Document Similarity Measure Using Selection of Present-Absent Feature Approach

Ms. Bhawna Gayakwad [1], Dr. S. D. Choudhari [2]
[1 & 2]M.Tech. CSE, Department, SBITM COE, Betul, India [1], Professor SBITM COE, Betul, India
Email - bhawnagayakwad@gmail.com[1], choudhari.sachin1986@gmail.com[2]

Abstract: In the text document processing field finding the similarity between several documents is a key operation. In this paper, we proposed an innovative similarity measure for text and pdf document clustering. To figure out the similarity between several text and pdf documents with respect to a feature, our implemented similarity finding measure takes the various following cases into account:

Case 1: The feature we selected may appear in both documents.
Case 2: The feature we selected appears in only one document
Case 3: The feature we selected appears in none of the documents.

As we know in the first case, the documents features are similar thus similarity actually increases and the difference between the selected features values are very less. In addition, the involvement of the feature difference is generally scaled by feature values. In the second case, a the similarity between different document is naturally less as feature we selected appears in only one document and thus feature difference is very high. In the last case, the selected features are totally absent between the documents and thus have no contribution to the document similarity.

Our proposed measure is comprehensively estimating the appropriate similarity between various document sets. The usefulness of our similarity measure is evaluated on several text document sets for the text classification and clustering problems. The results show that the performance obtained by the proposed measure is significantly better than that achieved by other measures.

Keywords: Document classification, document clustering, accuracy, entropy classifiers, clustering algorithms.

1. INTRODUCTION:

Text document processing plays a key role in data mining as well as web search for information retrieval. In text processing, the commonly used model is bag-of-words model [5]. In this model each document is typically represented in vector form in which each element indicates the value of the analogous feature in the document. The feature value can be selected by finding number of occurrences of a term in the document. However relative term frequency can be defined as the ratio between the term frequency and the total number of occurrences of all the terms in the document set. Frequently, the dimensionality of a document is large and the resulting vector is sparse, i.e., most of the selected feature values in the vector are zero. Such high-dimensionality and sparsity is a challenge for similarity measure and thus it is a very important operation in text processing algorithms.

A several measures have been proposed for computing the similarity between two document vectors. The Kullback-Leibler divergence [3] proposed a non-symmetric measure of the difference between the probability distributions associated with the two vectors. Euclidean distance [5] is a recognized similarity metric taken from the Euclidean geometry field. Manhattan distance [11], is very similar to Euclidean distance and also recognized as the taxicab metric, is another popularly used similarity metric. The Canberra distance metric [6] is frequently used in situations where vector elements are always nonnegative. Cosine similarity [2] is again significantly used measure taking the cosine of the angle between two vectors.

The Bray-Curtis similarity measure [3] proposed a city-block metric which is very sensitive to outlying values. The Jaccard coefficient [2] is a statistic generally used for comparing the similarity of two document sets, in this approach the size of the document feature intersection divided by the size of the union of the document

sets. The Hamming distance, [12] between two document vectors is the number of positions at which the related symbols are totally different.

In this paper we used novel measure for finding the similarity between two documents for that some characteristics are embedded in this proposed measure. It is a symmetric measure where the difference between absence and presence selected feature is considered essential than the difference between the feature values associated with a present terms. The similarity between documents increases as the difference between the two values associated with a present feature decreases. Furthermore, the contribution of the difference is normally scaled. The similarity decreases when the number of presence-absence features increases. An absent feature has no contribution to the document similarity measure. The measure is applied in several text data set applications, and use k-means like clustering, the results obtained exhibit the usefulness of the proposed similarity measure.

2. RELATED WORKS:

Some popular measures which have been used for finding the similarity between two documents sets are briefly described here.

