Relation Based Mining Model for Enhancing Web Document Clustering

M.Reka[1]
AP/MCA,
K.S.R College of Technology,
Tiruchengode, Namakkal (Dist.),
Tamilnadu,
India,
[1]rekamca@yahoo.com,

Dr.N.Shanthi[2]
Prof &Dean/ CSE
Nandha Engineering College
Erode (Dist.)
Tamilnadu
India
[2]shanthimoorthi@yahoo.com

Abstract

The design of web Information management system becomes more complex one with more time complexity. Information retrieval is a difficult task due to the huge volume of web documents. The way of clustering makes the retrieval easier and less time consuming. Thisalgorithm introducesa web document clustering approach, which use the semantic relation between documents, which reduces the time complexity. It identifies the relations and concepts in a document and also computes the relation score between documents. This algorithm analyses the key concepts from the web documents by preprocessing, stemming, and stop word removal. Identified concepts are used to compute the document relation score and clusterrelation score. The domain ontology is used to compute the document relation score and cluster relation score. Based on the document relation score and cluster relation score, the web document cluster is identified. This algorithm uses 2,00,000 web documents for evaluation and 60 percentas trainingset and 40 percent as testing set.

Keywords

Clustering, Semantic Ontology, Information Retrieval.

I .INTRODUCTION

Due to the rapid growth of Information systems, the web becomes huger and people started using World Wide Web for everything. The increased use of World Wide Web makes the information management as a difficult task. In order to provide exact web information to the user, the information management system must have well organized and efficient techniques for information storage and retrieval. The web page contains many contents like images, video, files etc. The textual content of the web page is only considered to cluster and retrieval.To provide the exact information to the web user, the information system must have an efficient clustering technique. Clustering is a technique, which divides the web pages into groups called clusters so that web pages in each cluster are more analogous to each other than the pages from different clusters. Clustering techniques are used in several application areas such as pattern recognition, data mining, and machine learning, and so on.

II .RELATED WORK

There exist various clustering methodologies like k-means clustering which is a basic clustering algorithm based on the distance between data objects.

Daniel Mullner[2] have proposed Modern hierarchical, agglomerative clustering algorithms, where the input data is given by pair wise dissimilarities between data points and the output is a stepwise dendrogram, a data structure which is shared by all implementations in current standard software.

Zhang Pei-ying and Li Cun-he[4] have introduced a sentence clustering and extraction technique for text summarization.

Odukoyo [1] hasproposed an Improved Data Clustering Algorithm for Mining Web Documents, where the algorithm was simulated using the fuzzy logic and statistical toolbox in Matlab 7.0. The simulated results were compared with the existing data clustering algorithm using accuracy, response time, adjusted rand index and entropy as performance parameters.

Wei liu[5]has presented a Vision-Based Approach for Deep Web Data Extraction, where the Deep Web contents are accessed by queries submitted to Web databases and the returned data records are enwrapped in dynamically generated Web pages.

Ashraf**[6]**haspresented a clustering technique for automatic IE from HTML documents containing semi structured data. Using domain-specific information provided by the user, the proposed system parses and tokenizes the data from an HTML document, partitions it into clusters containing similar elements, and estimates an extraction rule based on the pattern of occurrence of data tokens.

Sun Park, Dong Un an and Choi Im Cheon [3]have proposed that the quality of document clustering with weighted semantic features.

Tekir and Selma [7]have performed clustering tests with the k-means algorithm on the English Wikipedia hyperlinked data set with both traditional cosine distance and this proposed geodesic distance. The effectiveness of our approach is measured by computing micro-precision values of the clusters based on the provided categorical information of each article.

Anjali B. Raut and G. R. Bamnote **[9]** have proposed the document clustering, which was based on fuzzy equivalence relation that helps for information retrieval in the terms of time and relevant information. Their proposed technique for document retrieval on the web, based on fuzzy logic approach improves the relevancy factor. Their technique keeps the related documents in the same cluster so that the searching of documents became more efficient in terms of time complexity.

Hongchen Wu et al...[10]Have proposed improved k-means clustering to establish a framework which makes a modified collaborative filtering recommendation by building a basic model to show case the outline of current existing types of the recommendation platforms.

Azcarraga A., [11]has reported that the rules extracted and transformed into decision trees for performing almost as accurate as the network and to attain the benefit of being in an easily comprehensible format.

