Hierarchical Semantic Relational Coverage Measure Based Web Document Clustering Using Semantic Ontology

B.Selvalakshmi, M.Subramaniam

Abstract*: *The problem of web document clustering has been well studied. Web documents has been grouped based various features like textual, topical and semantic features. Number of approaches has been discussed earlier for the clustering of web documents. However the method does not produce promising results towards web document clustering. To overcome this, an efficient hierarchical semantic relational coverage based approach is presented in this paper. The method extracts the features of web document by preprocessing the document. The features extracted have been used to measure the semantic relational coverage measure in different levels. As the documents are grouped in a hierarchical manner, the method estimates the relational coverage measure in each level of the cluster. Based on the semantic relational measure at different level, the method estimates the topical semantic support measure. Using these two, the method computes the class weight. The estimated class weight has been used to perform document clustering. The proposed method improves the performance of document clustering and reduces the false classification ratio.

Index Terms*: *Web Semantics, Semantic Ontology, Clustering, Hierarchical Clustering. SRC, TSS.

I. INTRODUCTION

The growing size of web documents challenges the information retrieval and search engine systems. Every day lakhs and lakhs of web pages has been launched. This increasing phenomenon requires to be indexed by various information management systems and search engines. The modern users spends their most time in the web surfing. They look to learn through the web and for most things they approach the web to search about. The web user submits a query to the search engine and the search engine produces a result to the user. The search result has been produced in form of a web page which has number of hyperlinks represent a web page. Such web search result has numerous irrelevant pages in the result. This is due to the inaccurate clustering of web pages. What the search engine do in the reception of a query is, it search the appearance of the search term in the meta data being maintained. The search engine maintains lot of meta data of different web pages. The search has been performed on the Meta data and based on the occurrence of the search term; the search engine produces result to the user. Similar kind of work is performed in the retrieval of web documents also.

The reason for the irrelevant result of web search and irrelevant document at the retrieval phase is purely due to the inefficient clustering. The web document clustering is performed based on the detection of cluster. The group of any input document Di is performed based on the similarity of the terms present in the document. Number of clustering algorithms has been discussed earlier which uses various methods and measures. The popular KNN clustering algorithm uses the distance measure which is measured based on the terms available in the document and the cluster set. It produces higher overlapping in the result which introduces higher irrelevant results. Similarly, the support vector machine based approach estimates the document support based on the occurrence of the terms of document towards the corpus. However, there exist numerous techniques in web document clustering, they suffer to achieve higher performance.

The reason behind the irrelevant result is the lack of considering the semantic features. The previous algorithms consider only the textual features but misses the semantic features. The semantic feature help the document to represent the meaning of the document. For example, a web document would discuss about a hotel which in turn represent the boarding and lodging. Even though it has not been discussed, the semantic meaning represent the features. Similarly, the topic “Mining” represent semantically the terms grouping, clustering, classification. So to improve the performance of clustering of web documents, it is necessary to consider the semantic features and relations.

The web semantics are generated based on the relationship between the terms of any category. Consider the topic “computer”, it has several semantic meanings like “computing device”, “programming machine”, “processing Unit” and so on. Such like that you can define different meanings to the category. When you have such classes and features of different category, it can be used to perform web document clustering. Similar to the topic “computer”, the topic ‘network” can be generated for its semantic ontology. The relation between them is, the topic “network” would have few terms like “computing device” and “nodes and computers”. These two classes have few relations, but should be considered in proper manner to perform clustering of web documents related to many categories. Towards the scope, this paper present a semantic relational coverage measure based clustering algorithm is presented in this paper. In general the documents of web has been grouped based on certain specific features.

Revised Manuscript Received on August 02, 2019.

B.Selvalakshmi, Assistant Professor, Dept. of CSE, Tagore Engineering College, Chennai.

Dr.M.Subramaniam, Professor , S.A.Engineering College, Chennai – 600077.

There are methods which consider only the terms of the document and there are methods which uses only the meta data of the document to perform clustering. However, they suffer to achieve higher performance as they misses various other features. Identifying the document purely based on the frequency of any term in the document would not help to achieve higher performance. Identifying the category of the document is not essential and enough, it is necessary to identify the exact sub layer of the document. The hierarchical clustering is the process of grouping the web documents in different level. We extend this research to cluster the documents in multiple level where each class has been divided into few other levels and classes. This helps to identify the exact class of the document and helps to produce more informatics and more relevant results. Towards this, the author presented a semantic relational coverage measure (SRCM) and topical semantic support measures in this approach. The detailed approach of estimating such measures has been discussed in detail in the next section.

