A Similarity Measure for Text Processing

[1]Mirza Ruhi Masuma, [2]Prof. V. A. Losarwar

[1,2]Computer Science and Department, P. E. S College of Engineering Aurangabad, India.

[1]*ruhi_mm@yahoo.in*, [2]*v_a_losarwar@yahoo.com*

Abstract **Data mining concept deals with the mining of knowledge from a huge amount of data. Many techniques exists and are used for the classification and clustering of the data which already exists in an organized form which is based on the likeness or the similarities between the documents in the text processing field. Clustering is a technique that organizes large amount of unordered or unsorted text documents into lesser number of many meaningful and logical clusters. To specify how different two given documents are clustering mechanism needs a particular metrics. In this paper work, we worked on, Cosine similarity, Euclidean distance and Similarity measure for text processing distance measures. The effect of these three measures is evaluated on a real-world data set for the testing of text classification and text clustering problems. The performance obtained by the Similarity measure for text processing is better than that achieved by other measures is hence proved.**

Keywords — ***classification, clustering, accuracy, measure of similarity, text processing.***

I. INTRODUCTION

Text processing is a rapidly budding latest technology for discovery of knowledge. It plays a crucial role in, web search, data mining and information retrieval [1], [2], [3]. There are huge amount of texts that are flooding the internet, documents which are in large number of collection are stored in digital libraries apart from this the digitized personal information that is emails are getting piled up quickly with each passing day. In text processing, the bag of-words model is commonly used [4].

In text mining, a document is supposed or is associated as a vector and in this every component point outs a value. The value here is that of the component's corresponding feature in the document. The feature value can be term frequency, relative term frequency, or tf-idf. The definitions of the terminologies used are as follows. "Term frequency" is the number of times the term is repeating in the document, "Relative term frequency" is the ratio between the term frequency and the total number of occurrences of all the terms which is present in the document set, "tf-idf" is a combination of term frequency and the inverse document frequency [5]. Often, most of the feature values are zero in the vector, such high dimensionality and sparsely located becomes a major challenge for similarity measure. This similarity measure which is a crucial operation in text processing algorithms [6], [7]. Large numbers of similarity measures have been proposed and widely applied, naming the few such as cosine similarity, Euclidean distance, and the Jaccard correlation coefficient. The pair-wise distances can be calculated when relative entropy and Euclidean distance are applied in clustering. Also for computing the similarity between two vectors many other measures are also proposed. The Kullback-Leibler divergence [8] is a non-symmetric measure of the difference between the probability distributions associated with the two vectors. The Kullback-Leibler divergence on an average shows its efficiency in the clustering of text. To obtain accuracy in clustering it requires an exact term or the definition of the closeness between a pair of objects, in terms distance or pairwise similarity. An Euclidean distance [9] is a illustrious similarity metric taken from the Euclidean geometry field. Cosine similarity [10] is a measure taking the cosine of the angle between two vectors. Jaccard coefficient will be used for comparing the similarity between two set of samples. The Jaccard coefficient deals with the similarity between the finite sets of sample which is regarded as the size of the intersection which is then divided by the size of the union of the sample sets.

The measure for computing the similarity between two documents is shown as follows. Numerous characteristics are embedded in this measure and is symmetric. The difference between absence and presence of a feature is considered. This difference is more important than the difference between the values associated with a present feature. The similarity is inversely proportional to the difference between the values associated with a present feature. The similarity increases as the value associated with present feature decreases whereas the similarity decreases when the number of absence-presence features increases. An absent feature does not have any contribution towards the similarity. The similarity measure is applied widely in many text applications which include classification and clustering, the results obtained thereby exhibits the efficiency of the proposed similarity measure [11].

The rest of the paper is divided into sections. There is a complete and brief description of the paper in related work that is covered in the section 2. The similarity measures in the paper are introduced in the section 3. Results obtained from experimental evaluations are covered in the section 4. The paper is concluded in the section 5 which covers the remarks.

