



## A Survey on Classification and Clustering Techniques Using Similarity Measures

Nutan Bire\*

Department of Computer & Savitribai Phule, Pune University,  
Pune, Maharashtra, India

---

**Abstract**—Similarity measures have been used for document classification and clustering. To calculate similarity measure in two documents or document set with respect to the distinctive attribute the similarity measure takes the following three occurrences: a) The distinctive attribute appears in both documents. b) The distinctive attribute appears in only one document and c) The distinctive attribute appears in none of the documents. In the first occurrence, similarity increases as the dissimilarity between the two documents associated with a present feature decreases. In the first occurrence, similarity decreases when the number of presence-absence features increases. In the second occurrence, a fixed value is contributed to the similarity. In the third occurrence, an absent feature has no contribution to the similarity.

**Keywords**— entropy, classifiers, text mining, document similarity, feature vector

---

### I. INTRODUCTION

The analysis of data contained in natural language text is text mining. It is equivalent to text analytics, refers to the process of deriving high-quality information from text. The high quality information is derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent inserting into the database, deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text analytics usually refers to some combination of relevance, novelty, and interestingness. The text mining tasks include text categorization, text clustering, concept or entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modelling [2].

Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging or annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The goal of text mining is turn text into data for analysis, via application of natural language processing (NLP) and analytical methods [3].

The application is read from a set of documents written in a natural language. The model of the document set for predictive classification purposes or populates a database or search index with the information extracted. Document clustering is the application of cluster analysis to text documents. The automatic document organization, fast information retrieval and topic extraction is done by document clustering. The document visualization technique is an intuitive navigation and browsing mechanism by organizing documents into groups, where each group represents a distinguish document [4].

Techniques of clustering are based on the data representation model, similarity measure, clustering model, and clustering algorithm. The document clustering methods are based on the Vector Space Document (VSD) model. The starting of data model done with a representation of any document means the feature vector of the words that appear in the documents of data in document corpus. A distinct word appearing in the documents is usually considered to be an atomic feature word in the VSD model, so words are the basic units in natural languages to represent semantic concepts. In particular, the TFIDF (tf-idf, term-frequencies and inverse document-frequencies) of the words are also contained in each feature vector. The similarity between two documents is calculated with one of the several similarity measures based on the two corresponding feature vectors, e.g., cosine similarity measure, Jaccard coefficients, and Euclidean distance [4].

Document clustering is automatically group related documents into clusters. It is one of the most important tasks in machine learning and artificial intelligence and has received much attention in recent years. To handle document clustering a number of methods have been proposed on various distance measures. Euclidean distance measure is widely used. The k-means method is one of the methods that use the Euclidean distance, which minimizes the sum of the squared Euclidean distance between the data points and their corresponding cluster centers. Since the document space is always of high dimensionality, it is preferable to end a low-dimensional representation of the documents to reduce computation complexity [5].

## II. RELATED WORK

In this section we studied, previous methods of calculating similarity measures. The previous methods include Euclidian distance measure, Cosine similarity, Jaccard coefficient, Dice coefficient.

In Cosine Similarity  $d_1, d_2$  indicates dot product of feature vectors. It measures angles between  $d_1$  and  $d_2$  [6] i.e.

$$\text{Cosine Similarity } (d_1, d_2) = \text{Dot product } (d_1, d_2) / (d_1)^{1/2} \cdot (d_2)^{1/2}$$

Euclidian distance measure is defined as root of square differences between respective coordinates of  $d_1$  and  $d_2$  i.e.

$$\sqrt{[(d_1 - d_2) \cdot (d_1 - d_2)]}$$

Jaccard coefficient uses presence or absence data [6]. The data processing is

$$S_{Jc} = |A \cap B| / |A \cup B|$$

Where,

$S_{Jc}$  = Jaccard similarity coefficient

A = First document set, B = Second document set

Dice coefficient: It is similar to Jaccard's index. It also uses presence or absence data and is given as :-

$$S_s = 2|A \cap B| / |A| + |B|$$

Where,

$S_s$  = Dice similarity coefficient

A = First document set

B = Second document set

The KullbackLeibler divergence is a non-symmetric measure of difference between probability distributions associated with two feature vectors [7]. The Hamming distance between two feature vectors is the number of positions at which the corresponding symbols are dissimilar [10]. An IT-sim [Information theoretic] is a phrase-based measure to compute the similarity based on the Suffix Tree Document (STD) model [9]. Pairwise adaptive similarity dynamically selects a number of features between two documents [8].

