International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-1 Issue-6, August 2012

Comparative Study of K-means and Bisecting k-means Techniques in Wordnet Based Document Clustering

B.S.Vamsi Krishna, P.Satheesh, Suneel Kumar R.

Abstract*— Document clustering plays major role in the fast developing information explosion. It is considered as tool for performing information based operations. Document clustering generates clusters from whole document collection automatically and used in many fields. It is the process of grouping text documents into category groups. It has found applications in various domains in information retrieval and web information systems. Ontology-based computing is considered as a natural evolution of existing technologies to cope with the information onslaught. In current paper, background knowledge derived from Word Net as Ontology is applied during preprocessing of documents for Document Clustering. Document vectors constructed from WordNet Synsets is used as input for clustering. Comparative analysis is done between clustering using k-means and clustering using bi-secting k-means. Results indicate that the bi-secting k-means clustering technique is better than standard k-means clustering technique. These results based on the analysis of specifics of clustering algorithm and nature of document data.

Index Terms*—bisecting k-means, document clustering, standatd k-means, wordnet.

I. INTRODUCTION

Text document clustering is considered as machine learning task taking place in a high dimensional space of word vectors, where each word, i.e. each entry as input of a vector, is treated as a potential attribute for a text. However, has shown that in addition to computational inefficiencies clustering in dimensional spaces is associated with some sort of problem because every data point tends to have the same distance from all other data points.

To overcome from this drawback gather only the words that is very important for each document. But in some cases adding additional words (which may or may not contain in the document) for document vector gives better results when compared to vectors that do not contain additional words. Adding semantic information to each document using external sources like ontologies to document vector may increase the dimensionality, but it gives better clustering results. Efficient usages of this ontological information in preprocessing step are very crucial and plays important step in determining quality of the clusters. The quality of document clustering can be further improved by reducing the noise in the data in the pre-processing stage of data representation and also by applying some new clustering techniques.

The work is to perform document clustering using WordNet (Ontology) derived information better representation in the form of document vectors for documents can be obtained by Using WordNet's information,. The document vectors are considered as input for performing document clustering. The objective of the current work is to study the relevance of bisecting k-means algorithm for document clustering compared to standard k-means algorithm.

II. LITERATURE REVIEW

Initially, document clustering was developed to improve precision and recall of information extraction systems. Now a day's document clustering driven by the ever increasing amount of text documents available in corporate document repositories and on the Internet, the focus has transferred towards providing ways to efficiently browse large collections of documents and to reorganize search results for display in a structured, often hierarchical manner. [4]

Clustering can be broken down into two stages. The first stage has to preprocess the documents, i.e. converting the documents into appropriate needs of data schemes. The second stage has to analyze the available data from first stage and divide it into clusters. This process is carried out by clustering algorithm. [4]. Preprocessing has its importance as the choice of an algorithm, where an algorithm can be as good as the data it works on. At this stage important features are selected and features that do not add much information to the documents are discarded. The importance of incorporating semantic knowledge form external sources like WorldNet ontology can be know by understanding this example. Similarly, the two sentences "kalam has ability to do best" and "kalam has capability to do best" mean the same but are constructed using different synonymous words. Recent works has shown that incorporating ontology in processing step is useful to improve the performance of text clustering in these situations. [9]

Michael Steinbach George Karypis Vipin Kumar [5] in 2001 proposed two main approaches to document clustering techniques are agglomerative hierarchical clustering and standard K-means. Hierarchical clustering is often considered as the better quality clustering technique but it has limitation of its quadratic time complexity.

Manuscript published on 30 August 2012.
***** Correspondence Author (s)
B.S.Vamsi Krishna, Sr. Assistant Professor, CSE, MVGR College of Engineering, Chintalavalasa, Vizianagaram, Andhrapradesh, India.
P. Satheesh, Associate Professor, CSE department, MVGR College of Engineering, Chintalavalasa, Vizianagaram, Andhrapradesh, India.
Suneel Kumar R,CSE Department, MVGR College of Engineering, Chintalavalasa Vizianagaram, Andhrapradesh, India.

In contrast, K-means and its variants (Bisecting k-means) have a time complexity which has linear form in the number of documents, but are thought to produce clusters. Sometimes standard K-means and agglomerative hierarchical techniques are combined so as to get the best out of it. However, results indicate that the bisecting K-means technique is better than the standard K-means approach and hierarchical approach.

