A Comparative Analysis of Similarity Measures to find Coherent Documents

Mausumi Goswami

Computer Science and Engineering department, Christ (Deemed to be University), Bangalore

Advin Babu

Computer Science and Engineering department, Christ (Deemed to be University), Bangalore

B.S Purkayastha

Computer Science department, Assam University

Abstract*—*With ever increasing nature of text resources over world wide web and digitized libraries , it is a need of the hour to organize documents based on their practical need. Text document clustering is an interesting research problem. Clustering process automatically organizes documents into coherent groups. Significance of this process lies in the effective and efficient usage of digital documents for information retrieval and other natural language related tasks. Selection of a proper metric to quantify the extent of similarity between two documents play a very important role in this process. This work is an experimental study on investigating the same problem. Here, nine different distance measure and similarity measures are compared based on their pros and cons, application and experiments are conducted on the same.

Keywords—Natural language processing, document clustering, similarity measure, Euclidean distance, cosine similarity.(key words)

I. INTRODUCTION

Data has been increasing exponentially and many sources have predicted this growth towards 2020 and beyond that as well. The data generated by human and machine is experiencing a growth of 10x times faster than the normal traditional business data. Machine data is increasing at even more faster rate of 50x times. This exploding volume of data and ever increasing data growth has led to many challenges. But, it can be used to extract vital information as well. Thus, there is a need of some technique or measure which can be used to extract this information. Clustering is one such technique which can be used to organise this large quantity of data into small numbers of coherent clusters and these cluster can be used to extract valuable information from the dataset. Distance functions and similarity measures are required for clustering.

Similarity Measure is a measure that represents the similarity between a pair of objects or it can be defined has a function that computes the degree of similarity between any two objects. Organization of the paper is as mentioned below: section II describes the models for representing documents. Section III discusses few important distance measures and similarity measures used for text documents. Section IV compares the similarity measures and Section V discusses a methodology for implementation Section VI discusses application areas of various metrics. In Section VII experiments and results are discussed.

II. MODELS FOR REPRESENTING DOCUMENTS

A text document can be modelled in many ways. A text document is represented as a bag of words, where the document is the collection of words and its frequencies. The order of these words doesn't matter. Each document gets converted into a matrix, where the word represents the row of the matrix and the document becomes column vector. The frequency of every term is used as its weight that means the words with higher frequency is more descriptive about the document.

Let DR = dr_1, dr_2, dr_3..dr_N be N documents that are under study and TERMC = tc_1, tc_2, tc_3, .tc_M be the M unique terms which are present in these documents, then each document can be represented as a n by m dimensional vector. Although more frequency words are assumed to have more importance, but high frequency words such as a, is, the, are should be removed, since neither are they descriptive about the document nor are they important for the document. tfidf weights are used to discount the frequency of terms based on their relevance and importance to a particular document in the entire document set that is under consideration. This method is done as follows:

$tfidf(dr, tc) = tf(dr, tc) \times \log(|DR|/df(tc))$.

Here df(tc) is the number of documents in which term tc appears or document frequency of tc.

III. METRIC USED IN DOCUMENT CLUSTERING

A distance measure must be determined before clustering. The clustering algorithm requires a measure for making and organizing different groups. This measure gives a numeric value to the extent of difference between the pair of documents. Different distance measures work better in different cases. Selection of a distance measure depends on which measure is able to catch the essence of important distinguishing characteristics. These characteristics are contextual and varies from problem to problem. For a specific type of clustering selecting a proper similarity measures plays a very critical role. As an instance, we may take density based clustering such as DBSAN which is highly effected by similarity or dissimilarity computation.

Again, every similarity measure is not a metric. There are four conditions to be satisfied to consider any measure as a valid metric. Let us take a measure d and consider there exits two object a and b such that the distance between a and b are given by d.

1. Non Negativity : Distance between any two object must be non negative, that is, dist(a,b)>=0 if a !=b
2. Identical objects have zero distance. Only in case of identical object it is possible to have a zero distance value. Dist(a,b)=0 , if and only if a == b , that a and b are identical.
3. Distance from a to b and distance from b to a are considered to be same. This property is called symmetric.Dist(a, b) == Dist(b, a).
4. Triangular Inequality : If there is a non-negative distance between a and b given by Dist(a,b); if there is a is a non-negative distance between b and c given by Dist(b,c); If there is a non-negative distance between a and c given by Dist(a,c), then the following inequality holds:Dist(a, b) + Dist(b,c) >= Dist(a,c).

