International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195260

Hybrid SVD For Document Representation Using Different Vectorization

Kalpana P[1], Rosini B R[2], Sowmiya S[2], Sathya Priya K P[2]

[1]Assistant Professor, Department of Computer Science and Engineering, Sri Krishna College of Technology1, Coimbatore, Tamil Nadu, India

[2]UG Scholar, Department of Computer Science and Engineering, Sri Krishna College of Technology1, Coimbatore, Tamil Nadu, India

ABSTRACT

Document Clustering is the process of segmenting a particular collection of text into subgroups. Nowadays all documents are in electronic form, because of the issue to retrieve relevant document from the large database. The goal is to transform text composed of daily language in a structured, database format. In this way, different documents are summarized and presented in a uniform manner. The challenging problem of document clustering are big volume, high dimensionality and complex semantics. The objective of this paper is mainly focused on clustering multi-sense word embeddings using three different algorithms(K-means, DBSCAN, CURE). Among these three algorithm CURE gives better accuracy and it can handle large databases efficiently.

Keywords : SVD, Kmeans, DBSCAN, CURE

I. INTRODUCTION

With tremendous growth of Internet, more than quintillion(10^{18})data created each day. Among various types of data, text or document data accounts for large portion[1].To draw insights from unstructured text data, which is a goal of Natural Language Processing(NLP), the most important step is transforming these unstructured text data into numerical vectors. This problem is known as Representation Learning. In this paper we focus on representation learning of documents.

The goal of a document clustering scheme is to minimize intra-cluster distance between documents, while maximizing inter-cluster distance. A distance measure thus lies at the heart of document clustering. The large variety of documents makes it impossible to create a general algorithm which can work better in case of all kinds of datasets. One of the effective approach of document clustering is Fuzzy Bag of Words(FBoW).

A. Fuzzy Bag of Words

Fuzzy system has three modules fuzzification, inference, and defuzzification this also knows as knowledge based non-linear system. Fuzzy Bag of words is proposed to learn more dense and potent document representation encoding more semantics. FBoW replace hard mapping by fuzzy mapping[1], introduces vagueness in the matching between words and the basis terms. In FBoW word embedding technique is introduce to evaluate semantic similarity. Encode word meanings into vector and thus semantic similarity between towards can be evaluated using the cosine similarity. The fuzzy membership function is based on the similarity between words and the word clusters. Based on FBoW, Fuzzy Bag of Word cluster is proposed. FBoWC uses clusters of word as

the basis terms. Three variants named FBoWCmean, FBoWCmax, and FBoWCmin, respectively[1]. The main contribution of Fuzzy Bag of Words:

1. FBoW model reduces sparsity, improve potent and encode more semantic information, adopts Fuzzy mapping in which value can be determined according to word semantic similarity.
2. FBoWC produces representation with lower dimension than FBoW. It is constructed basedon similarity between words in corpus and clusters.

The cosine similarity can be interpreted as the degree of one word semantically matching another word. **d1** "the dog bark on sun" and **d2** "the cat licks a boy ". If the FBoW employs the following basis terms: {"cat", "bark", "dog", "boy"}, the two sentence d1=(1,0.7,0.8,1) and d2=(0.7,1,1,0.8) respectively as shown in Figurer 1.(a).For sentence d1 due to fuzzy mapping of BoW, the informative word **sat** is neglected. In summary fuzzy BoW model may not capture semantics of document and this lead to poor performance in classification and regression problem.

Representation Learning: The fuzzy membership function is adopted to count the number of occurrence of terms in a document. FBoW document representation is denoted by

x = [x1,x2,x3……..xi] where i-th element zi is the sum of

membership degrees that all words semantically match the i-th term i.e

$$zi=ci\ E\ Ati(wj)xj$$

where **w** denotes a set of all words in the document, ti is the i-th term and xj denotes the number of occurrence of wj[4]. Fuzzy Bag of Words clusters the data only based on three variants(min, mean ,max), it clusters small amount of data(min) or large amount of data(max) so the mean varies highly according to clusters.

In this paper, we propose a Hybrid SVD for document clustering. To overcome the limitation of original Fuzzy Bag of Words, we replace fuzzy mapping by Hybrid SVD (SVD-Kmeans,SVD-DBSCAN,SVD-CURE). Hybrid SVD introduce vagueness in matching between words and basis term.

