Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering

A. Anil Kumar
Dept of CSE
Sri Sivani College of Engineering
Srikakulam, India

S.Chandrasekhar
Dept of CSE
Sri Sivani College of Engineering
Srikakulam. India

`Abstract

Text mining refers generally to the process of extracting generally to the process of extracting interesting and non-trivial and knowledge from unstructured text data. Text mining is interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. Standard text mining and information retrieval techniques of text document usually rely on word matching. An alternative way of information retrieval is clustering. In which document pre-processing is an important critical step in the clustering process and it has a huge impact on the success extract knowledge.

Document clustering is a technique used to group similar documents. During the course of the project we implement tf-idf and singular value decomposition dimensionality reduction techniques. We proposed an effective pre-processing and dimensionality reduction techniques which helps the document clustering. Finally we have chosen one dimension reduction technique that performed best both in term of clustering quality and computational efficiency.

1. Introduction

All Before going to perform any operation on the text data, the data must be pre-processed. Because the text data often contains some special formats like number formats, date formats and the most common words that unlikely to help text mining such as prepositions, articles, and pro-nouns can be eliminated. Stemming or lemmatisation is a technique for the reduction of words into their root. Many words in the English language can be in multiple forms. So that, these words reduced to their roots. After applying stop word elimination and stemming the data is converted into vector space model. This is helpful to handle the data in terms of numeric values. Dimension reduction is an important step in text mining. Dimension reduction improves the performance of clustering techniques by reducing dimensions so that text mining procedures process data with a reduced number of terms. Singular value decomposition is a technique used to reduce the dimension of a vector. Finally clustering is introduced to make the data retrieval easy. K-means is an efficient clustering technique which is applied for clustering text documents.

2. Text Data Pre-processing

A database consists of massive volume of data which is collected from heterogeneous sources of data. Due to this heterogeneity, real world data tends to be inconsistent and noisy. If data is inconsistent, then there is possibility that mining process can lead to confusion which results in inaccurate results. In order to extract data which is consistent and accurate data pre-processing is applied on that data. The objective of this is that it enhances the quality of data and at the same time reduces the difficulty of mining process. For text data pre-processing in this work we used following methods for efficient text data pre-processing.

2.1 Tokenization

The first step of Morphological Analyses is the tokenization. The aim of the tokenisation is the exploration of the words in a sentence. Textual data is only a block of characters at the beginning. All following processes in information retrieval require the words of the data set. Hence, the requirement for a parser which processes the tokenisation of the documents. This may sound trivial as the text is already stored in machine-readable formats. Nevertheless, some problems are still left, like the removal of punctuation marks. Other characters like brackets, hyphens, etc. require a processing as well. Furthermore, tokenizer can cater for consistency in the documents. The main use of tokenization is identifying the meaningful keywords. The inconsistency can be different number and time formats. Another problem are abbreviations and acronyms which have to be transformed into a standard form.

2.2 Stopword elimination

The most common words are in any text document does not provide meaning of the documents; those are prepositions, articles, and pro-nouns etc. These words are treated as stopwords. Because every text document deals with these words which are not necessary for text mining applications. These words are eliminated. Any group of words can be chosen as the stopwords for a given purpose. This process also reduces the text data and improves the system performance. Example ‘the’,’in’,’a’,’an’,’with’etc.

In this work we followed a novel approach to eliminate stop words. The elimination is done based on ASCII values on ASCII values of each letter without considering the case (either lower case or upper case) and sum the each letter corresponding ASCII value for every word and generate the number. Assign number to corresponding word, and keep them in sorted order. But in this approach there is chance that the ASCII sum of the two word's values can be same as shown with the below example , the word "ask" sum value is 97+115+107=319 and the word "her" sum value is 104+101+111=31.

Solution for above mentioned problem is during the comparison we can compare with the ASCII sum value and in the corresponding array we can take stop words string. So that we can compare with the string and confirm so that there will be no loss of key words and also we should create a subset of strings with same ASCII sum so that it is enough to compare with only that subset. Our proposed algorithm is as follows.