Few important similarity measures are discussed below.

3.1 MINKOWSKI DISTANCE

Euclidean distance and Manhattan distance are a particular case of Minkowski distance. This distance measure performs well when all the datasets are compact and isolated. If the dataset is not able to fulfil this condition, then the large-set attributes will dominate the others. The largest-scale feature generally dominates the other. This is another problem that arises in case

of Minkowski distance measure. The solution to this problem is normalizing the continuous feature. The Minkowski distance measure is given as,

$d_{min} = (\sum_{j=1}^{l} |a_j - b_j|^p)^{\frac{1}{p}}$, p>=1

, where p is a positive real number and a_j and b_j are a pair of vectors in l-dimensional space.

For solving the clustering obstacles a modified version of Minkowski distance has been proposed.

3.2 EUCLIDEAN DISTANCE

Euclidean distance is defined as the normal distance between a pair of points. It is widely used in text clustering. The Euclidean distance measure is a very special case of Minkowski distance measure. Given two documents ra and rb, which can be represented by their term vectors ca and cb respectively. Thus, the Euclidean distance is:

$$Distance_{Euclidean}(\underset{c_a}{\rightarrow},\underset{c_b}{\rightarrow}) = (\sum_{c=1}^{m} |w_{ca} - w_{cb}|^2)^{1/2}$$

, where the term set is C = c1, ..., cm. Here , total number of terms is m, term wise difference between two vectors is calculated first to calculate the norm $|w_{ca} - w_{cb}|^2$. Summation of all such squared lengths are taken, next, square root of the same is computed.

3.3 COSINE SIMILARITY

When document is represented in the form of term vectors, the similarity between the pair of documents can be quantified as the cosine angle between the two vectors. This is called cosine similarity. This method of similarity is usually used in the context of text mining for comparing documents or emails.

Given two vectors tc_a and tc_b, the cosine similarity between them is,

$$SIM_C(\underset{tc_a}{\rightarrow},\underset{tc_b}{\rightarrow}) = \frac{\overrightarrow{tc_a} \cdot \overrightarrow{tc_b}}{\overrightarrow{|tc_a|} \times \overrightarrow{|tc_b|}}$$

, where tc_a and $tc_b$are vectorshaving m-dimensions over the term set C = c1, ..., cm. Here, m represents the dimension space for each coordinate. Cosine similarity lies between [0, 1]. If the cosine similarity between two document term vectors is higher, then both the documents have more number of words in common. Independent of document length is considered as a very significant property of cosine similarity.

3.4 MAHALANOBIS DISTANCE

Correlation of data set is not considered and it is not scale-invariant. This can be represented as:

$$r_{ac}^2 = (x_a - x_c)M^{-1}(x_a - x_c)'$$

, where M is considered as the covariance matrix.

3.5 JACCARD COEFFICIENTOF SIMILARITY

This coefficient measures the similarity as a ratio. Numerator of this ratio is the intersection of the pair of objects. Denominator of the ratio is the union of the same pair of objects which are considered for union. An alternative name given for Jaccard coefficient is Tanimoto coefficient. For normal text documents, the meaning of Jaccard Correlation is mentioned below. It compares the sum weight of all the shared terms present in both the documents to the sum weight of terms that are present in either one of the two documents but are not shared terms.

The Jaccard Coefficient's formal definition when it is expressed over a bit vector can be given by,

$$SIM_J(\underset{c_a}{\rightarrow},\underset{c_b}{\rightarrow}) = \frac{\overrightarrow{c_a}.\overrightarrow{c_b}}{\overrightarrow{|c_a|}^2 + \overrightarrow{|c_b|}^2 - \overrightarrow{c_a}.\overrightarrow{c_b}}$$

The Jaccard coefficient is a similarity measure and the coefficient's value ranges between 0 and 1. Equality of c_a and c_b will give a value of 1 (when c_a=c_b)and it will give a value of 0 when c_a and $c_b$are disjoint, where 1 means that the pair of objects are same and 0 means that the pair of objects are completely different. The corresponding distance measure can be computed by subtracting the similarity from 1. It is given as DISTANCE $_J$ = 1 − SIM$_J$.

