Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/070442

Fuzzy Ontology for Distributed Document Clustering based on Genetic Algorithm

Thangamani.M[1,*] *and P. Thangaraj*[2]

[1]Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Erode-638 052, Tamilnadu, India
[2]Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India

Received: 29 Nov. 2012, Revised: 14 Jan. 2013, Accepted: 22 Feb. 2013
Published online: 1 Jul. 2013

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 systems. But, present text clustering algorithms still suffer from problems of practical applicability. Recent studies have shown that, in order to improve the performance of document clustering, ontologies are useful. Ontology is nothing but the conceptualization of a domain into an individual identifiable format, but machine-readable format containing entities, attributes, relationships and axioms. By analyzing all types of techniques for document clustering, a clustering technique depending on Genetic Algorithm (GA) is determined to be better as GA is a global convergence technique and has the ability of determining the most suitable cluster centers without difficulties. In this paper, a new document clustering scheme with fuzzy ontology based genetic clustering is proposed. The experimental results reveal that the proposed approach increases the accuracy to a large extent and the clustering time is also highly reduced.

Keywords: Ontology, Genetic Algorithm, Document Clustering, Conceptual Clustering.

1 Introduction

Clustering of document is very important for the purpose of document organization, summarization, topic extraction and information retrieval in an efficient way. Initially, clustering is applied for enhancing the precision or recall in the information retrieval techniques [19,20]. In recent times, clustering technique is applicable in the areas which involves browsing a gathered data or in categorizing the outcome provided by the search engine for the reply to the query provided by the user. Document clustering can also be applicable in producing the hierarchical grouping of document. In order to search and retrieve the information efficiently in Document Management Systems (DMSs), the metadata set should be created for the documents with enough details. But, only one metadata set is not enough for the whole DMS. This is because the various document types need various attributes for distinguishing appropriately. So a novel approach is necessary for distinguishing the documents properly.

Presently, the effectiveness of ontology in document clustering scheme [21] has been realized as this technique is supposed to improve the accuracy of clustering. Whereas the role-played by ontologies [16], provides a position of legitimacy, research in this field leads to superior importance in the arising of different disputes featured in the modern digital situation. With the intention of providing solution to the various drawbacks related with present search techniques, ontologies are widely implemented for creating the capable document clustering techniques [22,23]. The example of research concern in the field of ontology is the Google which is a search engine for semantic web documents, terms and data found on the internet [24,26,28]. For retrieving information from the complex characteristic of the digital information available in digital libraries, ontologies can be successfully used for producing efficient information retrieval system.

In this paper, ontologies are introduced as a modeling technology for structured metadata definition within document clustering system. With the obtained metadata the clustering is processed using the Genetic Algorithms (GAs). GA is search technique that is based on natural genetic and selection merging the idea of survival of the fittest with a structured interchanges. These techniques

[*] Corresponding author e-mail: manithangamani2@gmail.com

involve the conservation of the attributes of the finest exponents of a generation in the next generation; additionally introducing the variations in the new generation composition with the help of crossing over and mutation function. This GA method attempts to resolve the technique for distributing N object in M clusters based on the minimization of several optimization measure additives over the clusters. Once the optimization measure is selected, the clustering difficulty is to offer an efficient technique for searching the space of the every potential categorization and to determine one on which the optimization function is minimized. The difficulty is to categorize a group of data [33]. These data form clusters of points in n-dimensional space. These clusters form groups of similar samples. Usually the methods will use an optimization measures such as decreasing the distance additions of every sample to the clusters centre, which can be taken as the gravitatory of a cluster. This indicates a unique point X which better characterize every point from this cluster. This optimization measures was used in this work and the minimization process is performed by GA. GA [14] is a famous technique to deal with complex search problems by implementing an evolutionary stochastic search because GA can be very effectively applied to various challenging optimization problems. The NP-hard nature of the clustering technique makes GA a natural choice for solving it [25,29,30,31]. A common objective function in these implementations is to decrease the square error. This paper clearly presents the ontology based document clustering methodology with GA [17]. The remaining of this paper is organized as follows. Section 2 discusses the related works on ontology generation and GA based clustering. Section 3 describes the motivation of the work. Section 4 describes complete methodology of the proposed clustering method. Section 5 discusses experiment results of the proposed clustering technique. Section 6 concludes the paper with some discussion.

2 Related Works

Lena Tenenboim *et al*, [1] proposed a novel technique on ontology based classification. They have discussed on classification of news items in ePaper, a prototype system of a future personalized newspaper service on a mobile reading device. The ePaper system comprises news items from different news suppliers and distributes to each subscribed user a personalized electronic newspaper,

making use of content-based and collaborative filtering techniques. The ePaper can also offer users standard version of chosen newspapers, besides the browsing abilities in the warehouse of news items. This deliberates on the automatic categorization of incoming news with the help of hierarchical news ontology. Based on this clustering technique on one hand, and on the users' profiles on the other hand, the personalization engine of the system is able to afford a personalized paper to every user onto the mobile reading device.

By considering the difficulty that classical Euclidean distance metric cannot create an suitable separation for data lying in a manifold, a GA based clustering method with the help of geodesic distance measure is proposed by Gang Li *et al*, [2]. In the proposed method, a prototype-based genetic illustration is used, where every chromosome is a sequence of positive integer numbers that indicate the k-medoids. In addition, a geodesic distance based proximity measures is applied to find out the similarity between data points. Simulation results on eight standard synthetic datasets with dissimilar manifold structure illustrate the effectiveness of the algorithm as a clustering technique. Evaluating with generic k-means method for the function of clustering, the proposed technique has the potential to distinguish complicated non-convex clusters and its clustering performance is obviously better than that of the K-means method for complex manifold structures.

Casillas *et al*, [3] put forth a novel concept on document clustering using GA. They have present a GA that deals with document clustering that computes an approximation of the optimum 'K' value, and resolves the best clustering of the documents into 'K' clusters. They have experimented the proposed technique with sets of documents that are the output of a query in a search engine. The simulation results show that the proposed GA attain better values of the fitness function than the well known Calinski and Harabasz stopping rule and performs in only lesser time.

Andreas *et al*, [4] discussed on the clustering technique for text data. Text clustering usually involves clustering in a high dimensional space that appears complex with considered to virtually all practical settings. Additionally, provided a scrupulous clustering outcome it is normally very tough to come up with a good clarification of why the text clusters have been created the way they are. In this paper, a novel technique is presented for applying background information during preprocessing for improving the clustering outcome and permit for selection between outcomes. They have preprocesses the input data supplied to ontology-based heuristics for feature selection and feature aggregation. Therefore, various choices for text illustrations are constructed. Based on these illustrations, they have calculate the multiple clustering outcomes using K-means. The achieved results by compared favorably with a sophisticated baseline preprocessing strategy.

