TEXT CLUSTERING USING INCREMENTAL FREQUENT PATTERN MINING APPROACH

A.AnandaRao[1], G.SureshReddy[2] and T.V.Rajinikanth[3]

[1]Professor of CSE, JNTU Anantapur, Hyderabad, India
[2]Associate Professor, Information Technology, VNR VJIET, Hyderabad, India
[3]Professor of CSE, SNIST, Hydeabad, India

ABSTRACT

Text mining is an emerging research field evolving from information retrieval area. Clustering and classification are the two approaches in data mining which may also be used to perform text classification and text clustering. The former is supervised while the later is un-supervised. In this paper, our objective is to perform text clustering by defining an improved distance metric to compute the similarity between two text files. We use incremental frequent pattern mining to find frequent items and reduce dimensionality. The improved distance metric may also be used to perform text classification. The distance metric is validated for the worst, average and best case situations [15]. The results show the proposed distance metric outperforms the existing measures.

KEYWORDS

frequent items, text mining, dimensionality reduction

1. INTRODUCTION

Text mining may be defined as the field of research which aims at discovering retrieving the hidden and useful knowledge by carrying out automated analysis of freely available text information and is one of the research fields evolving rapidly from its parent research field information retrieval [1]. Text mining involves various approaches such as extracting text information, identifying and summarizing text, text categorization and clustering. Text Information may be available either in structured form or unstructured form. One of the widely studied data mining algorithms in the text domain is the text clustering.

Text clustering may be viewed as an unsupervised learning approach which essentially aims at grouping all the text files which are of similar nature into one category thus separating dissimilar content in to the other groups. Clustering explores the hidden knowledge thus making it possible to perform statistical analysis [2, 15]. In contrast to the text clustering approach, the process of text classification is a supervised learning technique with the class labels known.

In this paper, we limit our work to text clustering and classification. Clustering is a NP-hard problem. One common challenge for clustering is the curse of dimensionality which makes clustering a complex task. The second challenge for text clustering and classification approaches

is the sparseness of word distribution. The sparseness of features makes the classification or clustering processes in accurate, in efficient and thus becoming complex to judge the result.

The third challenge is deciding the feature size of the dataset. This is because the features which are relevant may be eliminated in the process of noise elimination. Also deciding on the number of clusters possible is also a complex and debatable process.

2. LITERATURE SURVEY

Text mining spans through various areas and has its applications including recommendation systems, tutoring, web mining, healthcare and medical information systems, marketing, predicting, and telecommunications to specify a few among many applications [1].

The authors [2,9], study and propose various criteria for text mining. These criteria may be used to evaluate the effectiveness of text mining techniques used. This makes the user to choose one among the several available text mining techniques. In [3], the authors use the concept of text item pruning and text enhancing and compare the rank of words with the tf-idf method.

Their work also includes studying the importance and extending the use of association rules in the text classification. Association rule mining is playing an important role in text mining and is also widely studied, used and applied by the researchers in text mining community.

In [4] authors discuss the importance of text mining in the predicting and analyzing the market statistics. In short, they perform a systematic survey on the applicability of text mining in market research.

In [5], the authors work towards finding the negative association rules. Earlier in the past decade, the data mining researchers and market analysts were only interested in finding the dominant positive association rules. In the recent years, more research is being carried out towards finding the set of all possible negative association rules.

The major problem with finding negative association rules is the large number of rules which are generated as a result of mining. The negative association rules have important applications in medical data mining, health informatics and predicting the negative behavior of market statistics.
In [6], the authors use the approach of first finding the frequent items and then using these computed frequent items to perform text clustering. They use the method called “maximum capturing”.

With the vast amount of information getting generated in the recent years, many researchers started coming out with the extensive study and defining various data mining algorithms for finding association rules, obtaining frequent items or item sets, retrieving closed frequent patterns, finding sequential patterns of user interest [7].

All these algorithms are not suitable for their use in the field of text mining because of their computational and space complexities. The suitability of these techniques in text mining must be studied in detail and then applied accordingly. One of the important challenges in text mining is handling the problems of mis-interpretation and less frequency. The work of the authors includes proposing two methods namely

1. Pattern deploying and

2. Pattern evolving

which are used to refine the discovered patters for effective text classification. The experimental findings in [7] show their approach is far better than BM25 and SVM based models.

An extensive survey on dimensionality reduction techniques is per-formed in [8]. The authors discuss the method of principal factor analysis, maximum likely hood factor analysis and PCA (principal component analysis).

