International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7, Issue-6, March 2019

A Study on Dimensionality Reduction Methods for Finding Similarity in Indian English Authors Poetry

K. Praveen Kumar, T. Maruthi Padmaja

Abstract*: Due to application ranging from literature to product development companies, identifying a document similarity is one of the pivotal tasks in information retrieval systems. So far, most of the research in this area focused on identifying similarity across the normal documents of prose form. But a poem is different from a general prose text, as it consists stylistic (orthographic, phonetic and syntactic) features, further the data is also a high dimensional distinctiveness. This paper analyzed stylistic features of Indian English authors; using linear, nonlinear semantic and stylistic text semantic analysis methods. The computational methods used for semantic analysis are LSA, MDS, and ISOMAP. The similarity in structures across the poems are identified with Partitioning Around Medoid (PAM) algorithm. From the visualization of the results, it is observed that the poems feature space is linear and there is similarity structure. It was found that using stylistic features is better than the linear and nonlinear semantic methods.

***Index Terms*: Latent Semantic Indexing, TF, IDF, TF-IDF, Similarity, SVD, stylistic features, ISOMAP, MDS.**

I. INTRODUCTION

Finding document similarity is a key research issue at present. Due to its wide range of applications like near duplicate detection, automatic CV matching and searching for similar documents before patenting and security scrubbing [11]. So far, in this line of research most of the work reported on finding similarity across prose based documents like CV's, Articles and Novels. On the other hand, poetry from long ages is considered as powerful element in society whose impact is more on human development. When it comes to Indian poems, several famous poets whose poems influenced a great number of lives from the ages [14]. Therefore, computational analysis of the styles of these authors referring to their writings, help the amateur poets and linguistic researchers [13] in understanding the various styles of poetry to proceed further. In addition to these benefits, the analysis can recommend specific writers of a style in similarity to the readers who are interested to do a research or poetry lovers of a specific style.

Earlier the style similarity was analyzed manually, Josephine [1] had been analyzed the adjective-noun-verb-connectivity of American poetry, and examined the frequent adjectives of poems by hand, which was time intensive. However, with the help of modern computation techniques we can analyze the poetry on their latent structures and visualize the poems stylistic similarity on a vector space for better understanding. The research found the similarity in style of poems, it is referring to the American poetry which had reported the stylistic features of the poetry shows better similarity in structure [1]. However, According to Indian authors [15], "Indian English poetry has longer and more distinguished tradition and Indianized to Indian situations". Our work is a first of its kind to analyze the similarity across the Indian English poetry on stylistic grounds using few stylistic features like average word length per line, average number of lines per stanza, rhyming of the poems. Since our objective is to find out the style similarity we considered only, the other features phonetic are not considered that represents the sounds, and syntactic features represents order of the word usage Thus, it helps the academia to explore how Indian writer's styles are different with each other or similar to each other.

The main objective of this work is to analyze the significance of stylistic features in finding out the poetry similarity. To meet this objective, we have compared the similarities of the vector space of the stylistic features with the vector spaces of words method such as Term Frequency (TF) and Term Frequency and Inverse Document Frequency (TF-IDF). A high dimensional quality is the critical issue of the each poem both linear and nonlinear dimensionality reduction (DR) methods are used for dimensionality reduction. Further, initially the structure of the poetry data space was not known, both linear and nonlinear dimensionality reduction is considered in this study.

Because of non availability of standard Indian English Poetry Corpus, to carry out the analysis, we build a poetry corpus of 260 poems, from the poetry records [6] of 31 authors, available from poem hunter website. Here we projected the similarity among poems based on 2-dimensional vector space for easy visualization. The accuracy of the obtained results is provided using visualization aids of R software [12].

This paper is organized as follows, in section 2 we describe about work carried on poetry analysis. In section 3 we explain the background of this work, similarity measures and Latent Semantic Analysis with SVD, MDS, and ISOMAP. In section 4 we describe about methodology. In section 5 analyses of results. In section 6 we gave conclusion and in section 7 future projections.

