CHAPTER 7

CONCLUSION AND FUTURE WORK

This chapter summarizes the contributions of this thesis, and provides pointers to future work.

7.1 CONCLUSION

The primary aim of this research is to extract features from a document and to cluster the text documents. To achieve this, (i) Text clustering using the hierarchical algorithm, and the K-means algorithm based on association rule mining have been developed. (ii) A Genetic based keyword extraction technique has been designed and implemented for feature extraction. (iii) For clustering the extracted keywords, a Must Link and Cannot Link algorithm (MLCL) has been designed and implemented. (iv) Next, features are extracted based on sentences, and the extracted sentences are clustered using WordNet.

The Association Rule based hierarchical clustering algorithm described in this thesis, is an integration of the Association Rule and the Hierarchical algorithm. The main objective in developing this, is to improve the efficiency of the clustering algorithms even in a large document space. The documents that are associated with one another are identified using association rule mining, and then they are clustered using the Hierarchical algorithm. The algorithm is evaluated using F-measure, entropy values, cophenetic matrix, and the time taken to cluster the documents. The analysis
says that, the algorithm has reduced F-measure values and increased entropy values. The cophenetic matrix has shown that the quality of the clusters formed is very good, and the time taken to cluster the documents is also reduced.

An efficient approach for clustering the text documents, using the K-means algorithm has been proposed. In this, a method is proposed to locate the initial centroids. It is been analysed that locating the initial centroids at a farther distance will produce better clustering output. So, taking this into consideration, the Association Rule mining algorithm was used to find the initial centroids. The Association Rule mining algorithm finds the associated documents and then those documents are removed from the database. The remaining dissimilar documents are taken as initial centroids and then the documents are clustered using modified K-means algorithm. The clusters thus formed is evaluated, using the F-measure and Entropy values, and the values have shown better performance. The number of iterations taken to cluster the documents is also calculated, and it has been shown to be less.

A Genetic based keyword extraction technique has been developed for feature extraction. In this, the words in the document are assigned a numerical value using various formulae. The words are assigned a weight value based on parameters like ‘position of a term in a given sentence’ and ‘Average position in the whole document’. Average position of a word gives importance to a word which occurs many times at diverse positions in the document than a word which occurs only in the first sentence. The number of words in each sentence is also taken into consideration. Thus the words are weighted with respect to the document and each sentence which increases the precision and accuracy of the term weights. As a word is repeated, an incremental weight value is computed. The incremental weight value is calculated based on the number of words in the sentence in which it currently
appears and the improved average position of the term. This increases the weight of the word and thus the priority of the word also gets increased. As a result, a best set of keywords has been produced and are then passed into the genetic algorithm procedures. Then the genetic processes like probability crossover, mutation, and fitness based on new equations are applied to the words. The initial weights are sorted in descending order before the genetic procedures are applied. Then the words that withstand all the processes are extracted, as keywords for the entire document. The algorithm is evaluated using various metrics like the % of keywords extracted, F-Measure and Entropy. The values have shown better performance.

In the next work, the extracted keywords are then clustered using a Must Link and Cannot Link algorithm. The relationship between words is found, using the Must Link and Cannot Link equations. Then, the related keywords are clustered. Finally, the clustered outputs are optimized using Gaussian parameters. The algorithm is evaluated using Micro precision, Micro Recall, Micro F-Measure and Micro Accuracy values. The tabular values have shown better results than the other methods.

A sentence based text clustering, based on the WordNet has been developed. In this work, the features are extracted in the form of sentences. The sentences are extracted using various features like sentence position, positive keywords, negative keywords, number of nouns, adverbs, and adjectives etc. Each sentence in the document is ranked based on these variables and are arranged in descending order. The topmost ranking sentences are extracted based on size of the document and then the extracted sentences are clustered. The sentences are clustered based on semantics, using the WordNet. The documents having more than 80% of semantic similarity are clustered into one group and the process is repeated for the entire document set. Finally, the algorithm is evaluated using Micro Precision,
Micro Recall and Micro F-Measure. The values have shown better performance.

7.2 SCOPE FOR FUTURE WORK

The clustering process based on association rule can be enhanced, so that the clustered results can be used for text classification. This may enhance the accuracy of the classification.

The Keyword extraction technique presented in this thesis can be further developed where dimensions of the state space can be considered. This may improve the accuracy of the keywords being generated.

The WordNet based text clustering can be enhanced by considering various other features like polysemy, hypernymy, etc.