Samah Fodeh, Bill Punch and Pang-Ning [8] in 2010 proposed model based on polysemous and synonymous nouns that are present in the document collection. According to them ontology can be used to greatly reduce the number of features needed to do document clustering. Polysemous and synonymous nouns are both relatively prevalent and fundamentally important for document cluster formation. Nouns identification improves document clustering. These nouns may be polysemous and synonymous.

By disambiguating these nouns the quality of cluster can be increased. Thus core subset of semantic features represents entire text corpus. clustering implemented by using core semantic features reduce the number of features by 90% or more and still produce clusters that capture the main themes in a text corpus as showed in results.

Rekha Baghel [6] in 2010 has suggested a novel technique of document clustering based on frequent concepts. The technique FCDC (Frequent Concepts based document clustering) is a clustering algorithm works up on frequent concepts gathered rather than on frequent items used in traditional text mining techniques. Many of the clustering algorithms deal with documents as 'bag of word's representation and ignore the important relationships between words like synonyms. This proposed FCDC algorithm utilizes the semantic relationship between words to create concepts. It utilizes WordNet ontology concept in other way to create low dimensional feature vector which allows us to develop a efficient clustering algorithm.

III. METHODS

A. Word Net Assisted Document Clustering using K-means algorithm (WADC_KA)

The input for this algorithm is set of Documents that need to be clustered. For each document respected vector is generated. After generating vectors for all the documents, semantic based information is added to each vector. Once these preprocessing is finished, these vectors are given as input to k-means algorithm in turn specifying the number of clusters required. Once the algorithm finishes processing, it will give output clusters with documents that belong to that particular cluster. The efficiency of the algorithm can be analyzed using the Root Mean Square Error (RMSE).The following steps has to be carried out in preprocessing for generating vectors:

Preprocessing of Documents

a. Lexical analysis
b. Stopword removal
c. Collecting Nouns and Verbs
d. Stemming
e. Corpus Generation
f. Hypernym density representation
g. Pruning
h. Weighting

B. Document-term matrix

A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of text in documents. In a document-term matrix, rows represents collection of documents and columns represents terms. There are various techniques for getting the value that each entry in the matrix should take. In this matrix, each row represents a document and each column represents the tfidf value of the corresponding synset for that particular document. This matrix is given as input to the Clustering Algorithm. Distance between different document vectors is calculated using Document-Term matrix. Table depicts a document-term matrix for 'n'documents and 'm'synsets.

Tab illustration of document term matrix

File number	Synset1	Synset2	..	..	Synset m
Document1	23	42	25	34	45
Document2	12	25	86	65	75
…	…	…	…	…	…
Document n	12	87	45	95	63(=$tfidf_{ij}$)

C. K-means clustering Algorithm

Data mining has standard *k*-means clustering is a technique of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.

Algorithm

Input:
Document Vectors DV
Number of Clusters 'k'

Output:
'K'Clusters

Initially, the number of clusters should be known, or selected to be K say.

1. The initial step is the randomly choose a set of K instances as centers of the clusters.
2. Next, the algorithm considers each instance and assigns it to the cluster which is closest.
3. The cluster centroids are recalculated.
4. This process is iterated until there is not much change in the cluster centroids.

D. Distance Measure:

Generally for high dimensional data instead of using Euclidean distance using cosine similarity gives better results, because Euclidian distance may intercept all the objects to be of equal distance for high dimensional data.

- Cosine similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them.
- The cosine of 0 is 1, and less than 1 for any other angle.
- The cosine of the angle between two vectors thus determines similarity between the vectors.

Given two vectors of attributes, for example say A and B are considered.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Similarity can be represented using a dot product and magnitude as