Maria Alpuente and Daniel Romero [12] have presented a powerful optimization of their previous Web comparison algorithm that was based on similar memorization techniques and allowed us to compare documents that were extracted from real document collections and databases. They have also re-implemented the former Web comparison tool in Maude in order to improve both the functionality and the performance of the previous version.

Yanjun Li Congnan Luo[8]have proposed a new supervised feature selection method, named CHIR, which is based on the chi^2 statistic and new statistical data that can measure the positive term-category dependency.

Ruilong Yang *et al.* [13] have proposed a Weighted Suffix Tree Document (WSTD) to construct the Web document feature vector for computing the pair wise similarities of documents with weighted phrase. The weighted phrase-based document similarity was applied to the Group-average Hierarchical Agglomerative Clustering (GHAC) algorithm to develop a Web document clustering approach. First, different document parts were assigned different levels of significance as structure weighted. Second, the WSTD was built with sentences and their structure weighted. Third, each internal node and its structure weighted in WSTD model were mapped into a unique feature term in the Vector Space Document (VSD) model; the weighted phrase-based document similarity extended the term TF-IDF weighting scheme in computing the document similarity with weighted phrases. Finally, the GHAC algorithm was employed to generate final clusters.

Anil Kumar Pandey and T. Jaya Lakshmi [14] have presented mutually exclusive Maximal Frequent Item set discovery based K- Means clustering approach. It had been implemented in JAVA. The common text processing approach was to convert the downloaded web documents into vectors. It was done by extracting document features and it generated the document-feature data set. For a set of documents, the feature set was composed of all terms appearing in any one of the documents. They called this a document-feature data set. If document m contained feature n, then the corresponded value, in row n and column m of the table, was set to one. Otherwise, it was set to zero. Then, Apriori algorithm was applied to these document feature data set. The mutually exclusive frequent sets generated by Apriority algorithm were taken as initial points of K-Means algorithm. The output of the K-Means clustering algorithm will be the sets of highly related documents appearing together with same features. The proposed web document clustering method clusters the documents and presents to the researcher only those documents intended for them.

R.Subhashini and V.Jawahar Senthil Kumar[15] suggest that, clustering algorithm handles uncertainty in terms of cluster overlapping, an algorithm can be either crisp (or hard), which considers non-overlapping partitions, or fuzzy (or soft), with which a document can be classified to more than one cluster.

Takazumi Matsumoto and Edward Hung [16] have proposed that, Fuzzy logic is a form of many-valued logic. It deals with reasoning that is approximate rather than fixed and exact. In contrast with traditional logic theory, where binary sets have two-valued logic: true or false, fuzzy logic variables may have a truth value that ranges in degree between 0 and 1. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false.

Anna Huang [17] has compared and analyzed the effectiveness of these measures in partitional clustering for text document datasets. The experiments were done using the standard Kmeans algorithm and the results were reported on seven text document datasets and five distance/similarity measures that have been most commonly used in text clustering.

The clustering techniques explored the problem of time complexity and overlap, so that there is a necessity to develop a new clustering technique with reduced time complexity, reduced overlap and more efficiency.

III.PROPOSED SYSTEM

The proposed system has three phases:

Three different phases are introduced in the proposed system. In the first stage, the contents other than the textual elements of the web pages are removed in order to obtain the original textual features from the web page. In the second stage, the relational features are obtained by identifying the number of relations between the textual features in the web page. In the third stage, the web pages are clustered by computing the cluster relation score which is used to assign the webpage to the group.

A Preprocessing

In the preprocessing phase, the content from the web page is collected and all the presentation tags and images and videos are removed. The textual contents are extracted from the web page. The raw key terms are extracted by applying stemming and stop word removal process.

