Intelligent Document Clustering for Big Data Applications

Mr. A. SENTHIL KARTHICK KUMAR[1] , Dr. A.M.J MOHAMED ZUBAIR RAHMAN[2] &
Dr. M. THANGAMANI[3]

***Abstract*—** the increase in number of documents worldwide increases the difficulty for classifying those documents according to these needs. To retrieve the document traditional approaches has limited practical applicability. In order to improve the performance of document clustering, ontologies are playing important vital role in business intelligent. This research explores an approach to develop and use ontology in providing a semantically rich knowledge representation for intelligent information retrieval. The outcome of this approach is to establish common vocabulary to access library information intelligently by genetic fuzzy clustering.

***Keywords*—**Fuzzy ontology, clustering, distributed clustering, Peer to peer network.

I. Introduction

Genetic Algorithm (GA) is considered to provide better clustering results. The convergence time for the usage for GA is more and also the number of iterations required for GA is more when compared to other techniques. For enriching the performance measure, in this paper, the fuzzy ontology is applied to the database to reduce the convergence time and number of iterations before using GA. The usage of fuzzy ontology will provide better classification of a large, vague database. This has motivated the usage of fuzzy ontology and GA for clustering. Hence, in this approach, fuzzy ontology is combined with the GA to yield better classification accuracy for large databases.

In this paper, ontology [1] is introduced as a modeling technology for structured metadata definition within document clustering system. Documents can be clustered with the metadata obtained using the Genetic Algorithms (GAs) [2]. GA is a search technique based on natural genetic, selection and merging of survival of the fittest with structured interchanges. It conserves the attributes of finest exponents of a generation for use in the next generation; additionally introducing the variations in the new generation composition with the help of cross over and mutation function. GA [3] is a famous technique for handling complex search problems through implementing an evolutionary stochastic search because GA can be very effectively applied to various challenging optimization problems.

II. Related works

Lena Tenenboim et al, [4] proposed ontology based classification for document clustering. The author recommended classification of news items in an 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. As classical Euclidean distance metric could not create a suitable separation for data lying in manifold, a GA based clustering method based on geodesic distance measure was proposed by Gang Li et al, [5]. 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.

Casillas et al, [6] also put forth a concept of document clustering using GA. It deals with document clustering that computes an approximation of the optimum k value and resolves the best clustering of the documents into k clusters. It is experimented with sets of documents that are the output of a query in a search engine. Andreas et al, [7] advocated text data clustering technique. Text clustering, usually, involves clustering in a high dimensional space that appears complex with regard to all virtual practical settings. Additionally, a scrupulous clustering outcome is provided.

Word sets based document clustering algorithm for large datasets was proposed by Sharma et al., [8]. Document clustering is a significant tool for use in search engines and document browsers. It facilitates the user to have a better overall observation of the data available in the documents. There is also a strong requirement for hierarchical document clustering [9] where clustered documents can be browsed based on the increasing specificity of topics. Frequent Item set Hierarchical Clustering (FIHC) is used for hierarchical grouping of text documents. This technique does not provide consistent clustering results when the number of frequent sets of terms is large. In this paper, the authors proposed Word sets-based Clustering (WDC), an efficient clustering technique based on closed words sets. WDC makes use of hierarchical technique to cluster text documents having common words.

Mr.A.Senthil Karthick Kumar, Assistant Professor, Department of Computer Applications, Nehru Institute of Information Technology & Management, Coimbatore- 641105, Tamilnadu, India. (karthickmcamba@gmail.com)

Dr. A. M. J. Md. Zubair Rahman, Principal, Al-Ameen Engineering College, Erode, Tamilnadu, India. Pincode-638104. (mdzubairrahman@gmail.com)

Dr. M. Thangamani, Assistant Professor, Department of Computer Science, Kongu Engineering College, Perundurai -638 052, Erode District, Tamilnadu, India. (manithangamani2@gmail.com)

Cao et al., [10] 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. Named entities are introduced as objectives into fuzzy document clustering, the important elements of defining document semantics and in many cases are of user concerns. Zhang et al., [11] gave clustering aggregation based on GA for documents clustering. A technique based on GA for clustering aggregation difficulty, named as GeneticCA, is provided to approximate the clustering performance of a clustering division. In this case, clustering precision is defined and features of clustering precision are considered. Web document clustering using document index graph is put forth by Momin et al., [12]. Document clustering methods are generally based on single term examination of document data set. To attain more precise document clustering, more informative features like phrases are essential.

