Impact of Text Representation Techniques on Clustering Models

Ashwini K S (✉ ashwiniks.gemini@gmail.com)

Visvesvaraya Technological University

Shantala C P

Visvesvaraya Technological University

Tony Jan

Torrens University

Method Article

Keywords: Text Clustering, Text representation, Document clustering, Text mining, Document preprocessing

Posted Date: February 24th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1385057/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

In today's environment, an enormous amount of unstructured data is generated in an exponential manner. Understanding such complex unstructured data is imperative in the applications including analysis of social media data, image and video data, sensor data, medical data, and customer review data. Generally, clustering is a well-accepted model in classifying and analyzing such documents. An effective and efficient text clustering technique can significantly improve the task of document analysis and grouping with minimum human intervention. The main two factors in text clustering are the text representation model and the clustering algorithm. Firstly, unstructured data must be represented in a structured format for the analysis. Text representation models transform a large volume of text into vector representations by capturing the semantic information. In this paper, we are investigating the text document representation techniques, their limitations, and how text clustering is different from humans understanding of the text data. In this paper, we are aiming to focus on different text representation methods and clustering algorithms to demonstrate how the choice of representation techniques can impact on clustering results.

1 Introduction

Text data plays an important role in the information exchange and communication in today's organizations. A large amount of text data is generated by individuals, organizations, and governments. An estimate of more than 80% of the data generated is in textual format [1]. The enormous amount of complexity in text data make it difficult to analyze the textual data with minimum human interventions. The main aim of different natural language processing (NLP) methods is to achieve a human-like understanding of the text data. A successful such analytics can help to analyze large volume of unstructured and poor-quality data to discover the appropriate insights and hidden information [2]. Over the past years, text data has been used in numerous applications like Email Filtering, Documents organization, Sentiment Analysis, Opinion and Content mining, Speech Detection and many more.

Text Mining, also referred as Text Analytics [3] is the process of transforming unstructured data to structured format to discover the meaningful patterns, knowledge, and new insights. Text mining also has received great attention because of its high commercial potentials with the large amount of web data generated in a textual format [1]. Text Mining comprises several phases including pattern extraction, analysis of patterns, semantical interpreting and information extraction [3]. In the majority of text mining applications, text data must be represented in a structured format for analysis. Many text representation techniques have been introduced to meet the above requirements. The existing text representation techniques [4] are classified as Numerical, Graph-based, and Human cognition-based techniques. The unstructured content adds complexity to the model. Numerical based approaches are the most common techniques due to their simplicity and efficiency.

Among various text mining techniques, clustering is used most widely due to its applicability to categorize documents based on their content similarity. An efficient text clustering technique significantly

improves the performance of grouping, retrieval, and analysis. The performance of text clustering depends on text pre-processing and representation techniques. The language imperfections of text data may cause aberrations in applying appropriate pre-processing techniques. Besides understanding syntactical information, context and semantics is very important in text analysis. Many research have inferred structured data from text at character, word, sentence, or documents. This paper studies the impacts of the text representation on text document clustering and analysis. This paper investigates four different text representation methods including BOW, TF-IDF, Word2Vec [5] and Glove[4]. The experimental results showed that performance of document clustering was dependent on the selection of text representation technique that best represented the structure of the dataset.

The paper continues with the literature review of text representation methods followed by experimental results and analysis and conclusion in the subsequent sections.

2 Related Work

Several different text document representation techniques [15] [6] have been introduced and those techniques are classified as classical model, non-contextual model, and contextual model. For each of these models, various representation mechanisms are available. Classical model is the most popular and commonly used techniques because of its simplicity. The Bag-of-Words (BoW) [7][6] is one such a model where each document is represented as a set of words. In BoW, every term in the document is represented as binary variable. Index of corresponding word is represented by 1, whereas all other words are marked as 0. If number of terms increases, then vocabulary size would increase and thereby increasing the length of the vectors. The set of words generated by BoW ignores the semantic relation between the words in the document. This method is commonly used in the areas such as NLP, Information Retrieval (IR) and Computer Vision (CV). In BoW technique, weights can also be assigned to convey the importance of the relevant words in the document.