$$similarity = \cos\theta = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

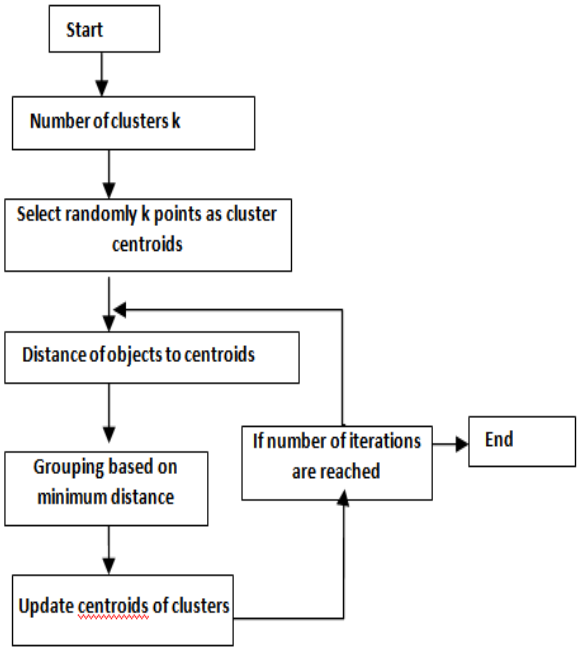

Fig 1 step by step process of vector generation

Initially select randomly k points from the available points as initial seed points. Then assign each object to one of the cluster with nearest seed point from among the k points. Once each point is assigned to the nearest cluster, compute the centroid of the clusters. After computing the centroid of the cluster, repeat the above process again until required number of iterations is reached or the change in the cluster centoid is very small. In WADC_KA k-means algorithm is applied on document-term matrix. Which is been generated by using WordNet. This Wordnet derives synset information. An description of all the stages of Document Clustering can be explained from the below figures

Stage 1: A brief illustration of all the steps have to be taken during preprocessing stage for generating Document Vectors

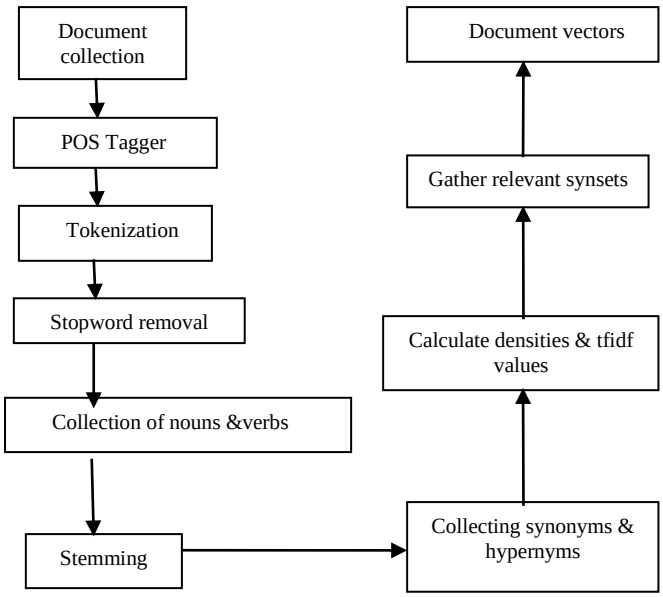

Fig 2 overview of processing stage of both sample vector representation and ontology based representation

Figure explains about all the stages during the pre-processing step i.e., during the first step of clustering. This figure shows how the Document Collection is converted into Documents Vectors using WordNet Ontology as the source of information. Once this stage is completed, Document Vectors are given as input to the clustering Algorithms.

$$rootMSE = \sqrt{\frac{1}{m}\sum_{t=1}^{m}(Y_t - \dot{Y}_t)^2} \quad (2)$$

The formula to calculate RMSE is given in Eq(2), where Y_t is the actual dimension of the document and is the centroid dimension and m is the number of the dimensions.

Stage 2: Performing Document Clustering using obtained Document Vectors

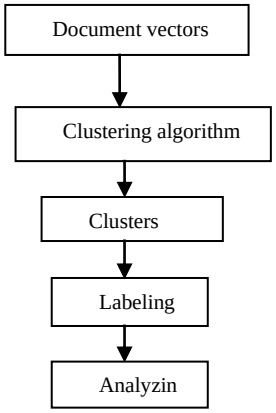

Fig3 Clustering Document Vectors

The Figure explains about the second stage during Document Clustering process. In this stage the clustering algorithm takes Document Vectors that are generated during the first stage, are considered as input to the K- means clustering Algorithm. Once clusters are generated they can be labeled and can be used for further analysis. Once the clusters are formed, labeling the cluster is important. Because once we label the cluster if will give a brief idea about the type of documents that are contained in that particular cluster.

E. Cluster Labeling:

To name the cluster following steps are followed:

a. For each cluster, collect all the synsets along with its term frequency, for all the documents that are present in that particular cluster.
b. Arrange all the synsets according to the descending order of term frequency for the collected synsets.
c. Now label the cluster with the top most synsets. This gives an idea about the type of documents that are contained in the particular cluster.
d. Repeat the same procedure for all the clusters.