## III. PROPOSED SYSTEM

### A. Similarity between Two Documents

To calculate the similarity between two documents method is called as SMTP (Similarity measure for text processing.)

SMTP, for  $d_1$  and  $d_2$  is

$$\text{SMTP } (d_1, d_2) = F(d_1, d_2) + \lambda / 1 + \lambda$$

Where,

$$F(d_1, d_2) = \sum_{j=1}^m N_*(d_{1j}, d_{2j}) / \sum_{j=1}^m N_U(d_{1j}, d_{2j})$$

$d_1$  = first document

$d_2$  = second document

$\lambda$  = value according to the training data.

To calculate similarity measure in two documents with respect to the distinctive attribute the Proposed measure takes the following three occurrences: a) The distinctive attribute appears in both documents. b) The distinctive attribute appears in only one document and c) The distinctive attribute appears in none of the documents. In the first occurrence, similarity increases as the difference between the two documents associated with a present feature decreases. For the first occurrence, we set a lower bound 0.5 and decrease the similarity as the divergence between the distinctive attribute values of the two documents increases, scaled by a Gaussian function where  $\sigma_j$  is the standard deviation of all non-zero values for distinctive attribute  $w_j$  in the training data set. For the second case, we set a negative constant  $-\gamma$  disregarding the magnitude of the non-zero feature value. For the last occurrence, the feature has no contribution to the similarity [1].

### B. Similarity between Two Document sets

This method is used to calculate the similarity between two document sets. The similarity between two document sets is designed to calculate an average score of the feature occurring in the two sets.

SMTP [Similarity measures for text processing], for  $G_1$  and  $G_2$  is

$$\text{SMTP } (G_1, G_2) = F(G_1, G_2) + \lambda / 1 + \lambda$$

Where,

$$F(G_1, G_2) = \sum_{k=1}^m \sum_{i=1}^{q1} \sum_{j=1}^{q2} N_*(d_{1j}, d_{2j}) / \sum_{k=1}^m \sum_{i=1}^{q1} \sum_{j=1}^{q2} N_U(d_{1j}, d_{2j})$$

$G_1$  = first document

$G_2$  = second document

$\lambda$  = value according to the training data.

## IV. CONCLUSIONS

Literature survey and design of the Document classification and clustering Technique has been done. From literature survey it is to be concluded that there are different similarity measure techniques Proposed until and emphasized their features and limitations. The Proposed methodology uses Term Frequency Inverse Document Frequency and Stylometry features which increases accuracy and Email Author Identification. By using Naive Bayes, ID3 classifiers correct the prediction and test the performance.

**REFERENCES**

- [1] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering" In IEEE Transaction on Knowledge and Data Engineering, Vol. 26, No.7, July 2014.
- [2] Umajancy .S, Dr. Antony Selvadoss Thanamani"An Analysis on Text Mining-Text Retrieval and Text Extraction" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 8, August 2013
- [3] Shaidah Jusoh and Hejab M. Alfawareh "Techniques, Applications and Challenging Issue in Text Mining" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012
- [4] Hung Chim and Xiaotie Deng, "Efficient Phrase-Based Document Similarity for Clustering," In IEEE Transaction on Knowledge and Data Engineering, Vol. 20, No.9, Sept 2008.
- [5] Taiping Zhang, Yuan Yan Tang, Bin Fang, and Yong Xiang, "Document Clustering in Correlation Similarity Measure Space," In IEEE Transaction on Knowledge and Data Engineering, Vol. 24, No.6, June 2012.
- [6] Ms.K.Sruthi, "Document Clustering on Various Similarity Measures," International Journal of Advanced Research in Computer Science and Software Engineering, Apr 2013.
- [7] S. Kullback, R.A.Leibler, "On information and sufficiency", *Annu. Math. Statist.* , Vol. 22, No. 1, pp. 79-86.
- [8] J. D'hondt, J. Vertommen, P.A. Verhaegen, D. Cattrysse & R.J. Duflou, "Pairwise-adaptive dissimilarity measure for document clustering", *Inf. Sci.*, Vol. 180, No. 12, pp. 2341-2358.
- [9] D. Lin, (1998) "An information theoretic definition of similarity", in *Proc. 15th Int. Conf. Mach.Learn.*, San Francisco, CA, USA.
- [10] R.W. Hamming, "Error detecting and error correcting codes", *Bell Syst. Tech. J.*, Vol. 29, No.2, pp. 147-160.