For each web page w_p from the training set $T_{w,}$ where T_w is the set of all web pages given for training, apply preprocessing which results in raw textual terms.

```
Algorithm: Preprocessing
    Input: Set of web pages for Training $T_w$.
    Output: Set of Key textual terms.
    Step1: Read all web pages given for training $T_{w.}$
    Step 2: For each web page $T_{wi}$ from $T_w$
                W= read content of the web page $T_{wi}$.
                W= remove html tags from w.
                $T_s$ = split w with pattern single space.
                For each term t in the term set $T_s$
                        t=remove “ing” if present.
                        t= remove “ed” if present.
                        If t present in stop word list sw then
                                Remove it from $T_{s.}$
                        End
                End
        End.
        Step 3: Return set of textual term set.
        Step 4: Stop.
```

B Document Relation Score

In this phase, the pre-extracted textual terms of each web page are collected and the document relation score DR_s is calculated for each term of a web page. Document Relation Score (DRS) is the relation between the key terms in the web page. For each term in the term set of the web page, the relation is identified with other terms in the term set of the web page. Some terms may have more number of relations and some may not. If a term doesn’t have relation with any other term, it will be simply ignored. If a term has relation with other terms, the number of relations is counted for future use. The term which has more number of relations with other term gets more importance and the term which has more impact will be sorted at first, when the top few terms are selected from the term set.

The Document Relation Score DR_s is calculated as follows.

$$DRS = \sum DRS_{i(Ts)} / \sum Ts$$

```
Algorithm: DRS Calculation
    Input: Set of textual terms extracted in preprocessing process $T_{s.}$
    Output: Document relation score DRS, top few terms S.
    Step1:   $T_s$ = read (preprocessed term set $T_s$).
    Step 2:  O = read semantic relation ontology owl.
    Step 3:  For each term $T_i$ from term set $T_{s.}$
                 $R_i$ = count number of relation between ($T_{i,}$ Ts)
             End
    Step4:   For each term $T_i$ from term set $T_s$
                 $DRS_{i =} R_i/N$.
                 N- total number of terms in $T_s$.
             End.
    Step5:   Sort $T_s$ based on DRS.
    Step 6:  For i=0: threshold //assign threshold based on accuracy 10
                 If $DRS_i$ > Threshold
                 $S(i) = T_s(i)$;
                 End
             End
    Step 7:  Stop.
```

C Clustering

In the clustering phase, the Cluster Relation Score CRS is calculated for the selected terms of each web page. For each term in the selected term set of each web page, therelation score is evaluated with the concept set of each cluster from the semantic ontology. From each web page, the top few terms are selected according to the document relation score of the web page. The Cluster Relation Score (CRS) is computed with the selected term set as follows.

$$CRS = \sum DRS_{i(STs)}/\sum Tc$$

Where,

Tc is the total number of concepts of each cluster

STs is the selected term set of each web page

For each term from the term set, the relation score is computed with the concept terms of particular cluster. The same will be repeated with the concept terms of the other cluster. The cluster which has high relation score is selected for particular web page based on the calculated relation score.

```
Algorithm: Cluster Relation Score (CRS)
    Input    : Selected term set S, computed DRS values of terms.
    Output   : Identified Cluster
    Step1:   CT = parse owl ();
             CT-concept terms from semantic ontology.
    Step2:   Read S.
             S-selected terms for a web page.
             For each term $T_i$ from term set S.
                 $R_i$ = count number of relation between ($S_{i,}$ $CT_i$)
             End
    Step3:   For each term $T_i$ from term set S
                         $CRS_i = R_i/N$
             N- Total number of concepts in CT.
             End.
    Step4:   For i=0: Number of cluster
                 For j=0: number of terms in S
```

```
                    CRS=CRS+CRS (j);
                End
                CRS (i) = CRS/$T_c$.
                $T_c$- number of concepts in a cluster.
            End.
        Step 5:   Sort the computed CRS value.
        Step 6:   Select the top CRS value and index the web page to the cluster.
        Step 7: End.
```

IV.EXPERIMENTAL SETUP:

The following settings are done to evaluate the proposed algorithm. Various measures were done to evaluate the proposed methodology.

A Dataset and Preprocessing:

Clueweb09 data set was used for the evaluation of our proposed methodology. It contains more than 1,040,809,705 web pages in 10 languages and 428,136,613 pages in English. Table I describes the details about the data set.

Data Set	No. of Pages	Total Sessions	Avg. Session Length
Clueweb09	428,136,613	4,02,10,023	7.8
NASA	1005	118,718	6.4
UOFS	5023	172,984	5.5
UAEU	5195	283,565	4.77

Table I: View of data set used

Preprocessing helped for identifying unique session and session id and removing noisy records. The proposed algorithmsstate the clear steps for preprocessing and session identification using the methods proposed in data preparation for mining World Wide Web browsing patterns.

