International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 7 (2016) pp 4770-4774
 https://dx.doi.org/10.37622/IJAER/11.7.2016.4770-4774

Text Document Clustering Using Dimension Reduction Technique

A. Sudha Ramkumar
Research Scholar, Bharathiar University, Coimbatore, India.

Dr. B. Poorna
Principal, SSS Jain College, Chennai, India

Abstract
Text document clustering is used to group a set of documents based on the information it contains and to provide retrieval results when a user browses the internet. Experimental evidences have shown that Information Retrieval applications can benefit from document clustering and it has been used as a tool to improve the performance of retrieval of information. Information retrieval is an interdisciplinary field of knowledge management and text mining. Dimensionality Reduction (DR) is a typical step in many text mining problems which involves transforming sparse data into a shorter and more compact one. DR can be done in 2 ways: feature reduction and feature selection. This study implements dimensionality reduction through feature selection with k-means algorithm. Feature Selection is implemented through the InfoGain DR technique. This paper presents an experimental analysis of the performance of the document clustering with the InfoGain technique and proves that this method significantly improves the performance in terms of Accuracy, Precision and Recall for the BBC Sports Dataset.

Keywords: Text document clustering, Dimensionality reduction, feature selection.

INTRODUCTION

Clustering is a common form of unsupervised learning and is widely used in many areas of research and science. According to Pankaj Jajoo[6], the following are the basic directions in which clustering is used,

- Finding similar documents,
- Organizing large document collections,
- Duplicate content detection,
- Recommendation system,
- Search optimization.

Clustering helps a lot in improving the quality and efficiency of search engines, as the query can be first compared to the clusters instead of comparing them directly to the documents and the search results can also be arranged easily [6]. Text document clustering is used to group a set of documents based on the information it contains and to provide retrieval results when a user browses the internet [3]. Experimental evidences have shown that Information Retrieval applications can benefit from document clustering and it has been used as a tool to improve the performance of retrieval of information [1]. Document clustering is being studied from many decades but still needs to be investigated. A fundamental goal of document clustering is the identification of a set of groups that accurately reflects the topics present in a corpus [4]. The similarity measures use cosine similarity functions or Euclidean distance.

The performance of clustering algorithms will decline dramatically due to the problems of high dimensionality and data sparseness [9]. Irrelevant features in the documents reduce the accuracy of information retrieval in terms of precision and recall as well as make it difficult for clustering algorithm to effectively cluster similar documents. Dimensionality reduction through feature selection selects a subset of original representation attributes based on quality metrics like information gain or chi-square [8]. K-Means Clustering is an efficient algorithm for text document clustering. The study presents the use of K-Means with feature selection in clustering a dataset of text documents and shows how it enhances the performance in terms of accuracy when compared to K-Means without feature selection.

As a first step of the research work, the preprocessing of the document collection is made; followed with the construction of term document matrix, perform Feature Selection using InfoGain DR Method and cluster with the most popular K-Means algorithm. Finally, this paper proves experimentally the effectiveness of clustering in terms of precision, recall and accuracy. This rest of this paper is organized as follows. Related work Section discusses some of the previous research work on the text document clustering, K-Means Clustering with Feature Selection Section presents the K-Means Clustering with Feature Selection InfoGain DR Method and describes the performance metrics that are applied to compare the two methods, Experimental results Section gives the results of original K-Means clustering algorithm, K-Means clustering algorithm with Feature Selection InfoGain DR and the comparison based on various metrics. Final Section concludes the paper and future work.

RELATED WORK

Text document clustering is used to automatically group the document that belongs to the same topic in order to provide user's browsing of retrieval results [3]. Experimental evidences prove that IR application can benefit from the use of document clustering [1]. Document clustering has always been used as a tool to improve the performance of retrieval and navigating large data. Recently, clustering has been proposed for use in browsing a collection of documents or in

organizing the results of search engine in response to a user's query.

Balabantaray et al.,[2] compares the K-Means Clustering with K-Medoids Clustering. K-Means was carried out using both Euclidean and Manhattan distance on WEKA tool and K-Medoids was carried out through java programming. Finally, it was observed that K-Means yields better result than K Medoids.

Greene Derek et al.,[4] introduced text clustering toolkit, a state of the art framework supporting the development of applications for unsupervised text mining tasks. This paper covers all phases of cluster analysis from the preprocessing of raw documents to the interpretation of a final clustering solution. Jain et al.,[5] discusses about document preprocessing, applications of text clustering, key methods for text clustering with their advantages and limitations and concludes with algorithms should be developed to allow the overlapping of clusters. Jajoo et al.,[6] discussed how to improve the efficiency and accuracy of document clustering. They discussed two clustering algorithms and the domains where these perform better than the known standard clustering algorithm. First approach is an improvement of the graph partitioning techniques used for document clustering and the second approach is a completely different in which words are clustered and then the word clusters is used to cluster the document. This reduces the noise in data and thus improves the quality of clusters.

