Accepted Manuscript

Classification of compressed and uncompressed text documents

S.N. Bharath Bhushan, Ajit Danti

PII: S0167-739X(17)32070-8
DOI: https://doi.org/10.1016/j.future.2018.04.054
Reference: FUTURE 4132

To appear in: *Future Generation Computer Systems*

Received date : 15 September 2017
Revised date : 17 April 2018
Accepted date : 19 April 2018

Please cite this article as: S.N.B. Bhushan, A. Danti, Classification of compressed and uncompressed text documents, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.04.054

Full Manuscript Title: **Classification of Compressed and Uncompressed Text Documents.**

First Author Name: S. N. Bharath Bhushan.

Affiliation: Jawaharlal Nehru National College of Engineering, Shivamogga – 577 204.

Second Author: Ajit Danti.

Affiliation: Jawaharlal Nehru National College of Engineering, Shivamogga – 577 204.

Corresponding Author Details:

S. N. Bharath Bhushan

Department of Computer Applications,

Jawaharlal Nehru National College of Engineering,

Shivamogga – 577 204.

Email: sn.bharath@gmail.com

Phone: +91 94807 66063.

Abstract: Computing the degree of closeness (similarity) between two sets of text documents is one of the core operations in many text mining applications like text classification, clustering and sentiment analysis. The efficiency such applications is mainly depends on the factors like selection of representation model, selection of the similarity metric and selection of learning algorithms. Among these three factors, selection of similarity measure is important since it contributes a lot to the efficiency of most of the text mining applications. In this research article, an efficient similarity measure is proposed for computing the closeness of two sets of text documents. The proposed measure has the capacity of considering different real time situations like presence of a feature or absence of features for computing the degree of similarity between the documents. Furthermore, a compression modeling similarity measure is also proposed for text documents. Two different sets of experiments are conducted to validate the efficacy of the proposed similarity measures. Experimental results demonstrate that, the f-measure score obtained from proposed similarity metric is better than the f-measure score of the existing state of the art techniques.

Keywords: Text Classification, Similarity measure, Text Compression.