A fuzzy approach for clustering features and text classification which involves soft and hard clustering approaches is discussed in [12].

An improved similarity measure overcoming the dis-advantages of conventional similarity measures is discussed in [10], their work also involves clustering and classification of text documents.

In [11], the concept of support vector machines, SVM is used for document clustering. Section 2 of this paper discusses the related works performed by various researchers

Section 3 outlines incremental frequent pattern mining algorithm. We use the same algorithm published in the literature [13, 14, and 15] without any modification. The objective here is to find the reduced dimensionality. Section 4 introduces the proposed approach with the algorithm pseudo code. For clustering we use the same process as used in [15].

We discuss a cases study in Section 5 and some preliminary results evaluated. Finally we conclude the paper in Section 6. This paper is extension of the work carried out earlier in [15] presented at ACM COMPSYSTECH 2014.

3. INCREMENTAL ALGORITHM

The common assumption that the frequent item finding algorithms make is that the database does not change. This is hypothetical. In reality, there are many instances where the database keeps changing. How can we find frequent item sets in this scenario?

This forms the basis for incremental approach. This is because as the existing databases changes, the frequent items and association rules corresponding to this database also changes.

So, the previously computed frequent item sets are no more valid and hence must be re-computed and updated w.r.t modified database.

This can be done in 2 ways.

1. Find the frequent items for the whole database once again. This is time consuming and also not efficient. Each time database changes, we must start the process of finding frequent items from start. So, this is not a better approach.

2. Alternately , if we can somehow make use of already computed frequent items and use these frequent items of the old database to find the frequent items of the updated database (old database + newly added database) then it would be better.

The second approach is the foundation for incremental approach. This incremental approach is used to find frequent item sets and perform clustering in this paper. We outline the incremental algorithm for finding frequent item sets.

3.1 Algorithm

Algorithm. Incremental approach using promoted border sets

{

// P pass

// I_i item sets of level i

// Δ_{min} user defined support value

// D_{old} , $D_{current}$ old and present database

P ← 1

i←1

I_i ← set of all item sets of level-i

Scan the database to find support of items in I_i

L_p ← store the frequent-1 item sets in the set denoted by L_p

P ← P+1

while (L_{p-1} is not empty)

{

I_p ← find candidate item sets from L_{p-1}

prune L_p

for each transaction

{

Increment count of each candidate item set in L_p contained in transaction, t

$L_p \leftarrow$ set of all candidate item sets satisfying user defined support

$B_p \leftarrow$ set of all candidate item sets below user defined support

$P \leftarrow P+1$

$L \leftarrow$ union of all the candidate item sets $\geq \Delta min$

$B \leftarrow$ set of all candidate item sets$< \Delta min$

}

// F frequent item sets of whole database

// B Promoted border set

Read newly added set of files and Increment support values of X such that $X \in L_{old} \cup B_{old}$ and place the item sets which support level w.r.t whole database in F and B respectively.

If (there exists border items)

then

$F \leftarrow$ frequent items of whole database

else

Generate candidate sets which are supersets of promoted border set, B

// end of algorithm

}

}

4. PROPOSED METHOD

In this section we first outline the proposed work and then give the pseudo code of the proposed approach. The objective is to perform clustering for a given set of text files taken as input. After selecting the input text files, the pre-processing phase applies stop-word removal which removes unnecessary words that doesn't have any meaning, followed by stemming to find the root/stem of a word.

We, then apply incremental algorithm to find frequent itemsets over the pre-processed data to further reduce the dimensionality of the terms, obtaining a document-by-word matrix. The document-by-word matrix contains data instances, where each data instance represents a binary value (0-1) to denote if the term is present or not in that particular document.

By applying our similarity measure over this matrix, we generate a similarity matrix that denotes the similarity between each document pair. Based on these similarity values, we perform clustering by grouping documents with same similarity value into a single cluster. We outline the pseudo code for the proposed approach below.

4.1 ALGORITHM

Clustering_text_files (input: Text files, output: Clusters)

{

Step 1:

Pre-process the set of input text files to eliminate unnecessary words. This may include elimination of stop words followed by stemming. In addition user may include additional stop words as per the requirement.

Step 2:

Form the feature vector, FV. // Here FV contains all unique words from the input files.

Step 3:

1. Using the FV as columns, transform the content of files in to equivalent matrix representation where rows include text files and columns include words.