Khadhim et al.,[7] implements TF-IDF and Singular Value Decomposition dimensionality reduction techniques and experimental results have shown that this method enhances the performance of text document clustering. Liu Luying et al.,[8] discusses about the four unsupervised feature selection methods such as document frequency(DF), Term Contribution(TC), term variance quality(TVQ), and a new proposed Term variance (TV). They evaluated the above said methods and found that the proposed term variance and term contribution are better than DF and TVQ. This paper also indicates that the performances of different cluster validity are not the same.

Liu Tao et al.,[9] shows the empirically that feature selection methods can improve the efficiency and performance of text clustering algorithm. This paper proposes a new feature selection method called term contribution and performs a comparative study on a variety of feature selection methods such as DF, term strength, Entropy based, Info Gain and Chi-square method. This paper concludes with term contribution are better than DF and Entropy based. Manoranjan Dash [10] presents the key aspect of feature extraction and feature selection and discussed about the some basic methods such as Principal Component Analysis and Latent Semantic Indexing and their importance in different application areas such as microarray gene expression data.

Mugunthadevi et al.,[11] surveyed different feature selection methods along with their advantages and limitations and concluded that feature selection remains and will continue to be active field to answer new challenges in the area of text mining. Tang bin et al.,[12] conducted a systematic study of six dimensionality reduction techniques in the context of text clustering problem using three different datasets. Finally, experiments with a selection technique followed by a transformation technique can substantially reduce the computational cost associated with the best transformation method such Independent component analysis ICA and Latent Semantic Indexing while preserving the clustering performance. Zhoa et al.,[13] proposes a new method which introduces the cloud model theory into feature selections, constructing the clouds filter for clustering documents. Experimental results with K-Means algorithm shown in this research had remarkable improvement in terms of accuracy of text clustering.

K-MEANS CLUSTERING WITH FEATURE SELECTION

The methodology of K-Means clustering with Feature Selection is shown in Figure 1, which involves six main stages: Document Collection, Preprocessing, Term Document Matrix Construction, Dimensionality Reduction with Feature Selection, Clustering with K-Means, Evaluation of Clusters and Comparison with K-Means Clustering Algorithm with the K-Means Clustering without Feature Selection.

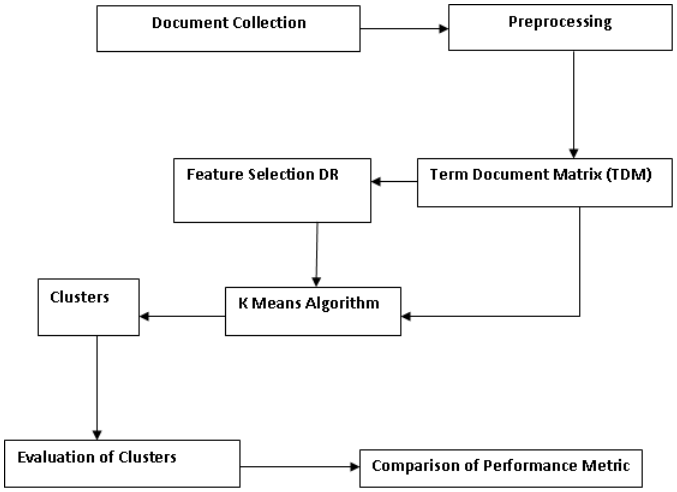

Figure 1: Methodology of K-Means with Feature Selection

Preprocessing

In order to form mathematical data model that the computer can deal with, Preprocessing techniques are applied to the document collection which greatly influence the outcome of any clustering process. Given a collection of text documents, the first task is to apply a once-off parsing process, where the documents are transformed into a data model which can be subsequently analyzed by a machine learning algorithm. The parsing task involves a tokenization, Stemming and Stop-word Removal. Tokenization will transforms the contents of a document into a sequence of terms which will be used to characterize the document.

Stemming is used to reduce the number of unique terms, it is generally useful to stem terms to their roots. For the English language, the standard procedure is to apply the porter stemming algorithm to eliminate common morphological and inflectional endings like "programming" into "program". Stop-word removal is used to remove the basic functional

words which occur so frequently in documents but do not have any discriminating power and these words are considered to be noise.

Term Document Matrix

After preprocessing, a term document matrix X is built with the table of frequencies of occurrences of terms in each document. Local and global weighting functions are applied to estimate the relative importance of a term within the document and within the whole collection. Since each document contains different words, this table is a high dimensional sparse m x n matrix, m is the number of unique terms in the document collection and n is the number of documents in the collection. The high-dimensional and sparse features bring great noise to the text clustering and make it difficult for clustering algorithms to effectively cluster similar documents.