Muflikhah et al. [13] proposed a document clustering based on concept space and cosine similarity measurement. This technique aims at incorporating the information retrieval and document clustering into concept space approach. It is known as Latent Semantic Index (LSI) because it uses Singular Vector Decomposition (SVD) or Principle Component Analysis (PCA). Its objective is to decrease the matrix dimension by identifying the pattern in document collection with reference to the terms. Affinity-based similarity measure for Web document clustering is presented by Shyu et al., [14]. Document clustering is extended into Web document clustering by establishing affinity based similarity measure, It 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. The experimental results illustrate the fact that similarity measure outperforms the cosine coefficient and the Euclidean distance technique under various document clustering techniques.

ELdesoky et al., [15] gave a similarity measure for document clustering based on topic phrases. In the conventional vector space model (VSM), researchers have used unique word available in the document set as the candidate feature. Currently, phrase based informative feature is considered because it contributes to enhancing the document clustering accuracy and effectiveness. Similarity measure of the traditional VSM is evaluated by considering the topics phrases of the document as the comprising terms instead of the conventional term. Thangamani et. al examined document clustering [16, 17] in individual and peer to peer environment and also developed the system for automatic extraction and classification of document using multi domain ontology. Senthil Karthick kumar et. al says Fuzzy Expert system (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. And evaluation result shows that the proposed approach is very significant in clustering the documents in the distributed environment [43].

A document clustering method based on hierarchical algorithm with model clustering is presented by Haojun et al., [18]. It analyzes and makes use of cluster overlapping to design cluster merging criterion. Document clustering with fuzzy c-mean algorithm is proposed by Thaung et al., [19]. Most traditional clustering technique allocates each data to exactly single cluster, therefore creating a crisp separation of the data provided. However, fuzzy clustering permits for degrees of membership to which data fit into various clusters.

In high dimensional data, clusters often exist in subspaces rather than in the entire space. One solution to this problem in text subspace clustering [20, 21], which aims to discovering the document clusters in different subspace of the original word space. Fuzzy clustering [22] in contrast to the usual (crisp) methods does not provide hard clusters, but returns a degree of membership of each object to all the clusters in Bezdek [23]. A feature-weighting algorithm combined with the fuzzy K prototypes algorithm was presented by [24].

Although K-means monitoring algorithm does not generate distributed clustering, normally, it helps centralized K-means process know when to recomputed the clusters by monitoring the distribution of centroids across peers and trigger reclustering in case the data distribution getting changed over time. On the other hand, P2P K-means algorithm Data et al., [25] updates the centroids at each peer as per the information received from their immediate neighbors. This process gets terminated when the information received does not result in significant update of the centroids of all peers. The P2P K-means algorithm finds its roots in a parallel implementation of K-means as proposed by [26]. Deise et al., [27], proposed a system that involved an increase in the semantic through considering both information search and storage. Chang et al., [28] introduced distributed document clustering for search engine. Qing et al., [29] propounded a text clustering algorithm based on frequent term sets for peer-to-peer networks. Wolff et al., [30] presented, threshold based Data mining in Peer-to-Peer Systems for local optimization. Yang Lu and Xue Wan, [31] introduced semantic-based P2P resource management system yet it supports only local optimization. Many other introduced centroids based algorithm for text categorization [32]. A new algorithm presented for fuzzy document clustering in [33].

Genetic algorithm [34] is a famous technique to deal with complex search problems by implementing an evolutionary stochastic search because genetic algorithm can be very effectively applied to various challenging optimization problems. The NP-hard nature of the clustering technique makes genetic algorithm a natural choice for solving it [35, 36, 37 and 38]. A common objective function in these implementations is to decrease the square error. This paper clearly presents the ontology based document clustering methodology with genetic algorithm [39]. Xindong et. al [40] investigated as data mining techniques can applied to big data set for information extraction.