3.6 PEARSON CORRELATION COEFFICIENT

Pearson's correlation coefficient is a measure to find the linear correlation between two variables. It can be used to measure to determine the relation between a pair of vectors. It is the ratio of covariance of the two variables and product of the standard deviation of the two variables.

$$\text{SIM Pearson} = \frac{\text{COVIARIANCE (X, Y)}}{[\text{ STD DEV(X) . STD DEV(Y)}]}$$

Given the term set, TC = tc$_1$, ...,tc$_m$. The formula used to calculate is,

$$SIM_P(\underset{c_a}{\rightarrow},\underset{c_b}{\rightarrow}) = \frac{\text{m}\sum_{c=1}^{m} w_{ca} \times w_{cb} - TCF_a \times CF_b}{\sqrt{\left[m\sum_{c=1}^{m} w_{c,a}^2 - CF_a^2\right]\left[m\sum_{c=1}^{m} w_{c,b}^2 - CF_b^2\right]}}$$

, where $CF_a = \sum_{c=1}^{m} w_{c,a}$ and $CF_b = \sum_{c=1}^{m} w_{c,b}$

This correlation coefficient is also a similarity measure. The value may range from +1 to -1. +1 is total positive correlation, -1 is total negative correlation and 0 indicates no correlation.

It 1 when two term vectors under consideration are the same.

Distance measure, DIST$_P$ = 1 − SIM$_P$ when SIM$_P$>= 0 and DIST$_P$= |SIM$_P$ | when SIM$_P$< 0.

3.7 HAMMING DISTANCE

Hamming Distance is considered as a practical metric for comparing data strings. In information theory, it calculates the number of position in which two strings have different values. Formally, if two strings of equal length are considered then hamming distance is equal to the number of positions for which the corresponding 1 and 0 are different.

Let, a and b M^n. Here the hamming distance between a and b, is the number of places (bits or characters) where a and b are different. It is denoted by dH(a,b).

The hamming distance can be defined as the number of bits or number of characters to be changed to turn a string into the other.

3.8 MANHATTAN DISTANCE

Manhattan distance measure a very special case of Minkowski distance measure, where the positive real number is equal to 1. It is the distance between a pair of points measured along the axes at 90 degrees or right angle. Manhattan distance measure is also used in clustering algorithms, where the shapes of all the clusters are hyper-rectangle. The Manhattan distance measure is given by,

$$d_{man} = \sum_{j=1}^{l} |a_j - b_j|$$

3.9 CHEBYSHEV DISTANCE

This measure examines the absolute value of the positive deviations between the pair of objects considering their coordinate positions. An alternative name used for Chebyshev

distance measure is Maximum value distance. The distance measure can be used for both quantitative and ordinal variables.TheChebyshev distance measure is a very special case of Minkowski distance measure, where the positive real number is equal to infinity (∞). The Chebyshev distance is given by,

$$d_{che} = \lim_{p\to\infty} (\sum_{j=1}^{l} |a_j - b_j|^p)^{\frac{1}{p}} = max_{j=1}^{l} |a_j - b_j|$$

IV. COMPARATIVE ANALYSIS

In this section a comparative analysis of the methods are done based on their relative advantage and disadvantages.

SI.	SIMILARITY MEASURE	PROS	CONS
1.	Minkowski Distance	Overflow is possible for large p values.	Only specific p values allow for proper consideration of overflow and underflow.
2.	Euclidean Distance	Easy to implement. Easy to test.	The variables which have the largest value greatly influence the result. Doesn't work efficiently with image data.
3.	Cosine Similarity	Both continuous and categorical variables may used	Doesn't work efficiently with nominal data.
4.	Jaccard Coefficient	Both continuous and categorical variables may used.	Doesn't work efficiently with nominal data.
5.	Mahalanobis Distance	Utilizes the group means and variances for each variable, so the problem of correlation and scale are solved.	Inverse of correlation matrix is needed, it can't be calculated if the variables are highly correlated.
6.	Pearson Correlation Coefficient	Accuracy of score increases when data is not normalized. Easy to compute.	Sensitive to outliers.
7.	Hamming Distance	Detecting and correcting errors	Code is k-errors correcting, if the minimum Hamming between the pair of codes is 2k+1
8.	Manhattan Distance	Easy to generalize into higher dimensions.	Doesn't work efficiently with image data Can't be used to classify documents
9.	Chebyshev Distance	Easy to implement. Easy to test.	Doesn't work efficiently with image data.