II. Related Works

Similarity measures have been largely used in text classification and clustering algorithms. Yung-Shen Lin, JungYi Jiang, and Shie-Jue Lee proposed a new measure for determining the similarity between two documents. The similarity is decreased when the number of absence-presence features increases. The spherical k means algorithm introduced by Dhillon and Modha adopted the cosine similarity measure for document clustering. Zhao and Karypis showed results of clustering experiments with seven clustering algorithms and twelve different text data sets, and showed that the objective function based on cosine similarity it leads to the best solutions irrespective of the number of clusters for most of the data sets. D'hondt et al. adopted a cosine-based pairwise adaptive similarity for clustering of documents. Zhang et al. used cosine similarity to calculate a correlation similarity between two documents in a low dimensional semantic space and performed clustering of documents in the correlation similarity measure space.

Ms.K.Sruthiet. al.[22] in her paper introduced multi-viewpoint based similarity measure and related clustering methods for text data. Using multiple viewpoints, we can get more informative assessment of similarity and performance is much better than Jaccard, Euclidean or Pearson coefficient similarity measures. Jayaraj Jayabharathy and Selvadurai Kanmani [23] in their papers shows how the emphasis of the work is Dynamic document clustering which is based on Term frequency and Correlated based Concept algorithms, using semantic-based similarity measure. Dr.R.V.Krishnaiah [24] approach in finding similarity between documents or objects while clustering is multi view based similarity. Measures such as Euclidean, cosine, Jaccard, and Pearson correlation are compared. The conclusion made is that Euclidean and Jaccard are best for web document clustering.

Many measures have been proposed for computing the similarity between two vectors. The Kullback-Leibler divergence is said to be a non-symmetric measure of the difference between the probability distributions which is related with the two vectors.

Let d1 and d2 be the two documents represented as vectors. The Euclidean distance is the distance between two points in the Euclidean space. Euclidean space becomes a metric space depending on this distance. The Euclidean distance [12] measure is given in the following

$$dEuc(\mathbf{d}1,\ \mathbf{d2}) = [(\mathbf{d}1 - \mathbf{d}2)\cdot(\mathbf{d}1 - \mathbf{d}2)]\ ½ \qquad (1)$$

where **A**·**B** denotes the inner product of and d1 and d2 are the respective co ordinates.

Cosine similarity[10] measures the cosine angle between d1 and d2 as follows:

$$Scos(\text{d1,d2}) = \frac{\text{d1.d2}}{(\text{d1.d1})½(\text{d2.d2})½} \qquad (2)$$

Pairwise-adaptive similarity [13] dynamically selects a number of features from **d**1 and **d**2:

$$dPair(\text{d1, d2}) = \frac{\mathbf{d}1{,}K\cdot\mathbf{d}2{,}K}{(\mathbf{d}1{,}K\cdot\mathbf{d}1{,}K)½(\mathbf{d}2{,}K\cdot\mathbf{d}2{,}K)½} \qquad (3)$$

where **d***i*,*K* is a subset of **d***i*, *i* = 1, 2, containing the values of the features which are the union of the *K* largest features appearing in **d**1 and **d**2, respectively.

The Extended Jaccard coefficient[14] is an extended version of the Jaccard coefficient[15] for data processing:

$$SEJ(\mathbf{d}1,\ \mathbf{d}2) = \frac{\mathbf{d}1\cdot\mathbf{d}2}{(\mathbf{d}1\cdot\mathbf{d}1) + (\mathbf{d}2\cdot\mathbf{d}2) - (\mathbf{d}1\cdot\mathbf{d}2)} \qquad (4)$$

III. SIMILARITY MEASURE

The similarity measures shows closeness or separation of objects and this should be determined before clustering. This should be associated to the characteristics or properties that are supposed to differentiate the cluster that is embedded in the data. These characteristics are dependent the data. There is no pre-determined measure that is suitable for all kinds of clustering problems. The density based clustering algorithms, like DBScan [19], depend on the computation of similarity. The closeness is nothing but similarity value. Similarity measure represent the similarity between symbolic description of two objects into single numeric value.