III. DISTRIBUTED ENVIRONMENTS

Clustering approach for P2P networks is built. Peer node information is created and distributed to other nodes. Each peer node upholds several document collections. Super nodes hierarchically control the peer nodes [41]. Each super node controls the peer node information. The peer node list illustrates details on the peer node. Peer node document details are registered in the super node. Users can view the documents in the chosen peer. The peer node updates super nodes with all information. Each super node is hierarchically linked to other similar super node. The main super node treats these super nodes as its peers. The peer nodes can dynamically join and leave network environment. This dynamic join and leave operations generated an impact upon the clustering process. Documents are clustered through extracting information from the peer.

IV. EXPERT SYSTEM FOR DISTRIBUTED TEXT CLUSTERING

Initially, ontology generation using fuzzy logic is implemented to the database containing a large amount of documents. This technique generates the ontology for the given database. With this ontology, the next step is the application of GA. GA is used for clustering the documents in the database with the help of ontology generated by fuzzy logic technique. The combination of expert system generation using Fuzzy Logic and GA helps to increase the accuracy of clustering. It consists of the following modules.

Formal concept analysis using fuzzy: In fuzzy formal concept analysis integrates fuzzy logic into formal concept analysis to represent vague data. A fuzzy formal context shown in Table 1 consists of three objects which denote three documents. Those objects are named as D1, D2 and D3. Moreover, it has three attributes such as Data Mining, Clustering and Fuzzy Logic indicating the three titles. A membership value between 0 and 1 denotes the relationship between an object and an attribute. To remove the relationships that have low membership values, a confidence threshold T is introduced. Table 2 represents the fuzzy formal context provided in Table 1 with confidence threshold T as 0.5. Usually, the attributes of a formal concept can be considered as the description of the concept. Thus, the relationships between the object and the concept must be the separation of the relationships between the objects and the attributes of the concept. A membership value in fuzzy formal context denotes all the relationship between the object and an attribute. Then based on fuzzy theory, the intersection of these membership values must be the minimum of these membership values. Figure 1 one shows the automatic generation of expert system for analyzing distributed textual clustering.

Table: Fuzzy formal context

	Data Mining	**Clustering**	**Fuzzy Logic**
D1	0.75	0.3	0.5
D2	1	0.75	0.25
D3	0.25	0.25	0.75

Table 2: Fuzzy formal context for table 1With T =0.5

	Data Mining	**Clustering**	**Fuzzy Logic**
D1	0.75	-	0.5
D2	1	0.75	-
D3	-	-	0.75

Expert system generation: While the formal concepts are also generated mathematically, distinct formal concepts are created on the basis of the difference in terms of attribute object and the traditional concept lattice. This produces the effect of concepts as interpreted by humans. Based on this observation, a cluster formal concept is infused into conceptual clusters of fuzzy conceptual clustering.

Class Mapping: In this process, the extent and intent of the fuzzy context are mapped into the extent and intent classes of the ontology. It requires supervised training to name the label for the extent class. Keyword attributes can be represented by appropriate names and they are used to label the intent class names also.

Taxonomy relation generation: With concept hierarchy in place, this phase produces the intent class of the ontology as a hierarchy of classes. The step can be considered as an isomorphic mapping from the concept hierarchy into taxonomy classes of the ontology.

Non-taxonomy relation generation: This step involves generating the similarities among the extent class and intent classes with no hierarchy between classes. The generation will mean an equivalent class with no sub class or super class.

Instances generation: In this process, instances for the extent class are generated. Each instance indicates an object in the initial fuzzy context. Depending on the data existing on the fuzzy concept hierarchy, instances attributes are automatically furnished with suitable values. For example, each instance of the class document, 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.

V. DESIGN OF CLUSTERING ALGORITHM USING GENETIC ALGORITHM

When considering that the number of category is k, this paper uses Genetic Algorithm [40] to find the better cluster centre. 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 centres. Each component ai presents a cluster centre and with the help of floating point number, every element of ai is encoded.