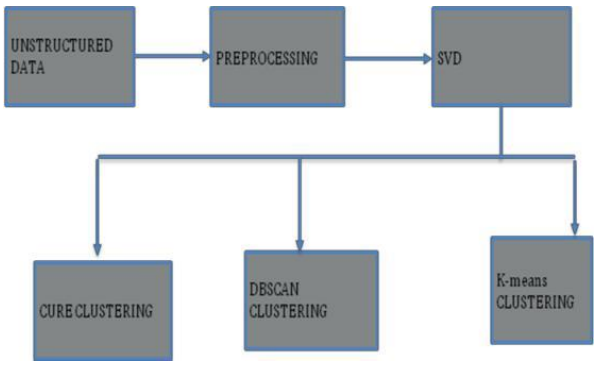


Fig (1) Block Diagram of Hybrid SVD

II. RELATED WORK

A. Document Preprocessing

The aim of preprocessing is to neglect all characters and terms with poor information from the document that affect the quality of group descriptions[5]. The first process is removing **Stop Words** that has no information and meaning such as pronoun, preposition etc… Stop words were removed automatically by **TF-IDF** vector which has natural capability of removing stop words.

B. Singular Value Decomposition

The singular value decomposition is commonly used in the solution of unconstrained linear least square problems, matrix rank estimation and canonical correlation analysis[2].A

large dataset is divided into smaller ones each contains data that are close in sense.SVD constructs n dimensional abstract semantic space in which term and document is represented as vectors.

$$A=T\sum D^{-1}$$

where

T= $m \times m$ unitary matrix.

Σ = m × n diagonal matrix

D = n × n unitary matrix .

D* is the conjugate transpose of the n × n unitary matrix.

C. Latent Semantic Analysis (LSA)

LSA is an extension of the vector space model that uses SVD which divide document into smaller dimension[4]. LSA can be used to reduce the dimension of document usingtruncated SVD. Let A be *m x n* **word** matrix, LSA decomposes matrix into m x p orthogonal matrix T, p x p matrix pseudo diagonal ∑ and p x n orthogonal matrix D^{-1}.T is word-latent semantic matrix, ∑ is latent semantic matrix, D^{-1} is latent semantic term.

Output of SVD

Performing dimensionality reduction using LSA

SVD data

[[0.9411481 0.33799447]

[0.92624091 -0.37693205]

[0.82101286 -0.5709097]

.....

[0.74088978 -0.67162664]

[0.97188458 -0.23545779]

[0.8110059 -0.58503797]

done in 0.514801s

Explained variance of the SVD step: 1%

D. K-means

K-means is an iterative algorithm where the results depend on the initial value of centroids. The goal of using K-means is to group a set of data into K-clusters with K-values[4]. K-means collects the vectors from the SVD and calculate the centroid values. It produce different results for every stimulation, we run the method several times until centroid values do not change[2]. It tries to make intra-clusters as similar as possible. K-means clusters large data and produce high clusters.

Output of SVD-Kmeans

Homogeneity: 0.168

Completeness: 0.194

V-measure: 0.180

Adjusted Rand-Index: 0.134

E. DBSCAN

DBSCAN(Density Based Spatial clustering of Applications with noise) is a process of grouping the points together based on a set of points that are close to each other based on a distance measurement (Euclidean distance) and a minimum number of points[12]. It also makes as outliers the points that are in low-density regions[3]. It is used to identify clusters of any shape in the dataset containing outliers and noise. DBSCAN requires two important parameters.

dbscan =(eps, minpts)

EPSILON (EPS)

eps defines the radius of the neighborhood around a point x called as neighborhood of x[8]. It means that if the distance between the two points is lesser or equal to the value (eps) then these points are said to be neighbors.

MINIMUM POINTS (MINPTS)

minPts is the minimum number of dense regions. The minimum values of minPts must be 3. If the dataset is larger then the value of the minPts should be chosen according to that.

