International Journal of Engineering & Technology, 7 (2.7) (2018) 1096-1099

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Finding author similarity by clustering probabilistic LSA factors in INDIAN english authors poetry

K Praveen kumar [1] *, Venkata Naresh Mandhala [2], Sudheshna Vempati [2], Dr. Subba Rao Peram [1]

[1] Department of IT, VFSTR University, Guntur, A.P., Guntur INDIA
[2] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur Andhra Pradesh
**Corresponding author E-mail: subbarao.peram@gmail.com*

Abstract

High dimensionality and sparseness is the big challenge to the data scientists to discover the similarity among the documents. In unsupervised learning data is unlabeled and there is no clear distance measures to discover the clusters among the data. In this paper we considered Indian English Authors poems to cluster them using Probabilistic Latent Semantic Analysis, using which we analyzed the authors similarity. We compared the results of clustering with Latent Semantic Analysis method, a word occurrence method. In this case, Results are shown that probabilistic methods are performing good clustering than the word occurrence method.

Keywords: *LSA; PLSA; Word Occurrence; High Dimensionality*

1. Introduction

In text Analytics often a document is represented as word vectors. As words are considered as features, the feature space will be very large, but a document containing words is a small subset in the feature space, therefore the document representation is very sparse i.e. a document vector contains many zeros[1]. High dimensionality representation causes big challenge to the data scientists to cluster the documents or to find the similarity among documents. One solution is to use dimensionality reduction techniques such as LSA. Latent Semantic Analysis reduces the dimensionality by preserving the greatest variation among the word features by generating latent concepts [2,3]. But LSA suffers from synonymy (two different words having same meaning) polysemy (a single word has multiple meanings and it is determined by context). Poetry is a literary art, to find the similarity among poems using LSA is unrealistic.

As we know, to express the same semantics, poets use similar kind of words in a similar style. This implies that different thematic poems are distributed in different subspaces spanned by the co occurrence words on related themes[7]. We can find the similar poems by finding themes, themes can be found using PLSA by reducing the dimensionality. PLSA is a classical aspect model for finding latent factors to perform dimensionality reduction.

In aspect model, themes are relevant to co-occurrence features and related to a particular aspect. To identify the themes one needs supervised learning strategy. But for unsupervised learning there are now obvious methods to find the number of themes directly to cluster the text. The idea is, for x number of themes, if co-occurrence among words is more and less overlapping of themes then that number of clusters are enough to make good clusters [4]. We did experiment using K-means and Partitioning around medoids on the data set using PLSA factors. When it compared to the LSA, PLSA performed well on clustering using PAM algorithm.

The remaining paper is arranged as follows. In section 2 we discuss about LSA and PLSA methodology. In section 3, we present the results. In section 4 we compare the results based on poets and poems. At last we conclusions are provided.

2. Motivation

2.1. Latent semantic analysis

Assume that we have been given a collection of documents. D= {d1, d2….dn}, with terms weighted using TF-IDF from words W= {w1, w2…wm}. By using Bag of words approach that ignores the ordering of words, we can summarize the documents in a N×M co –occurrence table.

In general N×M table entries can be term w how often occurred in document d, called TF or a better measure TF-IDF.

In this case N×M can be called as term document matrix, where rows are terms and columns are documents. A term document matrix represents each document as a term vector, where if term w appears in the document that cell will have a value 1 otherwise it is noted with a 0 value. Normally the corpus contains more number of words than a document, so every document vector contains more number of zeros and it is called as sparse data.

Idea of LSA is to map the documents to a reduced vector space called latent semantic space [5, 6]. The matrix is restricted to linear space and it is performed using Singular Value Decomposition.

SVD divides the matrix in to UEV matrices, where U&V are orthogonal matrices i.e. UTU=VTV=I and E matrix contains singular values of N.

LSA approximation is found by taking highest K values of E, which represents the maximum variance of matrix. UE2 is defined as coordinates of the document in the latent space. When we compare the latent vector space and original vector space the sparseness is low in latent vector space than high dimensional original space. Using the latent vector space we can compute meaningful associations among the documents

2.2. Probabilistic LSA

PLSA is a statistical model, called aspect model. Aspect model is a latent variable model for co-occurrence data which associates an unobservable class variable zϵZ={z1,z2…zk} with each observation[7,8,9].
A joint probability model over DxW is defined by

$$P(d,w) = P(d)P(w|d), P(w|d) = \sum P(w|z)P(z|d)$$

This model is well known for dimensionality reduction, PLSA can discover the latent semantic factors of text to express the documents in lower dimensional semantic space instead of higher dimensional space.