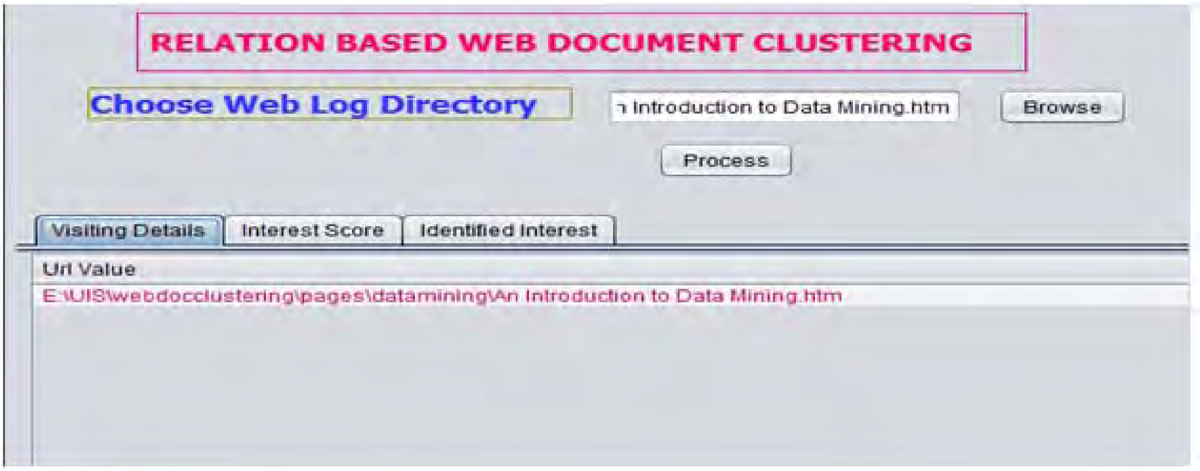
B Results and Discussion

The results show that thisalgorithm reduces the overlap and increases the clustering efficiency. Therefore, it is proved that the proposed algorithm reduces the overall clustering time and time complexity.

Algorithm	Number of Documents	Time in Sec
Online clustering	2,00,000	4200
k-means	2,00,000	2400
Term based	2,00,000	2200
Relation Based	2,00,000	600

Table II: Time Complexity Value in Seconds

The TableII shows the time taken by each algorithm to cluster the number of documents. Hence, it proves that the proposed algorithm takes very less time to generate the cluster.

