Comparative study of feature selection techniques and information retrieval

Mausumi Goswami

Christ University, Bangalore

, Bipul Syam Purkayastha

Assam Central University , Silchar

Abstract: It is a common practice to store majority of information related to on, business, official communication in governmental work, industrial interaction and so on in the electronic form using text databases. Representation and processing of this data is quite challenging due to the characteristics of natural language. In the paper few feature selection techniques are compared.

INTRODUCTION

Every aspect of society is influenced by digitization as the information age infiltrates the society. It has caused exponential growth of digital data. Unstructured texts are available in many different forms which can be categorized as text available on the world wide web, digital images, videos, sound, result of scientific experiments and user profiles for marketing. Representation and processing of text datasets is quite challenging due to high dimensionality and also it is challenging to retrieve useful information from them. There are certain characteristics of Natural Languages which may make the problem worse.

Data mining and text mining are not same. Finding hidden patterns in given data is one of the major objectives in data mining. But Data mining needs highly structured data. Text mining does not need an ordered set of numeric data. Text data is a collection of unstructured documents.

If we consider only text documents as data to perform a similar process to data mining Then the process is called text mining. .

A cleaning task involves throwing all unwanted words .The process of removing all unwanted words actually performs a cleaning task. This process of removing all unwanted words from the user input makes the process simpler. After throwing noise words the residual collection of words are looked up in every document. Using the search result a matrix is formed with the number of words and the corresponding document.. This matrix gives the frequency of each word in every document.

When a information is required from a huge collection of information repositories a specific process is used. This process is called information retrieval. I.e. IR process helps the user find the information that matches their information required. User's required information is expressed as queries. Automated Information retrieval systems are used to reduce the information overloaded.

I. BACKGROUND OF THE WORK

A. IMPORTANCE OF PREPROCESING

It is a common practice to store majority of information related to on, business, official communication in governmental work, industrial interaction and so on in the electronic form using text databases.. Before we apply computational techniques on documents it is important to make the documents ready for processing. Data mining can be defined as the process of extracting previously unknown data. This data may contain potentially useful information. A subset of data clustering which uses concepts from machine learning , uses domain knowledge of information retrieval and few concepts from NLP is called document clustering. Document Preprocessing is one such method applied for text documents. Document Preprocessing plays a vital role in document grouping.

B. IMPORTANCE OF DOCUMENT CLUSTERING

Machine Learning approach may be used to solve real life problems. A popular machine learning algorithm is unsupervised learning which is used to draw inferences from datasets consisting of input data without labeled responses. Cluster Analysis is an unsupervised learning method used for finding hidden patterns or used to categorize items group wise. Modeling clusters are performed by choosing a similarity measure. A distance measure is defined upon metrics such as Euclidean distance. Data mining uses this concept of cluster analysis for sequence and pattern matching. Image Processing uses this concept for image

segmentation and Computer vision for object recognition. Documents are arranged into different groups called as clusters, where the documents in each cluster share some common properties according to defined similarity measure. Unstructured text present in documents in the form of a natural language can be analyzed using document clustering. To illustrate the concept an example of web pages can be taken. Web pages discussing about the same text content can be automatically grouped. For example, if page 1, page 2, page 3 , page 4 and page 5 are five web pages ; page 1 , page 2, page 3 have similar text content then they will be grouped together. This grouping is performed in an unsupervised manner. Two core concepts required here are similarity measure and algorithm to perform this grouping. An algorithm uses the concept of similarity measure and does an optimization to ensure similar documents to fall in the same group.

C. Document Clustering vs Document classification

Document classification is much different than document clustering. A priori information about class labels and their properties are available in case of classification. This a priori information is not available in case of document classification. Hence, classification falls in the category of supervised learning and clustering falls in unsupervised learning.

D. Application

Document Clustering is applied in various fields of business , science and technology. In the initial period document clustering was more used for improvement of the precision and recall values of an information retrieval system. Automatic generation of hierarchical clusters of documents has also been done by document clustering. Few major applications of document clustering are the following:

- Identifying similar documents: Given a search document x to find matching documents with x from a given collection.
- Organizing large collection of documents: Large collection of uncategorized documents can be automatically organized in taxonomy identical to one human would create for easy retrieval.
- Duplicate content detection and maintaining integrity: It can be applied for plagiarism detection , grouping of related articles
- Recommendation System: An user is recommended articles based on the articles already read by user.
- Search optimization: Efficiency of search engines can be improved as the user query can be first compared to the representatives of the clusters instead of comparing it directly to the documents.

E. DOCUMENT REPRESENTATION AND CHARACTERISTICS

A text document can be represented by using vector space model. A vector is considered as a point in the n-dimensional space. If the presence or absence of a term t in document d is considered then the resulting vector for the representation of the document will contain only 1's and 0's. Such vectors are called Boolean vector.

The distinguishing characteristics of the text representation are the following:

- Although documents have high dimensional text data but underlying data is sparse.
- High variation among word count of different documents. Hence, Normalizing of the documents is important before doing clustering.