3. Methodology

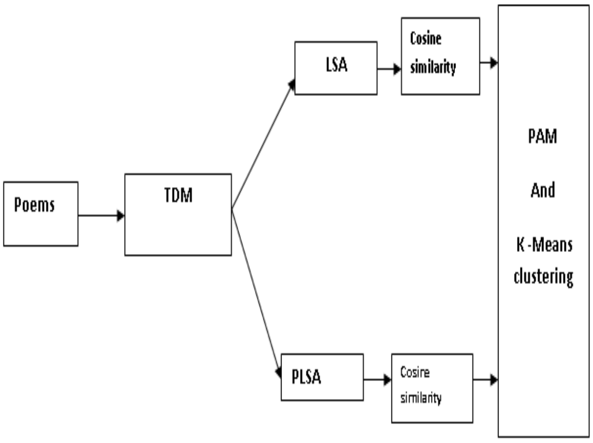

Fig.1: The Procedure Followed to Visualize the Poems Similarity on Vector Space.

Above show Fig.1 describes the procedure followed to perform experiment. First we collected 260 from 30 Indian English Authors poems from poemhunter.com website. Performed cleaning by removing stop words, then applied stemming to find the root word finally we created Term Document Matrix using TF-IDF weighting factor. TF-IDF is Term Frequency and Inverse Document Frequency it gives the important words by removing redundant and un important words. We performed 2 experiments one is finding LSA using Singular Value Decomposition (SVD) it gives the latent semantic space by reducing the higher dimension space in this case it is 5402 terms and 260 documents. After generating the lower dimension space of latent semantics applied cosine distance measure and finally performed clustering.
In second experiment on the same dataset we generated PLSA factors using topics 3, 4, 5 to find the less overlapping among topics we tried with these 3 numbers. Then cosine distance similarity calculated and finally applied clustering methods to find the similar documents.
In unsupervised clustering we do not know how many good clusters we can make; for this if we use subspace based clustering the results will be more appropriate. Here we followed 3 step procedure, First we calculated PLSA factors for various topic numbers, the we calculated correlative degree between topics and data finally for which number of topics correlative degree is high and overlap of subspace degree is low considered as appropriate topics.

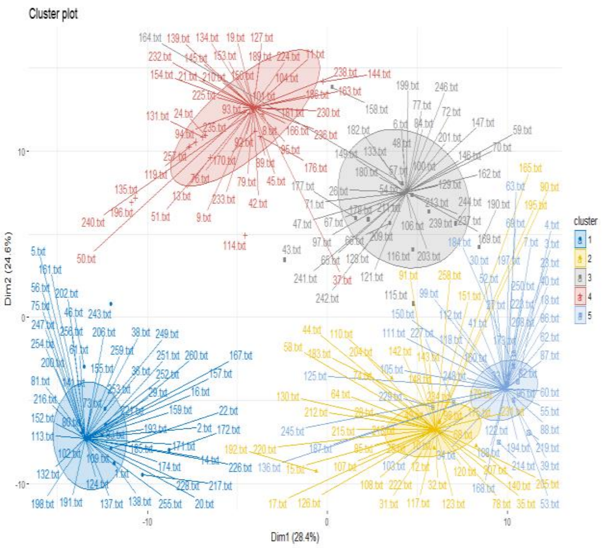

Fig. 2: Cluster Plot of five Topics Clustered in to five Clusters Using PAM.

Table 1: Cluster Information of five Topics Cluster.

S.No	Cluster Size	Max_Diss	Av_Diss	Diameter	Separation
1	58	3.737157	0.3390078	4.56232	4.532418
2	50	3.302203	0.219416	3.775485	5.136211
3	49	4.84917	0.5549877	5.997573	0.585485
4	52	5.23949	0.6881821	6.310162	0.585485
5	51	3.231748	0.3120559	3.840314	5.548014

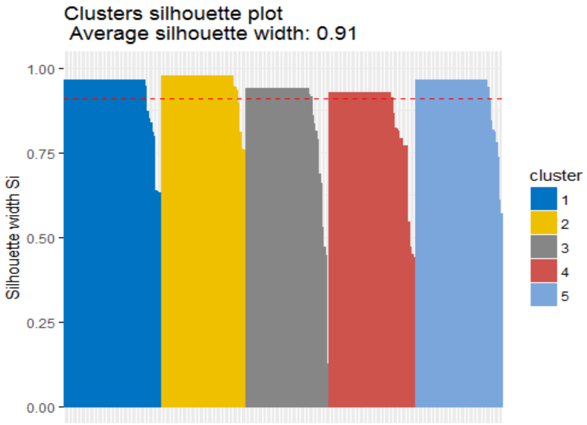

Fig. 3: Silhoutte Plot for 5 Clusters Using PAM.