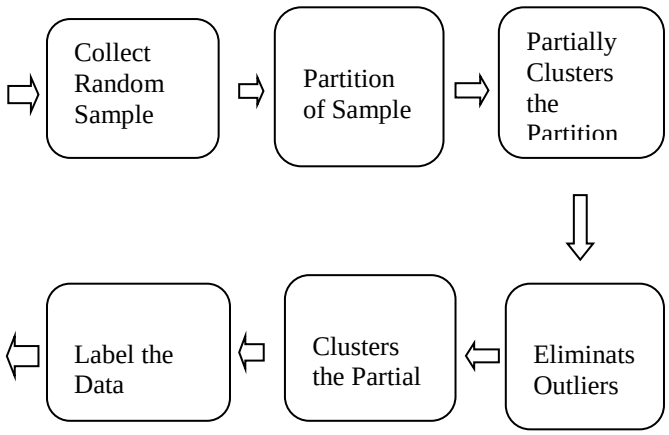
By using this parameters DBSCAN calculates the cluster value. After the process DBSCAN collects the dataset from SVD and performs its process. The dataset is clustered according to the condition in of the parameters. In the first partition the eps value is found by defining the neighborhood of eps and with the usage of minPts the clustering process takes place as the second partition. Unlike other algorithms[7], DBSCAN does not require the user to specify the number of clusters to be formed. DBSCAN can find any shapes of clusters and can identify outliers.

Output of SVD-DBSCAN

Homogeneity: 0.000
Completeness: 1.000
V-measure: 0.000
Adjusted Rand-Index: 0.000

F. CURE(CLUSTERING USING REPRESENTATIVES)

CURE is a hierarchical clustering algorithm that creates a balance between centroid and all point approaches. A constant number c of well scattered points in a cluster are chosen as representatives and those points catch all the possible form that could have the cluster[12]. The clusters with the closest pair are the cluster that are merged at each step of CURE. It is more robust to outliers and identifies clusters having non-spherical shapes and wide variances in size.



Fig(2) CURE Process

CURE is achieved by representing each cluster by certain fixed number of points that are generated by selecting well scattered points from the clusters and then shrinks them by a specified fraction towards the center of the cluster[6]. It allows CURE to adjust to the geometry of non-spherical shapes and shrink helps to dampen the effect of outliers if there are more than one representative point per cluster[13].

The above figure shows how the CURE process is performed. To handle large datasets, CURE uses a combination of random sampling and partitioning. In first partition a random sample is drawn from the dataset and each partition is partially clustered[11]. The desired cluster is yield from the partial cluster which are clustered in the second pass. The output confirms that the quality of clusters produced is much better than the existing algorithms. A combination of partitioning and random sampling enables CURE to not only outperform other existing algorithms but also to

scale well for large databases without out letting the clustering quality[10].

Output of SVD-CURE

Homogeneity: 0.008
Completeness: 1.020
V-measure: 0.00
Adjusted Rand-Index: 0.100

III. RESULTS

A .Dataset Description

In this paper, we took dataset from 20 Newsgroups which is a collection of 20,000 newsgroup documents, the collection of dataset as become popular dataset for experiments in machine learning techniques such as text classification and text clustering, we use dataset directly by logging into 20Newsgroups(from browser), we received 30,000 dataset in 3871 documents.

B. Data Characteristic

Nearly One thousands usenet articles were taken from 20Newsgroups.In this paper, we have used these four characteristic(alt.atheism,comp. graphics,talk.religion.misc, rec.sport.baseball).

C. Experimental Results

Measures of Hybrid SVD

1.Homogeneity

Homogeneity is the quality or state of having a uniform structure throughout the dataset (i.e) homogeneous. It is the main process of checking the homogeneous (identical) property of document

2.Completeness

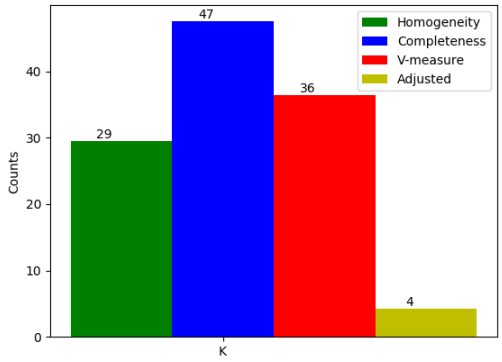
Completeness is the state or condition of having all the necessary and appropriate paths of documents. It checks all the paths are in correct positions or not.