Revised Manuscript Received on 30 March 2019.
* Correspondence Author
K. Praveen Kumar, Department of IT, VFSTR Deemed Tobe University, Guntur, A.P.,India.
T. Maruthi Padmaja, Department of CSE, VFSTR Deemed Tobe University, Guntur, A.P., India,

II. RELATED WORK

Heywards [4] analyzed style of poems by considering features such as prosody, meter and syntax. Kaplan & Blei (1996) has done influential work on style comparison of American poetry and visualized it on 2-dimensional scale. The authors mapped poems of different poets in a vector space based on stylistic elements of poetry. DanJurafsky [11] using computational methods presented an extensive study on various features of poetry such as diction, sound devices, affect and imagery to find most influential feature to depict the beauty of a poem. David J.Underhill [16] examined DR techniques on large corpora of text and by comparing the results he found that MDS is producing good results. MOHAMED-AMINE BOUKHALED[17] of their work they did a computational stylistic study of French classic literature they proposed an interesting measure to extract meaningful syntactic patterns of a specific author.

III. BACKGROUND

In this section, we have presented the description of the computation methods that we have adopted for the English poetry similarity analysis.

A. Linear Methods

Latent Semantic Analysis (LSA)

The best way to find the semantic structure is Latent Semantic Analysis sometimes called as latent semantic indexing [9]. The main idea of LSA is to derive a set of latent concepts (features) by studying the term co-occurrences in the document. A latent concept is nothing but set of words that frequently occur together, are assumed that they are more semantically related. Usually LSA component is a linear combination of coordinates of given n-dimensional vector, which is a scalar value. The LSA is realized by the following methods

Vector Space Model (VSM): It represents documents in n dimensional vector space, where each dimension represents a frequency of a term. Usually this frequency is calculated using Term Frequency (TF) or Inverse Document Frequency (IDF) methods [7].

Term Frequency (TF): It is the number how many times a term appeared in a document, the idea is if a document uses a term more frequently means the document talks about that topic, for example; in a medical journal if we found Breast cancer more times it means that article talks about breast cancer. But the problem is that few words are there called stop words such as the, and, is etc. may appear more number of times but these words do nor result the idea of the topic which the document is talking.

Inverse Document Frequency (IDF): It reduces the weight of the terms more frequently occurs in all documents and finds the terms which occur less frequently, this gives the idea of finding the terms which emphasizes the document. With this method, we can find the key terms which occur less frequently in document set which separates one document with others.

Term Frequency -Inverse Document Frequency (TF-IDF): it is the product of two statistical values TF and IDF. High term frequency and low Inverse Document Frequency together filter out the common terms. The feature space obtained by TF, TF-IDF methods is of high dimensional; to reduce the dimension Singular Value Decomposition (SVD) is used.

Singular Value Decomposition: It is a Linear Algebra technique, which decomposes a term-document matrix in to 3 matrices usually called u, d, v. In this u v are left and right singular value vectors and d is a diagonal matrix, which represents the concepts which separates the documents. This method reduces the dimensionality by preserving the uniqueness of document. From the singular matrices u and v if we pick k number of rows and columns then we get the k approximation of original matrix.

Once the latent semantic structure of the poem obtained the similarity across the poems structure is calculated using cosine similarity measure through (PAM) Partitioning Around Medoid algorithm

B. Non Linear Methods

This subsection depicts the brief description of the non linear dimensionality reduction techniques that are adopted in this work.

Multi Dimensional Scaling (MDS)

Multi Dimensional Scaling (MDS) [7] is one of the non linear dimensionality reduction techniques that work on the similarity/dissimilarity matrix. The MDS mainly applied when the actual data points are not available and instead the pair wise distances between them are available.

A simplified view of the algorithm is as follows:

a) Compute Euclidean distances among all pairs of points of m-dimensional space to form dissimilarity/similarity matrix D.

b) Evaluate the stress function

$$\sqrt{\sum\sum}(f(x_{ij})-d_{ij})2/\text{scale} \quad (1)$$

Where, the original data point and is the distance between i, j points. The smaller the stress value, the greater the similarity between them.

c) Adjust the coordinates of thedata points to the direction of the best maximally stress.

d) Repeat steps 2 through 4 until stress will not get any lower.

