DOCUMENT CLUSTERING USING AGGLOMERATIVE HIERARCHICAL CLUSTERING APPROACH (AHDC) AND PROPOSED TSG KEYWORD EXTRACTION METHOD

R. Nagarajan[1], S. Anu H Nair[2], N. Puviarasan[3], P.Aruna[4]

[1]*Assistant Professor/Programmer, Department of Computer Science and Engineering, Annamalai University, Tamilnadu, India*

[2]*Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Tamilnadu, India*

[3]*Associate Professor, Department of Computer Science and Engineering, Annamalai University, Tamilnadu, India*

[4]*Professor, Department of Computer Science and Engineering, Annamalai University, Tamilnadu, India*

Abstract

Document clustering is one of the emerging trends and it is an efficient approach for unorganized or unlabelled documents. Document clustering is a technique which finds out the concept of the documents using words presented in the documents. Dynamic growth of unorganized valuable documents increases the complexity of document organization. Keyword of the documents summarizes the documents. Keyword extraction techniques play an important role in document clustering. This research paper proposed a TSG keyword extraction method along with AHDC (Agglomerative Document Clustering approach) algorithm for document clustering. This proposed algorithm awards more than 90% of accuracy.

Keywords: *Keyword Extraction, Co-Word Construction, Statistical Method, Graph Based Method, Tagging Method, Document Clustering, AHDC.*

----------***----------

1. INTRODUCTION

Information is better utilized when it is processed, to be easier to find, better organized, or summarized for easier digestion. Areas dealing with such problems are at the cross-roads of information retrieval, machine learning and statistical analysis. Text and web mining problems in particular, use methodologies often spanning those areas. A complexity of the retrieval of relevant document from a large corpus of documents is the most common challenging problem. In addition, the growth of unlabelled and unsupervised documents also increases this complexity. Organized documents are very much useful to reduce the time factor as well as accurate retrieval of information against the user's request. A well organization of documents is achieved by grouping or clustering the documents based on their concepts. Document clustering algorithms play a vital role to reduce this problem.

2. LITERATURE REVIEW

This section reviews the literature review on document clustering algorithms and discusses the difference between them. [1] has evaluated the user interaction in the context of web-based group decision support systems. Similar user pattern groups were identified from information-rich server logs. The groups were derived through multiple sequence alignment and hierarchical cluster analysis based on varying user activity measures. [2] has proposed a new approach that exploits ontology hierarchical structure and relations to provide a more accurate assessment of the similarity between term for word sense disambiguation. [3] has aimed to improve the efficiency of the clustering of huge real-world data by introducing a set of agglomerative hierarchical clustering methods. This proposed approach builds a hierarchy based on a group of centroids instead of building cluster hierarchies based on raw data points. [4] has presented a novel down-top incremental conceptual hierarchical text clustering approach using CFu-Tree (ICHTC-CF) representation, which starts with each item as a separate cluster and term-based feature extraction is used for summarizing a cluster in the process. [5] has proposed a two-stage approach which involves clustering followed by Paraphrase Recognition for extraction of sentence level paraphrases from text collections. In order to handle the ambiguity and inherent variability of natural language a fuzzy hierarchical clustering approach which combines agglomeration based on verbs and division on nouns has been used. [6] has proposed a phrase based clustering scheme which based on application of Suffix Tree Document Clustering (STDC), and the proposed scheme has

four steps as Document collection as input, Document cleaning, Phrase Identification of clusters and Phrase merging clusters. [7] has proposed a general framework of hierarchical clustering working in any α-relaxed metric spaces. This proposed framework was able to achieve good approximation factors on all of the clustering levels and construct the clustering hierarchy only on meaningful clustering levels whose clustering costs were independent of lower levels. [8] has proposed a clustering algorithm based on the hierarchical structure, to form normal behavior profile on the audit records and adjust the profile timely as the program behavior changed. This proposed algorithm converts the problem to resolve the problem of massive data processing to the hot research point of anomaly detection. [9] has presented a memory efficient online hierarchical algorithm called SparseHC, scans a sorted and possibly sparse distance matrix chunk-by-chunk, a dendrogram was built by merging cluster pairs as and when the distance between them was determined to be the smallest among all remaining cluster pairs.