II. LITERATURE REVIEW

There are number of approaches has been discussed for the problem of web document clustering and information retrieval. Various methods of document clustering has been reviewed in this section. In [1], the author discusses a web search approach based on the semantic features and key words. The method uses NLP tools to extract the features towards estimation of similarity. Similarly, to retrieve documents from large set of document base ontology based document retrieval has been presented in [2]. The method uses the key terms of the document in measuring the similarity measure. An interactive algorithm for document retrieval is presented in [3]. The method uses the knowledge which has been organized in a structured form. The query or questions are received from the user and the system has been designed to produce result. In [4], the author prescribed an semantic network which cluster the nodes of web based on the semantic information. The method performs data retrieval based on the semantic relevancy. The method reduces the retrieval time and retrieval complexity. Semantic ontology has been used to perform web document clustering. Towards the development of document retrieval in web data set, an retrieve ability score based algorithm is presented in [5]. The method overrides the document versioning and removes the redundancy in clustering. The document similarity has been used to identify the redundant documents and improve the performance of web document clustering. In [6], the author neglects the relation towards the characteristic of query and retrieval bias. The correlations between the features are considered towards data retrieval. In [7], different methods of retrieval have been evaluated for their performance under standalone and combined way. Toward corpus retrieval of newspapers, an log based approach is presented in [8]. The retrieve ability measure is computed and evaluated towards performance. Similarly, in [9], a subject based approach for document retrieval is presented. The method combines probabilistic model which uses the key word extraction methods. Extracted keywords are used to retrieve the documents of Iranian papers.

In [10], the jaundice related medical documents has been retrieved with the use of semantic ontology. By maintaining ontology related to jaundice disease, the method extracts keywords from input documents and identifies the related documents from the data base. In [11], the combination of semantic ontology and text features in document retrieval is presented. The method uses, both low level features and top order concepts to improve the retrieval accuracy. In [12], an semantic based data retrieval approach is presented which performs retrieval based on the relations identified. The method generates co-occurrence matrix for input document using semantic ontology. Based on the matrix generated, the method estimates certain measure to perform data retrieval. An uncertain model using semantic ontology has been generated in [13], to support retrieval of spatial data. The method considers the incomplete data and imprecise information to present the ontology. Using these two the method estimates semantic relationship quantitative (SRQ) measure using possibility logic and probability statistic (PP). Based on these values, the data has been retrieved. The concept of semantic based data retrieval has been applied to the biological system in [14]. The method uses different gene ontology which classifies the genes into number of groups .Based on the ontology generated; the method extracts documents from the pool with higher efficiency. Similarly, the sports related document retrieval has been adapted using semantic ontology in [15], which uses the word net to measure the similarity. In [16], the indexing of documents has been performed based on the semantic information. The same has been used to perform document retrieval. All the above discussed methods suffer to produce higher efficient clustering and produce poor results in information retrieval.

III. SEMANTIC RELATIONAL COVERAGE MEASURE CLUSTERING

The proposed hierarchical semantic relational coverage based clustering algorithm preprocesses the web documents initially. The term sets being extracted from the document has been used to estimate the semantic relational coverage measure. The SRC measure has been measured for different sub class of various class or cluster documents. Finally a single topical semantic support (TSS) has been measured. Based on the value of TSS a single class and sub class has been identified for the document to be indexed. The detailed approach is discussed below:

The Figure 1, represent the general block diagram of proposed semantic relational coverage measure based web document clustering algorithm. The detailed stages of the algorithm have been discussed in detail in this section.

A. Preprocessing

The web documents would contain many features like text, images and other presentation tags. The preprocessing is performed to extract the required features and eliminate the unwanted features. First, the web document has been read and entire text content has been extracted. From the text feature extracted, the method removes the presentation tags of HTML. The remaining text feature has been split into single terms to produce term set. The generated term set has been used to compute different measures towards clustering.