Input: Text data

Output: Eliminate stop words from the given text file.

Noisy removal ()

Step 1: Extract ASCII value of character

Step 2: If it is >65 and <90 or >97 and <122

then

Step 3: Accept the character

Step 4: Else

Remove the character

Step 5: Go to next character

Step 6: Repeat until EOF

2.3 Stemming

Stemming or lemmatization is a technique for the reduction of words into their root. Many words in the English language can be reduced to their base form or stem e.g. agreed, agreeing, disagree, agreement and disagreement belong to agree. Furthermore are names transformed into the stem by removing the " 's". The variation "Peter's" in a sentence is reduced to "Peter" during the stemming process. The result of the removal may lead to an incorrect root. However, these stems do not have to be a problem for the stemming process, if these words are not used for human interaction. The stem is still useful, because all other inflections of the root are transformed into the same stem. Case sensitive systems could have problems when making a comparison between a word in capital letters and another with the same meaning in lower case. In this we applied standard Porter Stemming Algorithm for find the root words in the document.

3. Vector Space Model

In present document clustering after eliminating stop words and performing stemming keywords in the documents are remained. Vector space model with those keywords and documents is formed. Here every element in the vector space model indicates how much number of times a word occurs in the document. For example consider the following diagram which shows the term and document matrix. Rows represents words in each document, columns represents documents. Each cell represents the word count in the particular document. If the number of documents is increasing the size of the matrix will also be increased.

This leads to high-dimensionality. So for efficient clustering to take place this high-dimensionality must be reduced. Data mining is a way to find useful patterns from database. Clustering algorithms are mainly used to group these patterns from a large dataset. The algorithms must be prepared to deal with data of limited length. For this High-dimensionality of data must be reduced. To reduce high-dimensionality of data we are using Singular Value Decomposition (SVD).

Processing high dimensional data is very expensive in terms of execution time and storing the data. And the clustering is not done in an effective way. High dimensionality is challenging to perform a efficient clusters of the input documents, by using high dimensionality reduction techniques we can reduce the size of the vector space model. The dimensionality reduction techniques are Singular Value Decomposition

(SVD) , Principle Component Analysis (PCA) and Independent Component Analysis (ICA).

4. Dimensionality Reduction

Advances in data collection and storage capabilities during the past decades have led to an information overload in most sciences. Researchers working in domains as diverse as engineering, astronomy, biology, remote sensing, economics, and consumer transactions, face larger and larger observations and simulations on a daily basis. Such datasets, in contrast with smaller, more traditional datasets that have been studied extensively in the past, present new challenges in data analysis. Traditional statistical methods break down partly because of the increase in the number of observations, but mostly because of the increase in the number of variables associated with each observation.

The dimension of the data is the number of variables that are measured on each observation. High-dimensional datasets present many mathematical challenges as well as some opportunities, and are bound to give rise to new theoretical developments. One of the problems with high-dimensional datasets is that, in many cases, not all the measured variables are important" for understanding the underlying phenomena of interest. While certain computationally expensive novel methods can construct predictive models with high accuracy from high-dimensional data, it is still of interest in many applications to reduce the dimension of the original data prior to any modelling of the data.

High dimensionality reduction solves the problem of inefficient computation and increase the difficulty in detecting and exploiting the relationships among terms. After getting the relationship among the key words, clustering is done very effectively and easily. The processing time will be reduced.

4.1 Singular Value Decomposition

The purpose of singular value decomposition is to reduce a dataset containing a large number of values to a dataset containing significantly fewer values. SVD is based on a theorem from linear algebra which says that a rectangular matrix A can be broken down into the product of three matrices . Here A is n x m size matrix which indicates n documents and m terms. U is n x r matrix which indicates n documents and r concepts. S is r x r diagonal matrix rank of the matrix. V is m x r matrix contains m terms and r concepts. The definition of the SVD is as follows.

$$A_{m \times n} = U_{m \times m} S_{m \times n} V_{n \times n}^{T}$$

Here the U is an orthogonal matrix, S is a diagonal matrix, and V is the transpose of an orthogonal matrix. Calculating the SVD consists of finding the eigen values and eigenvectors of AA^T and A^TA. The eigenvectors of A^TA make up the columns of V, the eigenvectors of AA^T make up the columns of U. where $U^T U = I$; $V^T V = I$; The result of this work provides us to reduced format of the input matrix suitable for the document clustering.