Step 2: Group initialization: Assuming the amount of initialized group is M, matrix collection X= (A1, A2…An) indicates the group collection. Elements of each matrix are collection of k×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 centre, and then the relation among all documents and cluster centres 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:

$$f = \frac{1}{1+E} \qquad (1)$$

Where E is the Clustering objective function:

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

Where x_j^* denotes the centre of cluster c_j , n_j is the amount of documents in clusterc_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:

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

Where M is the group size, Rank(C) is the serial number following the sorting of individual, and Rank(C)∈{h+1,h+2,…,M }, a+b=2 and a∈{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

1. If r<pc, carry outs the crossover, 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 produces new individuals A' and B' by the following equation:

$$A' = rA + (1-r)B$$

$$\qquad (4)$$

$$B' = rA + (1-r)A$$

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

$$a_{ij}' = \begin{cases} a_{ij} + \Delta\left(t, a_{ij}^{max} - a_{ij}\right) & rand(0,1) = 0 \\ a_{ij} - \Delta\left(t, a_{ij} - a_{ij}^{min}\right) & rand(0,1) = 1 \end{cases} \qquad (5)$$

Where, $\Delta(t,y) = yr\left(1 - \frac{t}{T}\right)^b$, a_{ij}^{max} and a_{ij}^{min} j=1,2…m 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. Δ(t, y)∈[0, y], so the probability that Δ(t, 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.

VI. EXPERIMENT DISCUSSIONS

This technique avoids getting stuck into a local maximum from which one cannot escape reaching a global maximum. This is one of the main benefits of GA in opposition to the conventional search techniques as the gradient technique. Another advantage is the utility of GA for real time applications, in spite of its inability to offer the optimal solution to the problem. However, it provides almost a better solution in a shorter time, including complex problems.

6.1 Data set description:

Real time Data set: The text documents collected from the IEEE web site are used for experimentation. Data mining domain related journal collection is downloaded from the web. The journal abstract page is designed using HTML.

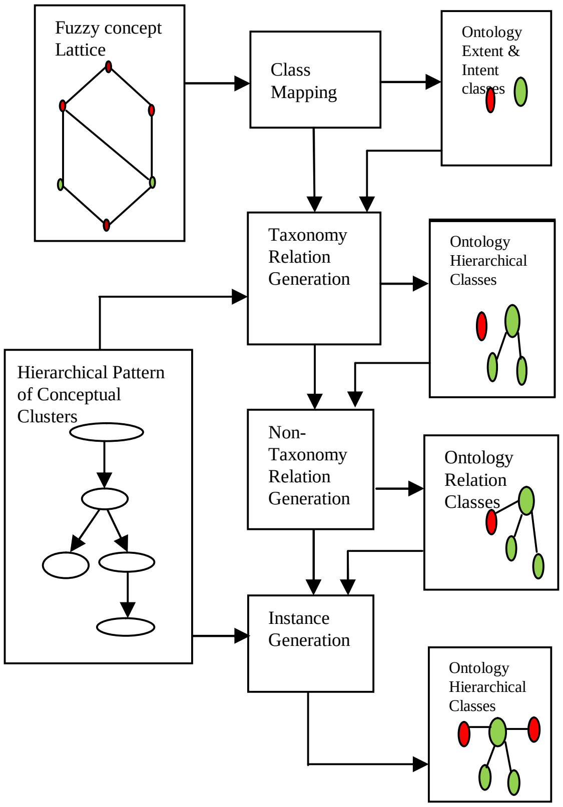
Fuzzy concept Lattice
Class Mapping
Ontology Extent & Intent classes
Taxonomy Relation Generation
Ontology Hierarchical Classes
Hierarchical Pattern of Conceptual Clusters
Non-Taxonomy Relation Generation
Ontology Relation Classes
Instance Generation
Ontology Hierarchical Classes

Figure.1: Expert system for distributed document clustering

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.