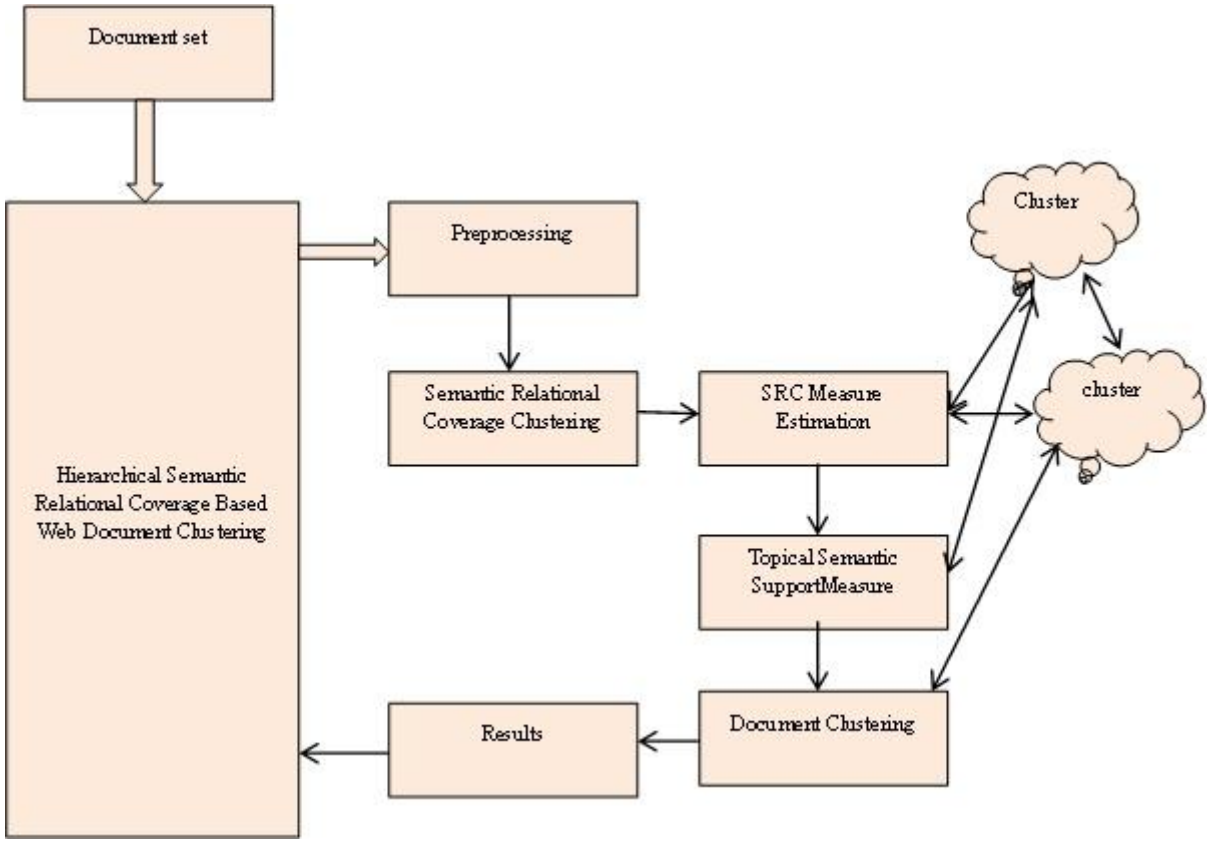

Figure 1: Architecture of proposed semantic relational coverage clustering

B. ALGORITHM

Input: Web Document Set Ws, Stop word set Sws.
Output: Term Set Ts
Start
Read web document Wds, and stop word set sws.
For each document Di
Text Feature Tf = $\int_{i=1}^{size(Wds)} TextFeature \in Di$
Eliminate presentation tags to produce raw text.
Raw text Rt = $\int \sum PresentationTags \cap Tf$
Term set Ts = $\int Split(Rt,'',' .')$
For each term Ti
Eliminate stop word.
If $\int_{i=1}^{size(Ts)} if(Ti \in Sws)$
Ts = $Ts \cap Ti$
End
End
End

The preprocessing algorithm extracts the textual features from the web document and eliminates the presentation tags from the text feature obtained. Finally, the pure terms of the document has been generated to the term set.

IV. SEMANTIC RELATIONAL COVERAGE ESTIMATION

The semantic relational coverage measure represent amount of relation the document covers. The document would speak about any topic but the relevancy of the document towards the category is highly questionable. To perform efficient clustering, the semantic relational coverage measure has been used in this approach. The SRC measure has been estimated based on the number of relations of the category being covered by the document. The method reads the terms set generated and semantic ontology of various classes to estimate the SRC measure.