3. KEYWORD EXTRACTION METHODS

Keywords play a vital part in document clustering. This research work proposed a novel keyword extraction algorithm. This proposed TSG keyword extraction method, integrates both statistical and graph based techniques with tagging method. Because of the use of impact of both statistical and graph based techniques, this proposed TSG method yields best keywords. Unlike the existing keyword extraction methods, this method calculates the term weight based on their physical position of document such us major title, subtitle, italic quotations, bold specification, author keywords etc. instead of their frequency alone. It is also found through our research that mere a single keyword is not enough for document clustering. Because of this reason, our second part of proposed algorithm constructs the co-words from the weighed keywords.

3.1 Statistical Methods

The statistical approaches used here are, 1. High Frequency (HF) 2. Term Frequency Inverse Sentence Frequency (TF-ISF) and 3. Co-occurrence Statistical Information (CSI). HF method extract important words based on their occurrences in a document. TF-ISF method is the combination of Term Frequency (TF) method and Inverse Sentence Frequency (ISF). Term Frequency (TF) measures the weight based on the number of occurrences of a word in a document. ISF method measures the weight based on the assumption that word importance is inversely proportional to the total number of sentences of the document to which the word is present. In this TF-ISF method, the weight of the word is less when it occurs more number of times in sentences. That is, it should be identified as a common word and it will not help us to identify the concept of the document. In CSI method, co-occurrence distribution of each word is the weighted average of the word distribution of all documents in which the word occurs. The co-occurrence distribution of a word can also be compared with the word distribution of a text. This gives us a measure to determine how typical (distinctive) a word is for a text. In addition, CSI method is used to construct co-words, which are the combination of two or more important keywords. These co-words contribute more to find out the concept of the documents.

3.2 Graph Based Method

In this graph based method, undirected graph is constructed for every document, the keywords of the documents are represented as nodes of the graph and co-occurrence relation between the words are represented as edges. Syntactic filters are used to filter the nodes of the graph. Edge weights are calculated as the co-occurrence count of the words in the order of connectivity. Finally, co-words are constructed by analyzing the graph constructed.

3.3 Proposed TSG Method

This proposed TSG method is the combinational method, which integrates the impact of statistical, graph based and tagging techniques. In tagging method, HTML tags are used extract the keywords from the documents. HTML tags are helped to identify the physical location of the words in the documents. By this method, words are extracted under three groups, 1. Title group 2. Property group and 3. Raw group words. Title group words are the words used in the Titles, subtitles, table captions, figure captions, images, author keywords of the documents. Property group words shows the properties of the words such that bold, bolditalic, italic, colored, big, strong, subscripts, superscripts, small, underlined, strikethrough, etc. The rest of the words are the Raw group words, usually used in the body of the documents.

In this proposed TSG method, initially keywords are extracted with their document ID, Sentence ID and frequencies, under three groups namely title group, property group and raw group. Then, co-words are constructed using statistical methods and graph based method. By comparing results obtained by these two methods, common co-words are identified and labeled as keyword/co-words of the documents.

4. DOCUMENT CLUSTERING BY AGGLOMERATIVE HIERARCHICAL DOCUMENT CLUSTERING APPROACH (AHDC)

Agglomerative hierarchical clustering method is a down-top method. This method start by assigning each document is a cluster and iteratively either merges clusters into larger clusters or splits the cluster. Merging processes implemented when two clusters are closed to each other's according to the similarity measures. Inversely splitting processes performs when two clusters are far away and this iteration process is continued until the termination constraint reached. Here, the iteration process stops, when all the documents are merged or split.

This method divides its process into two major parts, the first part is to find out the distance between the documents and second part forms clusters. The distance between the documents is calculated with the features of the documents derived. Here, the features are the keyword/co-words of the documents. The distance measure between the documents is done using Euclidean distance which is defined as,

$$Dist(D_i F_i , D_j F_j) = \sqrt{\sum_{k=1}^{m} |D_{ik} - D_{jk}|^2} \text{ where i } \neq \text{ j} \quad (1)$$

Where

- $D_i F_i$ represents the features of the documents D_i and
- $D_j F_j$ represents the features of the documents D_j

Initially, $D_i F_i$ and $D_j F_j$ are taken as two individual clusters C_i and C_j. If the distance between the two clusters is maximum value, then there is no common feature between them. Inversely, the distance between two clusters has minimum value when two clusters have common features. The range of values allowed for the distance between 0 to √2 to normalize the similarity measuring values. The similarity between the clusters i and j is defined as

$$Sim(F_i, F_j) = \sqrt{\frac{\sum_{k=1}^{m} |D_{ik} - D_{jk}|^2}{2}} \text{ where i } \neq \text{ j} \quad (2)$$