The vector space model (VSM) was originally proposed by Salton et al. [8] where documents were represented as a weighted term vector. In VSM, the documents are represented in the form of vectors, terms in the vectors represents words from the vocabulary. The dimension of vector depends on the vocabulary size of the document. The most popular techniques associated with this model are Term Frequency [TF] and Term Frequency-Inverse Document Frequency (TF-IDF). TF determines how often the term appears in the document. It calculates the weights based on the frequency of the relevant terms in the document. TF-IDF counts a term by considering frequency in the document and occurrence of relevant term in the document collection. TF-IDF is represented mathematically as:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \log\left(\frac{D}{df_t}\right)$$

Where t denotes the terms, d denotes each document, D represents the collection of input documents, df_t denotes sum of documents with term t .

The weight is inversely correlated to occurrence of the term in the document collection. The main advantage of using this method is that it is simple, versatile and suitable as a lexical feature [9].

Word2Vec [10] is a neural network model based vector representation of words. In this model, each word is mapped into unique vector which is represented by a column in a matrix [10]. The position of word in the dictionary is used to determine index in the column. This model is a two-layer neural network which embeds the words into continuous vector space. After training the input with this model, words with similar meaning group together in Vector space preserving the semantic similarities mathematically [11]. It is recommended to train the model with the large corpus to have a better representation of words. Therefore word2Vec have proved to be popular and useful in many NLP applications [12]. One of the key advantages of this model is it can represent documents with dense term vector which reduces the sparsity problem.

Glove [13] also referred as Global Vectors is an unsupervised learning model for obtaining dense vector representation of words. Training is done by aggregating the global word-to-word occurrence statistics and resulting representation shows the linear relation of word vector space. It is a count-based model because it counts how frequently the word appears in the corpus. The Glove based on matrix factorization where each element in the matrix X_{ij} depicts frequency with which the word w_i and word w_j co-occur in the given context window and it is extensively used in the task of text classification [12]. It is based on two step processes: initial step in creation of co-occurrence matrix using corpus then factorizing it to get all the vectors. The objective function is given below.

$$J = \sum_{i,j=1}^V f(X_{ij}) (w_i^T w_j' + b_i + b_j - \log X_{ij})^2$$

where V is size of vocabulary, X is co-occurrence matrix, X_{ij} is frequency of word i co-occurring with j in the corpus, w' is the context word embedding of dimension D . This model is simple, accurate and enforce to capture sub linear relationship in the vector space. In this paper, we use these methods to represent documents in document clustering. The experiment provides systematic research to compare the different representation techniques and its impacts on the final performance in document clustering.

3 Text Clustering

A massive amount of text data is generated at an exponential rate in many parts of online business. The goal of text clustering is to partition the similar text document into groups. Text documents that are grouped into single cluster shows high level of homogeneity and documents in separate cluster shows high level of heterogeneity [16]. The mechanism of clustering can be defined as follows:

1. Set of text document are selected as $D = \{d_1, d_2 \dots d_n\}$
2. A desired number of cluster value, K , is estimated
3. An objective function evaluates the quality of cluster. The objective function is to measure the similarity between documents.

Text clustering techniques can be classified as hierarchical clustering, agglomerative clustering, partition clustering and fuzzy clustering. Popular text algorithms include K-means, Density based spatial clustering of applications with Noise (DBSCAN), and hierarchical clustering .

K-means clustering is an unsupervised learning, and it is the most widely accepted clustering method. The K-means algorithm takes unlabeled text as input, group the data into k clusters and iterates the process until there is no better cluster. K-means algorithm perform the following two tasks:

1. Determines the initial k reference points.
2. Assign the data into closed reference point.

K-means algorithm is relatively simple to implement and handles the clustering of different shapes and size whereas it is very sensitive for the initial reference points and K value.

