ABSTRACT

In the last decades, the volume of text databases has rapidly grown due to the increasing amount of information available in electronic forms, such as WWW, emails, newsgroup messages, Internet news feeds, digital libraries, etc. Clustering can help in organizing the text collection for efficient browsing and searching. This has been a driving force for making clustering a highly active research area. Document clustering is a subset of the larger field of data clustering, which borrows concepts from the fields of Information Retrieval (IR), Natural Language Processing (NLP), and Machine Learning (ML), among others.

In this thesis, we propose a new document clustering algorithm where the concepts of fuzzy sets have been used. The proposed algorithm is agglomerative and at any given stage of the algorithm there are small clusters and the decision at the current stage is to merge the incoming document with the cluster that satisfies a user specified threshold. The clusters obtained are represented as fuzzy sets over a finite universal set which provides a compact representation of clusters. A similarity measure based on the fuzzy representation of the clusters is defined. The algorithm requires just one pass through the dataset and only the compact representations of the clusters are kept in the memory at any given time. Our algorithm is incremental and can deal with the dynamic nature of real world data.

With arbitrarily large datasets, the datasets cannot fit in memory. Several clustering algorithms are proposed for large datasets which follow a two-phase
approach. We propose a two-phase approach to the clustering problem of large dataset. In the first phase, a single pass over the database is used to produce an in-memory summary of the data set. In the second phase, the in-memory summary of the data set obtained in the previous phase is merged based on the concepts of neighbors and links.