By the similarity measure, the value 0 is assigned when the similarity measure between the two documents is far away and value 1 is assigned when the distance between the two documents are closer. The second part of this method forms the clusters based on their distances. Closet pair of clusters is merged together and form one larger cluster.

$$\text{Cluster}(C_i, C_j) \leftrightarrow \max(\text{Sim}(D_i F_i, D_j F_j)) \quad \text{i} \neq \text{j}$$

$$\text{and Sim}(D_i F_i, D_j F_j) \geq \emptyset \quad (3)$$

Where

- C_i and C_j are clusters that can be merged
- $\text{Sim}(D_i F_i, D_j F_j)$ is the similarity between C_i and C_j
- F_i and F_j are the features of C_i and C_j
- $\emptyset$ is the threshold value that the condition is satisfied.

Likewise, the far distance measure clusters are split from each other and kept as separate cluster.

$$\text{Split}(C_i, C_j) \leftrightarrow \min(\text{Sim}(D_i F_i, D_j F_j)) \quad \text{i} \neq \text{j}$$

$$\text{and Sim}(D_i F_i, D_j F_j) \geq \emptyset \quad (4)$$

Where

- C_i and C_j are clusters that can be split
- $\text{Sim}(D_i F_i, D_j F_j)$ is the similarity between C_i and C_j and
- F_i and F_j are the features of documents D_i and D_j.

The Algorithm 1 explains the steps involved in the agglomerative hierarchical document clustering method.

Algorithm 1 : Document clustering using AHDC method		
Input	:	Documents D_i
Output	:	Clusters Tree T
Step 1	:	Initialize T (Tree)
Step 2	:	**for** each D_i in a document set
		$C_i \leftarrow$ preprocessed(D_i)
		add C_i to T as a separate node
		repeat for each pair of clusters C_j and C_k in T
		if C_j and C_k are the closest pair of
		clusters in T
		merge(C_j, C_k)
		else
		split (C_j, C_k)
		compute cluster feature vectors
		changed
		end if
		Repeat until all clusters are not changed
		end for
Step 3	:	return tree T with clusters as nodes