A Wordsets based document clustering algorithm for large datasets was proposed by Sharma *et al*., [5]. Document clustering is a significant tool for applications like search engines and document browsers. It facilitates the user to comprise a better overall observation of the data contained in the documents. The available techniques of document clustering, however, do not actually consider the special difficulties of text document clustering: very high dimensionality of the document, very large size of the datasets and understandability of the cluster explanation. Also there is a strong requirement for hierarchical document clustering [15] where clustered documents can be browsed based on the increasing specificity of topics. Frequent Itemset Hierarchical Clustering (FIHC) is a novel data mining technique for hierarchical grouping of text documents. The technique does not provide consistent clustering results when the number of frequent sets of terms is large. In this paper they have proposed WDC (Wordsets-based Clustering), an efficient clustering technique based closed words sets. WDC makes use of hierarchical technique to cluster text documents having common words.

Cao *et al.*, [6] provided fuzzy named entity-based document clustering. Conventional keyword-based document clustering methods have restrictions because of simple treatment of words and rigid partition of clusters. In this paper, they have introduces named entities as objectives into fuzzy document clustering, which are the important elements defining document semantics and in many cases are of user concerns. Initially, the conventional keyword-based vector space representation is adapted with vectors defined over spaces of entity names, types, name-type pairs, and identifiers, alternative of keywords. Next, hierarchical fuzzy document clustering can be applied using a similarity measure of the vectors indicating documents.

Zhang *et al.*, [7] gives clustering aggregation based on GA for documents clustering. In this paper, a technique based on GA for clustering aggregation difficulty, named as GeneticCA, is provided to approximate the clustering performance of a clustering division, clustering precision is defined and features of clustering precision are considered. In the evaluation concerning clustering performances of GeneticCA for document clustering, hamming neural network is applied to make clustering partitions with fluctuant and weak clustering performances.

Web document clustering using document index graph is put forth by Momin *et al.*, [8]. Document clustering methods generally based on single term examination of document data set. To attain more precise document clustering, more informative feature like phrases are essential in this scenario. Therefore first part of the paper provides phrase-based model, Document Index Graph (DIG) that permits incremental phrase-based encoding of documents and capable phrase matching. It stress on efficiency of phrase-based similarity measure over conventional single term based similarities. In the second part, a Document Index Graph based Clustering (DIGBC) algorithm is provided to improve the DIG model for incremental and soft clustering. This technique incrementally clusters documents based on presented cluster-document similarity measure. It permits assignment of a document to more than single cluster.

Muflikhah *et al*. [9] proposed a document clustering technique using concept space and cosine similarity measurement. This paper aims to incorporate the information retrieval technique and document clustering technique as concept space approach. The technique is known as Latent Semantic Index (LSI) approach which used Singular Vector Decomposition (SVD) or Principle Component Analysis (PCA). The intention of this technique is to decrease the matrix dimension by identifying the pattern in document collection with refers to simultaneous of the terms. Every technique is employed to weight of term-document in vector space model (VSM) for document clustering with the help of fuzzy c-means technique. In addition to the reduction of term-document matrix, this research also utilizes the cosine similarity measurement as alternative of Euclidean distance to engage in fuzzy c-means.

Affinity-based similarity measure for Web document clustering is presented by Shyu *et al.*, [10]. In this paper, the concept of document clustering is extended into Web document clustering by establishing the approach of affinity based similarity measure, which makes use of the user access patterns in finding the similarities among Web documents through a probabilistic model. Various experiments are conducted for evaluation with the help of real data set and the experimental results illustrated that the presented similarity measure outperforms the cosine coefficient and the Euclidean distance technique under various document clustering techniques.

ELdesoky *et al.*, [11] given a novel similarity measure for document clustering based on topic phrases. In the conventional vector space model (VSM) researchers have used the unique word that is contained in the document set as the candidate feature. Currently a latest trend which uses the phrase to be a more informative feature has considered; the issue that contributes in enhancing the document clustering accuracy and effectiveness. This paper presented a new technique for evaluating the similarity measure of the traditional VSM by considering the topic phrases of the document as the comprising terms for the VSM instead of the conventional term and applying the new technique to the Buckshot technique, which is a combination of the Hierarchical Agglomerative Clustering (HAC) technique and the K-means clustering method. Such a method may increase the effectiveness of the clustering by incrementing the evaluation metrics values.

Nanas *et al*. [34] Introduce a document evaluation function that allows the use of the concept hierarchy as a user profile for Information Filtering. Zitao *et al.*, [36] proposed a feature selection method for document clustering based on part-of-speech and word co-occurrence. Wang *et al.*, [37] presents a document clustering algorithm based on Nonnegative Matrix Factorization (NMF) and Support Vector Data Description (SVDD). Fuzzy clustering of text documents using Naive Bayesian Concept is provided by Roy *et al.*, [38]. Web document clustering based on Global-Best Harmony Search, K-means, Frequent Term Sets and Bayesian Information Criterion is suggested by Cobos *et al.*, [39].

A document clustering method based on hierarchical algorithm with model clustering is presented by Haojun *et al.*, [12]. This paper involved in analyzing and making use of cluster overlapping technique to design cluster merging criterion. In this paper, they have presented a new method to calculate the overlap rate for improving time efficiency and the veracity. The technique is uses a line to pass across the two cluster's center as an alternative of the ridge curve. Depends on the hierarchical clustering technique, the expectation-maximization (EM) method is used in the Gaussian mixture model to count the parameters and formulate the two sub-clusters combined when their overlying is the biggest.

Document clustering with the help of fuzzy c-mean algorithm is proposed by Thaung *et al.*, [13]. Most traditional clustering technique allocate each data to exactly single cluster, therefore creating a crisp separation of the provided data, but fuzzy clustering permits for degrees of membership, to which a data fit various clusters. In this paper, documents are partitioned with the help of fuzzy c-means (FCM) clustering technique. Fuzzy c-means clustering is one of famous unsupervised clustering methods. But fuzzy c-means method needs the user to mention the number of clusters and different values of clusters corresponds to different fuzzy partitions earlier itself. So the validation of clustering result is required. PBM index and F-measure are helpful in validating the cluster.

