Machine Learning Approach by Document Clustering using Probability of Word Occurrences

Aranga Arivarasan[1] Dr.M.Karthikeyan[2]

[1,2]Department of Computer and Information Science,Annamalai University,Annamalai Nagar, Tamilnadu, India

***Abstract** - Now a day the rapid increase in the fields of internet, data science, big data and data mining the extraction of hidden information from the documents become a challenging task. The text document doesn't have the flexibility of easily understanding the context for which it was written like images. So it is necessary to extract the correct features and similarity measures to categorize the document to extract the information. In the proposed work the probability values of occurrences of similar words from a document is extracted to categorize it to its topic. The method also uses an elaborated preprocessing technique for the dimensionality reduction as well as removal of unnecessary vectors from the text documents. The proposed method uses three similarity measures to evaluate the categorization. The final results show that the spearman similarity yields a better result with an accuracy of 95.7.*

***Keywords** – Probability, Similarity Metrics, pre-processing, Clustering, K-Means*

I. INTRODUCTION

In the resent years the bigdata and data science become a fast developing area because of the ability of handling huge tens of terabytes of data. Generally the big data is classified into three categories such as structured, unstructured and semi-structured data. In bigdata the text plays a vital role because of its simplicity of representation and minimum storage capability. The rapid development of data science as well as the communication technology leads to very easy transmission and ability of accessing the text. One of the major setback in the text mining like human the machine cannot easily understand the text content. Because of this drawback the automatic text clustering is a little bit of concern. Since the www efficient usage for several decades made text document clustering very widespread as well as /implementation in numerous application like web mail spam filtering, web user emotion analysis, customer commodity searching requirements etc. Text clustering is executed by representing the documents as a set of terms of indexes associated with some numerical weights. The goal is always to cluster the given text documents, in a way that they get clustered by means of the similarity measures with certain accuracy. There are many approaches are available for classification of text documents Naïve bayes, Support Vector Machines, DBSCAN, K-medoids, k-means and expectation maximization. The performances of the above said algorithms highly rely on the datasets provided to them for training. Before going to execute the text clustering the document representation approaches suffix tree representation of document analysis of similarity or distance metrics and most importantly the correct clustering approach are to be considered very carefully.

In some cases Clustering is wrongly referred as automatic classification. The clusters formed are not known before processing. The distribution and the characteristics behavioural of data will be utilized to identify the cluster members. But through classification the classes are always pre-defined as well the classification algorithm learns the relationship towards the target output through analyzing from training set. The training set is nothing but a group of data labeled with human interpretation to its corresponding class then used to identify the learning activities of an unlabeled raw data. Many decades of study is going on document clustering but still it needs the researchers inventions of new technique in solving many problems in document clustering. The most important challenge lies in selecting relevant features from documents that are considered to perform clustering. Exact utilization of similarity measure between training and testing documents relevant clustering method for utilizing the selected similarity measure in an efficient way to make the clustering feasible and determining a way to associate the quality of clustering performance will increase the accuracy of achieving target output. To solve these issues several clustering techniques are available namely Distribution based methods**,** Centroid based methods**,** and Connectivity based methods Density Models and Subspace clustering

The rest of the paper is organized as seven sections. In Section 2 the referred related works regarding Document clustering is elaborately. The Section 3 describes the various distance metrics used in this paper. In Section 4, the system overview is elaborated briefly. Section 5, evaluates the experiment results in detail section 6 produces the conclusion of the paper and section 7 gives the references made.

II. SIMILARITY METRICS

In document clustering similarity is typically computed using associations and commonalities among features where features are typically words and phrases. If any two documents are determined as similar they should share similar topic or information. When clustering is performed with documents we are very much eager in grouping the component documents with reference to the type of information that the documents contain. Achieving the clustering accuracy require a precise definition of the closeness between a pair of objects in terms of either the pair wise similarity or difference. A variety of similarity or distance measures are proposed and widely applied such as Spearman similarity correlation similarity cosine similarity Jaeeard coefficient Euclidean distance and so on.