Fig. 1 shows the flow diagram of AHDC approach

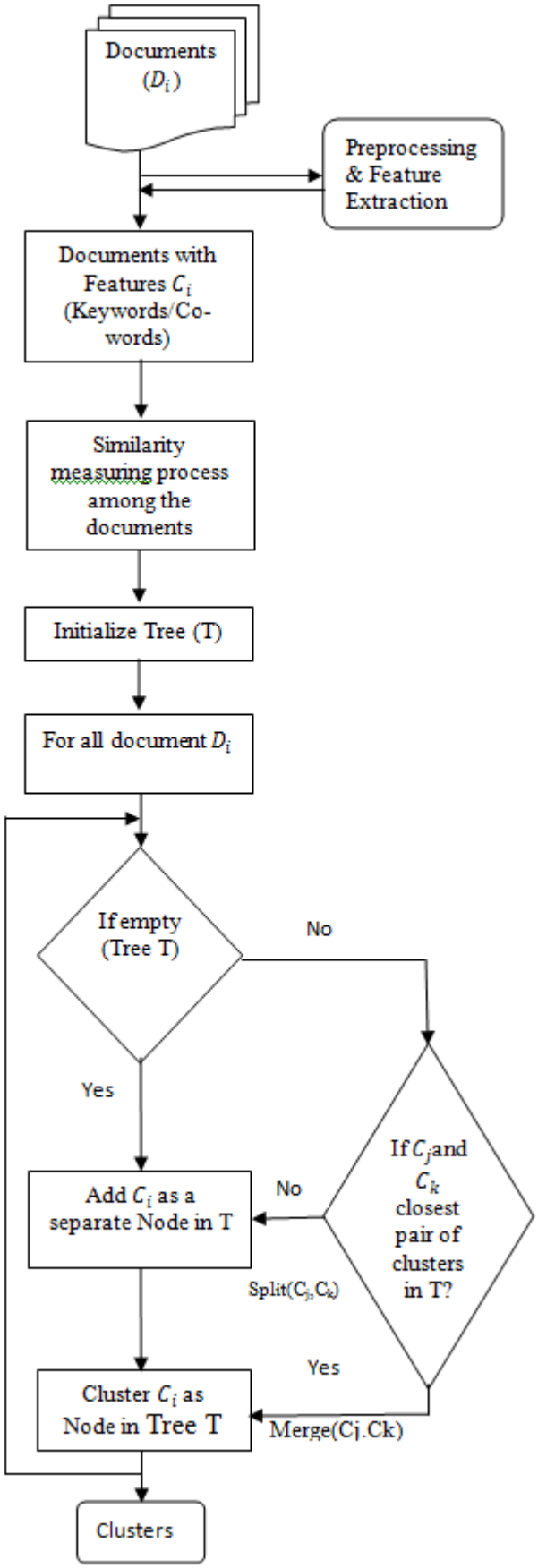


Fig. 1 Flow Diagram of AHDC Approach

5. EXPERIMENTAL ANALYSIS

To validate our proposed AHDC algorithm, sample of 5 documents with various filed of computer science were taken. Table 1 shows the sample documents with extracted keywords/co-words.

Table 1 Sample Documents with Keywords/Co-words

Doc.ID	Keyword/Co-words
D1	{data mining, dataset, preprocessing, information retrieval, machine learning}
D2	{image mining, graphics, recognition, segmentation}
D3	{database, data encryption, data compression, data mining}
D4	{data mining, preprocessing, dataset, machine learning}
D5	{database, data compression, data encryption}

By applying the distance measuring formulae in equation 2 to the above sample of documents, the distances between the documents were calculated. The Table 2 shows the distance between the sample documents.

Table 2: Similarity Measures among the Documents

D_j D_k	D1	D2	D3	D4	D5
D1	X	0.3	0.2	0.9	0
D2	0.3	X	0.3	0.2	0
D3	0.2	0.3	X	0.3	0.9
D4	0.9	0.2	0.3	X	0.3
D5	0	0	0.9	0.3	X

From the similarity measure values, initially D1 is chosen as our first node, the values of D1 and D2, D1 and D3, D1 and D4, D1 and D5 are compared, the values closer to 1 indicates that those documents deals same concepts. The distance value between D1 to D4 is 0.9, which is closer to value 1 and forms the first cluster C1{D1, D4}. Likewise, the value between D2 to other documents is compared and it is found that there no such distance values closer to 1, it shows that document D2 has not common discussions with other documents. The value between D3 and D5 is 0.9 which shows that there is a common discussion between D3 and D5 and forms the second cluster C2 {D3, D5}. Summarily, for the sample of documents two clusters were constructed C1{D1,D4} and C2{D3,D5}. Document D2 has no common concept with other documents, hence it was kept alone.

6. Results AND Discussions

Different performance measures used in this proposed work for document clustering are

i. Precision
ii. Recall
iii. F-measure
iv. Rand Index

6.1 Precision

Precision is another probability measure, which is also positive predictive value, that is, in clustering, precision is calculated by dividing documents clustered in a appropriate cluster(tp) with the sum of tp with documents clustered in not appropriate cluster (fp). Higher precision value closer to 1 shows the good clustering and vice versa. Precision is defines as,

$$Precision = \frac{tp}{tp+fp} \quad (5)$$

where

- tp is true positive, assigns similar documents in a same cluster
- fp is false positive, assigns dissimilar documents in a same cluster

6.2 Recall

Recall is another probability measure, which is also known as sensitivity, in clustering, recall is calculated by dividing documents clustered in a appropriate cluster(tp) with the sum of tp and missing documents in a cluster (fn). Higher recall value closer to 1 shows the good clustering and vice versa. Recall is defines as,

$$Recall = \frac{tp}{tp+fn} \quad (6)$$

where

- tp is true positive, assigns similar documents in a same cluster.
- fn is false negative, assigns similar documents in different clusters.

6.3 F-Measure

F-measure is a measure to test accuracy, which is also known as F-score. F-measure considers precision and the recall to calculate score. Precision is the number of correctly clustered documents divided by the number of all documents in a cluster, and recall is the number of correct positive clusters divided by the number of positive results that should have been clustered. The F-measure score can be interpreted as a weighted average of the precision and recall, where an F-measure score reaches its best value at 1 and worst at 0. F-measure is defined as,

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (7)$$

6.4 Rand Index

The rand index or rand measure is used to measure the similarity between two clusters, it is related to the accuracy of clusters formed. The higher rand index value closer to 1 indicates the good clustering and rand index value closer to 0 indicates bad clustering. Rand index can be defined as,

$$Rand\ Index\ (RI) = \frac{TP+TN}{TP+FP+FN+TN} \quad (8)$$

Where

- TP is true positive, assigns similar documents in a same cluster.
- TN is true negative, assigns dissimilar documents in different clusters.
- FP is false positive, assigns dissimilar documents in a same cluster.
- FN is false negative, assigns similar documents in different clusters.