3 Motivation

The major drawbacks of the existing text clustering techniques:

- Text clustering generally involves clustering in a high dimensional space, which is very difficult with regard to virtually all practical settings.
- Text clustering is often treated as an objective method, which offers one clearly defined result, which needs to be ”optimal” in some way.
- Text clustering is often ineffective, unless it is integrated with a clarification of why particular texts were categorized into a particular cluster.

It can be clearly observed that the existing techniques discussed earlier do not produce better accuracy. The time taken for the active clustering algorithms is more when the large databases are considered for clustering. Also in case of determining the initial clusters, the different clusters will be resulted for same dataset.

The early research shows that the usage of GA will provide better classification accuracy when compared to the other method. The usage of GA will increase the accuracy of classification for large database.

The major advantages of the GA are

- GA belongs to search techniques which could automatically exploit the optimal solution for objective or fitness function of an optimization problem.
- GA is the best-known evolutionary techniques.
- GA provides global optimal solution.
- GAs are powerful approaches to solving optimization problems.

But the convergence time for the usage for GA is more and also the number iterations required for GA is more when compared to other techniques. The usage of fuzzy ontology [27] will provide better classification of large vague database. Thus, the fuzzy ontology can be initially applied to the database to reduce the convergence time and number of iterations before using GA. This motivated the usage of fuzzy ontology and GA for clustering. Hence, in this paper fuzzy ontology is combined with the GA to yield the better classification accuracy for large databases.

4 Methodology

This section discusses about the methodologies implemented in the proposed technique. Initially, Ontology Generation using Fuzzy Logic (OGFL) framework is implemented to the database containing large amount of documents. The distributed architecture is used for this approach [41,42]. OGFL technique will generate the ontology for the given database. With this ontology, the next step is application of GA. This GA is used for clustering the documents in the database with the help of ontology generated by OGFL technique. The combination of OGFL and GA helps increasing the accuracy of clustering.

4.1 Ontology Generation Using Fuzzy Logic (OGFL)

The OGFL framework consists of the following components which are shown in figure 1:

- Fuzzy Concept Analysis using Fuzzy Set (FCAFS) [18]
- Conceptual clustering using fuzzy logic
- Fuzzy ontology creation.

Fuzzy Formal Concept Analysis: This process considers the database which contains ambiguity data and produces fuzzy formal concepts from it. Additionally, fuzzy formal concepts are created from the fuzzy formal context and categorize the generated concepts as a fuzzy concept lattice. Fuzzy conceptual clustering algorithm is shown in Figure 2.

Fuzzy Conceptual Clustering: This process generates conceptual clusters by clustering the fuzzy concept lattice. With the help of fuzzy logic and depending on fuzzy information integrated into the lattice, the clustering process is carried on. The algorithm for Fuzzy

Conceptual Clustering is represented in figure 2. This algorithm generates conceptual clusters from a concept C_S which is called the starting concept on a fuzzy concept lattice $F(K)$. The concepts are separated into different cluster based on the similarity threshold T_S. To generate all conceptual clusters of $F(K)$, C_S is selected as the supremum of $F(K)$, or $C_S = sup(F(K))$

Hierarchical Relation Generation: This process produces the concept hierarchy by generating the hierarchical relations among conceptual clusters. The detailed explanation for these components can be seen in [35].

4.2 Fuzzy ontology creation

In this step fuzzy ontology is generated from the fuzzy context with the help of concept hierarchy produced by fuzzy conceptual clustering. As both FCAFS [32] and ontology support formal definitions of concepts, this approach is carried out. Figure 3 shows the fuzzy ontology generation process.

Fuzzy ontology is created in this step from the fuzzy context with the help of concept hierarchy produced by fuzzy conceptual clustering. This is performed based on the fact that both FCA and ontology maintain the formal definitions of concepts.

Conversely, a concept defined in FCAFS constitutes both extensional and intensional data, whereas a concept in ontology just emphasized on its intensional characteristics. For generating the fuzzy ontology, it is required to convert both intensional and extensional data of FCAFS concepts into the equivalent classes and relations of the ontology.

Class Mapping: In this process, the extent and intent of the fuzzy context are mapped into the extent and intent classes of the ontology. Human participation is required to name the label for the extent class. Keyword attributes can be represented by appropriate names and it is used to label the intent class names also.

Taxonomy Relation Generation: With the help concept hierarchy, this phase produces the intent class of the ontology as a hierarchy of classes. The step can be regarded as an isomorphic mapping from the concept hierarchy into taxonomy classes of the ontology. Consider an illustration that the class Research Area can be developed into a hierarchy of classes, in which every class represents a research area, associated with the concept hierarchy.

Non-taxonomy Relation Generation: In this step, the similarity among the extent class and intent classes are generated. This is a very simple task to perform. Still the labeling of non-taxonomy relation is necessary to perform.

Instances Generation: In this process, instances for the extent class are generated. Each instance indicates to an object in the initial fuzzy context. Depends on the data existing on the fuzzy concept hierarchy, instances

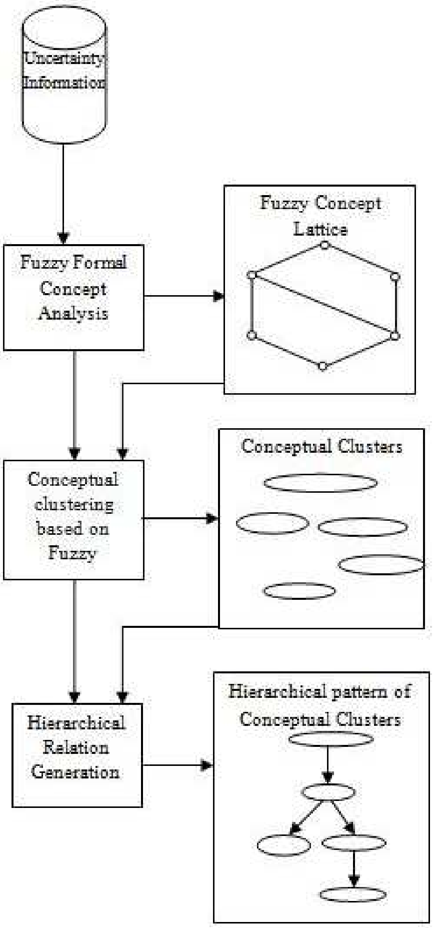
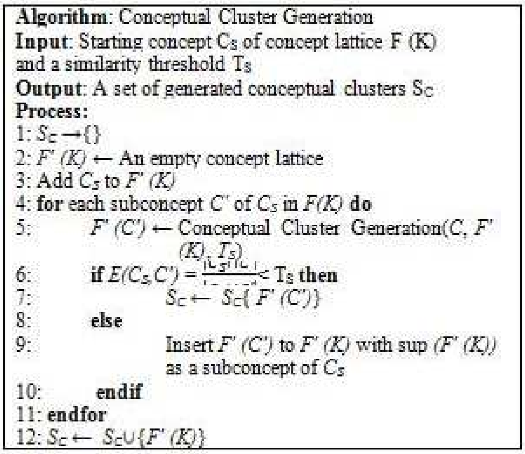

