matrix, a diagonal matrix S, and the transpose of an orthogonal matrix V.

\[ M = USV^T \] .................................(5)

M is an m*n matrix. Rows and Column represents those documents are projected in vector space. U is calculated as M * M^T (left singular matrix) . Take Eigen value and Eigen vectors and Orthonormalization process for the left singular matrix. S is a diagonal matrix . V is calculated as M^T * M (right singular matrix) and take V^T. At the end dimensions are reduced as discard the values which are 0 and below 0.

PROS and CONS

Reduce the sparse data.

2.4 Latent Semantic Indexing

LSI is one of the standard dimensionality reduction techniques in information retrieval. A collection of documents can be represented as a huge term-document matrix and various things such as how closely two documents is to a user issued query. In this matrix, each word is a row and each document is a column. Each cell contains the number of times that word occurs in that document. LSI transforms the original data in a different space so that two documents/words about the same concept are mapped close (so that they have higher cosine similarity). LSI achieved this by Singular Value Decomposition (SVD) of term-document matrix and embeds the original high dimensional space into a lower dimensional space with minimal distance distortion. q is query matrix.

\[ Q = q^TUS^{-1} \] .................................(6)

PROS and CONS

Finding relationships between terms in a document. The queries are projected into vector space for each and every user query

2.5 Latent Dirichlet Allocation

It’s a way of automatically discovering topics in documents. It represent documents as mixtures of topics that spit out words with certain probabilities. Topic modeling is a classic problem in information retrieval. LDA discovers the different topics used, classifies unseen documents into those topics. Topic modeling associates with each document a probability distribution over “topics”, which are in turn distributions over words. The task of parameter estimation in these models is to learn both what the topics are, and which documents use them in what proportions. The objective of LDA is to perform dimensionality reduction while preserving as much of the document discriminatory information as possible. It seeks to find directions along which the topics are best separated.

2.6 Linear Discriminant analysis

It does so by taking into consideration the scatter within-documents but also the scatter between -documents. The methodology is follows.

Suppose there are n documents. Let \( \mu_i \) be the mean vector of the documents i, i=1,2,3…n

\( M = \sum_{i=1}^{n} \mu_i \)

let \( M = \sum_{i=1}^{n} M_i \) be the total no of terms.

Within-class scatter matrix:

\[ S_w = \sum_{i=1}^{n} \sum_{j=1}^{M_i} (Y_{ij} - \mu_i)^T (Y_{ij} - \mu_i) \]

Between-class scatter matrix:

\[ S_b = \sum_{i=1}^{n} (\mu_i - \mu) (\mu_i - \mu)^T \]

LDA computes a transformation that maximizes the between-class scatter while minimizing the within-class scatter.

Maximize \[ \text{subject to} \]

Such a transformation should retain class separability while reducing the variation due to sources other than identity

PROS and CONS

Maximizes topic seperability.

2.7 CUR Decomposition

A CUR matrix is a set of three matrix that, when multiplied together. A CUR approximation can be used in the same way as the low-rank approximation of the Singular value decomposition (SVD). CUR approximations are less accurate than the SVD, but the rows and columns come from the original matrix.

\[ A = \text{CUR} \] .................................(7)

Matrix A is those documents are projected in vector space. Matrix C is columns of matrix A. Matrix U is pseudo inverse of intersection of C and R. Matrix R is rows of matrix A.
PROS and CONS

(+ )The vectors are original columns and rows.

(- )There are duplicated columns and rows. So the matrix is very large. So it takes more time to compute.

3. CONCLUSION

We conclude that the dimensionality reduction technique is efficient technique to cut back the inessential terms in documents. In this paper we provided major dimension reduction approaches which will be applied to document clustering. In this paper we provided major dimension reduction approaches that can be applied to document clustering. These techniques can be applied to linear datasets. It mainly reduces sparseness in data.

4. END SECTIONS

4.1 Acknowledgments

We would like to thank UCI repository of machine learning databases for the use of several of their public datasets.

5. REFERENCES


[6] Sabrina Simmons, Zachary Estes “Using latent semantic analysis to estimate similarity” Department of Psychology, University of Warwick Coventry CV4 7AL, UK


[8] L.J.P. van der Maaten , E.O. Postma, H.J. van den Herik “Dimensionality Reduction: A Comparative Review” MICC, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands

[9] Si Chen and Yufei Wang “Latent Dirichlet Allocation” University of California San Diego


[20] Arunasakthi. K, KartatchiPriya “A Review on Linear and Non-Linear Dimensionality Reduction Techniques” Assistant Professor Department of Computer Science and Engineering Ultra College of Engineering and Technology for Women, India. And Assistant Professor, Department of Computer Science and Engineering, Vickers College of Engineering, Enathu, Tamil Nadu, India

[21] Kresimir Delac 1, Mislav Gragic 2 and Sonja Gragic 2 “A Comparative Study of PCA, ICA AND LDA” Croatian Telecom, Savska 32, Zagreb, Croatia, e-mail: kdelac@ieee.org,University of Zagreb, FER, Unska 3/XII, Zagreb, Croatia

[22] David M. Blei, Andrew Y. Ng, Michael I. Jordan “Latent Dirichlet Allocation” Computer Science Division University of California Berkeley, CA 94720, USA, Computer Science Department, Stanford University, Stanford, CA 94305, USA, Computer Science Division and Department of Statistics, University of California, Berkeley, CA 94720, USA, Journal of Machine Learning Research 3 (2003) 993-1022


[24] Kresimir Delac, Mislav Gragic, Sonja Gragic “Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set” University of Zagreb, FER, Unska 3/XII, Zagreb, Croatia