3.V-measure

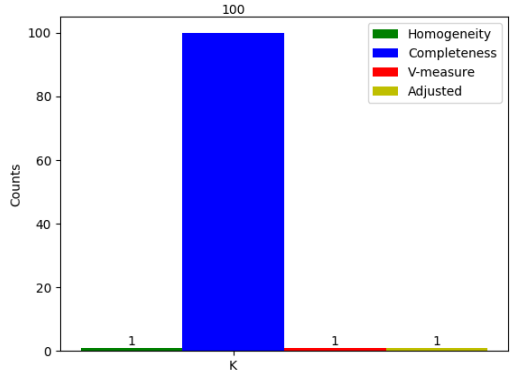
V-measure is the mean between homogeneity and completeness to group all the elements in single cluster.

4.Adjusted RandomIndex

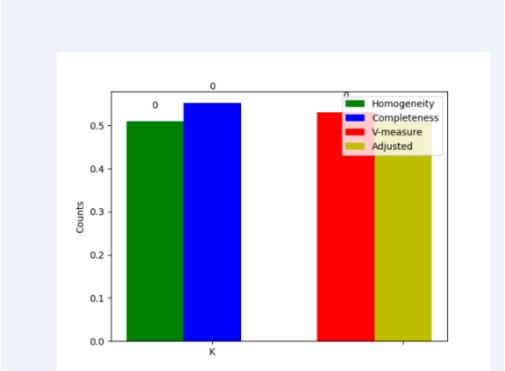
Rand index is the number of pairs that are either in same group or different group which lies between 0 and 1.



Fig(3) Hybrid SVD K-MEANS



Fig(4) Hybrid SVD DBSCAN



Fig(5) Hybrid SVD CURE

IV. CONCLUSION

In this paper, we have observed the effect of document clustering result using different vectorization method. Document representation gives more impact on classification and clustering results because it captures more semantics of document and also reduce the problem of high dimensionality. In above three algorithms used CURE gives better result in clustering (i.e) it gives more accuracy. As a next step work, the effect of multi sense word embeddings will be proposed in future.

V. REFERENCES

[1]. Rui Zhao and Kezhi Mao "Fuzzy Bag Of Words Model for document Representation" in IEEE transaction on Fuzzy System , vol.26,No. 2, April 2018.

[2]. Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, Madalina Persu" Dimensionality Reduction for k-Means Clustering and Low Rank Approximation" April 2015.

[3]. R.janai and Dr. S.Vijayarani "An Efficient Algorithm for document Clustering in Information Retrieval " Vol 4,Isuue XII, December 2016.

[4]. Michal Aharon, Michael Elad, and Alfred Bruckstein"K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation" IEEE Transaction On Signal Processing, Vol, 54, No. 11, November 2006

[5]. Paul S. Bradely and Usama M. Fayyad "Initial points for K-mean Clustering". In Proceedings of the 15th International Conference onMachine Learning(ICML98),1998.

[6]. Kiri Wagsta, Claire Cardie, Seth Rogers, Stefan Scroed "Constrained K-means Clustering with Background Knowledge." Proceedings of the Eighteenth International Conference on Machine Learning, pages 577584, 2001.

[7]. A.Hotho, S.Staab and G.Stumme,"Wordnet improves text document clustering" In Proceedings of the SIGIR Semantic Web Workshop, Toronto, 2003.

[8]. A.K.Jain, M.N.Murty and P.JFlynn. Dataclustering:Review. ACMcomputer surveys(CSUR,31(3):264-323,1999).

[9]. Bjornar Larsen and Chinatsu Aone "Fast and Effective Text Mining Using Linear-time" In Proceedings of the fifth ACM SICKDD International Conference on knowledge Discovery and Data Mining,1999.

[10]. D.D Lewis Reuters-21578 "Text Categorization text collection distribution" In proceedings of ACM SIGKIDD on 1999.

[11]. Lin "Divergence measures based on the Shannon entropy", IEEE transaction On information theory, 37(1):145-151-1991.

[12]. D. Arthur and S. Vassilvitsku" K-means – the advantage of careful seedings". In symposium on discerete algotithm, 2007.

[13]. D.Milne, O.Medelyan, and I.H.Witten. Mining domain-specfic thesauri from Wikipedia: A case study. In Proc. Of the International Conference on Web Intelligence (IEEE/WIC/ACM WI'2006),2006.

Cite this article as :

Kalpana P, Rosini B R, Sathya Priya K P, Sowmiya S, "Hybrid SVD For Document Representation Using Different Vectorization", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 388-393, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT195260

Journal URL : http://ijsrcseit.com/CSEIT195260