ISOMAP

The ISOMAP extends the MDS by replacing the pair wise Euclidean distance with the concept of geodesic distance [5]. The geodesic distance is calculated like this: initially the neighborhood graph is constructed next the shortest distance between any two pair of points in the neighborhood graph is calculated using Dijkstra algorithm. From then the rest of the low dimensional embedding procedure is similar to MDS.

C. Stylistic Features Method

This section describes about the stylistic features of a poem

Stylistic features of poem

According to [4] a poem can be characterized based on its style, imagery, affect and on semantic grounds. Further, poem similarity is computed based on its stylistic features, those are categorized in to 3 groups [4]

Orthographic Features: These features describe about physical structure of poem, includes word count, number of lines per stanza, average line length, average word length.

Syntactic Features: Deal with the connection among parts of speech (POS) elements of poem.

Phonemic Features: In this category Rhyme and meter are the features to be considered. As our concentration on style and style can be represented with many features but for this paper we considered average word-length of poem, average number of lines per stanza, rhyming of poem only.

Cosine similarity

In VSM most commonly used similarity measure is cosine similarity [11]. Vector space model represents documents in vectors; this method measures the similarity between two vectors in vector space. It value varies from -1 to 1, but when we consider positive relationship between 2 vectors we consider 0 and 1, if the angle between vectors is 90 degrees the two vectors are not similar, as the degree approaches to 0 means they are more similar.

Partitioning Around Medoids

It is a partition based clustering algorithm based on medoid computation. Here the medoid is the representative of a cluster [14]. This algorithm chooses representative objects arbitrarily, and improves cluster quality by replacing it with best suitable object. Cluster quality is measured by finding average dissimilarity between object and representative object of its cluster.

IV. METHODOLOGY

The flow diagram for proposed methodology is shown in Fig.1. We have used TDM based vector space model for linear dimensionality reduction (DR) and DTM based vector space model for non linear dimensionality reduction.

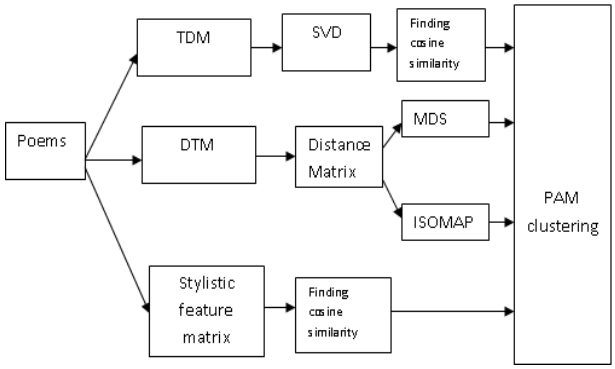

Fig. 1 The procedure followed to visualize the poems on vector space

A. LSA based on Linear DR

Initially we embedded each poem in to a vector space, then performed stop words removal, stemming to find the root words. Later Term Document Matrix(TDM) is calculated using TF-IDF weighting factor and to reduce the dimensionality we performed SVD on the TF-IDF matrix. On the reduced dimensionality matrix applied cosine similarity measure to find the similarity among vectors finally applied PAM clustering to find the clusters of poems.

B. LSA based on Non Linear DR

As we said in previous section we embedded poems in to vector space, then calculated Document Term Matrix (DTM), as these methods work on similarity or dissimilarity measures, we calculated cosine distance and passed it to MDS, ISOMDS and ISOMAP algorithms one after other and finally PAM is applied on resultant low dimensional data to find out the clusters of poems. Each method expects a k value, to find optimal k value we used scree plots.

C. Stylistic feature method

In this method we took 260 poems of 31 authors, and calculated few stylistic features using PHP programs later on this matrix we applied cosine similarity measure finally applied PAM to find the clusters based on stylistic similarity.

