ABSTRACT

For the past few years automated content-based document management tasks have become important in information systems field due to the availability of voluminous documents in digital form. Because of the large volume, information access becomes tedious. Text categorization evolved for arranging the documents and to provide an easier information access to the users. Text categorization (or classification) is the activity of assigning thematic label to the text documents from a predefined category labels. Text categorization is also vital in many other applications like summarization, question answering systems, document filtering etc. Thus improving the text classification in turn improves many related applications. This motivated us to work in the area of text categorization.

Text categorization is an area that is very well-researched and many methods have been suggested, each evolution brought in some extra complexity in the work. The aim of this research is to develop a categorization system that is computationally simple. Profile based classifiers are the best model for computational simplicity and hence this attracts us for the appropriate vehicle for doing categorization. There are two possibilities of improvement in such classifiers. The first is that the representations of profiles are far from perfect. We have identified the basic problem of the lower performance of the existing profile based classifiers as the lack of attention to inter-category relationship while obtaining class profiles. We
propose that if the distance between profiles is increased, the performance of such classifiers will surely improve. We have introduced the profile creation by using status based method and category based method. In status based method, profiles are built according to the status value of the terms in the profile. In category based method, a term’s importance is decided using category-related information.

The second improvement comes from the role played by the similarity measures. Classifier performance depends on the quality of the similarity measure used. Even with good profile representation, if the similarity measure does not perform well, it will affect the performance of the classifier. This is very much true for the profile based classifiers. To address this issue, we have introduced a new similarity measure called Fractional Similarity measure, which was further refined as Modified Fractional Similarity or MFractional Similarity method to perform well with various types of corpora.

Text classification is used in many applications. Story link detection is an application where the classification approaches fit very well as they share many common tasks like document representation and similarity measure between documents.

Link detection is a process of identifying link between any two random documents. In this context, link is used to mean whether two given documents are discussing the same event. In this regard, we have applied our MFractional Similarity measure for link detection system. Researchers have
proposed to use query expansion techniques in link detection system and reported to improve the performance. This motivated us to use query expansion approach in our efforts in link detection. Using query expansion technique, given documents are expanded and corresponding models are built for the given documents. For the purpose of expansion, we use a set of relevant documents retrieved from the corpus. Models for the two documents thus built are compared. For query expansion based link detection system the need of a good information retrieval system have been observed since models built from the relevant document sets play a very important role in the overall link detection system. For this purpose, some entity information and cluster similarity measure have been introduced in information retrieval system. For building models, we have proposed cohesion model. We have further experimented with SVM based systems, as SVM is a well established technique in classification. Both the models work better than the popular cosine similarity method.

We have used three standard corpora to experiment our proposed methods. Reuters 21578 and 20 Newsgroups are used for text classification and TDT4 corpus is used for story link detection system. We have used F1-measure and accuracy for evaluating the text classification system. For story link detection system we have used cost measure in addition to F1-measure and accuracy.

Our proposed status based method performs better than the centroid based profile, which use term frequency and tfidf. With the introduction of category information in the term weighting scheme, we have obtained
excellent classifier performance. Performance of category based classifier is better than the status based method. Proposed Fractional Similarity measure performs well in 20 Newsgroups, but its performance is not as good in Reuters corpus. However MFractional Similarity measure performs well in both the corpora. In link detection system, both the SVM based and Cohesion model worked much better than cosine similarity. Performance of link detection improved when improved information retrieval systems are used.