5. Clustering the Documents

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another with in the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group in many applications. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. The goal of clustering is to reduce the large amount of raw data by categorizing in smaller sets of similar items. In this work we used a partitioning clustering method K-means to cluster given N number of documents.

5.1 K-means Algorithm

The k-means algorithm takes the input parameter k, and partitions a set of n- objects into k-clusters so that the resulting intra-cluster similarity is high whereas the inter cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in the cluster, which can be viewed as the cluster’s “center of gravity”.

The data set is partitioned into k Clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same

number of data points. The procedure of the clustering perform as following steps.

Step 1: Partition objects into k nonempty subsets.

Step 2: Compute seed points as the centroids of the

clusters of the current partition(the centroid is the centre ,i.e., mean point, of the cluster)

Step 3: Assign each object to the cluster with the nearest seed point.

Step 4: Go back to step 2, stop when no more new assignment.

For each data point : Calculate the distance from the data point to each cluster. If the data point is closest to its own cluster, leave it where it is. If the data point is closest to its own cluster, move it into the closest cluster. Repeat the above step until a complete pass through all the data points results un no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.

6. Experimental Results

In our experiment we have taken standard documents, these documents contain all together 15809 words. These words reduced after elimination of stop words into 1840, after stemming remaining words are 323. In this section we show the experimental results. We conducted several experiments for document pre-processing, dimensionality reduction and classification. Standard documents were used for our experiments.