Uncertainty Information
Fuzzy Formal Concept Analysis
Fuzzy Concept Lattice
Conceptual clustering based on Fuzzy
Conceptual Clusters
Hierarchical Relation Generation
Hierarchical pattern of Conceptual Clusters

Fig. 1: The approach for automatic generation of concept hierarchy



```
Algorithm: Conceptual Cluster Generation
Input: Starting concept $C_S$ of concept lattice F (K)
and a similarity threshold $T_S$
Output: A set of generated conceptual clusters $S_C$
Process:
$S_C \rightarrow \{\}$
$F'(K) \leftarrow$ An empty concept lattice
Add $C_S$ to $F'(K)$
for each subconcept $C'$ of $C_S$ in $F(K)$ do
        $F'(C') \leftarrow$ Conceptual Cluster Generation($C$, $F'(K)$, $T_S$)
        if $E(C_S, C') =$ [illegible] $< T_S$ then
                $S_C \leftarrow S_C\{F'(C')\}$
        else
                Insert $F'(C')$ to $F'(K)$ with sup $(F'(K))$
                as a subconcept of $C_S$
        endif
endfor
$S_C \leftarrow S_C \cup \{F'(K)\}$
```

Fig. 2: The fuzzy conceptual clustering algorithm.

attributes are automatically furnished with suitable values. For examples, each instance of the class Document that is related to an actual document will be associated with the appropriate research areas.

After the ontology is generated, GA is used to cluster the documents. The usage of ontology helps in determining the best classification for clustering using GA.

4.3 Design of clustering algorithm using Genetic Algorithm

When considering that the number of category is k, this paper uses GA [40] to find the better cluster center. The steps involved in this algorithm are given below:

Step 1: Encoding: Adopting floating-point code. Individual data is indicated by the matrix $A = (a_1,\ a_2,\ldots,a_k)^T \subset R^{k\times m}$ that contains k cluster centers. Each component a_i presents a cluster center and with the help of floating point number, every element of a_i is encoded.

Step 2: Group initialization: Assuming the amount of initialized group is M, matrix collection $\mathrm{X} = (\mathrm{A}_1,\ \mathrm{A}_2\ldots\mathrm{A}_\mathrm{n})$ indicates the group collection. Elements of each matrix are collection of $\mathrm{k}\times\mathrm{m}$ random real numbers in the range of 0 to 1.

Step 3: Design of fitness function: At the start all component of each individual is considered as the cluster center, and then the relation among all documents and cluster centers are calculated. Then, depends on the minimum distance principle, the documents are grouped into most similar categories. Thus, all clusters are created. At last, the sum of mean square deviation of all intra class distance is determined. Depends on the design of objective function, the individual fitness function is defined as:

$$f = \frac{1}{1+E} \tag{1}$$

Where E is the Clustering objective function:

$$E = \sum_{j=1}^{k}\sum_{x_i \varepsilon c_j}(x_i - x_j^*)^2/n_j \tag{2}$$

Where x_j^* denotes the center of cluster c_j , n_j is the amount of documents in cluster c_j. The individual fitness is effective if the value of E is small. Moreover, for the effective clustering, the value of E should be very small.

Step 4: Selection: This step is to pick up several fine individuals from the present group and find out which individual can enter the next generation. The grouping of choiceness and sorting technique is implemented here. Initially, the individuals are size down in terms of fitness function and the first h individuals enter the next generation directly. Next the fitness of the remaining individuals in sequential order is calculated by the following equation:

$$P(C) = [b + (a-b)\frac{(M - Rank(C))}{M-h-1}]/(M-h) \tag{3}$$

Where M is the group size, $Rank(C)$ is the serial number following the sorting of individual, and $\mathrm{Rank}(\mathrm{C}) \in \{\mathrm{h}+1, \mathrm{h}+2,\ldots,\mathrm{M}\}$, $a+b=2$ and $\mathrm{a} \in \{1,5,2\}$. By stochastic universal sampling, $M-h$ individuals are chosen, cross and mutate them, and then create $M-h$ new individuals. Therefore, it is easier to maintain the best individuals and alter the worst ones, thus enhancing individual's capacity of fitness and guaranteeing a certain selection pressure.

Step 5: Crossover: Randomly take two individuals, cross them and create a symmetrical and random number r between 0 and 1. If $r < p_c$, carry outs the crossover, and produces new individuals $A^{'}$ and $B^{'}$ by the following equation:

$$A^{'} = rA + (1-r)B \tag{4}$$

$$B^{'} = rA + (1-r)A \tag{5}$$

Step 6: Mutation: In this step, generate a random number r between 0 and 1, if $r < p_m$, carry out the mutation. The nonsymmetrical mutation algorithm is implemented. For an individual A, if a_i is selected to be mutated, the equivalent component of a_i is changed as follows:

$$a_{ij}^{'} = \begin{cases} a_{ij} + \Delta\left(t,\ a_{ij}^{max} - a_{ij}\right) & rand\,(0,1) = 0 \\ a_{ij} - \Delta\left(t,\ a_{ij} - a_{ij}^{min}\right) & rand\,(0,1) = 1 \end{cases} \tag{6}$$

$$\mathrm{j} = 1, 2\ldots\mathrm{m}$$

Where, $\Delta(t,y) = yr\left(1-\frac{t}{T}\right)^b$, a_{ij}^{max} and a_{ij}^{min} are the maximum and the minimum elements in the row vector. T is the maximum iteration times and t is the current one. Usually $b = 2$, and it determines the non-symmetrical system parameter. $\Delta(\mathrm{t},\ \mathrm{y}) \in [0,\ \mathrm{y}]$, so the probability that $\Delta(\mathrm{t},\ \mathrm{y})$ is equal to 0 roughly increases with the growth of t. Such a characteristic facilitates the algorithm to search the global situation equably at the beginning and become convergence in the local.