Consider **d**1 and **d**2 be two documents and they can be represented as vectors. The Euclidean distance measure between these two documents is nothing but as a root of square differences between the respective coordinates of **d**1 and **d**2 which is shown below:

EucDistance(**d**1, **d**2) = [(**d**1 − **d**2)· (**d**1 − **d**2)] 1/2 ----------(1)

Cosine similarity measures is generally used to find the cosine of the angle between document **d**1 and **d**2 as shown:

SimCos(**d**1, **d**2) = **d**1· **d**2 (**d**1· **d**1)1/2(**d**2· **d**2)1/2----------(2)

Pair wise-adaptive similarity [17] uses dynamically selected number of features between **d**1 and **d**2 and is defined as follows:

dPair(**d**1, **d**2) = **d**1,K · **d**2,K (**d**1,K· **d**1,K)1/2(**d**2,K· **d**2,K)1/2---------- (3)

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

The Extended Jaccard coefficient [48], [49] is an extended version of the Jaccard coefficient [12] for data processing the formula used is as follows:

SEJ(**d**1, **d**2) = **d**1· **d**2 **d**1· **d**1 + **d**2· **d**2 − **d**1· **d**2----------(4)

In this study we found that all the measures have significant effect on partitioned clustering of text documents except for the Euclidean distance measurer. Jaccard correlation coefficient is slightly better as the resulting clustering solutions are more balanced and is nearer to the manually created categories. We can see that there are three components that affect the final results representation of the documents, distance or similarity measures considered, and the clustering algorithm itself.

3. IMPLEMENTED DOCUMENT SIMILARITY MEASURE:

The basic idea regarding document similarity measures have been popularly used in text classification and clustering approach. A measure for finding the similarity between documents which contain several characteristics that is it generally a symmetric measure, the difference between absence and presence of a feature is considered more important than the difference between the values associated with a present feature. As the similarity increases, difference between the two values associated with a present feature automatically decreases. The similarity decreases when the number of absence-presence features increases. It affects clustering quality. Here is need of better similarity measure to improve clustering outputs.

Implemented system architecture gives brief idea about what our system is. System architecture is the conceptual model that defines the structure, behaviour and more views of a system in a systematic way. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures of the system. System architecture comprises system components, the externally visible properties of those components, the relationships between them. It provides a plan from which system developed. In text mining and classification there are various steps involved.

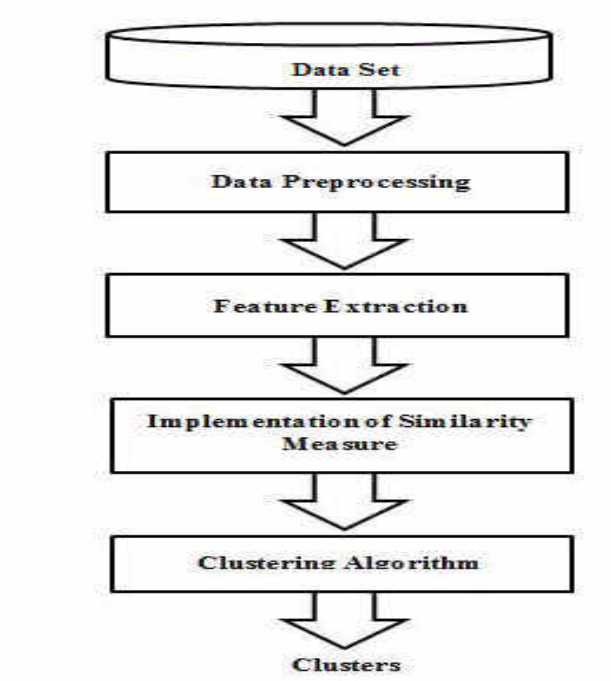

Figure 3.1: Summarizes the proposed system architecture

In our proposed system sample dataset containing various different domain data text and .pdf file is given as an input to our framework where first steps is to read the given file format data appropriately and remove the unwanted characters like stop words from the given dataset as well as steaming is performed this whole process is called as data pre-processing. After data processing step the next step is to identify the most frequently used words in the given dataset and create a feature file containing the feature words of different documents. Based on these feature word create a feature vector for each document and store them for further processing. Afterwards novel similarity measure, NSMM (Novel similarity measure) is used for document let **d**1 and **d**2 is shown below NSMM (**d**1, **d**2) = F(**d**1, **d**2) + λ1 where λ1 is constant value. The last step is to make a cluster based on similarity value calculated in previous step by using K-Means algorithm as shown below:

K-Means Algorithm:

Step1: Select K points as the initial centroids where K=N where N=1 any threshold value less than number of documents.