Benchmark data set: The Reuters text classification database [42] was derived from the original Reuters-21578 data set made publicly available as part of the Reuters Corpus, through Reuters, Inc., Carnegie Group and David Lewis. This data consists of 12,902 documents. Each document is a news article about some topic: e.g. earnings, commodities, acquisitions, grain, copper, etc. There are 5 category set in the dataset i.e., Exchanges, Orgs, People, Places and Topics. From the dataset articles were randomly chosen and used for evaluating the proposed clustering technique.

6.2 Experiment Measures:

F-Measure parameter is then considered for evaluating the proposed technique. The F-measure of a class $\mathbf{i}$ is defined as:

$$F(i) = \frac{2PR}{P+R} \qquad (6)$$

Where P is Precision and R is recall.

The overall F-measure for the clustering result C is the weighted average of the F-measure for every class i is given by

$$F_c = \frac{\sum_i (|i| \times F(i))}{\sum_i |i|} \qquad (7)$$

Where |i| is the number of objects in class i. When the overall F-measure is higher, the resulting clustering will be better because higher accuracy of the clusters mapping to the original classes.

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

$$E = \sum_{j=1}^{k} \sum_{x_i \in c_j} (x_i - x_j^*)^2 / n_j \qquad (8)$$

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

6.3 Experiment Result and Discussion:

MATLAB is used for evaluating the proposed clustering technique. This clearly indicates that the proposed fuzzy ontology with GA technique results in good clustering when compared to existing clustering techniques in Table 3 and figure 2 and figure 3 by calculating F-Measure using the following way.

Table 3 and Figure 2 show the result of F-measure for real time dataset. F-measure of fuzzy ontology using genetic algorithm is 0.69 than K-mean. High F-measure cluster more similar documents.

Table 3: Result of F-Measure for Real Time Dataset

Clustering Methods	**Fmeasure**
K-Mean	0.61
Fuzzy ontology with GA	0.69

Table 3: show the result of F-measure for real time dataset. F-measure of fuzzy ontology using genetic algorithm is 0.72 than K-mean. Hence relative measure of fuzzy ontology using genetic algorithm is 20% improvement compared to the existing system K-Means.

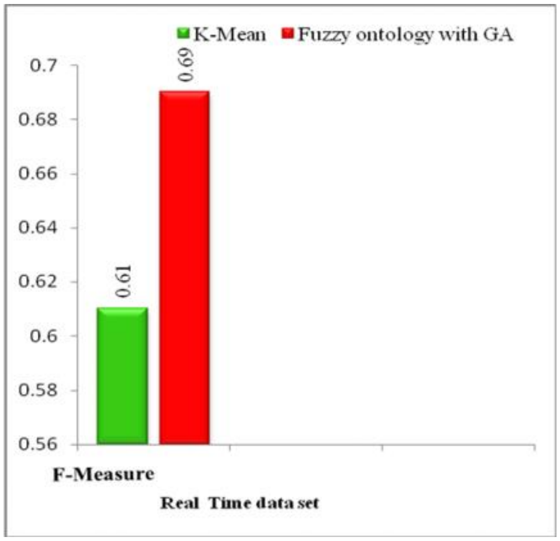

Figure 2: F-Measure Comparison of Fuzzy ontology using genetic algorithm with K-means for Real Time dataset

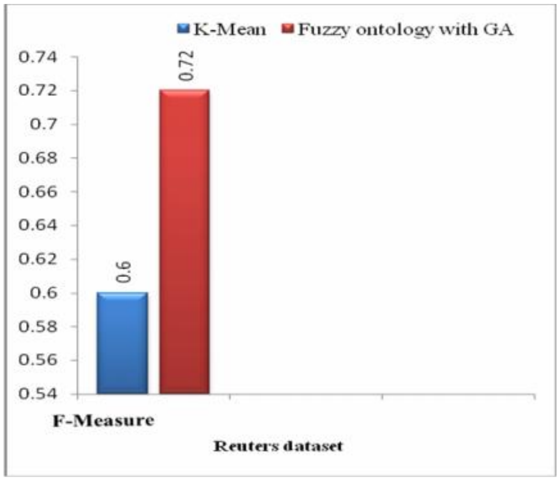

Figure 3: F-Measure Comparison of Fuzzy ontology using genetic algorithm with K-means for Reuter’s dataset

VII. CONCLUSION AND FUTURE WORK

In this work, expert system is created with help of bio inspired method for large data. In further investigation the expert system need to be applied towards different dataset to test how much this system will be effective. Also system can be deployed in Cloud Hadoop environment with e-Learning activities to handle the big data.