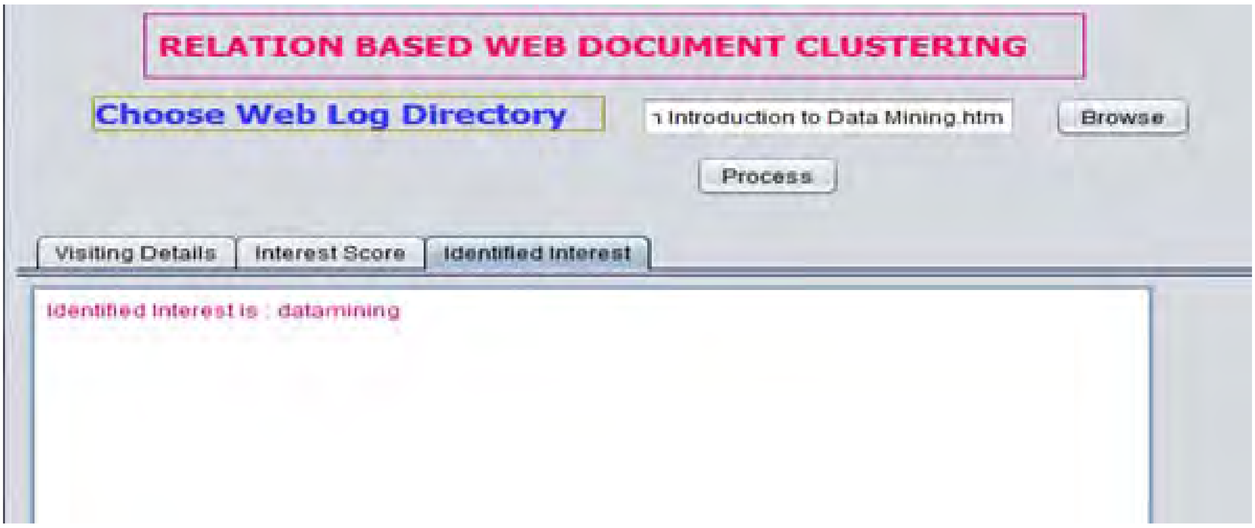


ı Introduction to Data Mining.htm

Url Value

E:\UIS\webdocclustering\pages\datamining\An Introduction to Data Mining.htm

Fig1: Details of pages visited



ı Introduction to Data Mining.htm

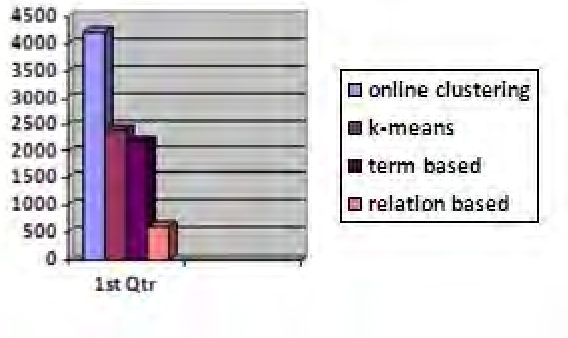
Identified Interest is : datamining

Fig2:Result of Identified Interest

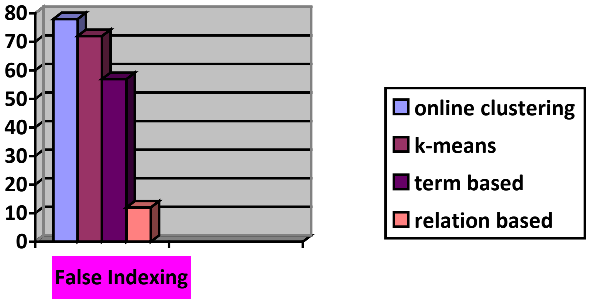
Concept / Interest	Score Generated
Data Mining	42.2
Networking	18.8
Image Processing	0.0
Natural Language Processing	19.2
Software Engineering	22.8

Table III: Result of calculatedInterest Score

Table III shows the result of calculated interest score and the Graph1shows the time taken by various algorithms in generating the cluster.



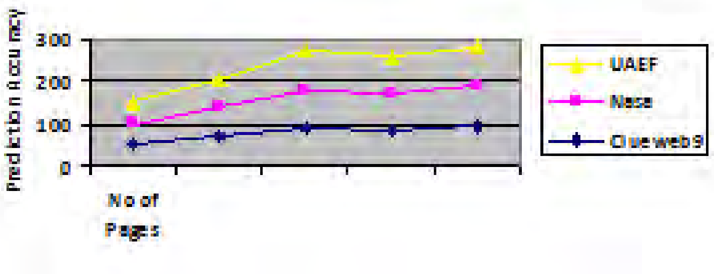
Graph1: Time Consumption by various algorithms for generating the cluster



Graph2: False indexing generated by various algorithms

Algorithm	No. of Pages	Prediction accuracy
Online clustering	10%	42.2
Term based	10%	61.4
K-Means	10%	48.8
Relation based	10%	89.8

Table IV: Comparison of Prediction Accuracy



Graph3: Comparison of Prediction Accuracy

V. CONCLUSION AND FUTURE ENHANCEMENT

The proposed clustering technique is based on the relation between the terms in the web page and the relation between the web pages in the cluster. The document relation score and the cluster relation score have been computed to identify that which cluster the particular document belongs to. The semantic ontology was used to compute both DRS and CRS. The taxonomy values were used to generate the domain ontology. The results show that the proposed algorithm produces good result and reduces the time complexity and false positive indexing. Further, the proposed system can be improved by increasing the number of concepts to cluster to reduce the overlap value. The value of the time complexity was computed for various algorithms. The result of identified interest and the result of calculated interest score were considered in generating the cluster by which the cluster efficiency has been increased. The betterment in the cluster efficiency has been proved against various algorithms through experimental results.This algorithm hasbeen done purely for text documents. The future enhancement can be done by incorporating the image along with text so as to bring out the betterment as web pages contained both text and image future.

REFERENCES

[1] Odukoyo, an Improved Data Clustering Algorithm for Mining Web Documents, IEEE 2010.

[2] Daniel Mullner, “Modern hierarchical, agglomerative clustering algorithms”, 2011

[3] Sun Park, Dong Un An, Choi Im Cheon, “Document Clustering Method using Weighted Semantic Features and Cluster Similarity,” digitel, pp.185-187,2010 Third IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning,2010.