The results of these performance indicators will help to find the best integrated clustering algorithm. Rand Index measures the accuracy of the clustering techniques. Table 3 shows the performance measures of AHDC approach with different keyword extraction methods.

Table 3 Performance Measures of AHDC approach with different keyword Extraction Methods

Keyword Extraction Methods	Performance measures of AHDC			
	Precision	**Recall**	**F-measure**	**Rand Index**
HFT	0.52	0.51	0.52	0.49
TF-ISF	0.62	0.58	0.60	0.51
CSI	0.72	0.67	0.69	0.63
GRAPH	0.71	0.64	0.67	0.63
Proposed TSG	0.92	0.89	0.90	0.84

Fig.2 Shows the precision, recall and F-measures of the combination of statistical, graph and proposed TSG methods with AHDC approach.

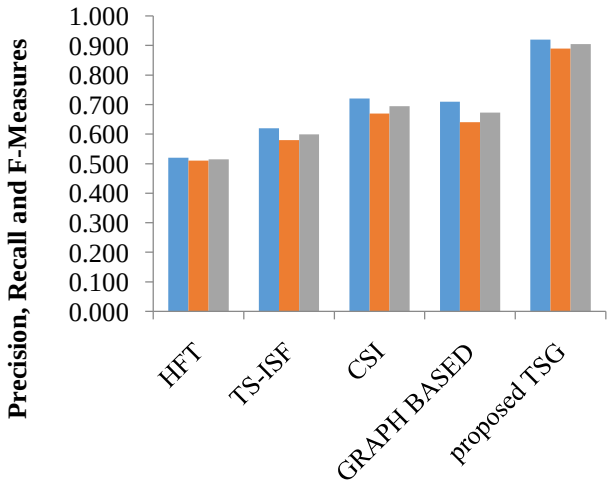
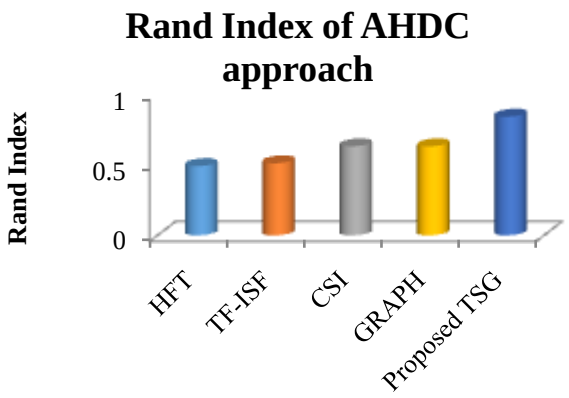


Fig.2 Precision, Recall and F-measures of AHDC Approach

Fig. 3 shows rand index (RI) value of the combination of statistical, graph and proposed TSG methods with AHDC approach.



Keyword extraction approaches

It is observed that the precision values for AHDC with HFT, AHDC with TF-ISF, AHDC with CSI, AHDC with graph based, AHDC with proposed TSG methods are 0.52, 0.62, 0.72, 0.71, 0.92 respectively, the highest precision value is obtained from the AHDC with proposed TSG approach. The recall values for AHDC with HFT, AHDC with TF-ISF, AHDC with CSI, AHDC with graph based, AHDC with proposed TSG methods are 0.51, 0.58, 0.67, 0.64, 0.89 respectively, here also the highest recall value is achieved from the AHDC with proposed TSG method. The F-measure values for AHDC with HFT, AHDC with TF-ISF, AHDC with CSI, AHDC with graph based, AHDC with proposed TSG methods are 0.52, 0.60, 0.69, 0.67, 0.90 respectively. It is found that AHDC with proposed TSG method has the highest F-measure value than other methods. The rand index values for AHDC with HFT, AHDC with TF-ISF, AHDC with CSI, AHDC with graph based, AHDC with proposed TSG methods are 0.49, 0.51, 0.63, 0.63, 0.84 respectively. The highest rand index values of 0.84 is obtained from the AHDC with proposed TSG method. Summarily, it is observed from the experimental results that the combination of proposed TSG keyword extraction method AHDC approach is better than the combination of other methods.