Hierarchical clustering [18] is another widely used unsupervised machine learning algorithm. As mentioned in K-means clustering, it does not require any predefined number of clusters and applicable to different types of data in the real-world text data. The two common types of hierarchical clustering are Agglomerative clustering and Divisive clustering. The agglomerative clustering [18] is also known as 'bottom-up' approach. In this approach, each document dataset is considered as a single cluster initially, these documents are combined with closest similar document and this process iterates until all the documents are merged into a single cluster or a pre-determined number of clusters. The divisive clustering also called 'top- down' approach starts with all documents in a single cluster and split it into new clusters. This splitting process is repeated until each document in single cluster or target number of clusters is reached.

DBSCAN [19] is a density-based clustering algorithm. DBSCAN outlines to discover the clusters for massive amount of data with different shapes and size which contains noise and outliers. The DBSCAN algorithm mainly uses two parameters: minimum number of points or threshold value clustered together and the distance measure used to locate the neighborhood point. DBSCAN can handle the clusters of arbitrary shape and is resistant of noise which is suitable to handle real world text data [19]. With DBSCAN it is easy to learn the dataset that exhibits wide differences with density, and it also shows good efficiency on large dataset.

4 Experimental Analysis

4.1 Dataset

In this paper, Amazon fine food review dataset [19] was used for experimental testing. The dataset contained the reviews of fine food with 568,454 reviews from amazon over 10 years of span till October 2012. Ratings, product, user information and plain text review were included in the reviews [19].

The data attributes are as follows: Product ID, User ID, Profile Name, Helpfulness Numerator, Helpfulness Denominator, score, time, summary, and text.

Table 1
Statistics of Dataset

Number of reviews	568,454
Number of Users	256,059
Number of Products	74,258
Timespan	Oct 1999 – Oct 2012
Number of Attributes	10

We trained various models with 6000 sample data points which has 12199 distinct words. To determine the positive or negative review, we considered the rating attribute. If the rating was between 1 or 2, the recommendation was considered negative, and the ratings of 4 or 5 were considered positive with rating 3 considered as neutral. Firstly, the basic pre-processing and data cleaning were completed on the corpus by removal of html tags, removal of punctuations and special character’s, convert the word to lower case, remove stop words and finally stemming the word. Next, text representative vectors were generated using BOW, TF-IDF, Word2Vec and Glove. Then, the data was clustered using K-Means, DBSCAN and hierarchical algorithm.

4.2 Experimental Results

Table 2. K-means Clustering Results

Clustering Algorithm	Text Representation Technique	SSE	Optimal K
K-Means	Bag-of-Words	436952.98	4
	TF-IDF	5794.53	3
	Word2Vec	4825.10	5
	Glove	6035.66	5

In this section, we compare the clustering performance between different text representation techniques given K-means clustering algorithm as shown in Table 2, in approaching an optimal K value. The results are presented in the following sub sections. Four representations techniques are compared using K-Means Algorithm for clustering and the results are shown in Table 2. As shown in the table, Word2Vec and Glove representation techniques showing the good results in terms of SSE and K value compared

with the other methods. We found the best K value using elbow method as shown in Fig. 1, Fig. 2, Fig. 3 and Fig. 4 respectively.

Second experiment we have considered is DBSCAN for clustering with the text representation techniques of TF-IDF and Word2Vec. In this work, the optimal Eps values were shown in Table 3. Using elbow method optimal Eps values are shown in Fig. 5 and Fig. 6.

Table 3
DBSCAN Clustering Results

Clustering Algorithm	Text Representation Technique	Optimal EPS
DBSCAN	TF-IDF	8
	Word2Vec	6

Next, hierarchical clustering is considered to start with multiple starting points and excluding outliers. We have experimented with different representation techniques and the results with 5 clusters are displayed in Fig. 7, Fig. 8, Fig. 9 and Fig. 10 respectively. The visualization of data clusters demonstrated that clear cluster boundaries were formed in some selected matches of cluster algorithms and text representation techniques which inform us the recommendation for future text document clustering.