A. Spearman Similarity

The documents are always represented as term vectors the similarity of two documents corresponds to the correlation between the vectors. The correlations between two sequences of term vector are measured through spearman Correlation. The two term vectors are ranked separately to determine the difference in rank between each position, *i*. The distance between sequences ***X*** = (***X****1*, ***X****2*, etc.) and ***Y*** = (***Y****1*, ***Y****2*, etc.) is computed using the following formula:

$$1 - \frac{6\sum_{i=1}^{n}(rank(X_i) - rank(Y_i))^2}{n(n^2 - 1)}$$

Where, ***Xi*** and ***Yi*** are the *i*th values of sequences ***X*** and ***Y*** respectively. The range of Spearman Correlation always lies between -1 to 1. Certain linear and non-linear correlations can be determined through spearman correlation.

B. Cosine Similarity

The wide spread common measure used in document clustering is the cosine similarity. Consider any two documents d_i and d_j, the similarity between them can be identified through the following formula.

$$\cos(d_i, d_j) = \frac{d_i \cdot d_j}{||d_i||\ ||d_j||}$$

where d_i, and d_j are m-dimensional vectors over the term set T= $\{t_i, \mathbf{t2}, \ldots t_m\}$ each dimension represents a term with its weight in the document which is non negative. Because of this always the cosine similarity is non-negative and lie between [0, 1]. The cosine similarity is irrespective of document length. When the document vectors are of unit length the above equation can be simplified as

$$\cos(d_i, d_j) = di.\ d_j$$

When the cosine value is 1 the two documents are identical and 0 if there is nothing in common between them. Since document vectors are orthogonal to each other.

C. Correlation Similarity

Correlation is techniques which explore the relationship between two quantitative continuous variables. There are many different forms of correlation coefficient. It is given by

$$(d_i, d_j) = \frac{m\sum_k d_{ik} - TF_i\, X\, TF_j}{\sqrt{[m\sum_k d_{ik}^2 - TF_i^2]\,[m\sum_k d_{jk}^2 - TF_j^2]}}$$

Where $TF_i = \sum_k d_{ik}$ and TF1 $=\sum_k d_{ik}$

The valuen ranges from +1 to -1. Positive correlations represent that both variables increase or decrease together. The negative correlations represent that as one variable increases the other variable decreases and vice versa. Two documents are identical when Pearson similarity is ±1. The spearman distance is a distance measure in other hand the cosine similarity and Pearson coefficient are similarity measures.

III. MATERIALS AND METHODS

A. Preprocessing

The clustering process depend on various preprocessing techniques to achieve quality and performance. The common preprocessing methods are described here. The main objective of preprocessing is to represent the data in a form that utilized to perform clustering process. Vector-Model, graphical model, TFIDF, Probability, keyword word count are some of the forms to represent the document. weighing of the documents and their similarities are measured by implementing various techniques. The importance of any word which appear within a document is always represented in a vector model through a numerical value is stored for each occurrences of word. The text mining approach highly relies with set of words a bag-of-words that the text document is efficiently represented. The text processing phase involves, after reading textual documents divides text document characteristic into tokens, words, terms, or attributes. The weight obtained from the occurrence of the terms in each text document followed by the removal of no informative attributes such as stop words, numbers and special characters. The rest of the terms are then minimized towards the root word during the error correction process. Despite removing no informative features, the size of a text document space may be too large. Certain constraints to minimize the size of character space of each document input text in addition with the occurrence of the features of each document. The purpose of this phase is to improve the quality of features extracted to represent the document and also to reduce the complexity of the mining procedure.

B. Tokenization

Tokenizations in our sense not only divide the tokens towards processing but also determine group of individual tokens to generate higher level interpretations. Tokenization converts a stream of characters into a sequence of tokens. A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. A type is the class of all tokens that containing the same sequence of characters. To perform the tokenization the document goes through three operations. The first operation is to convert the documents into the number of words equal to bag of words. The next operation is to remove an empty sequence appear in the document. This process is known as cleaning and filtering. Finally each text document is divided into a list of characteristics also called tokens. The tokens may also be called as words, terms, or attributes.

C. Stop words

Stop words is a list of frequently repeated tokens which emerge in every text document. The frequent tokens such as conjunction and pronoun need to be removed because it does not have any effect in the form of tokens and these words add a very little or in cases no value on the tagging process of a document representation. Some enormously common words that would come out to be of little value in helping documents matching to achieve the user need are excluded from the vocabulary lack. For some reason if the tokens are a special character or a number then those tokens should be removed. To identify stop words we can arrange our list of terms by frequency and pick the high common ones according to their absence of semantics value.