Step 7: Termination: If the average error of the fitness function of individuals between the new generation and the previous one is less than the given error parameter ε or if the iteration time has reached the maximum T, algorithm will terminate, or else go to step 4.

5 Experimental Results

The text documents collected from the IEEE web site are used for experimentation. Data mining domain related

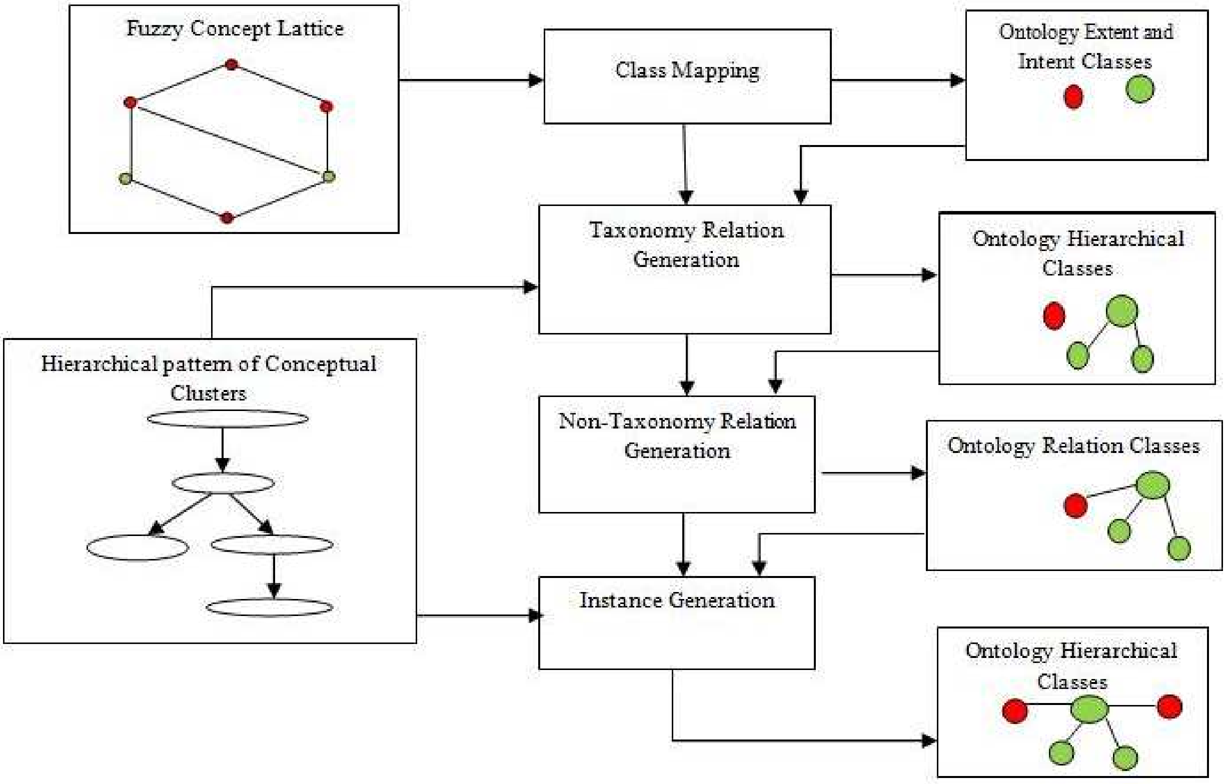

Fuzzy Concept Lattice
Class Mapping
Ontology Extent and Intent Classes
Taxonomy Relation Generation
Ontology Hierarchical Classes
Hierarchical pattern of Conceptual Clusters
Non-Taxonomy Relation Generation
Ontology Relation Classes
Instance Generation
Ontology Hierarchical Classes

Fig. 3: Fuzzy ontology generation process

journal collection is downloaded from the web. The journal abstract page is designed using HTML. The HTML pages are downloaded and transformed into text data. The text document conversion is performed by eliminating the HTML tag elements from the web documents. The text contents are maintained in separate text files. The list of journals from IEEE considered here are Biomedical Engineering, Circuits and Systems, Communications and Computer Graphics and Application. MATLAB is used for evaluating the proposed clustering technique. The platform used for this simulation is Windows XP. The processor used is Pentium IV. The experimentation needs a system RAM of 1 GB. The most efficient clustering approaches among the existing techniques are the fuzzy C-means and the clustering based on GA. The Fuzzy C-means clustering, clustering based on GA and the proposed fuzzy ontology clustering technique is applied for the considered data set.

The relevance factor is considered for different journals. When Biomedical Engineering journal is considered, many journals are with related topics and hence Fuzzy C-Means algorithm misclassifies the journals. When Circuits and Systems and Communication journals are considered, some of the journals are related to the communication purpose and thus existing techniques classify wrongly. So the proposed fuzzy ontology with GA method is implemented to it with better classification accuracy. This is clearly indicated in the Table 1 and figure 4. The same case is for Computer Graphics and Application journal and is showed clearly in Table 1 and 4.

Table 1 shows comparison of the accuracy of clustering classification for the proposed fuzzy ontology with GA method with the existing methods. From the Table 1, it can be observed that or IEEE Biomedical Engineering journal, the accuracy of clustering classification using Fuzzy C-Means algorithm is 91.6 %, for GA is 95.2 % and finally for the proposed fuzzy ontology with GA method, the classification accuracy is higher i.e. 97.2 %. When the Circuits and Systems journal is considered, better accuracy i.e. 98.3 % is achieved by the proposed fuzzy ontology with GA technique, whereas, the accuracy using Fuzzy C-Means algorithm is 89.3 % and the accuracy using GA is 96.7 %. When other journal such as Communications and Computer Graphics and Application is considered, better accuracy is achieved using the proposed fuzzy ontology with GA technique i.e. 98.3 % and 98.8 % respectively.

Figure 4 indicates the accuracy comparison of the proposed fuzzy ontology with GA technique with Fuzzy

C-Means and the GA. For different IEEE journals such as Biomedical Engineering, Circuits and Systems, Communications and Computer Graphics and Application are considered in the graphical comparison. It can be clearly observed from the graph that the proposed technique produces better accuracy than the existing techniques.

Then the clustering objective function is considered for experimentation. Clustering objective function is defined as:

$$E = \sum_{j=1}^{k} \sum_{x_i \in c_j} (x_i - x_j^*)^2 / n_j \tag{7}$$

Where x_j* indicates the center of cluster c_j, n_j indicates the amount of documents in cluster c_j. The clustering will be better when the value of objective function E is smaller.