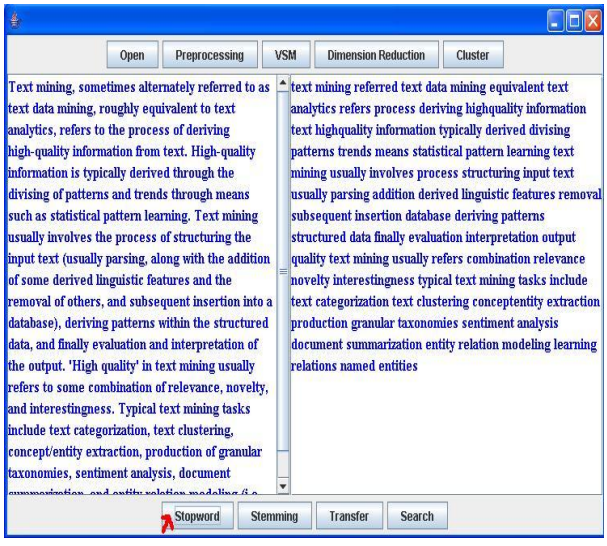

Figure1: Stopword elimination for the given documents

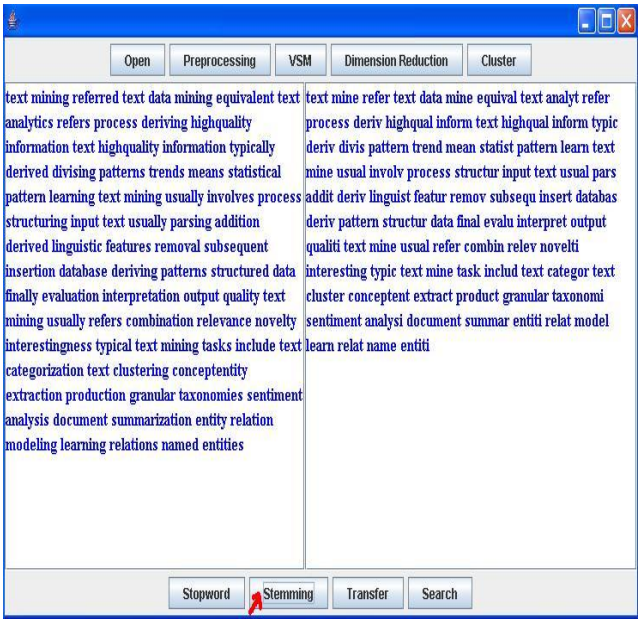

Figure2: Stemming the input documents

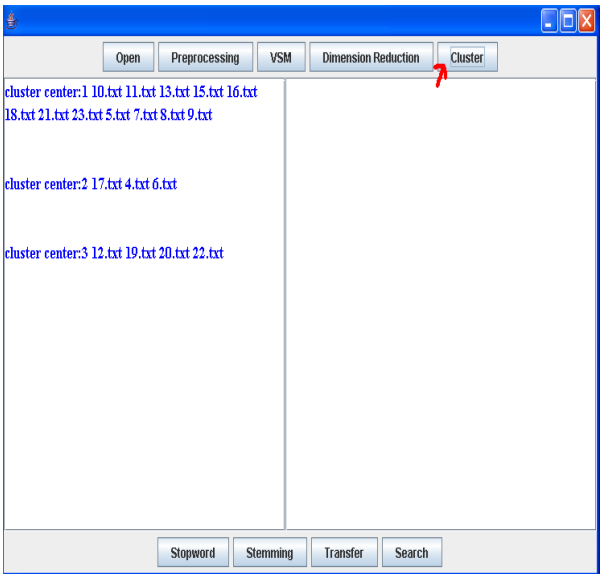

Figure 3: Clustering the documents into 3 categories

7. Conclusion and further work

In this work we have presented efficient pre-processing and high dimensionality reduction techniques. These pre-processing techniques eliminates noisy from text data, later identifies the root word for actual words and reduces the size of the text data. This improves performance of the

system. To deal with vectors, there is a need of dimensionality reduction. Singular value decomposition is the best technique to reduce the dimension. This can be computed easily for the vectors. Finally k-means clustering technique is used to make the data retrieval easy.

Further extension of this work, the following issues may be addressed: 1) Semantic query matching 2) Index based clustering method. For efficient data extraction, semantic search is also needed because it helps to retrieve relevant data. The synonyms of the words are identified and these words are replaced instead of actual word when they are not found in the content. This improves accuracy and performance of search. After performing clustering, index is needed for each cluster to identify relevant cluster. The most important keywords should be identified for each document. These keywords are used as index for each cluster. This procedure will help to identify the relevant cluster according to the given query.

Acknowledgement

My sincere thanks to my college management; they have been providing such environment to do research work in the campus. I also extend my sincere and hole hearted thanks to my head of the department and my guide to encourage me to this work.

10. References

[1] Durmaz,O.;Bilge, H.S "Effect of dimensionality reduction and feature selection in text classification " in IEEE conference ,2011, Page 21-24 ,2011.

[2] J Jun Yan; Benyu Zhang; Ning Liu; Shuicheng Yan; Qiansheng Cheng; Fan, W.; Qiang Yang; Xi, W.; Zheng Chen. "Effective and efficient dimensionality reduction for large –scale and Streaming data preprocessin" Vol18,Issue3 2006, pp.320–333.

[3] Lukui Shi; Jun Zhang; Enhai Liu; Pilian He" Text classification Based on Nonlinear Dimensionnality Techniques and support Vector Machines" IEEE Conference ,2007, pp.674-677.

[4] M.Ramakrishna Murty,JVR Murty,PrasadReddy PVGD" Tect document classification based on least square support vector machine with singular value decomposition " published IJCA, Vol 27,No-7,Aug-2011.

[5] D. S´anchez, M.J. Mart´ın-Bautista, I. Blanco C. Justicia de la Torre "Text Knowledge Mining: An Alternative to Text Data Mining" IEEE International Conference on Data Mining Workshops 2008.

[6] Vishal Gupta , Gurpreet S. Lehal "A Survey of Text Mining Techniques and Applications" Journal of Emerging technologies in web intelligence, vol,1 no1 Auguest 2009.