A. ALGORITHM

Input: Term set Ts, Ontology O
Output: SRC
Start
Read input term set Ts, ontology O.
For each term Ti
Count number of relations in class C
Class Relational count CRC = $\int_{i=1}^{Size(Ts)} \sum Relations(C(O)) \in Ts$
Compute number of relations the term sets has with other class ORC.
Compute ORC = $\int_{i=1}^{Size(Ts)} \sum Relations(OC(O)) \in Ts$
Compute semantic relational coverage measure SRC.
SRC = $\frac{CRC}{\sum Relations \in C(O)} \times \frac{ORC}{\sum relations \in OC(O)}$
End
Stop

The above discussed algorithm computes the relational coverage measure on semantic ontology. The method has been used to perform web document clustering later.

V. TOPICAL SEMANTIC SUPPORT ESTIMATION

The topical semantic support represents the strength of document between the documents of the class. Even though the document have enough semantic relation, it is necessary to consider the bonding of documents within the class. The topical semantic support measure has been computed based on the occurrence of the terms of corpus. First, the method estimates the topical support based on the term set and the taxonomy being used. Second, the topical support on semantic features is estimated. Using these two, the method estimates the topical semantic support measure which has been used to perform clustering later.

A. *Algorithm*

Input: Term set Ts, Ontology O, Taxonomy T.
Output: TSS
Start
Read input term set Ts.
Compute topical support Topsup.
TopSup = $\int_{i=1}^{size(Ts)} \frac{\sum Ts(i) \in Taxonomy(C)}{size(Taxonomy(C))}$
Compute semantic support SemSup.
SemSup = $\int_{i=1}^{size(Ts)} \frac{\sum Ts(i) \in O(C) completely}{size(O(C))} \times \frac{\sum Ts(i) \in O(C) Partially}{Size(O(C))}$
Compute Topical Semantic Support TSS = TopSup×SemSup
Stop

The above discussed algorithm estimates the topical semantic support measure based on the text features and semantic ontology.

VI. SEMANTIC RELATIONAL COVERAGE CLUSTERING

The proposed clustering algorithm reads the input web document set and performs preprocessing to extract the features. From the features extracted as term set, the method estimates the semantic relational coverage measure and topical semantic support. Using these two measures, the method estimates the class weight for each hierarchy of the cluster. Based on the class weight, a single class has been identified to perform clustering of the document.

B. ALGORITHM

Input: Document set Ds, taxonomy T, Ontology O.
Output: Cluster C
Start
Read Ds, T, O.
For each document Di
Term set Ts = preprocessing (Di)
For each class C
For each sub class sc
Compute semantic relational coverage SRC.
Compute topical semantic support TSS.
Compute class weight CW = SRC×TSS
End
End
End
Class C = Choose the class with higher class weight.
Index the document with the class selected.
Stop.

The above discussed algorithm estimates the class weight towards each level of the cluster. Based on the class weight the method identifies the class of the document.

VII. RESULTS AND DISCUSSIONS

The proposed semantic relational coverage measure based clustering method has been tested for its efficiency in different factors. The proposed SRCM algorithm has been developed in Advanced Java. The method has been evaluated for its performance in various parameters. The method has produced the following results. The performance of the method has been validated under varying number of documents and classes.

Table 1: Simulation Details

Parameter	**Value**
Tool Used	Advanced Java, Pos Tagger, Word Net
Number of Classes	100
Number of Documents	1 million
Data Set	Twitter
Algorithm	SRCM

The details of data set being used for the evaluation of proposed semantic relational closure measure (SRCM) based web document clustering algorithm. The details of data set have been presented in Table 1. The methods are evaluated under varying number of class of documents. The performance has been measured in different parameters of clustering.

Table 2: Performance Analysis on Clustering Accuracy

Method	30 Classes	50 Classes	100 Classes
SRQPP	77	72	67
SOR	79	76	73
SIRSD	82	79	77
SCRS	86	92	95
SRCM	91	95	98

The performance of different methods on clustering accuracy at varying number of document class has been measured. The result of analysis has been presented in Table 2. The result recorded pinpoint that the SRCM algorithm achieved higher clustering accuracy compare to existing algorithms.