The objective function value obtained for clustering the different IEEE journal using the proposed fuzzy ontology with GA clustering technique and existing clustering techniques is shown in Table 2. When considering the Biomedical Engineering journal, the objective function obtained by using the proposed fuzzy ontology with GA technique is 10.66 which is lesser than the objective function obtained by Fuzzy C-Means clustering and GA i.e. 10.9 and 10.79 respectively. This clearly indicates that the proposed fuzzy ontology with GA technique results in better clustering when compared to existing clustering techniques. When Circuits and Systems journal is considered, the objective function for existing methods are 11.15 and 11.01, whereas, for the proposed fuzzy ontology with GA clustering technique the objective function is 10.98 which are much lesser than conventional methods. The objective function obtained for the Communications and Computer Graphics and Application journal using the proposed fuzzy ontology with GA technique is 9.8 and 9.7 respectively that are lesser when compared to the usage of Fuzzy C-Means and GA techniques i.e. 10.24 and 10.11, 9.98 and 9.76 respectively for Communications and Computer Graphics and Application journal. From these data, it can be clearly seen that the proposed fuzzy ontology with GA technique will produce better clusters when compared to the existing techniques.

The performance of the proposed fuzzy ontology with GA and existing techniques in terms of comparison with their objective function is shown in figure 5. It can be clearly observed that the proposed fuzzy ontology with GA clustering technique results in lesser objective function for the considered IEEE journals when compared to the existing techniques. This clearly indicates that the proposed fuzzy ontology with GA clustering technique will produce better clusters for the large database when compared to the conventional techniques.

The convergence behavior of the proposed clustering algorithm and the existing algorithms (Fuzzy C-Means and GA) with the number of iterations are shown in figure 6. By using the proposed technique, the average value of objective function decreases from 13.65 to 10 suddenly only in 20 iterations. After sharp decrease, the average value of objective function becomes steady and converges only in 20 iterations. This is clearly shown in figure 5. But the number of iterations taken by the Fuzzy C-Means algorithm and GA for convergence is more when compared to the proposed technique. From figure 6, the average value of objective function decreases from 14.4 to 11.1 only in 80 iterations for using Fuzzy C-Means algorithm i.e. the number of iterations required for convergence is 80 iterations. Also, by the usage of GA, the average value of objective function decreases from 14.1 to 13.8 in 60 iterations i.e. the number of iterations required for convergence is 60 iterations. From these observations, it can be said that the proposed technique converge in lesser iterations when compared to the Fuzzy C-Means and GA. This shows that the time taken for clustering by the proposed technique is lesser when compared to the conventional techniques.

6 Conclusions and future enhancement

In this paper, Ontology based document clustering using GA is proposed. Initially, the fuzzy ontologies are generated for the available vague documents. Then with the help of those generated ontology, the clustering of documents are carried with the help of GA. This paper utilized a document clustering algorithm based on GA to search the best cluster center in the global situation. OGFL consists of the following process: Fuzzy Formal Concept Analysis, fuzzy conceptual clustering and fuzzy ontology creation. For experimentation on the proposed technique, documents are collected form the IEEE web site. The IEEE journals considered for experimentation are Biomedical Engineering, Circuits and Systems, Communications and Computer Graphics and Application. From the experimental results, it can be clearly observed that the proposed technique results in better accuracy when compared to the existing techniques such as Fuzzy C-Means and GA. The objective function for the proposed fuzzy ontology with GA method is lesser which indicates that the proposed technique can be able to perform better clustering. Also, the convergence time for the proposed technique is lesser when compared to the conventional techniques. Thus the proposed approach which uses fuzzy ontology with GA effectively classifies the documents in large databases. Moreover, even when the size of the database increases the performance of the proposed approach is effective thus scalability of the approach is efficient as it provides optimal solution. The problem that still occur in this clustering and also in the real world is how to determine exactly how many concepts actually presences in document collection. The problems included are a) for realistic instances hundreds of unique keywords are resulted, so each individual is a vector of several hundreds real numbers. And it is known

Table 1: Accuracy of Classification

Clustering Method	Biomedical Engineering	Circuits and Systems	Communications	Computer Graphics and Application
Fuzzy C-Means	91.6 %	89.3%	89.4%	90.00%
GA	95.2%	96.7%	93.6%	96.3%
Proposed	97.2%	98.3%	98.3%	98.8%

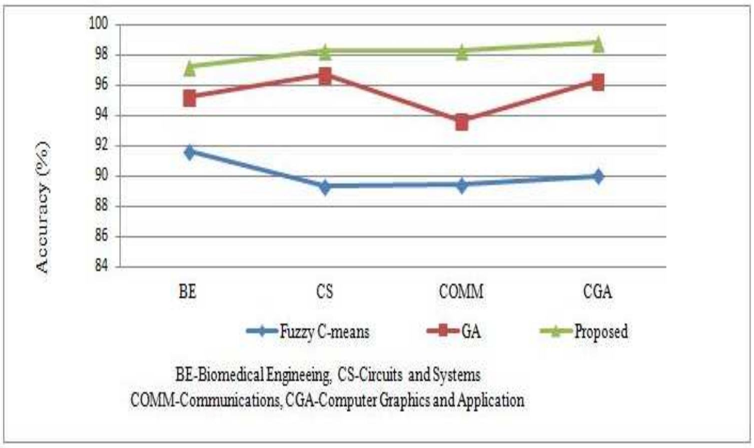


Fig. 4: Accuracy Comparison of the Proposed Technique and Existing Techniques

Table 2: Objective Function for Different Clustering Methods

Clustering Method	Biomedical Engineering	Circuits and Systems	Communications	Computer Graphics and Application
Fuzzy C-Means	10.9	11.5	10.24	9.98
GA	10.70	11.01	10.11	9.76
Proposed	10.66	10.98	9.8	9.7