V. RESULTS AND DISCUSSION

To study the similarity of poems, we have collected 260 poems of 31 authors from poem hunter [6] website. We have considered authors from award winning to amateur writers. From each authors' writings, we have considered almost similar number of poems of similar sizes to avoid domination of one poet poetry. For our analysis, the TF-IDF, SVD, MDS, ISOMAP and PAM clustering methods are considered from NLP (Natural Language Processing) module of the R software. Concern with the experimental results of LSA, Fig 2 and Fig 3 depicts the scree plot and cluster plot for k=25 From these plots it can be observed that MDS method is appears to be good on the basis of average dissimilarity and diameters of clusters, but when it come to intra cluster similarity SVD is tightly coupling the poems than MDS.ISOMAP and MDS are showing better results for inter cluster similarity with low values.

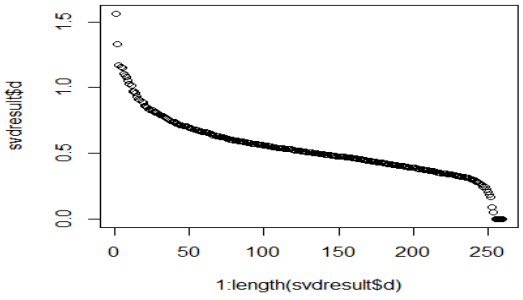

Fig. 2 Scree Plot for SVD Method

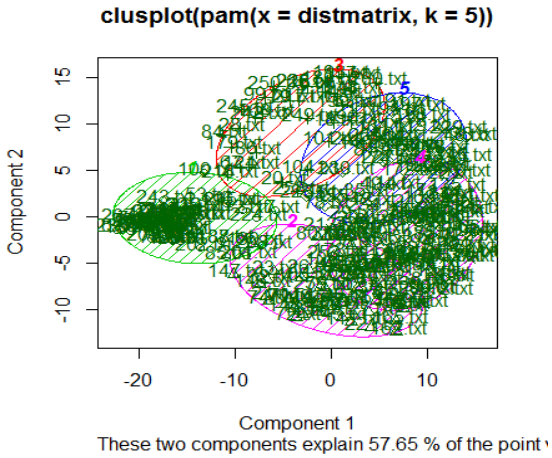

Fig. 3 Cluster Plot of SVD Result for K=25

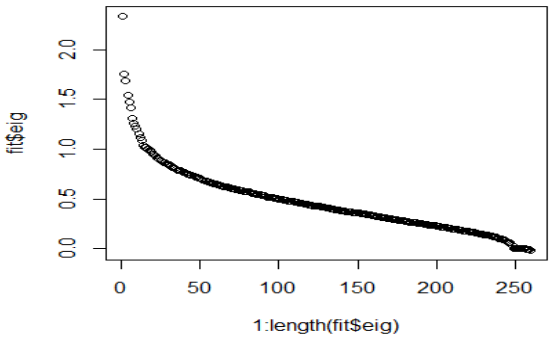

Fig. 4 Scree Plot for MDS

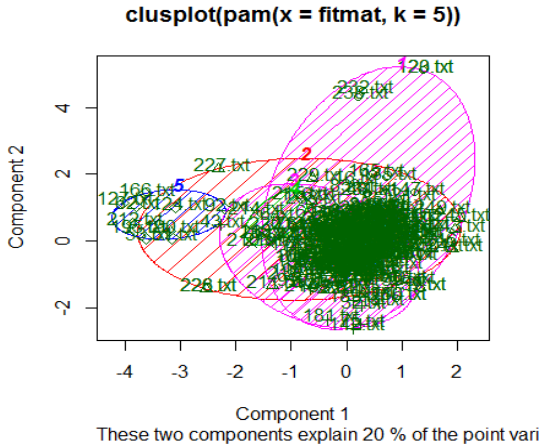

Fig. 5 Cluster Plot for MDS method

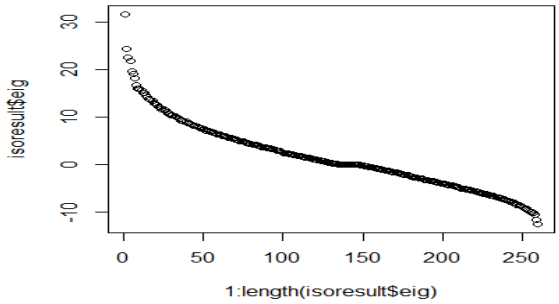