Dimension Reduction

Feature Selection dimensionality reduction selects a subset of the original representation attributes focusing on the word importance based on the evaluation function. The process of Feature Selection consists of 2 steps, first calculating the importance value for each word, then selecting the words whose importance values are greater than the predefined threshold value. There are many Feature Selection methods such as Document Frequency, Information Gain, Mutual Information, Chi Square Statistics etc.

Information gain (IG) is an effective Feature Selection method and is widely used in text data mining. Information gain does not concern the relation between a certain feature word and certain class, but treat all classes in training set as a whole. And the importance of a certain word is measured by calculating the information amount that each class takes. Information gain of the feature word t refers to the D-value between the information amount of the whole training set without regard to feature word t and that of the training set with regard to feature word t.

K-Means Clustering Algorithm

The K-Means clustering algorithm is an efficient unsupervised learning algorithm and was developed by MacQueen to solve the well known clustering problem. The K-Means algorithm aims to partition a set of objects based on their features into k clusters, where k is a predefined constant. The main idea is to define k centroids, one for each cluster. The centroid of a cluster is formed in such a way that it is closely related in terms of similarity function in which similarity is measured by Euclidean distance to all objects in that cluster.

The steps of K-Means Algorithm are as follows,

1. Initialize the parameters for the K-Means algorithm.
2. Define K, the number of clusters.
3. Define the initial set of centroids and apply the K-Means clustering algorithm to the Document collection.
4. To identify the relation among cluster and classes:

- Compute the Euclidean distance among cluster and class centroids.
- Select the Classes to Clusters evaluation.

5. Consider all possible assignments.
6. Compute overall distance for all cases.
7. Get the best assignment.
8. Go back to step 3, stop when no more new assignment.

Clusters Performance

Being unsupervised machine learning, the clustering has no test data and whole document collection is preprocessed with the help of StringToWordVector filter of WEKA Tool. The preprocessing step involves Tokenization, Stemming and Stop-word removal. Then the K-Means clustering algorithm is applied on this which takes the Euclidean distance to measure the similarity between documents. It is considered that the closer two texts have high similarity between them.

The confusion matrix is generated as the result of Classes to Clusters evaluation method of WEKA tool. A confusion matrix [14] is a table that allows the visualization of the performance of an algorithm and in unsupervised learning it is called as matching matrix as shown in Figure. 2. In the confusion matrix, all the diagonal elements are true positives and it is the relevant document to that particular class, where as the number of documents retrieved are true negatives and true positives. To evaluate the effectiveness of the clustering, Recall, Precision and Accuracy metrics are considered. These metrics are widely used in data mining and information retrieval.

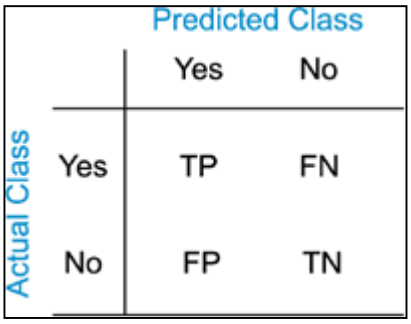

Figure 2: Confusion Matrix

$$\text{Precision} = \frac{\text{Number of relevant documents Retrieved (TP)}}{\text{Number of documents Retrieved(TP+FP)}}$$

$$\text{Recall} = \frac{\text{Number of documents Retrieved (TP)}}{\text{Number of relevant documents(TP+FN)}}$$

EXPERIMENTAL RESULTS

To verify the effectiveness of the Feature Selection DR method, we compared the K-Means algorithm after the Feature Selection DR with the K-Means algorithm. In the experiments, the BBC Sport Dataset has been used for both the methods and is downloaded from the BBC website. The confusion matrix generated from the Classes to Clusters evaluation method is used for the comparison. The BBC Sport

dataset consists of 737 text documents and 5 natural classes: Athletics, Cricket, Football, Rugby and Tennis.
For the K-Means algorithm without DR method, the overall recall is 83%, overall precision is 85% and the accuracy is 83.4%. The total numbers of incorrectly clustered documents are 123 out of 737 documents. The confusion matrix for the K-Means without DR Method is shown in Table 1.