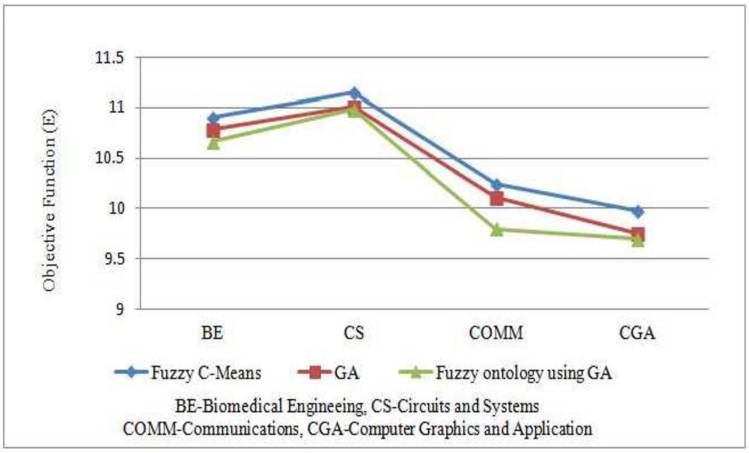


Fig. 5: Objective Function Comparison for the Proposed Technique and Existing Technique

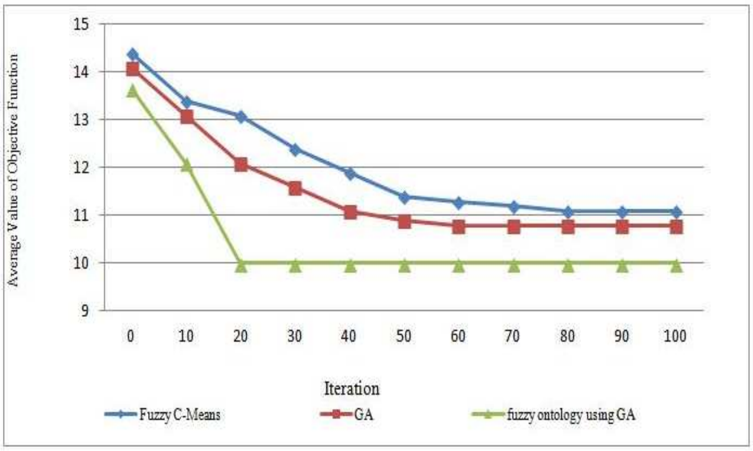


Fig. 6: The Convergence Behaviors of Proposed and Existing Techniques

that the size of an individual needed for GA to evolve satisfactory solutions grows exponentially with the length of the representation. So, it searches a way to reduce the dimension of clustering space so that the algorithm can be applied to large dataset. b) A constraint is given that the longest continuous common subsequence shorter than 4, maybe it is better fit for Chinese characters rather than for other languages. Future study to solve this problem by using statistical method applied to find the optimal cluster number might be the answer of this problem. Furthermore, comparative study with other reduction dimension method need to be done.

References

[1] Lena Tenenboim., Bracha Shapira. and Peretz Shoval., “Ontology-Based Classification of News in an Electronic Newspaper”, International Book Series Information Science and Computing, Pp: 89-98, 2008.

[2] Gang Li., Jian Zhuang., Hongning Hou. and Dehong Yu., “A Genetic Algorithm based Clustering using Geodesic Distance Measure”, IEEE International Conference on Intelligent Computing and Intelligent Systems, Pp: 274 - 278, 2009.

[3] Casillas, A., Gonzalez de Lena, M.T. and Martnez, R., “Document Clustering into an Unknown Number of Clusters Using a Genetic Algorithm”, Lecture Notes in Computer Science, Vol. **2807**, Pp. 43-49, 2003.

[4] Andreas Hotho., Alexander Maedche. and Steffen Staab., “Ontology-based Text Document Clustering”, Journal on Kunstliche Intelligenz, Vol. **4**, Pp. 48-54, 2002.

[5] Sharma, A. and Dhir, R., “A Wordsets based Document Clustering Algorithm for Large datasets”, Proceeding of International Conference on Methods and Models in Computer Science, 2009.

[6] Cao, T.H., Do, H.T., Hong, D.T. and Quan, T.T.; “Fuzzy Named Entity-Based Document Clustering”, IEEE International Conference on Fuzzy Systems, Pp. 2028 - 2034, 2008.

[7] Zhenya Zhang., Hongmei Cheng., Shuguang Zhang., Wanli Chen. and Qiansheng Fang., “Clustering Aggregation based on Genetic Algorithm for Documents Clustering”, IEEE Congress on Evolutionary Computation, Pp. 3156 - 3161, 2008.

[8] Momin, B.F., Kulkarni, P.J. and Chaudhari, A., “Web Document Clustering Using Document Index Graph”, International Conference on Advanced Computing and Communications, Pp. 32 - 37, 2006.

[9] Muflikhah, L. and Baharudin, B., “Document Clustering Using Concept Space and Cosine Similarity Measurement”, International Conference on Computer Technology and Development, Vol. **1**, Pp. 58-62, 2009.

[10] Shyu, M.L., Chen, S.C., Chen, M. and Rubin, S.H., “Affinity-based similarity measure for Web document clustering”, IEEE International Conference on Information Reuse and Integration, Pp. 247 - 252, 2004.

[11] ELdesoky, A.E., Saleh, M. and Sakr, N.A., “Novel Similarity Measure for Document Clustering based on Topic Phrases”, International Conference on Networking and Media Convergence, Pp. 92-96, 2009.

[12] Haojun Sun., Zhihui Liu. and Lingjun Kong., “A Document Clustering Method Based on Hierarchical Algorithm with Model Clustering”, 22nd International Conference on Advanced Information Networking and Applications, Pp. 1229 - 1233, 2008.

[13] Thaung Win. and Lin Mon., “Document clustering by fuzzy c-mean algorithm”, 2nd International Conference on Advanced Computer Control (ICACC), Pp.239 - 242, 2010.

[14] Murthy, C.A. and Chowdhury, N., “In Search of Optimal Clusters using Genetic Algorithms”, Pattern Recognition Letters, Pp. 825-832, 1996.

[15] Koller, D. and Sahami, M., “Hierarchically Classifying Documents using Very Few Words”, Proceedings of the 14th International Conference on Machine Learning (ML), Pp. 170-178, 1997.

[16] A. Maedche and S. Staab, “Ontology Learning for the Semantic Web”, IEEE Intelligent Systems, Special Issue on the Semantic Web, Vol. **6**, No.2, 2001.

[17] Banerjee, A. and Louis, S.J., “A Recursive Clustering Methodology using a Genetic Algorithm”, IEEE Congress on Evolutionary Computation, Pp. 2165-2172, 2007.

[18] Peici Fang and Siyao Zheng; “A Research on Fuzzy Formal Concept Analysis Based Collaborative Filtering Recommendation System”, Second International Symposium on Knowledge Acquisition and Modeling, KAM ’09, Pp. 352-355, 2009.

[19] C. J. Van Rijsbergen, “Information Retrieval”, Buttersworth, London, second edition, 1989.