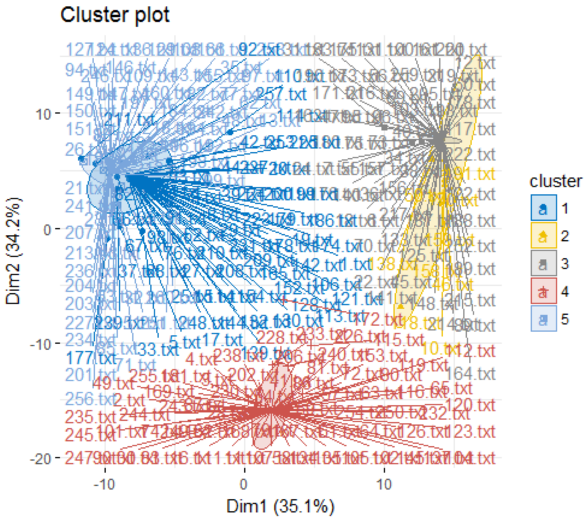

Fig. 4: Cluster Plot of 4 Topics Clustered in to 5 Clusters Using PAM.

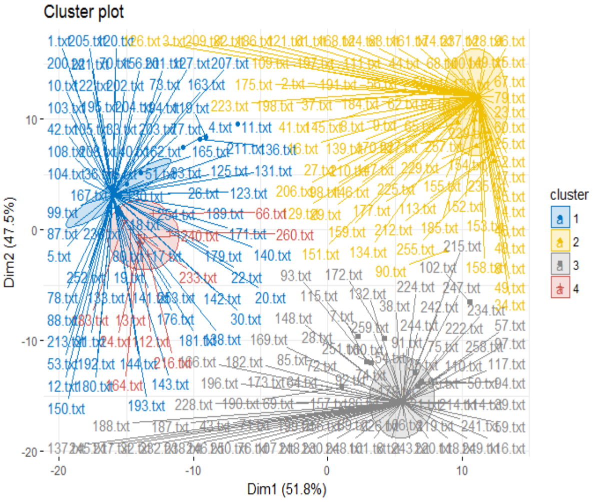

Fig. 5: Cluster Plot of three Topics Clustered in to five Clusters Using PAM.

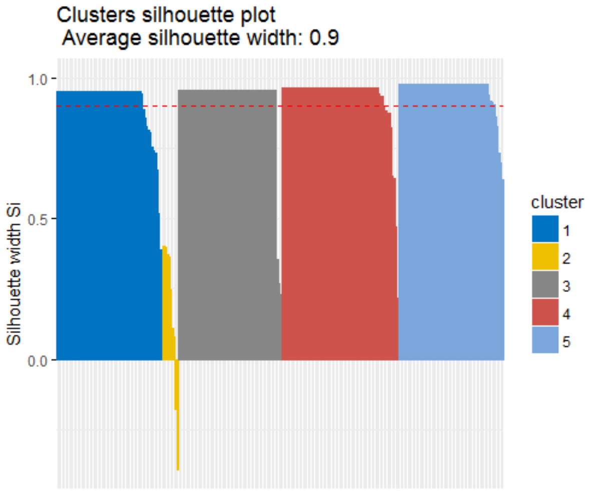

Fig. 6: Silhoutte Chart for Topics 4 PAM Clustering.

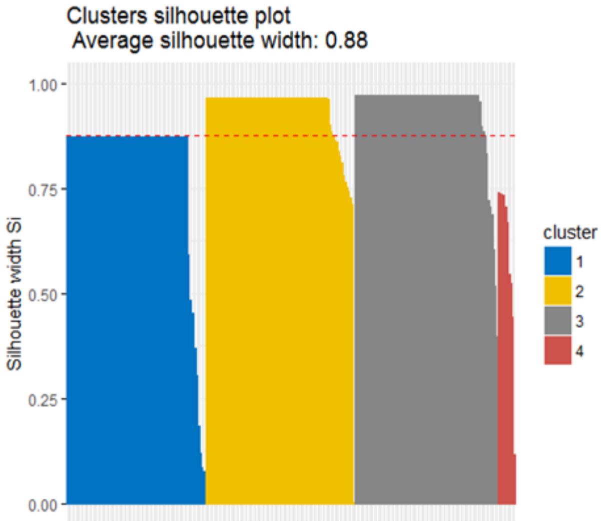

Fig. 7: Cluster Silhoutte Chart for Topics 3 Clustering.