5 Conclusion

In this paper, we have compared four different text representation techniques in the task of clustering low quality text. Word2Vec and Glove methods showing good results with K-means clustering algorithm for the given dataset. DBSCAN clustering algorithm is not performing well as with little change in eps values, number of core points in the cluster are changing drastically. DBSCAN is not appropriate clustering algorithm while clustering high dimensional data. Hierarchical clustering performed better with TF-IDF and Glove representation techniques. But performance showed different trends when changing the corpus data points and dimensions.

Our experiment investigated the performance of representation methods in clustering process. Future work can choose the data sets of different sizes and dimensions. Further different clustering algorithms can be used to test our conclusion about the behavior of different techniques.

References

1. Gupta, V., and G. Lehal, A Survey of Text Mining Techniques and Applications. Journal of Emerging Technologies in Web Intelligence, 2009. 1.

2. Blitsas, P. and M. Grigoriadou. Knowledge Representation for an Automated Normalized Answers Assessment System, Based on Text Comprehension Theories. in 2008 Eighth IEEE International Conference on Advanced Learning Technologies. 2008.
3. Ramanathan, V. and T. Meyyappan. Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism. in 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC). 2019.
4. Li, Z. and M. Zhou. Use semantic meaning of coreference to improve classification text representation. in 2010 2nd IEEE International Conference on Information Management and Engineering. 2010.
5. Le, Q. and T. Mikolov, Distributed Representations of Sentences and Documents. 31st International Conference on Machine Learning, ICML 2014, 2014. 4.
6. Gizlenmistir, Y. Comparison of text representation methods for author recognition. in 2018 26th Signal Processing and Communications Applications Conference (SIU). 2018.
7. Hotho, A., A. Nürnberger, and G. Paass, A Brief Survey of Text Mining. LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 2005. 20: p. 19-62.
8. Salton, G., A. Wong, and C. Yang, A vector space model for automatic indexing. Commun. ACM, 1975. 18: p. 613-620.
9. Bernotas, M., et al., THE PECULIARITIES OF THE TEXT DOCUMENT REPRESENTATION, USING ONTOLOGY AND TAGGING-BASED CLUSTERING TECHNIQUE. Information Technology and Control, 2015. 36.
10. Mikolov, T., et al. Distributed Representations of Words and Phrases and their Compositionality. in NIPS. 2013.
11. Zhang, Y. and B. Wallace, A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820, 2015.
12. Collobert, R., et al., Natural Language Processing (Almost) from Scratch. J. Mach. Learn. Res., 2011. 12: p. 2493-2537.
13. Manning, J.P.a.R.S.a.C.D. GloVe: Global Vectors for Word Representation. 2014.
14. Leskovec, J.J.M.a.R.P.a.J., Inferring Networks of Substitutable and Complementary Products. CoRR, 2015. abs/1506.08839.
15. Ashwini, K.S and C. Shantala, DOCUMENT CLUSTERING TECHNIQUES USING HADOOP FRAMEWORK: A SURVEY. IJRAR-International Journal of Research and Analytical Reviews (IJRAR), 2020. 7(1): p. 663.
16. Aggarwal, Charu C., and ChengXiang Zhai. "A survey of text clustering algorithms." Mining text data. Springer, Boston, MA, 2012. 77-128.
17. Mehrotra, Shashi, Shruti Kohli, and Aditi Sharan. "To identify the usage of clustering techniques for improving search result of a website." International Journal of Data Mining, Modelling and Management 10.3 (2018): 229-249.

18. Bouguettaya, Athman, et al. "Efficient agglomerative hierarchical clustering." Expert Systems with Applications 42.5 (2015): 2785-2797.
19. Ester, M., H. P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226-231. 1996
20. J. McAuley and J. Leskovec. [From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews](#). WWW, 2013.

Figures

Figure 1

K-means with BOW

Figure 2

K-means with TF-IDF

Figure 3

Figure 4

Figure 5

DBSCAN with TF-IDF

Figure 6

Figure 7

BoW with 5 clusters

Figure 8

TF-IDF with 5 clusters

Figure 9

Word2Vec with 5 clusters

Figure 10

Glove with 5 Clusters