D. Stemming

Stemming is the progression to remove prefixes and suffixes from tokens. The process is conceded for reducing the resultant words to their stem. The stem need to be identified to the original morphological root of the word and it is usually suitable related through words map to the similar stem. This procedure is carried out to reduce the count of tokens in the feature space and advance the performance of the clustering when the different forms of features are stem into a single feature. The streaming process is carried using the following algorithm

Step 1: Eliminate plurals (-s) and suffixes (-ed or -ing).

Step2: If the vowel occurs in the previous step, replace y to i on the next word.

Step 3: From the step 3, Map double suffixes to single ones (-ization,-ational).

Step 4: Additionally, reduces the suffixes like (-full, -ness) etc.

Step5: Deducts (-ant, -ence) etc.

Step 6: If a word ends with a grammatical verb ending, then it has been removed.

Step 7: Finally, removes a (-e).

E. Bag of Words (BOW)

BOW is a simplified illustration used in data mining for information retrieval and document clustering. Bag of Word is a simplest method for feature identification and representation of text document. BOW process consists of the following steps,

Step 1: All documents is titled with the bag of terms, by a vector with one document for each term taking place in the whole gathering of tokens in document. Each vector has a corresponding value representing the occurrence of the term in the document.

Step 2: All document represent as a point in a vector space with one quantity for every term in the dictionary.

Step3: If a word does not emerge in a feature vector, that vector is set to the value null.

F. Word frequency count

An important set of metrics in text mining relates to the frequency of word count (or any token) in a certain corpus of text documents. However, one can also use an additional set of metrics in cases where each document has an associated numeric value describing a certain attribute of the document. One will first go through the process of creating a simple function that calculates and compares the absolute and weighted occurrence of words in a corpus of documents. This can sometimes uncover hidden trends and aggregates that aren't necessarily clear by looking at the top ten or so values. They can often be different from the absolute word frequency as well. Then It is simple to do the basic analysis and find out that your words are split 50:50 to measure the absolute frequency of words, and try to infer certain relationships. In this case, you have some data about each of the documents. The key word exists: in which case the assignment is done (adding one). Now the key exists, its value is zero, and it is ready to get assigned an additional 1 to its value.

Although the top word was in the first table, after counting all the words within each document we can see that other words are tied for the first position. This is important in uncovering hidden trends, especially when the list of documents you are dealing with, is in the tens, or hundreds, of thousands. With counted occurrences of each word in the corpus of documents, the weighted frequency can be obtained. This reflects how many times the words appeared to readers; compared to how many times used them.

G. Probability

The important contribution in this

proposed method is to find the probability distribution of similar documents to perform the clustering process. The proposed methods determine a unique probability distribution equation to achieve the most significant accurate clustering process. In the document clustering phase each document is selected from the dataset and by using the probability distribution function the corresponding probability of each unique word in that document is calculated for the purpose of clustering. For each word in the document the relationship between the selected document and the corresponding cluster is determined. Depending on the probability values the document which has the overall maximum probability value is assigned to that cluster.

$$P(D_i,C_j) = \sum_{i=1}^{k} P(D_{wt}) * P(w_i|C_j)$$

$$where, P(D_{wt}) = \frac{no{:}of\ wt\ count\ in\ \ selected\ Doc}{Total\ word\ count\ in\ selected\ Doc}$$

$$P(w_i,C_j) = \frac{No\ of\ wt\ count\ in\ cluster}{Total\ word\ count\ in\ cluster}$$

IV. EXPERIMENT RESULTS

For our experimental purpose the proposed system collected 300 documents for the five categories Business, Entertainment, Politics, Sports and Technology. Initially the proposed system splits the entire documents in the corpus in to individual tokens. Then all tokens are calculated to find the keyword occurrence. The calculated numeric values are tabulated in the Table 1. The context of document clustering relies on the commonly used similarity representing documents as normalized vectors. Each dimension of the vector corresponds to a distinct word in the collection of all words.