F. DATA PREPROCESSING STEPS

The document is parsed through to find out the list of all the words. Document preprocessing can be categorized into following stages:

1. Parse sentences into terms to have a parsed collection.
2. Removal of Stop words: A customized list called Stopword List is used that contains the words to be excluded. The Stopword list is applied to remove terms that do not exhibit discrimination for topics
3. Word Stemming: A stemmer stems the different morphological forms of the word. A stemming algorithm such as porters is used to reduce a word to its stem or root form.
5. Global Unique words (very significant role) and frequent word (in a particular document) sets are generated.

Efficient and effective document clustering helps in the following activities.

- To navigate
- To summarize, and
- To organize the information
- To recommend
- To optimize search in search engines
- To find similar document
- To find plagiarism

G. COMPARISON OF BOOLEAN AND TF-IDF MODEL

A comparison between the two representations Boolean and frequency method is shown below:

Boolean model	TF-IDF model
This is based on Boolean logic and set theory	Frequency based model
Importance of the terms are not considered	Importance of terms are considered.
In the worst case it may reject all terms	Depends on the definition of the frequency
Data based representation	Content based

	representation is possible.
While doing clustering we may use hamming distance	While doing clustering we can not use hamming distance
The term document matrix can become a sparse matrix	It may not be sparse.

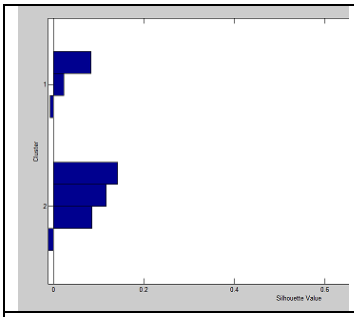
The tf-idf model is more advantageous than the simple Boolean model, as this model shows how important a word is, in a collection of keywords or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Hence, the efficiency of the information retrieval is enhanced by using the tf-idf model.

II. RESULTS

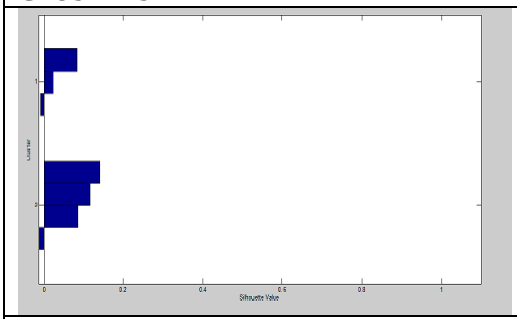
The evaluation of various clustering done on us TDM, TF-IDF, LOGARITHMIC and AUGMENTED frequency based data are shown in this section clustered by using Kmeans and

(I)
TDM FREQUENCY AND EVALUATION OF CLUSTERS

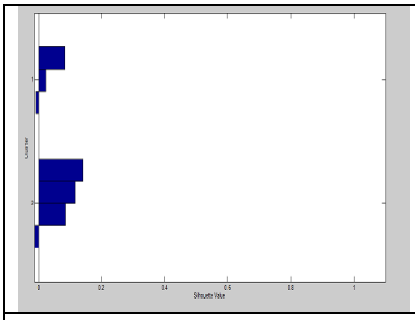
(2) tf-idf frequency evaluation clusters
3) logarithmic frequency based evaluation of clusters.
(4) augmented frequency based evaluation of clusters



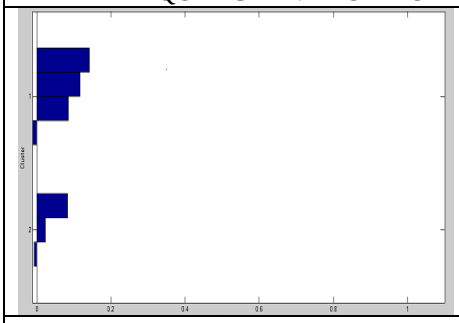
TDM FREQUENCY AND EVALUATION OF CLUSTERS



3) LOGARITHMIC FREQUENCY BASED EVALUATION OF CLUSTERS.



TF-IDF FREQUENCY EVALUATION CLUSTERS



AUMENTED FREQUENCY BASED EVALUA OF CLUSTERS

TABLE: Results of Comparison of evaluation of various term frequency and document frequency representation and clustering technique with different values of k.

Matrix created Using different types of frequency	K=2	K=3	K=4	K=5	K=6
TDM	0.2183	0.4963	0.5379	0.4904	0.7439
TF-IDF	0.7439	0.4837	0.6229	0.6200	0.7369
LOGARITHMIC	0.5154	0.4298	0.5426	0.6512	0.7480
AUGMENTED	0.5353	0.4252	0.4087	0.6666	0.7983

REFERENCES

[1] Mausumi Goswami, Gowtham, Balachandran, B. Purkayastha, "An Approach for Document Pre-processing and K Means Algorithm Implementation", IEEE conference held at Kochi in August,2014

[2] Mausumi Goswami, B.Purkayatha,"Term frequency and inverse document frequency based preprocessing" , international conference IAETSD: ICDER – 2015

[3] Stuti Karol, Veenu Mangat, " Evaluation of text document clustering approach based on particle swarm optimization," *Central European Journal of Computer Science, Springer research article on* , vol.3, issue.2, pp.66-90, 29th June 2013

[4] Huang JZ , Ng MK , Rong H and Li Z., " Automated Variable Weighting in k-Means Type Clustering," Pattern Analysis & machine Intelligence,, *IEEE Transactions on* , vol.27, no.5, pp.657-668, May 2005