ADICHUNCHANAGIRI INSTITUE OF TECHNOLOGY
CHIKMAGALUR-577102

PROJECT REPORT

PROJECT TITLE:

"A SIMILARITY MEASURE FOR CLASSIFICATION AND CLUSTERING IN TEXT BASED BANKING AND IMAGE BASED MEDICAL APPLICATIONS"

Submitted to:

The Secretary,
Karnataka State Council for Science and Technology (KSCST)
Indian Institute of Science Campus,
Bangalore- 560012.

From,

1. Mr. Ramesh Patel (USN 4AI10IS042)
2. Ms. Sahana Madpur (USN 4AI10IS044)
3. Ms. Sreenitha A.S (USN 4AI10IS050)
4. Mr. Subhash S.S (USN 4AI10IS051)

(Signature of Guide)	(Signature of HOD)	(Signature of the Principal)
Mr. Sandesh K.P.	Mr. Sampath.S.	Dr. C.K. Subbaraya.
Assistant Professor	Assistant Professor & HOD	Principal
Dept of IS & E, A.I.T.	Dept of IS & E, A.I.T.	A.I.T.
Chikmagalur.	Chikmagalur.	Chikmagalur.

13/06/2014

ABSTRACT

Finding similarity between documents is a vital issue, since making machine to understand the meaning of the content.An approach has to be developed to make the best similarity match.In this project we propose a technique for document matching text based similarity matching.The effectiveness of our measure is evaluated on several real-world data sets for text classification and clustering problems.

Measuring the similarity between documents is an important operation in the text processing field. In our paper, a new similarity measure is proposed. To compute the similarity between two documents with respect to a feature, the proposed measure takes the following three cases into account: a) The feature appears in both documents, b) the feature appears in only one document, and c) the feature appears in none of the documents. For the first case, the similarity increases as the difference between the two involved feature values decreases. Furthermore, the contribution of the difference is normally scaled. For the second case, a fixed value is contributed to the similarity. For the last case, the feature has no contribution to the similarity. The proposed measure is extended to gauge the similarity between two sets of documents. The effectiveness of our measure is evaluated on several real-world data sets for text classification and clustering problems. The results show that the performance obtained by the proposed measure is better than that achieved by other measures.