REFERENCES

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

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

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

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

[5] 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, 2009, Pp: 274 – 278.

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

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

[8] 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.

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

[10] 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, 2008, Pp. 2028 – 2034.

[11] 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, 2008, Pp. 3156 – 3161.

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

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

[14] 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, 2004, Pp. 247 – 252..

[15] 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, 2009, Pp. 92-96.

[16] Thangamani .M and Thangaraj .P,“Survey on Text Document Clustering”, International Journal of Computer Science and Information Security, Vol.8(4),2010.

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

[18] 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, 2008, Pp. 1229 – 1233.

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

[20] Thangamani .M and Thangaraj P,“Ontology Based Fuzzy Document Clustering Scheme”, Modern Applied Science,vol.4(7), 2010, pp.148-153.

[21] Hung J. Z, Ng. M. K, Rong .H and Li .Z, “Automated variable weighting in K-means type clustering,” IEEE Transactions on knowledge and Data Engineering, Vol. 27(5), 2005, pp.657-668.

[22] Bezdek, J., C., Pattern Recognition with Fuzzy Objective Function Algoritms, Plenum Press, New York, Vol.2, 1988.

[23] Li .J, X.Gao, and L.Jiao, “A novel feature weighted fuzzy clustering algorithm,” LNAI, vol. 3641, 2005, pp. 412–420.

[24] Jing. L, M.K.Ng, and J.Z.Huang, (2007) “An entropy weighting kmeans algorithm for subspace clustering of high-dimensinoal sparese data,” IEEE Transactions on Knowledge and Data Engineering, vol. 19(8),2007, pp. 1–16.

[25] Datta .S, Giannella .C, and Kargupta .H, “K-means Clustering over a Large, Dynamic Network,” Proc. Sixth SIAM Int’l Conf. Data Mining (SDM ’06), 2006, pp. 153-164 .

[26] Datta .S, Giannella .C, Bhaduri .K, Wolff .R, and Kargupta .H, "Distributed Data Mining in Peer-to-Peer Networks," IEEE Internet Computing, vol. 10, no. 4, 2006, pp. 18-26.

[27] Deise de Brum Saccol, Nina Edelweiss, Renata de Matos Galante, Márcio Roberto de Mello, "Managing Application Domains in P2P Systems", IEEE international conference on Information Reuse and Integration, 2008, Pp.451-456.

[28] Chang Liu, Song-Nian Yu and QiangGuo, "Distributed document clustering for search engine", Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition, 2009, Pp. 454 – 459.

[29] Qing He, Tingting Li, Fuzhen Zhuang and Zhongzhi Shi, "Frequent Term based Peer-to-Peer Text Clustering", 3rd international symposium on knowledge Acquisition and Modeling, 2010, Pp.352-355.

[30] Wolff .R, K. Bhaduri, and H. Kargupta, "Local L2-Thresholding Based Data Mining in Peer-to-Peer Systems," Proc. Sixth SIAM Int'l Conf. Data Mining (SDM '06), 2006, pp. 430-441.

[31] Yang Lu, Xue Wan, " A semantic-based P2P Resource Management System", 2nd International Asia Conference on Informatics in Control, Automation and Robotics, 6-7 March 2010, Pp. 188-191.

[32] Chen.L, Jiang.Q, and Wang .S, "A probability model for projective clustering on high dimensional data," Proceeding of the IEEE ICDM, 2008, pp. 755–760.

[33] Lifei Chen, Shengrui Wang and Qingshan Jiang, "A Robust Algorithm for Fuzzy Document Clustering", 2009, Pp. 679-684.

[34] Murthy, C.A. and Chowdhury, N., "In Search of Optimal Clusters using Genetic Algorithms", Pattern Recognition Letters,1996, Pp. 825–832.