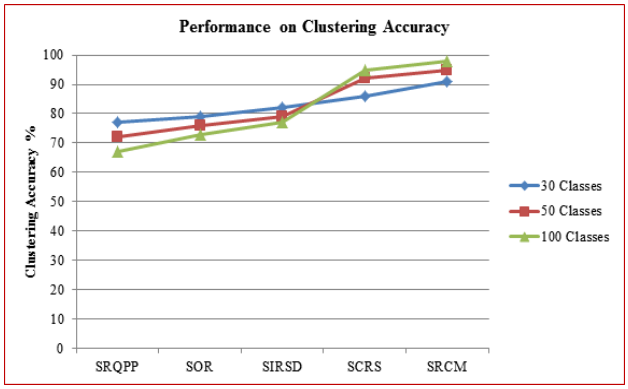

Figure 2: Comparison on clustering performance

The performance on clustering accuracy produced by different methods has been measured and compared with the results of other methods. The proposed SRCM algorithm has achieved higher performance in clustering than other methods.

Table 3: Performance Analysis in False Ratio of Classification

Method	30 Classes	50 Classes	100 Classes
SRQPP	7.7	9.2	11.4
SOR	7.9	8.6	10.2
SIRSD	7.2	5.9	2.3
SCRS	6.6	4.2	1.8
SRCM	4.1	2.5	0.7

The performance analysis on false classification ratio has been measured and presented in Table 3. The proposed SRCM algorithm has produced less false ratio in all the test cases considered.

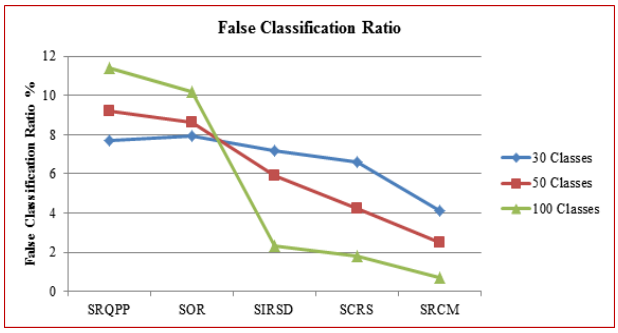

Figure 3: Comparison on false classification ratio

The methods have been evaluated for their performance in false ratio. The proposed SRCM algorithm has produced less false ratio than other methods. The complexity on time produced by different methods on document clustering has been measured and compared. The proposed SRCM algorithm has produced less time complexity in all the test cases considered. The time complexity introduced by different method have been measured and compared with the result of proposed algorithm. The proposed algorithm has achieved less time complexity than other methods.

Table 4: Comparison on Time complexity

Method	30 Classes	50 Classes	100 Classes
SRQPP	23	35	73
SOR	19	32	69
SIRSD	11	21	61
SCRS	6	4	3
SRCM	4	3	2

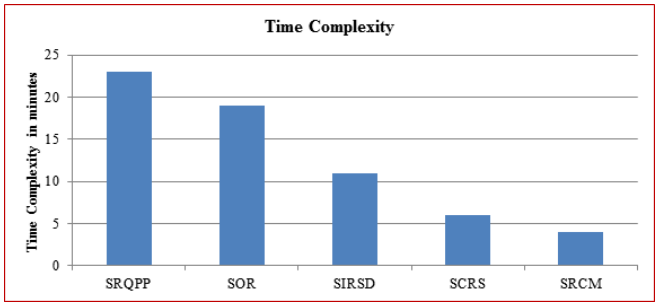

Figure 4: Comparison on time complexity

VIII. CONCLUSION

In this paper, en semantic relational coverage measure based hierarchical web document clustering algorithm is presented. The method preprocesses the document set to extract the textual features. Using the term set generated, the method computes the semantic relational coverage and topical semantic measures. Using these two, the method computes the class weight for each class with each class and sub class. Based on the weight, the method identifies the class of the document. The method introduces higher accuracy in document clustering and reduces the false ratio.

ACKNOWLEDGMENT

This work was done in "***Big Data and Cloud Computing Lab***" at Dept. of Information Technology, S.A. Engineering college. The authors would like to thank, *Department of Science & Technology, Ministry of Science & Technology, Govt. of India, for granting the fund under "Fund for Improvement of S&T Infrastructure in Universities and Higher Educational Institutions (FIST) Program – 2014", Grant Sanction order vide: SR/FST/COLLEGE-239/2014, dated: 21st Nov 2014*, for establishing "Big Data and Cloud Computing Lab" for strengthening the existing institutions S&T infrastructure and support for advancement in research works.