TABLE 1. TOTAL WORD AND UNIQUE WORD COUNT

	Distance Measures	Spearman Similarity	Cosine Similarity	Correlation Similarity
Probability	Accuracy	95.7	88.4	86.8
	Precision	9.45	8.84	8.77
	Recall	9.33	8.66	8.33
	F-Measures	9.39	8.75	8.54

Table 2. Results using probability of word occurrences

Category	Total Number Of Unique Words	Total Number of words
Business	6544	60786
Entertainment	8929	64622
Politics	7422	76277
Sports	6919	59930
Technology	8207	87550

With table. 1. Results probability of keyword occurrence was calculated. The determined values are formed in to clusters by means of K-Means clustering algorithm. To perform the clustering the K-means uses three Similarity metrics Spearman Similarity, Cosine Similarity and the Correlation Similarity. From the clusters the confusion matrix is determined. The confusion matrix is used to find the accuracy, precision, recall and F-measures.

Figure.1. Shows the word cloud chart of the five categories Business, Entertainment, Politics Sports and Technology. The word cloud is an image consist of words appeared in a selected document, in that image the occurrence of each word is indicated as its frequency and that is given the priority to perform the clustering. The tag cloud is a new technique of representing the text data as to visualize keywords and metadata in a document corpus. Tags are always single words, and the importance of each tag is shown with different font size and color.

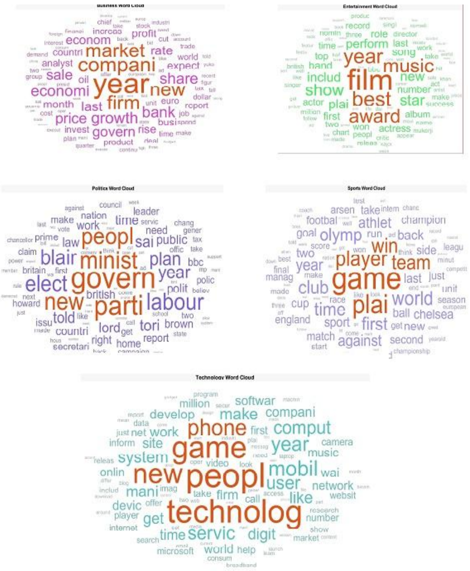

Figure1. Word Cloud of Business, Entertainment, Politics Sports and Technology

The precision, Recall and F-measures were calculated by using the following formulas

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ positive}{True\ Positive + False\ Negative}$$

$$F1 = 2\ X\ \frac{Precision * Recall}{Precision + Recall}$$

The calculated values are shown in Table 2. The Figure.2. Graphically represent the accuracy of the three similarity metrics for the Probability values. It clearly shows that the clustering operation gives better results by using the probability values for all the three similarity metrics. Among the similarity metrics the Spearman similarity gives better results than the other two Cosine Similarity and Correlation Similarity for Probability values.

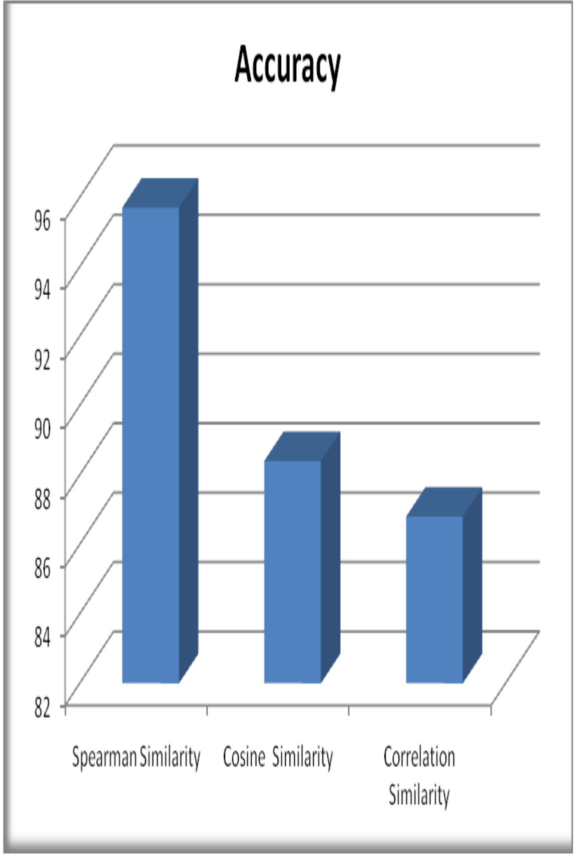

Figure.2. Bar chart showing the Accuracy

For our better understanding and demonstration purpose all the calculated results were shown in Figure.3. Our proposed system calculates Accuracy, Precision, Recall and F-Measures by using Probability values through K-Means clustering. The clusters are determined by using the three similarity metrics Spearman Similarity, Cosine Similarity and the Correlation similarity.

