Scientific Journal of Impact Factor (SJIF): 3.134

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 1, January -2016

A Similarity Measure for Text Classification and Clustering

Swati Hatekar[1], Snehal Pagar[2], Ashish Tayde[3], Vikas Lande[4] ,Mrs.Poonam Sawdekar[5]

[1,2,3,4,5]*B.E. Computer Engineering, D.Y.P.I.E.T., Pimpri, Pune*

Abstract —*Clustering is one of the necessary techniques in machine learning and data mining techniques. Similar data grouping is performed using clustering techniques. In document vector each component indicates the value of the corresponding feature in the document. The characteristic measure could be term frequency, are similar to relative term frequency. Similarity Measurement for Text Process (SMTP) is used to measure the similarity between two documents with respect to a feature. Presents and options of the features in both documents are used to estimate the similarity values. The SMTP is extended to estimate similarity between two set of documents. The SMTP scheme is used with text clustering and classification task. K means algorithm is used for the clustering techniques.*

Keywords*- Document classification, document clustering, entropy, accuracy, classifiers, clustering algorithms .*

I. INTRODUCTION

The text processing plays an essential role in information retrieval, data mining, and web search In this technique, the bag-of-words model is commonly used. A document is usually represented as a vector in which each element shows that the value of the corresponding feature in the document. The feature value can be term frequency (the number of occurrences of a term appearing in the document), relative term frequency (the ratio between the term frequency and the total number of occurrences of all the terms in the document set), or tf-idf (a combination of term frequency and inverse document frequency). Usually, the dimensionality of a document is large and the resulting vector is sparse, i.e., most of the measure values in the vector are zero. Such high dimensionality and sparsity can be a severe challenge for similarity measure which is an essential operation in text processing algorithms. A lot of measures have been approach for computing the similarity between two vectors. The Kullback-Leibler divergence is a non-symmetric measure in the difference between the probability distributions assigned with the two vectors. Euclidean distance is a well-known similarity metric will be taken from the Euclidean geometry field. Manhattan distance, similar to Euclidean distance and also known as the taxicab metric, is different similarity metric. The Canberra distance metric is used in situations, then the elements in a vector are always non-negative. Cosine similarity is a measure taking the cosine of the angle between two vectors. The Bray-Curtis similarity technique is a city-block metric which is sensitive to outlying values. The Jaccard coefficient is a statistic and it will be used for comparing the similarity of two sample sets, and is defined as the size of the intersection divided by the size of the only one of the sample sets. The Hamming distance between two vectors is the number of positions at the time corresponding symbols are different. The extended Jaccard coefficient and the Dice coefficient retain the property of the cosine similarity measure technique while allowing discrimination of collinear vectors. An information-theoretic calculate the document similarity, named IT-Sim, says that in. Chim et al. This phrase-based measure to compute the similarity are related to the Suffix Tree Document (STD) model. Similarity measures have been extensively used in text classification and clustering algorithms. The spherical k -means algorithm introduced by Dhillon and Modha adopted the cosine similarity measure for document clustering. Zhao and Karypis reported results of clustering experiments with 7 clustering algorithms and 12 different text data packets, and decided that the objective function based on cosine similarity "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 set of adaptive similarity for document clustering. Zhang et al. Used cosine to measure the correlation between two similar projected documents in a low-dimensional semantic space and performed document clustering in the correlation similarity measure space. Kogan et al. says that, two step clustering procedure in which the PDDP is used to0 generate starting partitions in the first step and a k-means clustering algorithm using the Kullback-Leibler divergence is applied in the second step. Dhillon et al. says that the divisive information-theoretic feature clustering algorithm for text classification on using the Kullback-Leibler divergence. Euclidean distance is usually the default choice of similarity-based methods, e.g. k-NN and k-means Kogan et al. combined squared Euclidean distance with relative entropy in a k-means like clustering algorithm. Chim et al. Performed document clustering based on the proposed phrase based similarity measure. The extended Jaccard coefficient can be used for document data and it reduces to the Jaccard coefficient in the case of binary attributes. We propose a new measure for computing the similarity between two documents. Several characteristics are embedded in this measure. It is a symmetric measure. The difference between presence and absence of a feature is considered more essential than the difference between the values associated with a present feature. The similarity increases as the difference between the two values assigned with the present feature decreases. Furthermore, the contribution of the difference is normally scaled. The similarity exceeds when the number of presence-absence features reduced. An absent feature has no contribution to the similarity. The proposed measure is extended to gauge the similarity between two sets of documents. The measure is applied in indefinite text applications, adding with single label classification, multi-label classification, k-means like clustering, and hierarchical clustering are collected, and the results will be prove the effectiveness of the proposed similarity measure.

