CHAPTER 8

CONCLUSION AND FUTURE WORK

8.1. CONCLUSION

The proposed approach presents document clustering techniques using partition-based clustering and frequent item clustering. This research uses Ontology for effective document clustering. Feature selection using Term Frequency / Inverse Document Frequency algorithm techniques provides better results than the existing term clustering technique. The proposed “Enhanced Feature Weight Techniques with Ontology clustering”, provides significant clusters and reduces the problems like modularity, flexibility, and scalability. Moreover, clustering techniques such as k-means, EM, sGEM, Apriori, FP-Growth, and FP-Bonsai have been incorporated with Ontology generation framework for effective clustering in a large document corpus such as Reuters-21578, 20 NG, and TDT-2 document corpus.

The main contribution of this research is the implementation of Ontology-based document clustering for categorizing and clustering to achieve high performance by using domain ontologies. Ontology generation framework and its results include all the aspects of document clustering, namely feature extraction, Ontology formation, and efficient cluster. The hybrid schemes of “Ontology and partition-based clustering” techniques and “Ontology and frequent item clustering” technologies combine the comfort of efficient
document clustering and this process, high performance and better document clusters are achieved. Evaluation results of this research proved that the proposed approach can easily outperform the traditional (TF/IDF) approach.

Finally, this research work provides a qualitative improvement over document clustering by using Ontology. The document clustering is done through k-means and it is combined with the Ontology model for clustering documents. When compared to traditional TF/IDF the proposed method overcomes the existing problem in k-means of clustering. With techniques such as EM, sGEM, Apriori, FP-Growth and FP-Bonsai, the proposed Ontology framework provides improved performance and better clustering.

The Experimental result evaluates the performance of the proposed approach. The experimental evaluation has been carried out through the dataset Reuters-21578 – a set of approximately 21,578 documents, 20 NG – a set of approximately 20,000 newsgroup documents and TDT-2 document corpus - a set of approximately 61,000 documents.

The proposed system clusters the documents, and thus improves the precision, recall, f-measure and purity substantially. Finally, the comparison result describes that the existing k-means, EM, sGEM, Apriori, FP-Growth, and FP-Bonsai produce lower precision, recall, f-measure, and purity. After introducing Ontology to each technique, the document clusters are formed efficiently. The results obtained are drawn and discussed with respect to six parameters. From the experimental results, it is clear that the OFPBC (Ontology based FP-Bonsai Clustering Algorithm) shows significant improvement in terms of purity of clustering and f-measure. Hence, OFPBC can be considered as an efficient clustering algorithm for clustering text documents.
8.2 FUTURE ENHANCEMENT

Several interesting research lines can be studied to enhance the existing document cluster. Document clustering techniques can include future lines within the areas of personalization, contextualization, and recommendation. The future research directions include applying Ontology, indexing in clustering and classification techniques. The research work can be extended to the social networking web site to perform document analysis. A Clustering approach using Ontology has been applied in a wide variety of fields such as pattern recognition, image segmentation, web mining, privacy, remote sensing, and marketing business especially in personalized multimedia content and in the filtering process.