Fig. 6 Scree Plot for ISOMAP Method

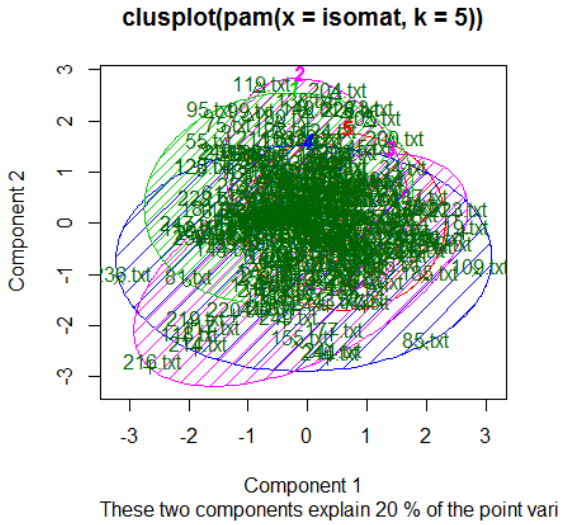

Fig. 7 Cluster Plot for ISOMAP Result

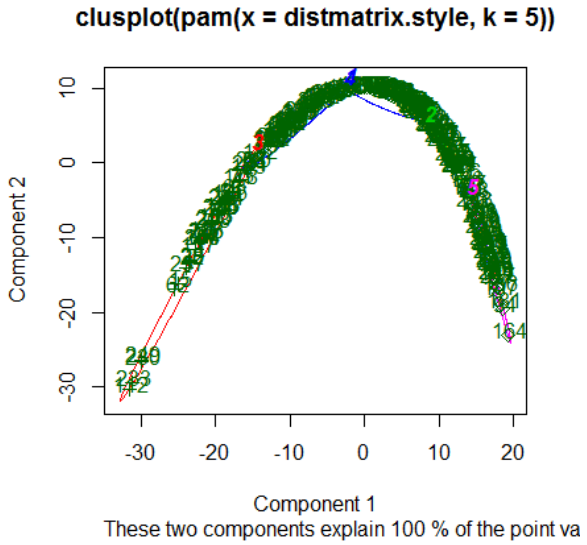

Fig. 8 Cluster Plot for Stylistic Feature Method.

Table. 1 Cluster Information of LSA Method

Cluster No	Cluster Size	Maximum Dissimilarity	Average Dissimilarity	Cluster Diameter	Cluster Separation
1	96	1.446298	0.458811	1.650912	0.069449
2	46	1.802027	0.581554	1.903346	0.069449
3	94	3.007365	0.594464	3.282341	0.076676
4	16	4.68763	1.797953	4.907747	0.478707
5	8	4.956817	3.398781	7.52537	2.380725

Table. 2 Cluster Information of MDS Method

Cluster No	Cluster Size	Maximum Dissimilarity	Average Dissimilarity	Cluster Diameter	Cluster Separation
1	132	0.565473	0.174018	0.749808	0.055356
2	50	0.528667	0.226387	0.755497	0.087255
3	60	0.319754	0.15916	0.476757	0.055356
4	8	0.215009	0.12493	0.287658	0.257995
5	10	0.285438	0.167052	0.41322	0.193773

Table. 3 Cluster Information of ISOMAP Method

Cluster No	Cluster Size	Maximum Dissimilarity	Average Dissimilarity	Cluster Diameter	Cluster Separation
1	98	1.397958	0.773684	2.181228	0.326069
2	36	1.230912	0.709047	1.948493	0.340631
3	63	1.382003	0.771293	2.236538	0.326069
4	49	1.873261	0.741279	2.578574	0.350856
5	14	1.073779	0.60518	1.30835	0.454037