Table 2: Cluster Information for three Topics Clustering

S.No	Cluster Size	Max_Diss	Av_Diss	Diameter	Separation
1	81	5.57507	0.298514	5.615585	0.3
2	86	7.086204	0.420669	7.086204	2.5
3	83	5.145401	0.325346	5.594416	2.5
4	10	1.471686	0.59245	2.158019	0.3
5	81	5.57507	0.298514	5.615585	0.3

4. Results and discussions

Fig 2, Fig3, Fig4 are the cluster plots for topics 5, 4, 3 respectively. These cluster plots will tell how the poems are clustered based on distance of their probability of occurrence. For this we used cosine distance metric. In order to find the best number of clusters we can use elbow curve method, this curve is drawn on the data on which we want to apply clustering method. Elbow curve will give us the point where the data diversity is preserved. Using that point we can choose the optimum number of clusters. In our case it has given that 4 clusters are optimum. Table 1, table 2 describes about the cluster characteristics, by studying these parameters we can come to a conclusion that how best the clusters are formed.

Categorizing the large number of documents in some meaningful way is a challenging task. If documents are predefined in to some classes, then based on those classes we can cluster them, but if such classification scheme is not available and the document corpus is very large then we need to take the help of clustering algorithms to classify the documents automatically based on their structure and content. After performing the clustering visualizing the cluster is important. For visualization of clusters one of the best ways are using cluster plots, these plots will reduce the dimensionality of the data and shows in a 2 dimensional scale.

Cluster information table contain the entities such as Maximum distance, Diameter, separation. Maximum distance will represents the maximum distance between any 2 vectors in the same cluster. Diameter will represent the largest dissimilarity between any 2 pairs of observations with in a cluster. Separation will tell about the minimum dissimilarity between 2 cluster observations.

Our observations are 5 topics based PLSA factors are clustered in a good manner when we compare it with 4 and 3 topics. And Fig3,Fig6, Fig7 shows the silhouette plots, these plots will tell the separation among the clusters, if its value is high then those are good clusters. We performed the clustering with LSA and PLSA using k-means and Partitioning Around Medoids (PAM) algorithms the results are better for PLSA and PAM than other methods.

When we look at clusters of Fig2, the cluster 1, contains poems still life, Chicago zen of AK Ramanujan and Land of Agha shahid ali and further amol redjis lost glory, Aravind mehrotras inscription and continuities. These all poems are having close meaning; they mostly talk about past life of a human being. In cluster 2, poems On the death of a poem and shaving describes similar feelings which are written by AK ramajujan and Agha shahid ali respectively. AK ramajujan describes the innermost corner of a poet, who has a close relation with the poetry. He describes how a poem feels bad when it is unnoticed and it is considered as the death of that poem. Shahid ali describes In shaving the hair dropped in the wash basin describes as the unwritten poetry or the thoughts which have not taken a form of poem.

Arundhathi subrahmaniam unveils her past feelings in the poems small questions and the city and I. In poems luminous, for the record, enemy and curious the author unveils his dark feelings of heart, In poem enemy he describes his he has no human heart and that is replaced with a grenade.

5. Conclusion

In this paper, we have experimented with Indian English author’s poetry with help of LSA and PLSA factors. The results shown that PLSA factors method outperform LSA in clustering the poems by using the probability values as weights than the TF-IDF weights. We also concluded that to find best number of clusters we can use elbow method and PAM method is showing good clustering results than the K – means clustering method. This work can be extended by changing the weighting factors from TF-IDF to word vectors, and with the help of deep learning methods, we can achieve better learning of poems and further grouping of poems.

References

[1] Capasso, Vincenzo, and David Bakstein. "Fundamentals of Probability." In An Introduction to Continuous-Time Stochastic Processes, pp. 3-76. Birkhäuser, Boston, MA, 2012.

[2] Chua, Freddy Chong Tat. "Dimensionality reduction and clustering of text documents." Singapore Management University (2009).

[3] Hornik, Kurt, and Bettina Grün. "topicmodels: An R package for fitting topic models." Journal of Statistical Software 40, no. 13 (2011): 1-30.

[4] Hofmann, Thomas. "Probabilistic latent semantic analysis." In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pp. 289-296. Morgan Kaufmann Publishers Inc., 1999.

[5] Hosseinia, M., K. Badie, and A. Moeini. "Aspect-Oriented Document Clustering for Facilitating Retrieval Process." International Journal of Computer Theory and Engineering 4, no. 5 (2012): 707

[6] Lancia, Franco. "Word co-occurrence and similarity in meaning." Mind as infinite dimensionality. Charlotte, NC: Information Age Publishers (2007).

[7] Scheaffer, Richard L., Madhuri Mulekar, and James T. McClave. Probability and statistics for engineers. Cengage Learning, 2010.

[8] K. Praveenkumar., T. M. Padmaja, “An Analysis on Computational Approach for Finding Similarity in Indian English Authors Poetry”, International Conference on SMART DSC-2017, Vignan Institute of Information Technology, Visakhapatnam, on November 30 to December 02, 2017, Advanced Science and Technology Letters,Vol.147 (SMART DSC-2017), pp. 193–203, 2017.