[20] G. Kowalski, “Information Retrieval Systems - Theory and Implementation”, Kluwer Academic Publishers, 1997.

[21] A. Faatz and R. Steinmetz, “Ontology enrichment with texts from the WWW”, in Proceedings of Semantic Web Mining second Workshop at ECML/PKDD-2002.

[22] Stekh Yu, Sardieh. F.M.E, Lobur. M and Dombrova. M, “Algorithm for clustering web documents”, Proceedings of VIth International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH), Pp. 187, 2010.

[23] Berkhin. P, “Survey of clustering data mining techniques”, Accrue Software Research Paper, 2002.

[24] Hongwei Yang, “A Document Clustering Algorithm for Web Search Engine Retrieval System”, International Conference on e-Education, e-Business, e-Management, and e-Learning, Pp. 383-386, 2010.

[25] R. Cucchiara, “Genetic algorithms for clustering in machine vision. Machine Vision and Applications”, **11**: 1-6, 1998.

[26] S. Kampa, T. Miles-Board and L. Carr, “Hypertext in the Semantic Web”, In Proceedings ACM Conference on Hypertext and Hypermedia, Aarhus, Denmark, pp. 237-238, 2001.

[27] C. S. Lee, Y. J. Chen, and Z. W. Jian, ”Ontology-based fuzzy event extraction agent for Chinese E-news summarization,” Expert Syst. Appl., vol. **25**, no. 3, pp. 431-447, Oct. 2003.

[28] Ding, Li et al. 2004. Swoogle: A Search and Metadata Engine for the Semantic Web. In Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management. New York: ACM Press, pp. 652-659. http://ebiquity.umbc.edu/paper/html/id/183/Swoogle-A-Search-and-Metadata-Engine-for-the-Semantic-Web.

[29] A.M. Bensaid, L.O. Hall, J.C. Bezdek, and L.P. Clarke. Partially supervised clustering for image segmentation. Pattern Recognition, **29**(5): 859-871, 1996.

[30] M. Sarkar, B. Yegnanarayana, and D. Khemani. A clustering algorithm using an evolutionary programming-based approach. Pattern Recognition Letters, **18**: 975-986, 1997.

[31] C.A.Murthy and N. Chowdhury. In search of optimal clusters using genetic algorithms. Pattern Recognition Letters, **17**: 825-832, 1996.

[32] B. Ganter and R. Wille, “Formal Concept Analysis: Mathematical Foundations”, Springer, Berlin - Heidelberg, 1999.

[33] King-Ip Lin, Ravikumar Kondadadi, “A Similarity Based Soft Clustering Algorithm for Documents”, 7th International Conference on Database Systems for Advanced Applications, 2001.

[34] N.Nanas, V.Uren and A. de Roeck, “Building and Applying a Concept Hierarchy Representation of a User Profile”, In Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, 2003.

[35] Thanh Tho Quan, Siu Cheung Hui and Tru Hoang Cao, “FOGA: A Fuzzy Ontology Generation Framework for Scholarly Semantic Web”, Proceedings of Knowledge Discovery and Ontologies Workshop, 2004.

[36] Zitao Liu, Wenchao Yu, Yalan Deng, Yongtao Wang and Zhiqi Bian, “A feature selection method for document clustering based on part-of-speech and word co-occurrence”, Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Vol. **5**, Pp. 2331-2334, 2010.

[37] Ziqiang Wang, Qingzhou Zhanga and Xia Sun, ”Document clustering algorithm based on NMF and SVDD”, Second International Conference on Communication Systems, Networks and Applications (ICCSNA), Vol. **1**, Pp-192-195, 2010.

[38] Roy. R.S. and Toshniwal. D, “Fuzzy Clustering of Text Documents Using Nave Bayesian Concept”, International Conference on Recent Trends in Information, Telecommunication and Computing (ITC), Pp. 55-59, 2010.

[39] Cobos. C, Andrade. J, Constain. W, Mendoza. M and Leon. E, “Web document clustering based on Global-Best Harmony Search, K-means, Frequent Term Sets and Bayesian Information Criterion”, IEEE Congress on Evolutionary Computation (CEC), Pp. 1-8, 2010.

[40] Jones, Gareth, Robertson, Alexander. M, Santimetvirul, Chawchat and Willett Peter, ”Non-hierarchic document clustering using a genetic algorithm”, Information Research, **1**(1), 1995.

[41] Thangamani, M. and Thangaraj, P. “Effective Fuzzy Ontology for Distributed Document Using Non-Dominated Ranked Genetic Algorithm”, International Journal of Intelligent Information Technologies, Vol. **7**, Issue 4, pp. 26-46, 2011.

[42] Thangamani.M and Thangaraj.p, “Effective fuzzy semantic clustering scheme for decentralized network through multidomain ontology model”, International Journal of Metadata, Semantics and Ontologies, Inderscience publication Vol. **7**, Issue 2, Pp. 131-139, 2012.

M. Thangamani completed her B.E.,(Electronic and Communication Engineering) from Government College of Technology, Coimbatore, India. She completed her M.E.,(Computer Science & Engineering) from Anna University, Chennai, India. Now she is doing research in the field of Fuzzy concepts with soft computing. Currently, she is working as Asst. Professor in the Department of Computer Science and Engineering, Kongu Engineering College, Tamil Nadu, India. She has published 15 articles in International journals and presented papers in 42 National and International conferences. She has published 8 books for polytechnic colleges and also guided many UG projects. She has organized many self supporting and sponsored National Conference and Workshop in the field of Data Mining. She also seasonal reviewer in IEEE Transaction on Fuzzy System and International journal of advances in Fuzzy System.

P. Thangaraj received the Bachelor of Science degree in Mathematics from Madras University in 1981 and his Master of Science degree in Mathematics from the Madras University in 1983. He completed his M.Phil degree in the year 1993 from Bharathiyar University. He completed his research work on Fuzzy Metric Spaces and awarded Ph.D degree by Bharathiyar University in the year 2004. He completed the post graduation in Computer Applications at IGNOU in 2005. His thesis was on ”Efficient search tool for job portals”. He completed his Master of Engineering degree in Computer Science in the year 2007 from Vinayaka Missions University. His thesis was on ”Congestion control mechanism for wired networks”. Currently he is a Professor and Head of Computer Science and Engineering at Bannari Amman Institute of Technology, Sathyamangalam. His current area of research interests are in Fuzzy based routing techniques in Ad-hoc Networks.