Table 1 : Comparison of pros and cons of different measures.

V. PROPOSED METHODOLOGY

In this section the steps followed to apply a few of the measures on real life data set is described.

MODULE A

1. Read the Data set
2. Pre processing the data by applying techniques of stop words removal, stemming, converting text to lower case, removal of punctuation symbols etc.
3. Select a Feature Selection Technique
4. Computation of Term frequency and inverse document frequency .
5. Construction of Term Document Matrix
6. Sparsification of the data to save storage.

MODULE B

1. Selection of Similarity Measure
2. Computation of Similarity Matrix for finding coherent Text

MODULE C

1. Select a clustering Algorithm to find coherent groups of documents
2. Evaluation of clustering results
3. Preparing Conclusive reports

In this proposed methodology the different stages of clustering a collection of text documents are depicted. In this work during implementation focus is given on Module B and few steps of Module C. Mainly two similarity measures are chosen for finding the patterns. The steps followed are given in the next Section.

VI. APPLICATION AREAS

SI	METRIC	BASIC MEANING	SIGNIFICANT APPLICATION AREA
1.	Minkowski Distance	Distance between a pair of vectors. Generalized form of Manhattan and Euclidean distance.	Applied in machine learning to find the distance similarity between a pair of vectors.
2.	Euclidean Distance	Ordinary distance between a pair of objects which can be measured with the help of a ruler.	Application involving Interval Data DNA Analysis Health Psychology Analysis
3.	Cosine Similarity	Cosine angle between a pair of vectors of a dimension.	Text Mining
4.	Jaccard Coefficient	Size of the intersection by the size of the union of a pair of sets.	Document Classification
5.	Mahalanobis Distance	Distance between a pair of points in multivariate space.	Find multivariate outliers.
6.	Pearson Correlation Coefficient	Normal measure to determine relation between a pair of vectors.	Information Science
	Hamming Distance	Examines the number of characters/bits to be changed to change a string into other.	Coding Theory Cryptography
8.	Manhattan Distance	Distance between a pair of points which is measured along the axes at 90 degrees or right angles.	Integrated Circuits
9.	Chebyshev Distance	Examines the magnitude of difference between the coordinates of a pair of objects.	Chess Board

Table 2: Significant areas of Application

VII. EXPERIMENTS AND RESULTS

Implementation is done using MATLAB2018 R using an intel i3 processor and 4 GB RAM . Experiments are done by using 6 different documents. Experimental Evidences are described here.

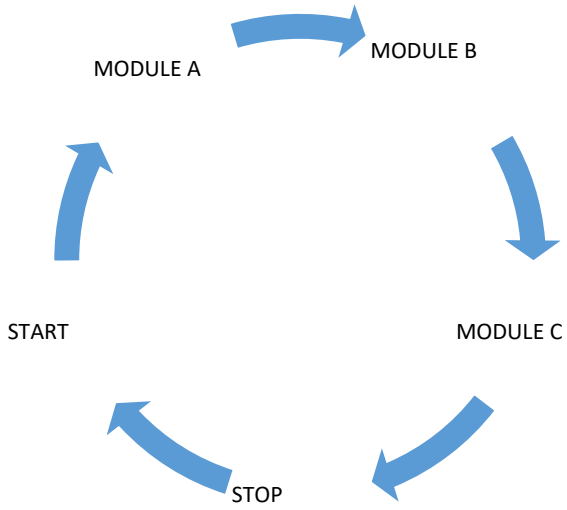


Figure 1 : Proposed Methodology

It is found that the storage space consumed is reduced by a substantial amount after applying sparsification. The matrix containing 345 terms and 6 documents had consumed 16560 bytes but after applying sparsification it is reduced to 10368 bytes .The first raw of the table below shows results before applying sparsification. the second raw shows results after sparsification for the same term document matrix.

Name	Size	Bytes	Class	Attributes
S1	6x345	16560	double	
Name	*Size*	*Bytes*	*Class*	*Attributes*
sparse_S1	*6x345*	*10368*	*double*	*sparse*
ysim	*6x6*	*288*	*double*	

Table3: Results after Sparsification

A good quantitative approach to find the optimum number of clusters is to compare average silhouette values in each cluster by varying the number of clusters. A quality measure called silhouette indicates the similarity of a point with the members of its own group (the same cluster where it belongs to) Vs. members in other groups. Its value ranges from -1 to +1. Let us consider b is the lowest average distance between two clusters i and j ; a is the average distance within the cluster I. Then Silhouette value is given by the following:

$$S(i) = \frac{\mathrm{b} - \mathrm{a}}{\max(\mathrm{a}, \mathrm{b})}$$

Experiments are conducted with varied number of clusters to find the optimum number of cluster.