A. *Metric*

To certify as a metric, a measure must satisfy four conditions. Let x and y be any two objects. The objects x and y are present in a set. The distance between the two objects is given by d(x,y). The following are four conditions:

1. The distance between two points must be zero or more than zero.
2. The distance between two objects must be zero iff the two objects are exactly the same.
3. Distance should be symmetric, that is, the distance from x to y is same as the distance from y to x.
4. The measure must always satisfy the triangle inequality.

B. Euclidean distance

It is the distance between two points. Euclidean distance is widely used in clustering problems. Since Euclidean satisfies all the four conditions it is considered as a true metric. The K-means algorithm uses Euclidean distance as a default distance measure.

da and db represents distance and $\overline{ta}$ and $\overline{tb}$ as term vectors. The Euclidean distance of the two documents is defined as

$$de(\overline{ta},\overline{tb})=(\sum |Wa-Wb|^2)^{1/2} \qquad (5)$$

Where T = {t1, ... , tn} stands for term set.

C. Cosine Similarity

In the case when documents are mapped as term vectors then similarity arising between two documents is same in character to the correlation between various vectors. Cosine similarity is popular similarity measure that is applied to text documents, mainly in information retrieval applications and clustering [16].

Given two documents *ta* and *tb*, then their cosine similarity is,

$$SIMc(\overline{ta},\overline{tb})= \frac{\overline{ta}.\overline{tb}}{|\overline{ta}|\ X\ |\overline{tb}|} \qquad (6)$$

Where $\overline{ta}$ and $\overline{tb}$ are vectors over the term set T = {t1, ... , tm}. Dimension represents a term with its weight and it is zero or more than zero. Resulting in the cosine similarity as non-negative and bounded between [0, 1].

Cosine similarity is independent of document length. For example, when two copies of document d are combined then we get a new pseudo document d` for this the cosine similarity between d and d` is 1,hence the documents are regarded as similar. When the documents results the same but difer in totals they will still be treated identically. This will henceforth not satisfy the second condition of a metric. However, when the term vectors are normalized to 1 the representation of d and d' is the same.

D. Similarity Measure for Text Processing

Consider a document d with m features w1, w2, . . . , wm be represented as an m-dimensional vector, i.e., d = . If wi, 1 ≤ i ≤ m, is not present in the document then di= 0. Otherwise, di>0. The following properties among other ones are desirable for a similarity measure between two documents.

The absence or presence of a feature is necessary than the difference between two values associated with a present feature. Here we consider two features wi and wj and two documents d1 and d2.

Let wi does not appear in d1 but it does appears in d2, then wi have no relationship with d1 while it has some relationship with d2. If case d1 and d2 are dissimilar in terms of wi. And if wj appears in both document d1 and d2 then wj has some relationship with d1 and d2 simultaneously. Here in this case d1 and d2 are similar to some degree in terms of wj. For the above two cases it is reasonable to say that wi carries more weight than wj in determining the similarity degree between documents d1 and d2.

The similarity degree increase when the difference between two values (that are non zero) of a specific feature decreases. For example the similarity that is involved with d13 = 2 and d23 = 15 should be smaller than that involved with d13 = 2 and d23 = 4.

The similarity degree should decline when the number of absence-presence features increases.

Two documents are considered to be least similar to each other if none of the features have non-zero values in both documents.

Similarity measure should be symmetric. The similarity degree between d1 and d2 should be same as that between d2 and d1. The standard deviation of the feature is taken into account for its input to the similarity between two documents feature with a superior spread offers more involvement to the similarity between d1 and d2.

Based on the properties discussed, a similarity measure, called Similarity Measure for Text Processing, for two documents d1 = <d11, d12, d13 ... ,d1m> and d2 = <d21, d22, d22 ... , d2m> defines a function F as follows[18]:

$$F(d1,d2) = \frac{\sum_{j=1}^{m} N_*(d_{1j},d_{2j})}{\sum_{j=1}^{m} N_\cup(d_{1j},d_{2j})} \qquad (7)$$

Then the similarity measure, S_{SMTP}, for $d_1$and d_2 is

$$S_{SMTP}(d_1,d_2) = \frac{F(d_1,d_2)+\lambda}{1+\lambda} \qquad (8)$$

This measure considers following cases-

1) The feature that we are considering should be present in both the documents,

2) The feature we are considering should be present in only one of the document, and

3) The feature we are considering should be present in none of the documents.