7. CONCLUSION

In this paper, integrated proposed TSG keyword extraction method with AHDC approach has been introduced. This proposed TSG keyword extraction method integrates the impact of statistical, graph based methods with tagging method and similarity measures between the documents were calculated by AHDC approach. Finally, documents were clustered using merging and splitting process of the AHDC approach based on their similarity measures.

REFERENCES

[1] Martin Swobodzinski, Piotr Jankowski, (2015), Evaluating user interaction with a web-based group decision support system: A comparison between two clustering methods, *Decision Support Systems (Elsevier)*, Vol.77, pp.148-157.

[2] Tingting Wei, Yonghe Lu , Huiyou Chang , Qiang Zhou , Xianyu Bao, (2015), A semantic approach for text clustering using WordNet and lexical chains, *Expert System with Applications (Elsevier)*, Vol.42(4), pp. 2264-2275.

[3] Athman Bouguettaya, Qi Yu, Xumin Liu, Xiangmin Zhou, Andy Song, (2015), Efficient agglomerative hierarchical clustering, *Expert Systems with Applications (Elsevier)*, Vol.42(5), pp.2785-2797.

[4] Tao Peng, Lu Liu, (2015), A novel incremental conceptual hierarchical text clustering method using CFu-tree, *Applied Soft Computing (Elsevier)*, Vol. 27, pp. 268-278.

[5] A. Chitra, Anupriya Rajkumar, (2015), Paraphrase Extraction using fuzzy hierarchical clustering, *Applied Soft Computing (Elsevier)*, Vol.34, pp. 426-437.

[6] Anoop Kumar Jain, Satyam Maheshwari, (2013), Phrase based Clustering Scheme of Suffix Tree Document Clustering Model, *International Journal of Computer Applications*, Vol.63(10), pp. 30-37.

[7] Ruichu Cai, Zhenjie Zhang, Anthony K.H. Tung, Chenyun Dai, Zhifeng Hao, (2014), A general framework of hierarchical clustering and its applications, *Information Sciences (Elsevier)*, Vol.272, pp. 29-48.

[8] Zhenguo Chen, Dongmei Zhu, (2011), Hierarchi Clustering Algorithm used for Anomaly Detecting, *Advance in Control Engineering and Information Science-Procedia Engineering (Elsevier)*, Vol.15, pp. 3401-3405.

[9] Thuy-Diem Nguyen, Bertil Schmidt, Chee-Keong Kwoh, (2014), SparseHC : a memory-efficient online hierarchical clustering algorithm, *ICCS 2014-14th Int. Conference on Computational Science, Procedia Computer Science (Elsevier)*, Vol.29, pp. 8-19.

BIOGRAPHIES

R. Nagarajan was born in 1970 in India. He received his M.C.A and M.Phil (Computer Science) from Bharathidasan University, Tiruchirapalli. Presently he is working as a Assistant Professor/ programmer in Annamalai University. His area of research is Data Mining and Document Clustering. He has fifteen years of experience in Application Programming. He is involved in research activities for the past Nine years.

Dr.S.Anu H Nair is an Assistant Professor in the Department of Computer Science and Engineering of Annamalai University. She received her Bachelor degree in 2002, Master Degree from Manonmaniam Sundaranar University in 2005 and the Ph.D degree from Annamalai University in 2016. Her main research areas include image processing, pattern classification techniques and neural networks. She has published ten research papers in International conferences and national conferences. She has published several research papers in International journals and a book chapter.

N. Puviarasan was born in 1963 in India. He received his M.S(Soft Sys) from BITS, Pilani, M.S (Engg) from Annamalai University. Presently he is working as an Associate Professor in the Department of Computer Science and Engineering of Annamalai University. He has published 25 research papers in International journals and conferences and 13 research papers in National journals and conferences. He has twenty seven years of teaching experience and thirteen years of research experience. He has published one book chapter. His area of specialization includes Neural networks and Fuzzy systems, Data mining and Image Processing.

Dr.P.Aruna was born in 1968 in India. She received her B.E. from Madras University, M. Tech from IIT Delhi and the Ph.D degree from Annamalai University. Presently she is working as a Professor in the Department of Computer Science and Engineering of Annamalai University. She has published 80 research papers in International Journals and Conferences and 26 research papers in National Journals and Conferences. She has twenty five years of teaching experience and sixteen years of research experience. She has published 3 book chapters. Her area of specialization includes Neural networks & Fuzzy systems, Data Mining and Image processing. She has guided 6 Ph.D scholars.