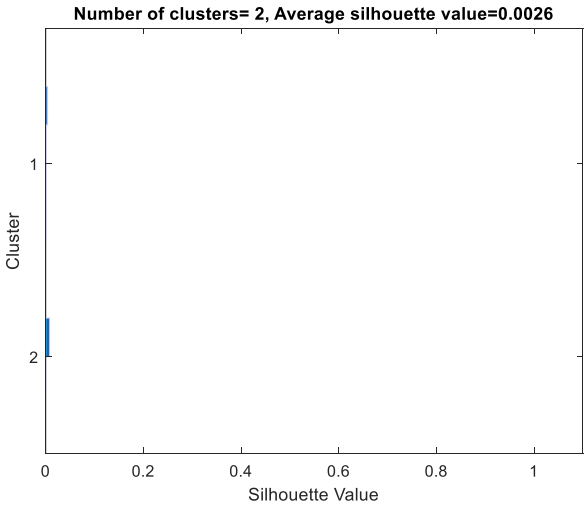


Figure 2 : Two clusters , Cosine similarity

Formally, the definition may be given as follows:

$$S(i) = \frac{min(MeanD\_BET(i,j)\ ) - MeanD\_WITHIN(i)}{max[\ MeanD\_WITHIN(i)\ ,\ min(MeanD_{BET(i,j)})]})$$

where$a = MeanD_WITHIN(i)$ is the mean distance from the i-th point to the other points in its own cluster, and b= minimum value of *MeanD_BET(i,j) across different clusters.*