Step2: Assign all points to the closest centroid.

Step3: Recompute the centroid of each cluster.

Step4: Repeat steps 2 and 3 until the centroids don't change.

The applied k-means algorithm has been shown to be effective in producing good clustering results for many practical applications. As k-means finding centroid in iterative manner this set of information will help us to find manual justification to output. Further, with the help these results document clusters are formed.

4. CONCLUSION:

We have described an innovative similarity measure of different domain document. There are several important properties are embedded in this. The absence or presence of a selected feature is considered to be very significant than the difference between the values associated with a present feature. The application of document clustering to information retrieval has been motivated by the possible effectiveness gains postulated by the cluster hypothesis. Hence, the implementation of similarity measure for clustering is initially motivated by a research on automated text categorization. By optimizing similarity measures the best possible clusters can be formed thus performance is improved. Overall aim is organizing data in such a way that to improve data availability and to fasten data access, so that text information retrieval and content delivery should be improved.

REFERENCES:

1. Yung-Shen Lin, Jung-Yi Jiang and Shie-Jue Lee, \A Similarity Measure for Text Classi cation and Clustering", IEEE Transactions On Knowledge And Data En-gineering, 2013.
2. Gaddam Saidi Reddy and Dr.R.V.Krishnaiah, \Clustering Algorithm with a Novel Similarity Measure", IOSR Journal of Computer Engineering (IOSRJCE), Vol. 4, No. 6, 2012, pp. 37-42.
3. Shady Shehata, Fakhri Karray, and Mohamed S. Kamel, \An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 10, 2010.
4. Anna Huang, Department of Computer Science, The University of Waikato, Hamilton, New Zealand, \Similarity Measures for Text Document Cluster-ing", New Zealand Computer Science Research Student Conference (NZCSRSC), Christchurch, New Zealand, April 2008.
5. H. Chim and X. Deng, \E cient phrase-based document similarity for clustering",
6. IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 9, 2008, pp. 1217-1229.
7. Yanhong Zhai and Bing Liu, \Web Data Extraction Based on Partial Tree Align-ment", International World Wide Web Conference Committee (IW3C2), 2005,1-59593-046.
8. J. Kogan, M. Teboulle and C. K. Nicholas, \Data driven similarity measures for k-means like clustering algorithms", Information Retrieval, Vol. 8, No. 2, 2005,331-349.
9. S. Dhillon, J. Kogan and C. Nicholas, \Feature Selection and Document Clus-tering", In Berry MW Ed. A Comprehensive Survey of Text Mining,2003.
10. Syed Masum Emran and Nong Ye, \Robustness of Canberra Metric in Computer Intrusion Detection", IEEE Workshop on Information Assurance and Security United States Military Academy, West Point, NY, 2001, pp. 5-6.
11. Alexander Strehl, Joydeep Ghosh, and Raymond Mooney, \Impact of Similarity Measures on Web-page Clustering", Workshop of Arti cial Intelligence for Web Search, 2000.
12. S. Kullback and R. A. Leibler, \On information and su ciency", Annuals of Mathematical Statistics, Vol. 22, No. 1, 1951, pp. 79-86.
13. Mei-Ling Shyu, Shu-Ching Chen, Min Chen and Stuart H. Rubin, \A nity-Based Similarity Measure for Web Document Clustering", Distributed Multimedia In-formation System Laboratory, School of Computer Science Florida International University Miami, FL 33199, USA., 4 (5), 2011, pp. 79-83.
14. N. Sandhya, Y.Sri Lalitha, Dr.A.Govardhan and Dr.K.Anuradha, \Analysis of Similarity Measures for Text Clustering", GRIOET, Hyderabad, India.