II. LITERATURE SURVEY

1. In this text processing, the similarity measure is the important process. It calculate the similarities between the two documents. The computation of similarity measurement is based on the feature of two documents. This system contains three cases to compute the similarity. The three cases are, both two documents contains features, only one document contains feature, there is no feature into the two documents. In first case, the similarity is exceeds when the differences of feature value is reduced between the two documents, then the given differences are measured. In second case, fixed value is given to the similarity. In third cases there is no contribution to the similarity. Finally this method achieves the better performance compared than other measurement.

2. A novel document clustering method which aims to cluster the documents into different semantic classes is introduced. The document having high dimensionality and clustering in such a high dimensional space is often infeasible due to the fault of dimensionality. In Locality Preserving Indexing (LPI), the documents can be projected into a lower-dimensional space when in which the documents related to the all semantics are close to each other. Different from previous document clustering methods are to be based on Latent Semantic Indexing (LSI) or Nonnegative Matrix Factorization (NMF), this method help to discover both the geometric and discriminating structures of the document space.

3. As one of the most fundamental yet important methods of data clustering, centre-based partitioning approach clusters the dataset into k subsets, each of which is represented by a centered or medoid. A new medoid -based k-partitions approach called Clustering around Weighted Prototypes (CAWP), which works in the similarity matrix. In CAWP, each cluster is featured by many objects with different respective weights. The new cluster representation scheme, CAWP aims to simultaneously formed clusters of modified quality and a packet of ranked representative objects for each cluster. An efficient algorithm is derived to different update to the clusters and the representative weights of objects with respect to each cluster. A hardening like optimization procedure is enhanced to alleviate the local optimum problem for better clustering results and same time to make the algorithm less reactive to parameter setting. Experimental results on benchmark document datasets show that, CAWP achieves favorable effectiveness and efficiency in clustering, and also provides useful information for cluster-specified analysis.

4. The text mining is the analysis of data contained in natural language text. It is the process of extracting information from text. Text analysis includes information retrieval, lexical analysis, pattern recognition, information extraction. The main demand in text mining is to find the similarity between documents in order to group the similar documents. The word frequency distributions are identified to find the similarity between various documents. Vector space model was used to classify the related documents. Long documents are poorly represented because they have poor similarity values and keywords must precisely match the document terms. Documents might have similar context but different term vocabulary won't be associated which leads to less accuracy. Spectral based approach was used and it will be helpful for the short queries.

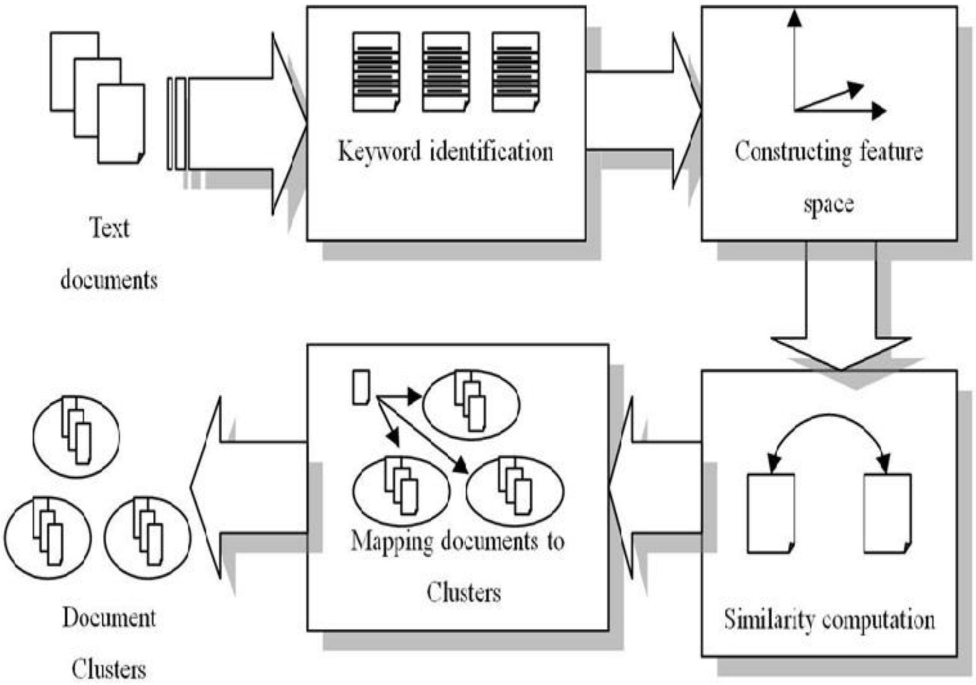
5. The incredible progress of computer technology in the few decades has attended to large supplies of powerful and popular computers. Increase in the number of electronic documents it is hard to visualize these documents efficiently by putting human application. These have brought challenges for the effective and efficient organization of web page documents automatically. Extracting features from web pages is initial task found in mining. On the basis of extracted features similarity between web pages are going to be calculated. There is various similarity measures are pointed out for work. To implement the efficient similarity measure one has to do survey on outcomes.