2. Generate index for each word for making search process simple and efficient. The cells include frequency of each word in corresponding text files

Step-4:

Form binary matrix from frequency matrix obtained

Step-5:

Apply Incremental frequent pattern mining algorithm and find the final frequent items

Step-6:

These frequent items form the reduced dimensions. This reduced dimension is taken as the input for the clustering phase.

Step-7:

Apply clustering algorithm for the matrix with reduced dimensions.

Step-8:

The result is set of clusters which are similar to a threshold

}

4.2 SIMILARITY MEASURE

For Clustering we use the proposed distance metric for computing document similarity between two text files.

In this section, we explain the proposed distance measure , $\boldsymbol{Doc_{SIM}}$ for finding the document similarity between any two text documents. The Table.1 represents the functional table of the function $f_c < w_{ik}, w_{jk} >$ which maps the particular word combination to one of the values 0, 1, -1. Here w_{ik}, w_{jk} represents k[th] word in documents f_i and f_j

Table 1. Truth Table

combination	w_{ik}	w_{jk}	$f_c < w_{ik}, w_{jk} >$
1	0	0	-1
2	0	1	1
3	1	0	1
4	1	1	0

To define the proposed similarity function, $\boldsymbol{Doc_{SIM}}$, we define a function F_{avg} as given by the equation 1.

The function N is defined as

$$F_{avg} = \frac{\sum_{k=m}^{k=1} F1\,(F_{ik}, F_{jk})}{\sum_{k=m}^{k=1} F2\,(F_{ik}, F_{jk})} \tag{1}$$

where

$$F1 = \begin{cases} -e^{-(\frac{f_c < w_{ik}, w_{jk} >}{\sigma})^2} & ;\ f_c < w_{ik}, w_{jk} > = 1 \\ 0 & ;\ f_c < w_{ik}, w_{jk} > = -1 \\ e^{-(\frac{f_c < w_{ik}, w_{jk} >}{\sigma})^2} & ;\ f_c < w_{ik}, w_{jk} > = 0 \end{cases} \tag{2}$$

$$F2 = \begin{cases} 0 & ; \ f_c < w_{ik}, w_{jk} > = -1 \\ 1 & ; \ f_c < w_{ik}, w_{jk} > \neq -1 \end{cases} \quad (3)$$

The similarity measure is defined as

$$Document\ Similarity, \boldsymbol{Doc_{SIM}} = \frac{(1 + F_{avg})}{(1 + \lambda)} \quad (4)$$

The value of lamda, λ is fixed to 1 to normalize similarity value. $F1$ and $F2$ are distribution factors which indicate the statistical distribution and the features considered to evaluate the similarity measure.

The similarity value lies between 0 and 1. The parameter, F_{avg} gives the normalized distribution effect of all the features over the documents being considered.