REFERENCES

1. Vajenti Mala ; D. K. Lobiyal, Semantic and keyword based web techniques in information retrieval, Computing, Communication and Automation (ICCCA), 2016
2. Weiguang Fang ; Yu Guo ; Wenhe Liao,Ontology-based indexing method for engineering documents retrieval, (ICKEA),2017.
3. Avani Chandurkar ; Ajay Bansal ,Information Retrieval from a Structured KnowledgeBase, ICSC, 2017

4. Sanjib Kumar Sahu ; D. P. Mahapatra ; R. C. Balabantaray, Analytical study on intelligent information retrieval system using semantic network, Computing, ICCCA, 2016.
5. Thaer SamarMyriam C. TraubJacco van OssenbruggenLynda HardmanArjen P. de Vries, Quantifying retrieval bias in Web archive search, Springer, International Journal on Digital Libraries, Vol 19, Issue 1, pp 57–75,2018.
6. Bashir, S., Rauber, A.: On the relationship between query characteristics and IR functions retrieval bias. J. Am. Soc. Inf. Sci. Technol. **62**(8), 1515–1532 (2011)
7. Klein, M., Nelson, M.L.: Moved but not gone: an evaluation of real-time methods for discovering replacement web pages. Int. J. Digit. Libr. 14(1–2), 17–38 (2014)
8. Traub, M.C., Samar, T., van Ossenbruggen, J., He, J., de Vries, A., Hardman, L.: Querylog-based assessment of retrieve-ability bias in a large newspaper corpus. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, pp. 7–16. ACM (2016)
9. Azadeh Mohebi, Subject-based retrieval of scientific documents, case study: Retrieval of Information Technology scientific articles, Library Review, Vol. 66 Issue: 6/7, pp.549-569.
10. Hany M Harb, Khaled M. Fouad , Nagdy M. Nagdy, Semantic Retrieval Approach for Web Documents, International Journal of Advanced Computer Science and Applications(IJACSA), Volume 2 Issue 9, 2011.
11. Yanti Idaya Aspura M.K., Shahrul Azman Mohd Noah, (2017) "Semantic text-based image retrieval with multi-modality ontology and DBpedia", The Electronic Library, Vol. 35 Issue: 6, pp.1191-1214.
12. Wen Lou, Junping Qiu, (2014) "Semantic information retrieval research based on co-occurrence analysis", Online Information Review, Vol. 38 Issue: 1, pp.4-23.
13. Shengtao Sun, Semantic analysis and retrieval of spatial data based on the uncertain ontology model in Digital Earth, International Journal of Digital Earth Volume 8, 2015 - Issue 1.
14. Mohamed Marouf Z. Oshaiba, Enas M. F. El Houby, and Akram Salah, Semantic Annotation for Biological Information Retrieval System, Hindawi, Advances in Bioinformatics Volume 2015 (2015).
15. M. Uma Devi and G. Meera Gandhi, Wordnet and Ontology Based Query Expansion for Semantic Information Retrieval in Sports Domain, Journal of Computer Science, 2015.
16. Kara, Soner, et al. "An ontology-based retrieval system using semantic indexing." Information Systems 37.4 (2012): 294-305.

AUTHORS PROFILE

B. Selvalakshmi is an Assistant Professor in Tagore Engineering College, Chennai. She received her B.E. Degree in Computer Science and Engineering from Madras University in 1998, M.B.A. Degree from Periyar University, Salem in 2001 and M.E. Degree in Computer Science and Engineering from Anna university, Chennai in 2013. She once worked as Sr. Lecturer in Vinayaga Mission Kirupananda Variyar Engineering College, Salem during 2001 to 2006 and System Analyst in L3 Info Solution during 2007 to 2010. She Joined Tagore Engineering College in 2013. Her research interest include big data, cloud computing and Networking. She has published around 5 academic papers.

M. Subramaniam (1974) is a Professor & Head for the Department of Information Technology at S.A. Engineering College affiliated to Anna University, Chennai, (INDIA). He obtained his Bachelor's degree (B.E) in Computer Science and Engineering from University of Madras (1998), Master degree (M.E) in Software Engineering and Ph.D from College of Engineering Guindy (CEG), Anna University Main Campus, Chennai -25 in the year 2003 and 2013 respectively. His research focuses are Computer & Mobile Networks, Cloud, Big-data and Software Engineering. He is an active life member of the Computer Society of India (CSI) and the Indian Society for Technical Education (ISTE). He has six Research scholars perusing Ph.D under his guidance. He published many research papers in reputed journals. He is also reviewer in IEEE- International Journal of Communication Systems.