IV. EXPERIMENTAL RESULT

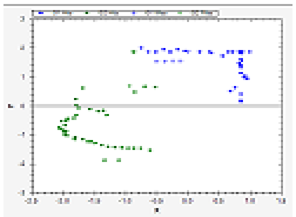
In this section results obtained from experimental evaluations is covered. We probed the efficiency of our similarity measure SMTP.

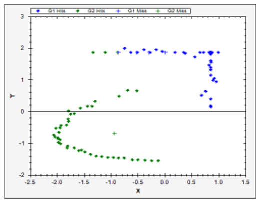
Here we are applying the measures in one or more text applications such as naïve bayes classification and k-means clustering[21] and k-NN classification. The performance of SMTP is compared with , Euclidean [12], Cosine [5]. We are selecting random documents from data sets such as WebKB [26] and Reuters-8 [25].

a)Classification and clustering Dataset

We have selected some documents randomly. Randomly selected documents are categorized into training documents and the testing documents. Training or validation is performed on training documents and testing is performed on testing documents. In every case the data for training and testing is separated. The training set is used for building up a model and this model is validated by a test set. We are removing class labels from document corpus. One third of the document is selected for training and the remaining for testing.

The performance of various distance measures is compared then the performance is checked for accuracy. The predictable label is provided by the document corpus which is later compared with that of every document. Figure 1 shows the Depiction of performance measures and Table I shows the accuracy results for Euclidean, Cosine and SMTP.


(a) Euclidean


b)Cosine

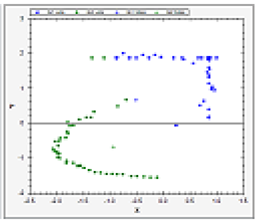
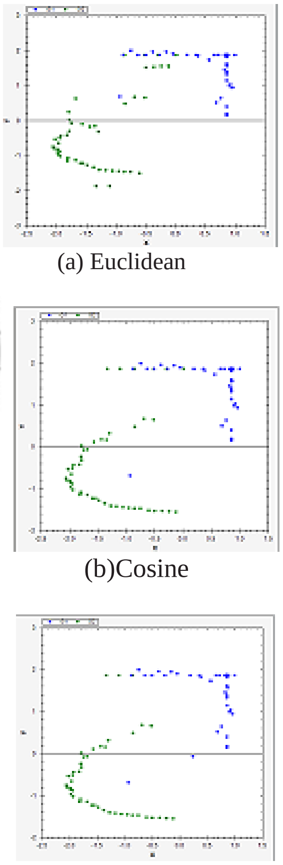

(c) SMTP

Figure 1. Depiction of Performance Measures

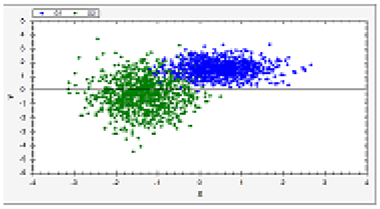
In figure 2 shows scatter plot for training data. Figure 3 shows class probabilities whose values are obtained for group G1 and G2 and this is illustrated.


(a) Euclidean

(b)Cosine

(c) SMTP

Fig. 2. Scatter Plot


(a)Euclidean

(b)Cosine

(c) SMTP

Fig. 3. Class Probabilities

TABLE I: ACCURACY RESULTS

Distance Measure	True Positives	False Negatives	True Negatives	False Positives	Sensitivity	Specificity	Efficiency	Accuracy
Euclidean	39	9	41	1	0.81	0.97	0.89	0.88
Cosine	43	5	40	1	0.89	0.97	0.93	0.93
SMTP	43	2	41	1	0.95	0.97	0.96	0.966

V. CONCLUSION

We have presented a similarity measure between two documents. Quite a few wanted or desirable properties are entrenched in this measure. For example, the similarity measure is symmetric. Numerous characteristics are embedded in this measure and is symmetric. The difference between absence and presence of a feature is considered. This difference is more important than the difference between the values associated with a present feature. The similarity degree increases when the numbers of presence-absence feature pair's decreases. If none of the features have non-zero values in both the documents then the two documents are said to be least similar to each other. It is desirable to consider the value distribution of a feature.