The figure.1 below shows the workflow

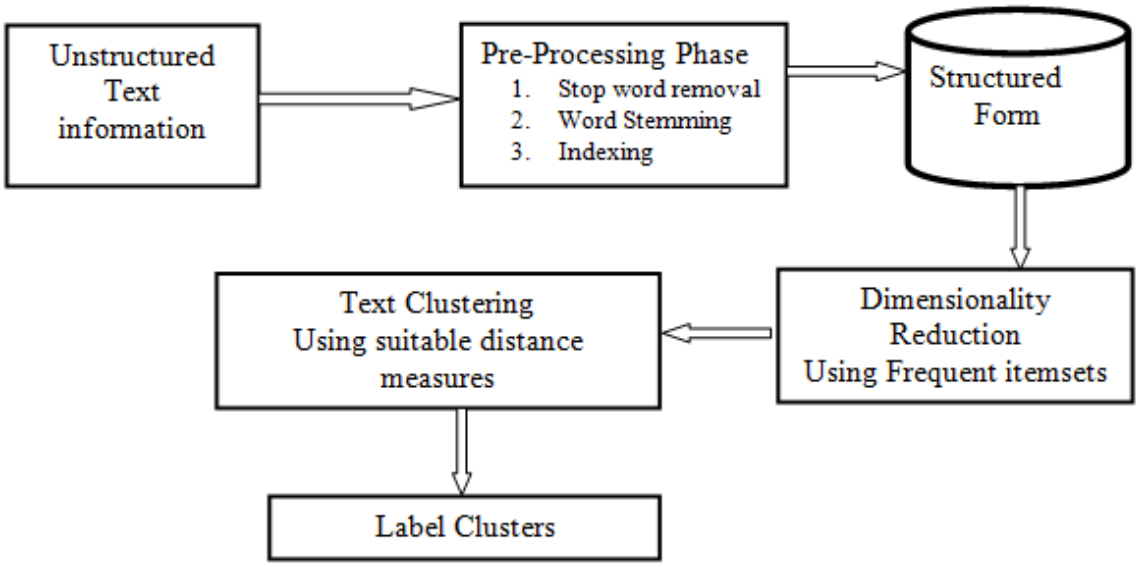

Fig. 1 Workflow of proposed approach

5. CASE STUDY AND RESULTS

Consider the document word matrix shown below in Table.2. Here File-1 to File-9 are text files with feature vector containing 7 features.

Table.2: Matrix in Binary Form

File/Word	W1	W2	W3	W4	W5	W6	W7
File1	1	0	0	1	0	0	0
File2	1	0	0	0	0	0	1
File3	0	0	0	0	0	0	0
File4	1	1	0	0	0	0	1
File5	1	0	0	0	1	1	0
File6	1	1	1	1	0	1	0
File7	1	1	1	1	1	1	1
File8	1	1	1	0	1	0	0
File9	0	1	1	1	1	0	0

The Table.3 shows the similarity matrix. The clusters obtained using the proposed similarity measure are shown in Table.4

Table.3 Similarity Matrix

	F2	F3	F4	F5	F6	F7	F8	F9
F1	0.64	0.45	0.60	0.56	0.65	0.58	0.51	0.52
F2	0	0.48	0.83	0.58	0.53	0.58	0.52	0.42
F3	0	0	0.48	0.43	0.44	0.43	0.40	0.39
F4	0	0	0	0.56	0.62	0.65	0.62	0.50
F5	0	0	0	0	0.59	0.67	0.65	0.53
F6	0	0	0	0	0	0.83	0.71	0.71
F7	0	0	0	0	0	0	0.77	0.78
F8	0	0	0	0	0	0	0	0.78
F9	0	0	0	0	0	0	0	0

Table.4 Clusters Obtained Using Proposed Measure

Clusters	Documents
Cluster-1	6,7
Cluster-2	2,4
Cluster-3	8,9
Cluster-4	1,5
Cluster-5	3

The Silhouette Plot for Proposed Similarity Measure is shown in the Figure.2 below. The distribution of files with in clusters is found even using proposed measure as seen in the plot below.

Also, the silhouette plot has maximum positive values compared to negative values. In this case, we have 8 positive and 1 negative values. The files with in the cluster are not much separated and are evenly distributed.

This may be deduced from clusters 1, 3. Also, Cluster-5 has silhouette value of 1, which means that the document is correctly placed and separated w.r.t other clusters.

This is not true w.r.t silhouette plot obtained in Fig.3 where the distribution of files is not well.