Table 1: Confusion Matrix for K-Means

	Athletics	Cricket	Football	Rugby	Tennis	Predicted
Athletics	**100**	1	6	2	24	133
Cricket	0	**112**	34	14	0	160
Football	0	7	**221**	14	11	253
Rugby	1	4	4	**117**	1	127
Tennis	0	0	0	0	**64**	64
Original Classes	101	124	265	147	100	737

The same data representation is implemented by Feature Selection DR Method with InfoGain is carried out which returns a more compact and small subset of original data representation. With this reduced data, once again K-Means clustering algorithm is applied. For that reduced data, the Table 2. Shows the confusion matrix generated from the Classes to Clusters evaluation,

Table 2: Confusion Matrix for K-Means with InfoGain Feature Selection

	Athletics	Cricket	Football	Rugby	Tennis	Predicted
Athletics	**100**	0	0	0	1	101
Cricket	0	**106**	0	0	0	106
Football	1	18	**265**	23	17	324
Rugby	0	0	0	**124**	0	124
Tennis	0	0	0	0	**82**	82
Original Classes	101	124	265	147	100	737

For the K-Means clustering algorithm with Feature Selection DR Method, the overall recall is 90%, overall precision is 96% and accuracy is 91.9%. The total numbers of incorrectly clustered documents are 60 out of 737 documents.

Table 3: Comparison of K-Means and K-Means with InfoGain Feature Selection

Metric	K-Means in %	K-Means with InfoGain FeatureSelection in %
Recall	83	90
Precision	85	96
Accuracy	83.4	91.9

The Figure.3 shows the significant improved performance in terms of Recall, Precision and Accuracy for the K-Means algorithm without Feature Selection and the K-Means algorithm with InfoGain Feature Selection DR method.

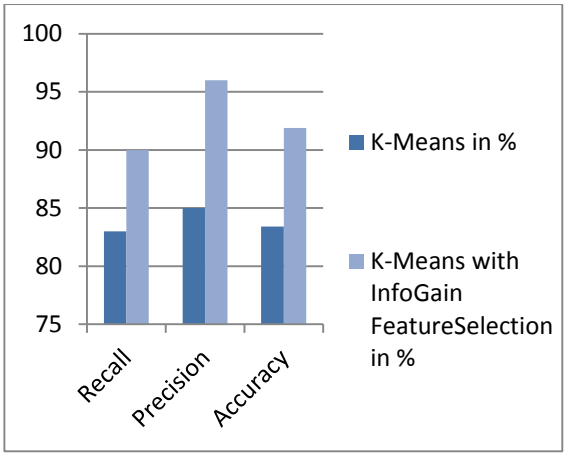

Figure 3: Comparison of K-Means and K-Means with InfoGain Feature Selection

CONCLUSION

The goal of text document clustering is to minimize the intra cluster distance between documents while maximizing the inter cluster distance using an appropriate distance measure between documents. With the full data representation, the K-Means algorithm cannot effectively clusters the document collection. By applying the Feature Selection DR method with K-Means algorithm improves the clustering quality significantly. Finally, the experimental results have been shown how Feature Selection DR method with K-Means is superior to K-Means algorithm in performance. In future, Feature Reduction DR technique Latent Semantic Indexing through Singular Value Decomposition can be used to analyze the hidden semantics and to produce better clustering results.

REFERENCES

[1] Andrews, Nicholas O., and Edward A. Fox. "Recent developments in document clustering." (2007).
[2] Balabantaray, Rakesh Chandra, Chandrali Sarma, and Monica Jha. "Document Clustering using K-Means and K-Medoids." arXiv preprint arXiv:1502.07938 (2015).
[3] Gao, Jing, and Jun Zhang. "Clustered SVD strategies in latent semantic indexing." Information processing & management 41.5 (2005): 1051-1063.
[4] Greene, Derek. A State-of-the-art Toolkit for Document Clustering. Diss. Trinity College, 2007.
[5] Jain, Yogesh, and Amit Kumar Nandanwar. "A Theoretical Study of Text Document Clustering."
[6] Jajoo, Pankaj. Document clustering. Diss. Indian Institute of Technology Kharagpur, 2008.
[7] Kadhim, Ammar Ismael, Yu-N. Cheah, and Nurul Hashimah Ahamed. "Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering." Artificial Intelligence with Applications in Engineering and Technology (ICAIET), 2014 4th International Conference on. IEEE, 2014.

[8] Liu, Luying, et al. "A comparative study on unsupervised Feature Selection methods for text clustering." Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on. IEEE, 2005.

[9] Liu, Tao, et al. "An evaluation on feature selection for text clustering." Icml. Vol. 3. 2003.

[10] Manoranjan dash ―Dimensionality Reduction‖ Department of Information Systems,School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798.

[11] Mugunthadevi, K., et al. "Survey on feature selection in document clustering." International Journal on Computer Science and Engineering 3.3 (2011): 1240-1241.

[12] Tang, Bin, et al. "Comparing and combining dimension reduction techniques for efficient text clustering." Proceedings of SIAM International Workshop on Feature Selection for Data Mining. 2005.

[13] Zhao, Junmin, Kai Zhang, and Jian Wan. "Research of feature selection for text clustering based on cloud model." Journal of Software 8.12 (2013): 3246-3252.

[14] R. E. Banchs. “Text Mining with MATLAB” . Springer, 2012.