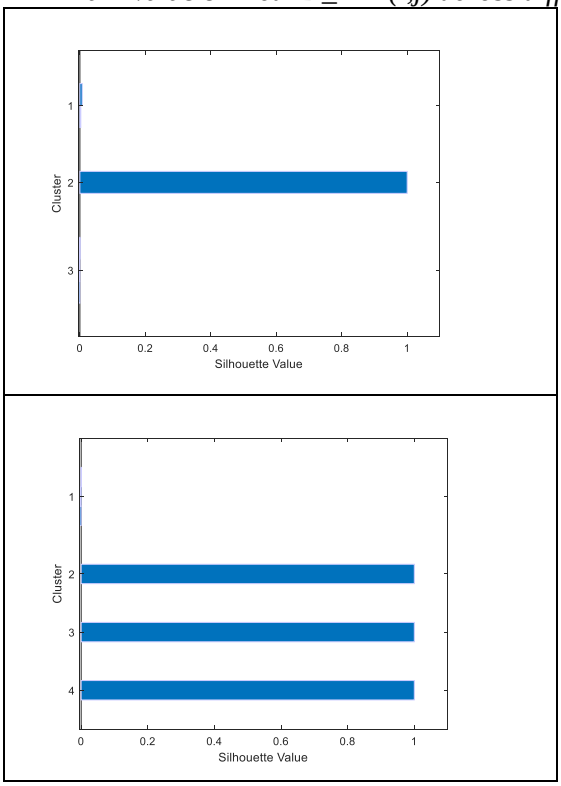


Figure 3 : Three and four clusters, Cosine Similarity

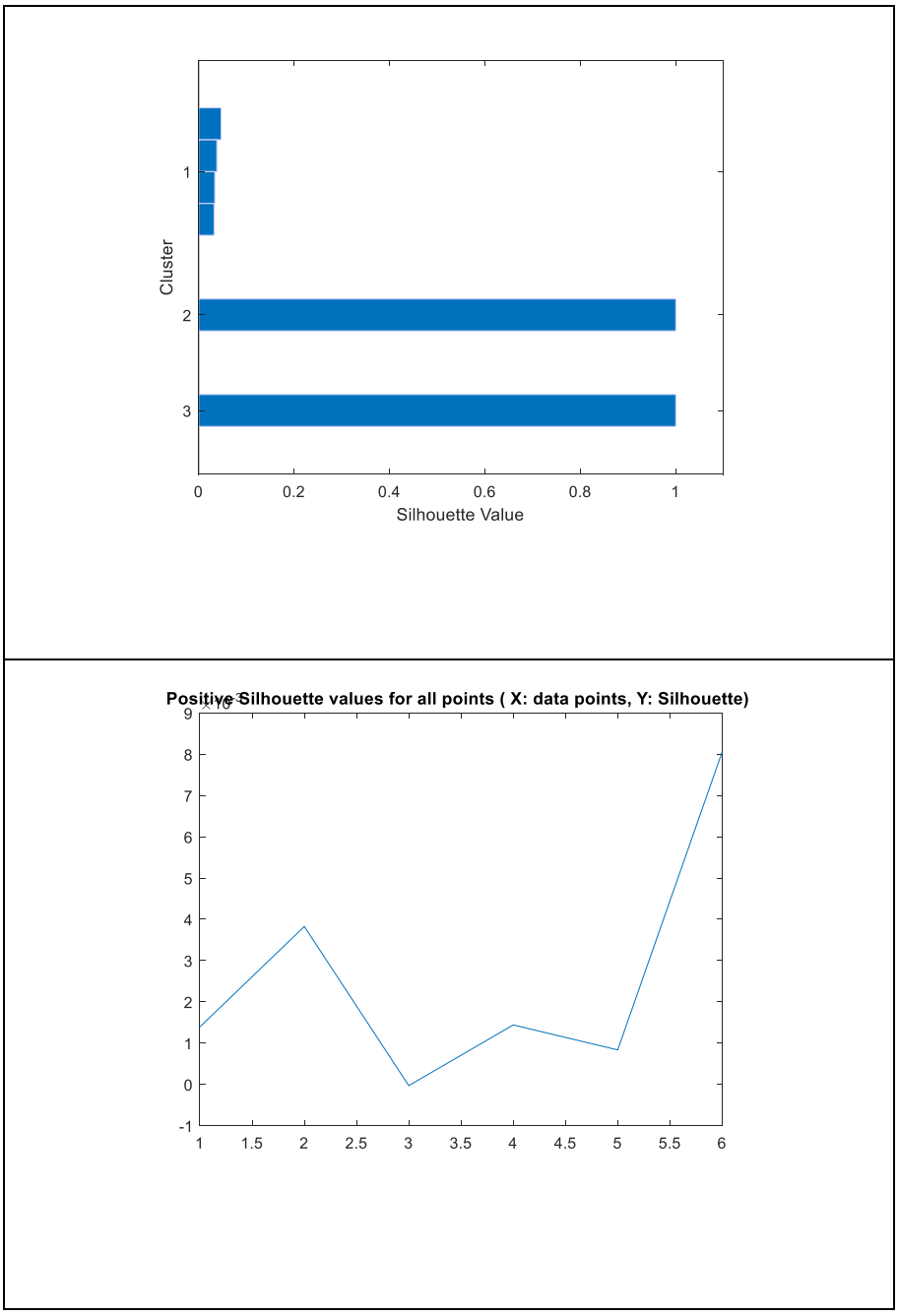


Figure 4 : a) Three clusters, Cosine Similarity b) Silhouette values correlation metric

Figure 5.a) shows the clusters formed and the silhouette values for different clusters. It is found that there are two clusters with lesser number of documents but silhouette values are found very high for them. The third cluster is having more number of documents in it but the silhouette values are positive , which indicate correct clustering , but less than 0.5 . Figure 5.b) shows documents in x axes and silhouette values in y axis. It reflects that all silhouette values are equal to or greater than 0.197 ,which is approximately 0.2. It indicates clustering results are correct. That is, intra cluster similarity is high and inter cluster similarity is low. In figure 6,a) correlation based clustering is evaluated using silhouette values. In figure 6.b) silhouette values are plotted. It is found that few clusters are showing silhouette values below zero. It is observed that clusters are tight enough to identify coherent documents.