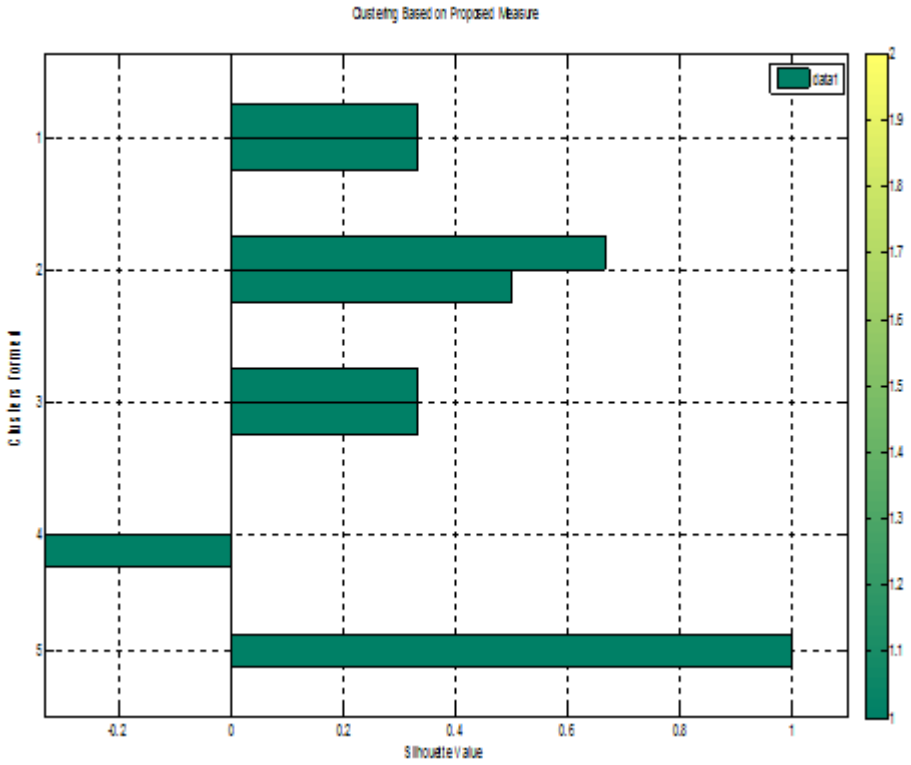

Figure 2: Silhouette Plot for Proposed Similarity Measure

Fig.3 below shows the Silhouette Plot for Euclidean distance measure. Cosine and City block distance measures were found infeasible to obtain the silhouette plots.

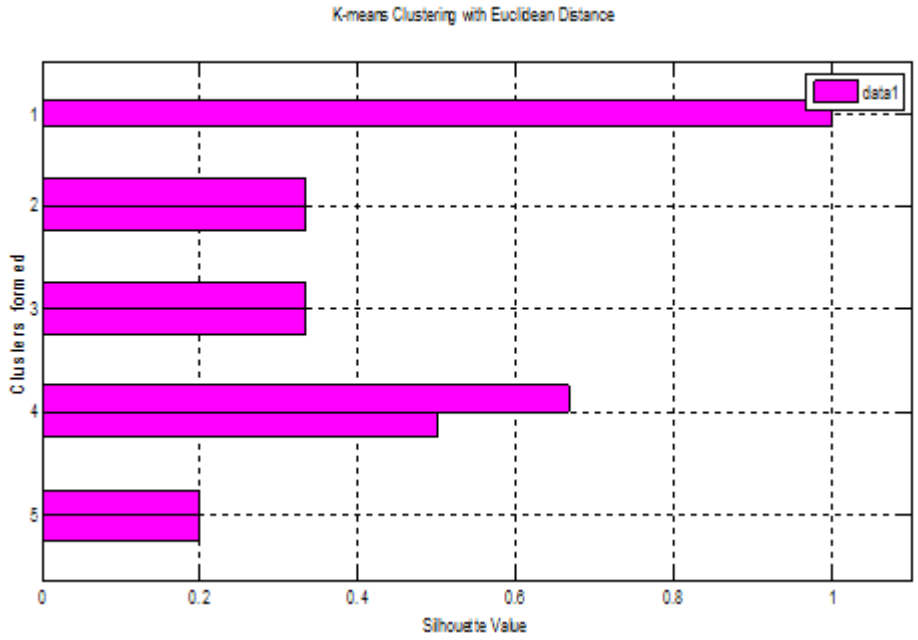

Figure 3: Silhouette Plot for k=5 Using Euclidean distance metric

6. CONCLUSION

Text clustering has been extensively studied in various research areas some of which include bio-informatics, business intelligence, text mining, web mining, and security. In this work, we use the concept of frequent itemsets to perform dimensionality reduction and use this reduced dimensionality to perform clustering. For frequent patterns, we use incremental approach as discussed. To perform text clustering, we make use of the distance metric which is an improved version of our previous measure [15]. The clustering approach in [15] is used to cluster the text files with the similarity matrix replaced by the proposed measure.

REFERENCES

[1] Information and retrieval. Andrew Stranieri, John Zeleznikow. Knowledge Discovery from Legal Databases Law and Philosophy Library Volume 69, 2005, pp 147-169

[2] Hussein Hashimi, Alaaeldin Hafez, Hassan Mathkour: Selection criteria for text mining approaches. Computers in Human Beha-vior.2015

[3] Yannis Haralambous and Philippe Lenca: Text Classification Using Association Rules, Dependency Pruning and Hyperonymization. Proceedings of DMNLP, Workshop at ECML/PKDD, Nancy, France, 2014.

[4] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo: Text mining for market prediction: A systematic review. Expert Systems with Applications 41 (2014) 7653–7670.

[5] Sajid Mahmood: Negative and Positive Association Rules Mining from Text Using Frequent and Infrequent Itemsets. The Scientific World Journal. Volume 2014(2014).

[6] Wen Zhang, Taketoshi Yoshida, Xijin Tang, Qing Wang: Text clustering using frequent itemsets. Knowledge-Based Systems 23 (2010) 379–388.