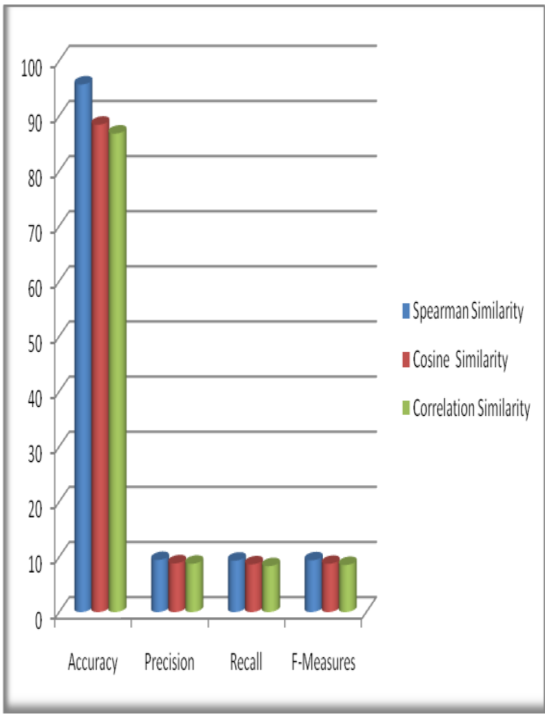

Figure.3. Accuracy, Precision, Recall and F-Measures.

From the Figure3. Results we can clearly understand that the Spearman Similarity measures achieves the best overall performance than the other two similarity measures.

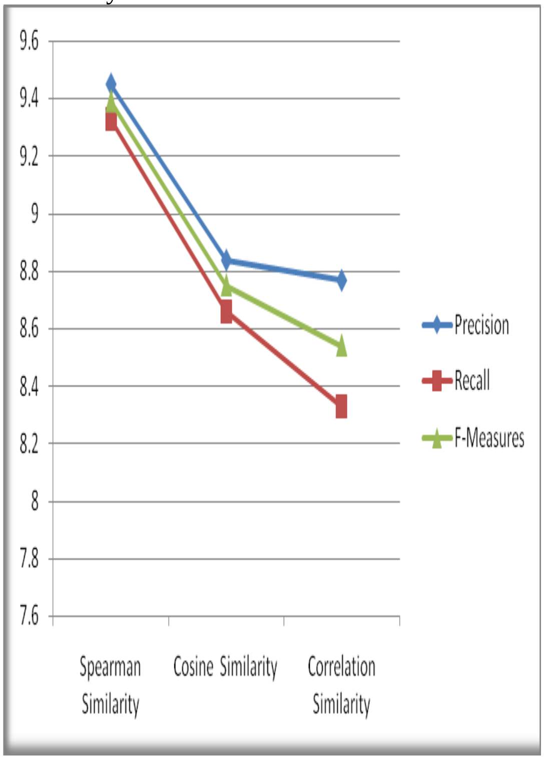

Figure.4. Performance Analysis chart.

Figure.4. describes the line chart of Precision, Recall and F-Measures determined by our proposed system. Precision, Recall and F-Measure are the external measures to analysis the performance of the system. Precision retrieves the number of correct assignments out of the number of total assignments made by the system. Recall retrieves the number of correct assignments made by the system, out of the number of all possible assignments. F-measure is a combination of the precision and recall measures used in system. From Figure4 shows our system yields better performance.

V. CONCLUSION

In this paper, the proposed Novel K-Means clustering approach is performed by the Probability of occurrences of words within the documents of five different categories. For each category we have taken 300 documents. The proposed system retrieves 6544 key words out of 60786 unique words for the business category. The Entertainment category documents retrieve 8929 key words out of 64622 unique words. The Politics category retrieves 7422 key words out of 59930 unique words. The sports category retrieves 6919 key words out of 59930 unique words. The Technology category retrieves 8207 key words out of 87550 unique words. The proposed model uses three similarity measures to compute the clustering operation. The Spearman Similarity measure yields an accuracy of 94.80.. Cosine Similarity measure yields an accuracy of 80.80. Correlation Similarity measure yields an accuracy of 74.40. The Spearman similarity gives better results than the other two Cosine Similarity and Correlation Similarity for the Probability of word occurrences.

