CHAPTER 7
CONCLUSION AND FUTURE WORK

7.1 INTRODUCTION

The proposed approaches presented a novel document clustering techniques using fuzzy ontology. This research uses Ontology for effective document clustering. Feature selection using semantic clustering techniques provide better results than the term clustering technique. SEHP2PC architecture clustering provides significant clustering to the problems like modularity, flexibility and scalability. Moreover, optimization techniques such as GA, NDRGA and ACO have been incorporated with ontology generation framework for effective clustering of vague documents in a large database.

The datasets used for the experimental observations are

- 20 newsgroups data set
- Reuters information retrieval data set
- Real time data set

7.2 PERFORMANCE EVALUATION

The performances of the proposed approaches were evaluated based on the following measures:
(i) **Classification accuracy**: The clustering approaches are evaluated using their classification accuracy of the document. From the experimental results, it is observed that FOACO approach shows very high clustering accuracy in all the data sets. From the experimental results, it is observed that the relative classification accuracy improvement of the proposed FOACO over K-means and HP2PC is found to be very significant when compared to the relative measure improvement of the other approaches. The highest relative percentage accuracy of the proposed FOACO over the K-means is 14.35% in circuits and systems and communications journal and HP2PC is 10.29% in orgs categories.

(ii) **Objective function**: The objective function is considered for the performance evaluation of document clustering. Clustering will be better when the value of objective function $E$ is smaller. The experimental results make it clear that the objective function of the FOACO is less when compared with other clustering approaches. FOACO has the least objective function of about 8.23 in misc.forsale category when compared with the other approaches.

(iii) **Classification time**: The time taken by the clustering approaches to classify the documents is used to evaluate the performance of the approaches. The relative improvement measures of the approaches are compared with standard K-means and HP2PC approaches. The experimental results, testify that the time taken by FOACO is less when compared with the other clustering approaches like FONDRGA,
GFOM, SEHP2PC, HP2PC and K-means. The relative classification time of FOACO over K-means and HP2PC is found to be significant with 56.98% in rec.motorcycles category and 54.44% in place categories respectively.

(iv) Algorithm complexity and speed up: Algorithm complexity is seen that does not increase linearly. It is seen that with 65 nodes the speed up increased. FOACO technique has the high performance when compared with the other techniques.

(v) **Convergence behavior**: The convergence behavior of the clustering algorithms and the standard algorithm (K-means) with the number of iterations is considered for performance evaluation. It is observed from the experimental results that the convergence behavior of the NDRGA technique is very significant. FOACO and FONDRGA approaches converge in just 20 iterations where as the standard K-means takes 80 iterations and GFOM and SEHP2PC takes 50 and 70 iterations respectively.

(vi) **Silhouette coefficient**: The silhouette coefficient is used for measuring the quality of clustering that is not dependent on the number of clusters. The experimental results show that the relative silhouette coefficient measure of FOACO approach over K-Means and HP2PC is found to be very significant in all the three datasets considered for experimentation.

(vii) **F-measure**: The experimental results show that the FOACO approach has very high F-measure in all the datasets considered. The highest relative F-measure obtained for
FOACO over the K-means and HP2PC is 57.14% and 53.84% in real time data set respectively.

(viii) **Entropy:** The experimental results establish that FOACO approach has the least entropy value when compared with other techniques. The highest relative entropy measure of the proposed FOACO approach obtained over K-means and HP2PC is 65.52% and 58.33% in 20 newsgroups data set respectively.

(ix) **Separation index:** The experimental results ensure that the FOACO technique has the least separation index performance when compared with the other techniques.

From the experimental results, it is found that the proposed FOACO is very effective when compared to other approaches.

In the thesis, classification accuracy has been analyzed for different methods discussed and is found that the accuracy has invariably increased. SCFS, method is compared to TCFS, and it was found that accuracy is marginally increased as only 3.3% in compu.graphics category on 20 newsgroups. But the classification time has been susceptibility reduced. People category in reuters data set classification time of SCFS has 45.25% improvement. SEHP2PC technique is compared to K-means and HP2PC, in which accuracy is marginally increased as 5.15% and 3.29 % respectively in reuters data set. But the classification time of SEHP2PC is susceptibility reduced to 16. 49% and 12.09% compared to K-means and HP2PC in exchanges categories. The proposed methods FONDRGA, and FOACO also take less number of iterations to converge, due to usage of fuzzy ontology, GA and ACO. Time complexity is not crucial matter in this place. Converge
behavior of proposed techniques have lesser error values compare to existing techniques.

7.4 FUTURE WORK

The clustering approaches provide effective document clustering using the fuzzy ontology technique. FOACO provides very good performance, when compared with the other proposed clustering approaches. This research utilized a document clustering algorithm to search the best cluster center in the global situation.

The approach can be applied for further enhancement for the following:

- The ontology based approaches along with statistical methods can be adopted for finding the optical cluster number in other languages like Chinese.

- Better optimization and swarm intelligence techniques (Tang Xian-lun et al, (2011)) can also be used for enhancing the performance of the document clustering approach.

- Distributed document clustering can be applied to reduce map type (Chandra and Ajitha, 2011).

- Fuzzy can be applied to clustering of more web and XML documents using distributed data mining.

- The Portable Document Format (PDF) and Rich Text Format (RTF) documents can be clustered as the future version of fuzzy document clustering system.