[7] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu: Effective Pattern Discovery for Text Mining. IEEE Transactions on Knowledge and Data Engineering. Volume 24, No. 1, Jan 2012
[8] Imola K. Fodor : A survey of dimension reduction techniques.
[9] Christopher J. C. Burges: Dimension Reduction: A Guided Tour. Foundations and Trends R in Machine Learning Vol. 2, No. 4 (2009) 275–365.
[10] Yungshen Lin, Jung-Yi Jiang et.al. A similarity measure for text classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 2013.
[11] Sunghae Jun et.al. Document clustering method using dimension reduction and support vector clustering to overcome sparseness, Expert Systems and Applications, 2014, Volume 41, Pages 3204-12
[12] Jung-Yi Jiang et.al A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification, IEEE Transactions on Know-ledge and Data Engineering,Vol.23,No.3, 2011
[13] Freddy Chong Tat Chua: Dimensionality Reduction and Clustering of Text Documents.
[14] Hui Han, Eren Manavoglu, C. Lee Giles, Hongyuan Zha: Rule-based word clustering for text classification. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval.
[15] G.SureshReddy, T.V.Rajinikanth, A.AnandaRao: Design and Analysis of Novel Similarity Measure for Clustering and Classification of High Dimensional Text documents, CompSysTech'2014, Ruse, Bulgaria.

AUTHORS

Dr. Ananda Rao Akepogu received B.Tech degree in Computer Science & Engineering from University of Hyderabad, Andhra Pradesh, India and M.Tech degree in A.I & Robotics from University of Hyderabad, Andhra Pradesh, India. He received PhD degree from Indian Institute of Technology Madras, Chennai, India. He is Professor of Computer Science & Engineering Department and currently working as Director Academic and Planning , of JNTUA College of Engineering, Anantapur, Jawaharlal Nehru Technological University, Andhra Pradesh, India. Dr.Rao published more than 100 publications in various National and International Journals/Conferences. He received Best Research Paper award for the paper titled "An Approach to Test Case Design for Cost Effective Software Testing" in an International Conference on Software Engineering held at Hong Kong, 18-20 March 2009. Received Best Paper Award :"Design and Analysis of Novel Similarity Measure for Clustering and Classification Of High Dimensional Text Documents" in the Proceedings of 15th ACM-International Conference on Computer Systems and Technologies (CompSysTech-2014),pg:1-8,2014, Ruse, Bulgaria, Europe. He also received Best Educationist Award, Bharat Vidya Shiromani Award, Rashtriya Vidya Gaurav Gold Medal Award, Best Computer Teacher Award and Best Teacher Award from the Andhra Pradesh chief minister for the year 2014. His main research interest includes software engineering and data mining.

G.Suresh Reddy received B.Tech Degree in Computer Science & Engineering from Bangalore University, Bangalore, Karnataka , India and M.Tech Degree in IT from Punjabi University, Punjab, India. Persuing Ph.D at JNTUA, Anatapuramu, Andhra Pradesh, India. Working as Associate Professor and Head of Department in Department of Information Technology, VNR Vignana Jyothi Institute Of Engineering and Technology, Hyderabad, Telangana, India..Research areas include Data Mining, Networking. Published several papers in various International Journals/ Conferences. Received Best Paper Award :"Design and Analysis of Novel Similarity Measure for Clustering and Classification Of High Dimensional Text Documents" in the Proceedings of 15th ACM-International Conference on Computer Systems and Technologies (CompSysTech-2014),pg:1-8,2014, Ruse, Bulgaria, Europe.

Dr.T.V.Rajinikanth received M.Tech degree in Computer Science & Engineering from Osmania University Hyderabad, Andhra Pradesh, India and he received PhD degree from Osmania University Hyderabad, Andhra Pradesh, India. He is Professor of Computer Science & Engineering Department, SNIST, Hyderabad, Andhra Pradesh, India. Published more than 50 publications in various National and International Journals/Conferences. Organised and Program Chaired 2 International Conferences,2 grants received from UGC,AICTE . Editorial Board Member for several International Journals. Received Best Paper Award :"Design and Analysis of Novel Similarity Measure for Clustering and Classification Of High Dimensional Text Documents" in the Proceedings of 15th ACM-International Conference on Computer Systems and Technologies (CompSysTech-2014),pg:1-8,2014, Ruse, Bulgaria, Europe. His main research interest includes Image Processing, Data Mining, Machine Learning.