[35] Cucchiara, "Genetic algorithms for clustering in machine vision. Machine Vision and Applications", 1998, 11:1–6..

[36] 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.

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

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

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

[40] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE, "Data Mining with Big Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, 2014, Pp. 97- 107.

[41] Thangamani M, "Mining intelligence and knowledge exploration for automatic classification in distributed environment ", International journal of Computer and Informatics, published by Slovak Academy of Sciences, Bratislava. (Forthcoming)

[42] Wai Lam, Kon-Fan Low , "Automatic document classification based on probabilistic reasoning: model and performance analysis, IEEE International Conference on Computational Cybernetics and Simulation, proceeding of Systems, Man, and Cybernetics, 1997 Vol. 3 , pp- 2719 – 2723.

[43] Senthil Karthick Kumar A, Zubair Rahman A.M.J.Md and Thangamani M, "Bio-Inspired Fuzzy Expert System for Mining Big Data", WSEAS Proceedings of the 16th International Conference on Mathematical and Computational Methods in Science and Engineering , Vol 24, PP 222-227, 2014

A.Senthil Karthick Kumar is working as an Assistant Professor in Nehru Institute of Information Technology and Management, Coimbatore Affiliated to Anna University. He enrolled his Research at the year of 2012 as a part time Scholar in Bharathiar University, Coimbatore. He has an experience of 5 years in Academics and 3 years in Information Technology field. He completed his B.Sc in Information Technology from Madurai Kamaraj University in 2003. Did his MCA from Bharathiar University in 2006; Completed his M.Phil in Computer Science in 2009 and E.M.B.A in Human Resource Management, from MS University 2012.

Prior to joining in NIITM he worked for 3years as a Human Resource Executive (Technical) in various companies like Perot Systems, Bangalore. He enrolled his Life time Membership with ISTE, Member in CSI and IAENG. He is Editorial Board member in Techno Publications, and Technical Board member in Science Publication.org.

He has published and presented around 10 papers in National and International level Conference, Seminars and Journals. He visited countries like Malaysia, Thailand for his research work. His area of interest in research includes E-Learning, Cloud computing, Software Engineering and Data mining.

Dr. A.M.J Mohamed Zubair Rahman, Principal, Al-Ameen Engineering College, Erode. He is a Person with 22 Years of Teaching Experience and He was awarded with Ph.D from Anna University, Chennai in the year 2009. Add on to his academics excellence he completed his M.S. Software Systems from BITS-Pilani in the year 1996, He completed his B.E Computer Science Engineering from IRTT in 1989, further in continuation of his education he did his M.E Computer Science Engineering from Bharathiar University in 2002.

To his credit he has attended several National and International Seminars and presented more than 20 papers in various conferences. He enrolled his Life time Membership with ISTE, and Member in CSI. He has published and presented around 20 papers in National and International Journals. His area of interest in research includes Data mining, Network Security, Software Engineering and E-learning.

Dr. M. Thangamani completed her B.E., from Government College of Technology, Coimbatore, India. She completed her M.E., and PhD (Computer Science and Engineering) from Anna University, Chennai, India. Currently, she is working as Assistant Professor in the Department of Computer Science and Engineering, Kongu Engineering College, Tamil Nadu, India.

She has published 23 articles in International journals and presented papers in 48 National and International conferences. She has published 11 books for polytechnic colleges and also guided many UG projects. She has delivered more than 34 Guest Lectures in reputed engineering colleges on various topics. She has organized many self supporting and sponsored National Conference and Workshop in the field of Data mining and Cloud computing.

She also seasonal reviewer in IEEE Transaction on Fuzzy System, International journal of advances in Fuzzy System and Applied mathematics and information journals. She is also the editorial member for many International Journals and organizing chair for International conferences in India and other countries. She received best paper award in International conference on Advances in Science and Technology held in Thailand conducted by Science publication. She is keynote speaker in International conference on Trends and Innovation in Science and Technology-2014 in Thailand. Her research interests include Fuzzy, Data mining; Cloud computing, Ontology development, Web Services and Open Source Software.