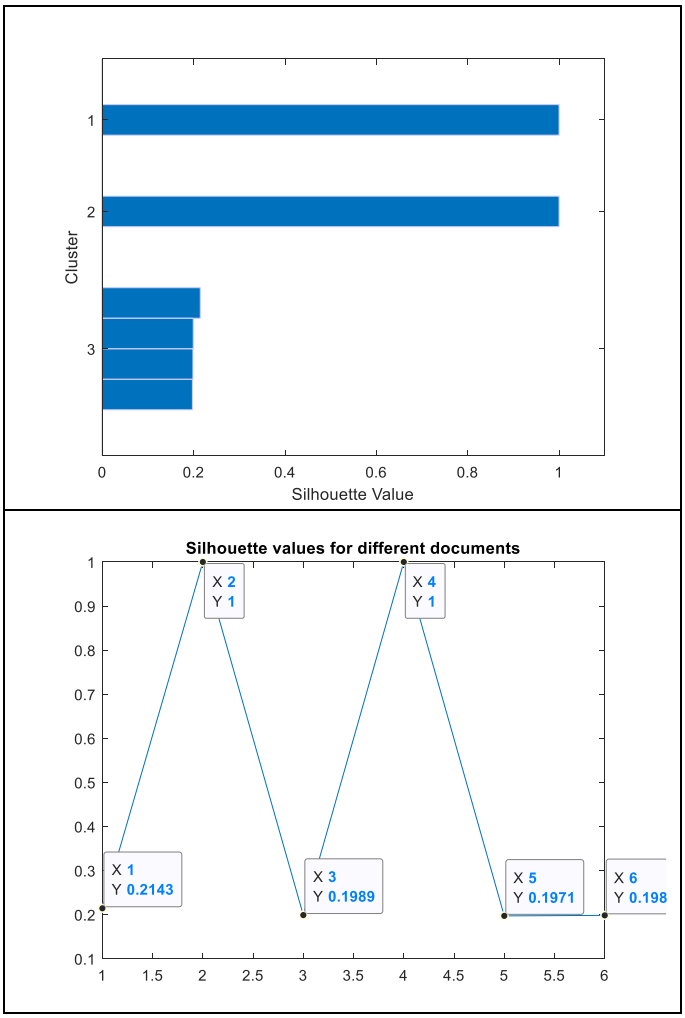
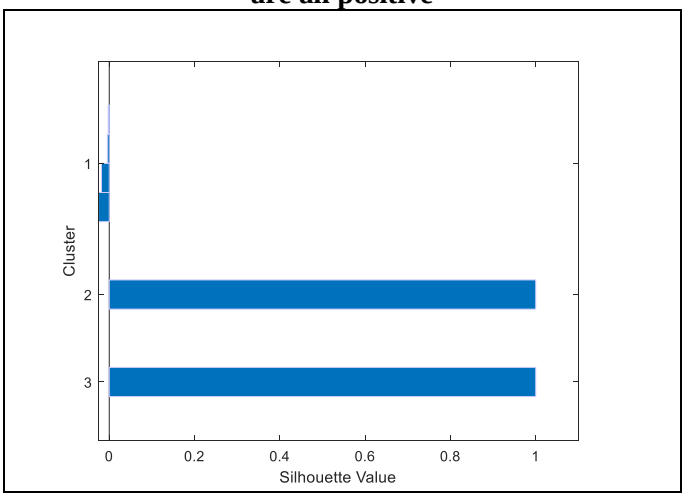


Figure 5 : a) Three clusters, Squared Euclidean distance b) Silhouette values obtained are all positive



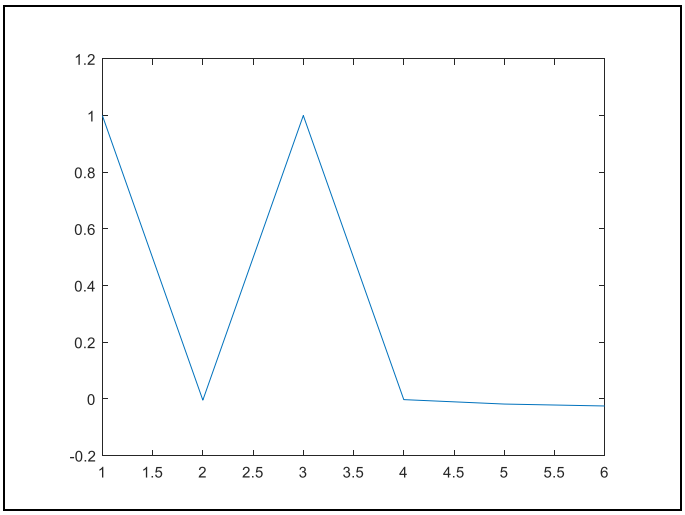


Figure 6 : a) Three clusters, Correlation based clustering b) Silhouette values obtained are not all positive