[4] Zhang Pei-ying, Li Cun-he,"Automatic text summarization based on sentences clustering andextraction,"2nd IEEE International Conference on Computer Science and Information Technology, p.167-170, 2009.

[5] Wei Liu, "VIDE: A Vision-Based Approach for Deep Web Data Extraction,"IEEE Transactions on Knowledge and Data Engineering, 2010.

[6] Ashraf "Employing Clustering Techniques for Automatic Information Extraction from HTML Documents, “IEEETransactions on Systems, Man, and Cybernetics, 2008.

[7] Tekir, Selma, “Geodesic distances for web document clustering, Computational Intelligence and Data Mining (CIDM), 2011 IEEE.

[8] Yanjun Li Congnan Luo, “Text Clustering with Feature Selection by Using Statistical Data” IEEE 2008.

[9] Anjali B. Raut and G. R. Bamnote, "Web Document Clustering Using Fuzzy Equivalence Relations" Journal of Emerging Trends in Computing and Information Sciences, Vol.2,pp.22-27, Feb 2011.

[10] Hongchen Wu et al., “Actively building collaborative filtering recommendation in clustered social data”, IEEE16th International Conference on Computer Supported Cooperative Work in Design, 2012.

[11] Azcarraga A., “Keyword extraction using backpropagation neural networks and rule extraction”, IEEE International Joint Conference on Neural Networks (IJCNN), 2012.

[12] Maria Alpuente and Daniel Romero, "A Tool for Computing the Visual Similarity of Web Pages" 10th Annual International Symposium on Applications and the Internet, pp. 45-51, October 2010.

[13] Ruilong Yang, Qingsheng Zhu, and Yunni Xia, "Weighted Suffix Tree Document Model for Web Documents Clustering" Proceedings of the Second International Symposium on Networking and Network Security, pp. 165-169, April 2012.

[14] Anil Kumar Pandey and T. Jaya Lakshmi, "Web Document Clustering for Finding Expertise in Research Area" International Journal of Information Technology, Vol. 1, No. 2, 2009.

[15] R.Subhashini and V.Jawahar, Senthil Kumar, "The Anatomy of Web Search Result Clustering and Search Engines" Indian Journal of Computer Science and Engineering, Vol. 1, No. 4, pp. 392-401, 2010.

[16] Takazumi Matsumoto and Edward Hung, "Fuzzy Clustering and Relevance Ranking of Web Search Results with Differentiating Cluster Label Generation" IEEE International Conference on Fuzzy Systems (FUZZ), pp.1 - 8, September 2010.

[17] Anna Huang, “Similarity Measures for Text Document Clustering” Computer Science Research Student Conference 2008.

ACKNOWLEDGEMENT

Mrs.M.Reka received the B.Sc. degree in Computer Science from Periyar University in 2002, the M.C.A. degree from Periyar University in 2005, the M.Phil. Degree in Computer Science from Periyar University in 2006 and pursuing Ph.D. degree in Anna University from the year 2010.She has been working as an Assistant Professor at K.S.Rangasamy College of Technology, Tiruchengode, since 2008. Her research interests fall in the areas of Data Mining, Networking, Data Communications and ComputerGraphics. She is a life member of ISTE (Indian Society for Technical Education).

Dr.N.Shanthi received the B.E. degree in Computer Science and Engineering from Bharathiyar University, Coimbatore, Tamil Nadu, India in 1994 and the M.E. degree in Computer Science and Engineering from Government College of Technology, Coimbatore, Tamil Nadu, and India in 2001. She has completed the Ph.D. degree in Periyar University, Salem in offline handwritten Tamil Character recognition. She worked as a HOD in department of Information Technology, at K.S.Rangasamy College of Technology, Tamil Nadu, India since 1994 to 2013, and currently working as a Professor Dean in the department of Computer Science and Engineering at Nandha Engineering College Erode. She has published 8 papers in the reputed international journals and 9 papers in the national and international conferences. She has published 2 books. She is supervising 10 research scholars under Anna University,Chennai. She is the Single Point of Contact for KSRCT for Infosys campus connect program. She acts as the reviewer for 4 international journals. Her current research interest includes Document Analysis, Optical Character Recognition, and Pattern Recognition and Network security. She is a life member of ISTE.