6. The conventional feature selection classifiers work with known and actual data values. In recent data collection methods, definite amount of attributes are inconsistent. The inconsistent attributes, in almost all applications, have more influences on the data set on information classification and element selection constructs. Uncertainty needs to be handled properly reasons for uncertainty are due to measurement errors, quantization errors, weak data and multiple repeated measurements. The inconsistent of a document item is represented in terms of multiple values. Usually uncertain document are abstracted by statistical derivatives Complete information of the data item improves the accuracy of feature selection algorithm. Clustering techniques increase the speed of feature selection construction and minimize the reduced time to greater extent. Distance boundary clustering technique, works based on the criteria of lower and upper bounds distances of the uncertain attributes values.

Summary of Literature Survey

No	Method Name	Advantages	Disadvantages
1	Lazy learning: Special issue editorial [3].	low computational load and fast.	Low accuracy.
2	Probabilistic models of information retrieval based on measuring the divergence from randomness[4].	The advantage of having a nonparametric approach is that the derived models do not need to be supported by any form of data-driven methodology, such as the learning of parameters from a training collection, or using data smoothing techniques.	This experiments should be performed to assess the effect on performance of word stemming, document pruning, and word pruning and to include these factors as explicit variables within the framework.
3	A clustering techn-ique for summarizing multivariate data[6].	nats-model performance more closely matched the performance of the Dice measure.	Worst performers were the cosine measure and the corpus independent model.
4	Document clustering using locality preserving indexing [7].	It is capable to work highly non-linear data .	some empirical estimation on the dimensionality using LPI. However, it lacks a strong theoretical foundation.
5	Concept Decompo-sitions for large sparse text data clustering [8].	These models are very high-dimensional and sparse, and present unique computational and statistical challenges not commonly encountered in low-dimensional dense data. Computational and statistical challenges not commonly encountered in low-dimensional dense data.	any proposed statistical model for text data should be consistent with this fractal behavior. In fact, it might be meaningful to seek maximum entropy distributions subject to such empirical constraints fractal behaviour. In fact, it might be meaningful to seek maximum entropy distributions subject to such empirical constraints.

III. SYSTEM ARCHITECTURE



- Initially take input text document, after that system read the all the input documents.
- Apply stemming:
  A stemming algorithm is a process of linguistic normalization, in which the variant forms of a word are reduced to a common form, for example, Connection, connected, connecting, connections, connective, ---> connect removing stopword ::
- Read stopword file.
- Remove stopwords from file and find out unique keywords.
  Keyword identification::
- Check frequency of each keyword and select those keywords whose frequency is greater than threshold value and add in final keywords list.

Similarity Computation:

- Use here SMTP formulae for finding similarity between two documents.
  Clustering:
  K means algorithm use for clustering documents and we use here SMTP similarity measure at the time of clustering.

Condition to generate vector:

- Generate the vector of each word in tf-idf value is calculated. tf–idf, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.
- After that similarity between the words are evaluated.
- Mapping Document to Clusters: **Clustering** is the task of grouping a set of objects in such a way that objects in the same group is more similar to each other than to those in other groups.

IV. CONCLUSION

Thus, we have studied and applied techniques like spherical k-means, k-NN, Locality Preserving Indexing(LPI), Latent Semantic Indexing (LSI) and Information Retrieval model for similarity computation.

REFERENCES

[1] P. K. Agarwal and C. M. Procopiuc, "Exact and approximation algorithms for clustering," in *Proc. 9th Annu. SODA, Philadelphia*, PA, USA, 1998, pp. 658– 667.

[2] D. W. Aha, "Lazy learning: Special issue editorial," *Artif. Intell. Rev., vol. 11, no. 1–5, pp. 7–10, 1997.*

[3] G. Amati and C. J. V. Rijsbergen, "Probabilistic models of Information retrieval based on measuring the divergence from randomness," *ACM Trans.Inform. Syst., no. 4, pp.357–* 389,2002.

[4] J. A. Aslam and M. Frost, "An information-theoretic measure for document similarity," in *Proc. 26th SIGIR, Toronto, ON, Canada*, 2003, pp. 449–450.

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

[6] D. Cai, X. He, and J. Han, "Document clustering Using locality preserving indexing," *IEEE Trans. Knowl. Data Eng., vol.17, no. 12,* pp. 1624–1637, Dec. 2005.

[7] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Mach Learn., vol. 42, no. 1*, pp. 143–175, 2001.

[8] H. Chim and X. Deng, "Efficient phrase-based document similarity for clustering," *IEEE Trans Knowl. Data Eng., vol. 20, no. 9*,pp. 1217–1229, Sept. 2008.

[9] 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.

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

[11] M. B. Eisen, P. T. Spellman, P. O. Brown, and D.Botstein, "Cluster analysis and display of genome-wide expression patterns," *Sci.*, vol. 95, no. 25, pp. 14863–14868, 1998.