The probing is done for knowing the effectiveness of Euclidean distance, Cosine similarity and similarity measure for text processing. The results shows better performance by the SMTP measure compared to other measures.

REFERENCE

[1] T. Joachims and F. Sebastiani, "Guest editors' introduction to the special issue on automated text categorization," J. Intell. Inform. Syst., vol. 18, no. 2/3, pp. 103–105, 2002.

[2] K. Knight, "Mining online text," Commun. ACM, vol. 42, no. 11, pp. 58–61, 1999.

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

[4] G. Salton and M. J. McGill, "Introduction to Modern Retrieval." London, U.K.: McGraw-Hill, 1983.

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

[6] P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining," Boston, MA, USA: Addision-Wesley, 2006.

[7] M. L. Zhang and Z. H. Zhou, "ML-kNN: A lazy learning approach to multi-label learning," Pattern Recognit., vol. 40, no. 7, pp. 2038– 2048, 2007.

[8] S. Kullback and R. A. Leibler, "On information and sufficiency," Annu. Math. Statist., vol. 22, no. 1, pp. 79–86, 1951.

[9] T. W. Schoenharl and G. Madey, "Evaluation of measurement techniques for the validation of agent-based simulations against streaming data," in Proc. ICCS, Kraków, Poland, 2008.

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

[11] Anna Huang, "Similarity Measures for Text Document Clustering", NZCSRSC 2008, Christchurch, New Zealand, 2008.

[12] T. W. Schoenharl and G. Madey, "Evaluation of measurement techniques for the validation of agent-based simulations against streaming data," in Proc. ICCS, Kraków, Poland, 2008.

[13] J. D'hondt, J. Vertommen, P.-A. Verhaegen, D. Cattrysse, and J. R. Duflou, "Pairwise-adaptive dissimilarity measure for document clustering," Inf. Sci., vol. 180, no. 12, pp. 2341–2358, 2010.

[14] A. Strehl and J. Ghosh, "Value-based customer grouping from large retail data-sets," in Proc. SPIE, vol. 4057. Orlando, FL, USA, Apr. 2000, pp. 33–42.

[15] C. G. González, W. Bonventi, Jr., and A. L. V. Rodrigues, "Density of closed balls in real-valued and autometrized boolean

spaces for clustering applications," in Proc. 19th Brazilian Symp. Artif. Intell., Savador, Brazil, 2008, pp. 8–22.

[16] B. Larsen, C. Aone, "Fast and e?ective text mining using linear-time document clustering," Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.

[17] R. B. Yates, B. R. Neto, "Modern Information Retrieval," ADDISON-WESLEY, New York, 1999.

[18] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee" A Similarity Measure for Text Classification and Clustering" EEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 7, JULY 2014 1575.

[19] M. Ester, H.-P.Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," Proceedings of 2nd International Conference on KDD, 1996.

[20] G. H. Ball and D. J. Hall, "A clustering technique for summarizing multivariate data," Behav. Sci., vol. 12, no. 2, pp. 153–155, 1967.

[21] R. O. Duda, P. E. Hart, D. J. Stork, "Pattern Recognition," New York, NY, USA: Wiley, 2001.

[22] K. Sruthi, B. Venkateshwar Reddy, "Document Clustering on Various Similarity Measures", in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2013 ISSN: 2277 128X.

[23]Jayaraj Jayabharathy and Selvadurai Kanmani, ― Correlated concept based dynamic document clustering algorithms for newsgroups and scientific literature‖, Decision Analytics, Springer open journal, Puducherry, India, 2014

[24]GaddamSaidi Reddy and Dr.R.V.Krishnaiah,‖ Clustering Algorithm with a Novel Similarity Measure‖, IOSR Journal of Computer Engineering (IOSRJCE),Vol. 4, No. 6, pp. 37-42, SepOct. 2012

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

[26] [Online]. Available: http://www.cs.technion.ac.il/ ~ronb/thesis.html.