ABSTRACT

The availability of large quantity of text documents from the world wide web and business document management systems has made the dynamic separation of texts into new categories, as a very important task for every business intelligence system. Text document clustering is one of the emerging and most needed clustering techniques used to cluster documents with regard to similarity among documents. It is used widely in digital library management system in the modern context. Likewise, document clustering is widely applicable in areas such as search engines, web mining, information retrieval, and topological analysis.

The increase in the number of documents worldwide leads to the difficulty of classifying those documents according to the needs. Clustering analysis is one of the main research areas of artificial intelligence and data mining. Its fundamental task is to utilize characters to compute the degree of related relationship between objects and to accomplish automatic classification without any prior knowledge. Document clustering utilizes clustering technique to gather the documents of high resemblance collectively by computing the documents resemblance.

There are several clustering approaches available in the literature to cluster the document in a distributed environment. But most of the existing clustering techniques suffer from a wide range of limitations. The existing clustering approaches face the issues like practical applicability, lower accuracy, less scalability and more classification time etc. Thus a novel
approach is very much needed for providing significant accuracy with less classification time.

Recent studies have shown that ontologies are very efficient in improving the performance of document clustering. So, this research mainly focuses on the application of ontology in document clustering. Ontology algorithm plays a vital role in knowledge and semantic web management. Fuzzy technique is also integrated with the ontology for the significant performance of the clustering approaches. Fuzzy ontology is especially significant for the databases with very large and vague documents. The research in focus mainly deals with five proposed approaches for document clustering. They are:

1. Distributed clustering with feature selection for text documents based on ontology
2. Ontology based hierarchical distributed document clustering
3. A Genetic Fuzzy Ontology Model (GFOM) for distributed document clustering
4. A novel clustering approach using Fuzzy Ontology with Non-Dominated Ranked Genetic Algorithm (FONDRGA)

The underlying objective of this research work is to develop an effective and efficient clustering technique that can achieve good accuracy and performance.
In the first proposed research work, the distributed clustering with feature selection for text documents for better cluster performance in Peer to Peer (P2P) systems is illustrated by a semantic ontology mechanism. The performance of the proposed semantic clustering is compared with the term clustering approach.

The Semantic Enhanced Hierarchically Distributed P2P (SEHP2PC) clustering algorithm dealing with the problems like modularity, flexibility and scalability is proposed in the second work by replacing weight with semantic weight for each document. The relative measure of the proposed SEHP2PC approach over the standard K-means and existing Hierarchically Distributed P2P (HP2PC) shows a significant improvement in the performance of the proposed SEHP2PC approach.

In the third proposed work, GA is used to improve the overall performance of document clustering. Initially, Ontology Generation using Fuzzy Logic (OGFL) was implemented to the database, containing a large amount of documents. OGFL will generate the ontology for the given database and GA is applied with this ontology generation for performance improvement.

The NDRGA which has several more advantages than the traditional GA is applied in the fourth contribution for better convergence and less computational complexity.

ACO algorithm is used as the next proposed work for the clustering of documents in a peer to peer environment for further improvement than the earlier contributions. OGFL generates the ontology for the given database. ACO is used for clustering the documents in the database with the help of
ontology generated by OGFL technique. The combination of OGFL and ACO helps to increase the accuracy of clustering.

All the contributions are evaluated using the three datasets namely 20 newsgroups data set, reuters information retrieval data set and real time data set. The measures used for the experimental evaluations are classification accuracy, objective function, classification time. The efficiency of the proposed approaches is evaluated by the algorithm complexity and speed up measures. Moreover, the quality of the resulting document clustering is evaluated using convergence behavior and silhouette coefficient, F-measure, entropy and separation index. All the techniques have been implemented using MATLAB 7.10. In addition General Architecture for Text Engineering (GATE) is also used for text extraction and protégé opens source software is used for ontology generation.

From the experimental results, it is observed that the final work, namely FOACO technique produces very good accuracy of about 99.70% in reuters and real time data sets; the least classification time of about 0.39 seconds in 20 newsgroups data set; a better convergence in only about 20 iterations and a better efficiency when compared with SEHP2PC, GFOM, FONDRGA, HP2PC and K-means. Moreover, FOACO has the least objective function of about 8.23 in 20 newsgroup data sets. FOACO has the highest silhouette coefficient and F-measure of 0.91 for the 20 newsgroups and real time data sets. The least entropy of the proposed FOACO is 0.1 for the 20 newsgroups data set.