Table. 4 Cluster Information of Stylistic Features

Cluster No	Cluster Size	Maximum Dissimilarity	Average Dissimilarity	Cluster Diameter	Cluster Separation
1	69	0.928421	0.392811	1.709954	0.032607
2	42	0.534355	0.253005	1.045217	0.081732
3	38	2.189328	0.531792	2.962461	0.032607
4	61	0.91951	0.319796	1.374636	0.081732
5	44	1.041995	0.298197	1.625524	0.117134

The above shown tables 1,2,3,4 explain the cluster information about LSA, MDS, ISOMAP and Stylistic Features methods. Each table consists size of cluster, i.e. how many poems are comes in to each cluster, maximum dissimilarity i.e. maximum distance among the objects in each cluster, average dissimilarity i.e. the average dissimilarity of each cluster, diameter it describes about the diameter of cluster where the objects are spread through the cluster and cluster separation. The findings from these tables are MDS method makes slightly better clusters than stylistic Features method, but to find style of a poet stylistic feature based clustering is more suitable.

VI. CONCLUSION

In this paper, we have analyzed the adaptability of LSA, MDS, ISOMAP based features and style based features to find out the similarity across the poetry of Indian English authors. The analysis is carried out on vector spaces of TF-IDF, and stylistic features of the poem. From the experiments, we identified that style based features expressed better similarity across the poems than linear and nonlinear based semantic analysis methods, further we identified the obtained similarity of linear space is better is nonlinear space thus it can be concluded that the poem's vector space is linear in nature.

VII. FUTURE DIRECTIONS

One area of future scope is to explore other features like prosody, stress words, etc. to find style similarities of poems and comparing the poetic styles of American poets and Indian poets. Another area is to check the other methods to reduce the dimensionality and also to perform author identification.

REFERENCES

1. J. Miles. Major adjectives in English poetry: from Wyatt to Auden. University of California Publications in English,12(3):305–426, (1946).
2. J. Miles. Eras & Modes in English Poetry. University of California Press, Berkeley, CA, (1957).
3. J. Miles. Style and Proportion: The Language of Prose and Poetry. Little, Brown and Co., Boston, (1967).
4. Raghavan, V. V. and Wong, S. K. M. A critical analysis of vector space model for information retrieval. Journal of the American Society for Information Science, Vol.37 (5), p. 279-87, (1986).
5. Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science, 290, 2319-2323.
6. Jiawei Han, Michelinekamber, Jianpei, Data Mining Concepts and Techniques, 3rd edition, Text Book, Morgan Kaumann Publishers.
7. https://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/chapter3.pdf
8. Davi M. Kaplan, David M.Blei. A computational Approach to Style in American Poetry. Seventh IEEE International Conference on Data Mining, DOI 10.1109/ICDM.(2007).76, 1550-4786/07.
9. ArashHeidarian and Michael J.Dinneen " A Hybrid Geometric Approach for Measuring Similarity Level Among Documents and Document Clustering" International Conference on Big Data Computing Service and Applications 978-1-5090-2251-9/16 (2016)IEEE
10. ArashHeidarian, Michael J.Dinneen. A Hybrid Geometric Approach for Measuring Similarity Level Among Documents and Document Clustering. 2016 IEEE Second International Conference on Big Data Computing Service and Applications, DOI 10.1109/BigDataServce.(2016).14
11. Justin Kao, Dan Jurafsky A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry Workshop on Computational Linguistics for Literature, pages 8-17, Montre`al, Canada, June 8, (2012) Association for Computational Linguistics.
12. https://www.poemhunter.com
13. R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. ACM Press, New York, 1999.
14. https://matpalm.com/lsa_via_svd/intro.html
15. Grossman and Frieder's Information Retrieval, Algorithms and Heuristics
16. https://www.searchtechnologies.com/document-similarity-analysis
17. R Studio, R language A statistical analytical tool
18. SunitaRana. A study of Indian English Poetry, Hindustan Institute of Technology & Management Dheen, Ambala.
19. Pankaj Mishra Wikipedia.