REFERENCES

[1] Saqib Alam, Nianmin Yao, "Big Data Analytics, Text Mining and Modern English Language" journal of Grid Computing 2018

[2] Vladimer B. Kobayashi1, Stefan T. Mol1, Hannah A. Berkers1, Ga´bor Kismiho´k1and Deanne N. Den Hartog, "Text Mining in Organizational Research", Organizational Research Methods,21(3), 733-765.2018.

[3] Robert Wing Pong Luk, Kam-Fai Wong, Kui-Lam ACM Kwok "Interpreting TF-IDF Term Weights as Making Relevance Decisions", Transactions on Information Systems, Vol. 26(3) 2008.

[4] Dibyendu Mondal Pushpak, Raksha Sharma, "Comparison among Significance Tests and Other Feature Building Methods for Sentiment Analysis: A First Study", International Conference on Computational Linguistics and Intelligent Text Processing, pp.3-19, 2017.

[5] Kasula Chaithanya Pramodh, Dr.P.Vijayapal Reddy, "A Novel approach for Document Clustering using concept extraction", International Journal of Innovative Research in Advanced Engineering, 05(3),pp.59-65, 2016.

[6] Charu C. Aggarwal, ChengXiang Zhai. "A survey of text classification algorithms", Mining text data. pp.163–222, (2012)

[7] Borovikov, E. "A survey of modern optical character recognition techniques", Computer Vision and Pattern Recognition (2014)

[8] Bsoul, Q., Salim, J., Zakaria, L. Q. "An intelligent document clustering approach to detect crime patterns", Procedia Technology, 11, pp.1181–1187, 2013.

[9] Cohen Priva, U., Austerweil, J. L., "Analyzing the history of cognition using topic models", Cognition, 135, pp.4–9, 2015.

[10] Aranzabe, M. J., A. D. de Ilarraza & I. Gonzalez-Dios . "TransformingComplex Sentences using Dependency Trees for Automatic Text Simplificationin Basque", SEPLN, pp. 61–68. 2012

[11] Matthew Honnibal and Ines Montani. spacy " Natural language understanding with bloom embeddings", convolutional neural networks and incremental parsing. 2017

[12] Sowmya Vajjalla and Detmar Meurers "Readability assessment for text simplification: From analysing documents to identifying sentential simplifications", International Journal of Applied Linguistics, 165(2)pp.194–222, 2015.

[13] Yuqiang Tong, authorLize Gu, "A News Text Clustering Method Based on Similarity of Text Labels", Advanced Hybrid Information Processing,279 pp.496-503, 2018.

[14] Marzieh Oghbaie, Morteza Mohammadi Zanjireh, " Pairwise document similarity measure based on present term set", Journal of Big Data, 5:52,2018.

[15] Marmar MoussaIon, I. Măndoiu , "Single cell RNA-seq data clustering using TF-IDF based methods",BMC Genomics 19(Supl 6) : 569 ,2018

[16] Yehang Zhu,Mingjie Zhang, Feng Shi, "Application of Algorithm CARDBK in Document Clustering", Wuhan University Journal of Natural Sciences, 23:6, pp.514-524, 2018.

Aranga Arivarasan is a Research Scholar who is working as Assistant Professor in Division of Computer and Information Science, Annamalai University, India. He completed his B.Sc[Computer Science] and M.Sc[Computer Science] From Madras university in 1998 and 2000 respectively, the M.B.A and M.Phil[Computer Science] from Annamalai University in 2005 and 2007 respectively.

Dr.M.Karthikeyan is an assistant professor in Division of Computer and Information Science, Annamalai University, India. He completed his M.Sc[Computer Science] from Bharather University in 1993 and M.Phil[Computer Science] and Phd from Annamalai University in 2005 and 2014 respectively. His area of interest is Data Mining, Digital Image Processing, and Artificial Neural Networks.