VIII. CONCLUSION

In this paper nine different similarity measures are compared based on their pros and cons. Also, they are compared based on their significant areas of application. It is found that similarity measures discussed here are effective to find the similarity between the documents. In this work, three such models of distance measure and similarity measures are implemented. It is found that sparsification process reduces storage requirements and increases space complexity. It is to be noted that cosine ,jaccard and pearson correlation are used to find similarity between objects. Euclidean, Mahalanobis, Hamming, Manhattan, Chebychev are used to find distance between objects. Overflow for large values of p is possible in case of Minkowski Distance. Cosine Similarity and Jaccard Coefficient both works well with continuous and categorical variables. Pearson Correlation is easy to compute but is sensitive to outliers. Hamming Distance used in cryptography can to detect and correct errors. In our future work ,we are going to implement the remaining measures and investigate it further to find variation of model which are competent enough to identify coherent documents.

IX. REFERENCES

1) Anna Huang, “Similarity Measures for Text Document Clustering”, In Jay Holland, Amanda Nicholas, and DelioBrignoli, editors, New Zealand Computer Science Research Student Conference, pages 49–56, April 2008.

2) Pranjal Singh, Mohit Sharma, “Text Document Clustering and Similarity Measures”, Dept. of Computer Science &Engg., November 2013.

3) Michael Steinbach, George Karypis and Vipin Kumar, “A Comparison of Document Clustering Techniques”, Dept. of Computer Science &Engg.

4) Kaufman L., and P. J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: John Wiley & Sons, Inc., 1990.

5) Peter J. Rousseeuw (1987). “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis”. Computational and Applied Mathematics 20: 53-65.*doi:10.1016/0377-0427(87)90125-7*.

6) R.C. de Amorim, C. Hennig (2015). "Recovering the number of clusters in data sets with noise features using feature rescaling factors". Information Sciences. 324: 126&ndash, 145. arXiv:1602.06989. doi:10.1016/j.ins.2015.06.039.

7) Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, Member, IEEE,"A Similarity Measure for Text Classification and Clustering", inIEEE Transactions On Knowledge and Data Engineering, Vol. 26,No. 7, July 2014, 1575

8) H. Chim and X. Deng, "Efficient phrase-based document similarityfor clustering," IEEE Trans. Knowl. Data Engg., vol. 20, no. 9, pp.1217–1229, Sept. 2008.

9) Kalaivendhan. K, Sumathi. P, "An Efficient Clustering Method toFind Similarity Between The Documents", in International Journalof Innovative Research in Computer and CommunicationEngineering ,Vol.2, Special Issue 1, March 2014.

10) B Sindhiya and N Tajunisha1,"Concept And Term Based Similarity Measure For Text Classification And Clustering", in Int. J. Engg. Res. & Sci. & Tech. 2014, Vol. 3,No. 1, February 2014.

11) K. Sruthi, B. Venkateshwar Reddy, "Document Clustering onVarious Similarity Measures", in International Journal of AdvancedResearch in Computer Science and Software Engineering, Volume3, Issue 8, August 2013 ISSN: 2277 128X.

12) VenkataGopalaRao S. Bhanu Prasad A., "Space and CosineSimilarity measures for Text Document Clustering", in InternationalJournal of Engineering Research & Technology (IJERT),Vol. 2Issue 2, February- 2013,ISSN: 2278-0181.

13) P.Sowmya Lakshmi, V.Sushma, T.Manasa, "Different SimilarityMeasures for Text Classification Using Knn", inIOSR Journal ofComputer Engineering (IOSRJCE), Volume 5, Issue 6 (Sep-Oct.2012), PP 30-36.

14) Anil Kumar Patidar,JitendraAgrawal,Nishchol Mishra, "Analysis ofDifferent Similarity Measure Functions and their Impacts on SharedNearest Neighbor Clustering Approach", in International Journal ofComputer Applications ,Volume 40– No.16, February 2012.

15) F. Sebastiani, "Machine learning in automated text categorization," ACM CSUR, vol. 34, no. 1, pp. 1–47, 2002.

16) P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to DataMining," Boston, MA, USA: Addision-Wesley, 2006.

17) J. Han and M. Kamber, "Data Mining: Concepts and Techniques,"2nd ed. San Francisco, CA, USA: Morgan Kaufmann; Boston, MA,USA: Elsevier, 2006.

18) [Online]. Available: http://web.ist.utl.pt/~acardoso/ datasets/