For clustering, two measures of cluster goodness or cluster quality are used. first type of measure specify to compare different sets of clusters without reference to external knowledge and is named as an internal quality measure.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The second type of measures specify to lets us calculate how well the clustering is working by comparing the groups produced by the clustering techniques to determined classes. This type of measure is named as an external quality measure. One internal quality measure to check the cluster quality is RMSE (Root Mean Square Error), can be calculated as shown in Equation

F. WordNetAssisted Document clustering using Bisecting K-means algorithm (WADC_BK)

This algorithm resembles to a hierarchical clustering algorithm: in fact, hierarchical clustering algorithms have the advantage of not requiring a priori the number of clusters, since the clusters are bisected at each step. In these algorithms however, the problem is in defining a stopping rule, i.e., deciding if and which clusters have to be still bisected. To this aim, two main approaches are used: the first one applies the simple strategy of bisecting the greatest cluster and the second one is to split the cluster with greatest variance with respect to the centroid of the cluster.

Bisecting K-Means Algorithm

Input:
Document Vectors DV
Number of Clusters 'k'
Number of iterations of k-means ITER
Output:
'K'Clusters

1. Pick a cluster to split (split the largest)
2. Find 2 sub-clusters using the basic K-means algorithm
3. Repeat step 2, the bisecting step is doing for ITER times and takes the split process that results in the clustering with the highest overall similarity
4. Repeat steps 1, 2 and 3 until the desired number of clusters 'k'are reached.

In the above procedure ITER must be sufficiently large so that the change in the cluster centroid from its previous iteration is almost negligible.

IV. RESULTS &DISCUSSION

A. Results of k-means algorithm

Results for Reuter's top10 dataset:

It contains around 200 documents. These 200 documents are taken from 10 different topics. The minimum length of file in the dataset has around 10 words while the maximum length has around 600 words in that document. Total size of the corpus is 13,200 words. The results for Reuters top10 is show in the *Table 4.1*. Here, number of clusters is fixed as 10 and the RMSE is calculated for different combinations of Height of Generalization (h) and Threshold (α).

Tab 4.1 results for Reuter's top 10 dataset

Number of Clusters	Height of Generalization (h)	Threshold (α)	RMSE
10	0	10	90
10	0	13	88
10	2	9	79
10	2	11	76
10	3	13	79
10	4	10	81
10	4	12	81

Synthetic Dataset: It contains around 250 documents manually collected. All these documents are .txt files. The contents of the documents include abstracts of rough sets papers, space research, economics and physical fitness. All these documents are of length from 150 to 600 words. Total words of the corpus are around 18,000 words. The *Table 4.2* shows RMSE for the synthetic dataset when number of clusters is 2 and for varying Height of Generalization (h) and Threshold (α).

Tab 4.2 RMSE results when no. of clusters 2

Number of Clusters	Height of Generalization (h)	Threshold (α)	RMSE
2	0	25	80.2
2	3	25	76
2	4	25	77.2
2	2	20	68.3
2	3	20	70.2

The *Table 4.3* shows RMSE for the synthetic dataset when number of clusters are 3 and for varying Height of Generalization (h) and Threshold (α).

Tab 4.3 RMSE results when no. of clusters 3

Number of Clusters	Height of Generalization (h)	Threshold (α)	RMSE
3	2	20	68.3
3	3	20	70.2
3	4	20	69.6
3	4	25	68.2

The *Table 4.4* shows RMSE for the synthetic dataset when number of clusters is 4 and for varying Height of Generalization (h) and Threshold (α).

Tab 4.4 RMSE results when no. of clusters 4

Number of Clusters	Height of Generalization (h)	Threshold (α)	RMSE
4	3	20	68.4
4	3	25	65.3
4	4	20	69.7
4	4	25	70.6

The *Table 4.5* shows RMSE for the synthetic dataset when number of clusters is 5 and for varying Height of Generalization (h) and Threshold (α).

Tab 4.5 RMSE results when no. of clusters 5

Number of Clusters	Height of Generalization (h)	Threshold (α)	RMSE
5	3	20	70.9
5	4	20	71.6
5	4	25	70.4
5	3	20	70.9

Problems of k-means Algorithm

- When number of features is more, k-means may not give good results.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

- The user needs to specify *k*.
- K-means, the results vary quite a bit from one run to another.
- The algorithm is has problem when dealing with outliers.

1. Outliers are data points that are very far away from other data points.
2. Outliers could be errors in the data recording or some special data points with very different values.

B. Results of bisecting k-means

Reuters top 10 dataset: The results for Reuters top10 is show in the *Table 4.6* Here, number of clusters is fixed as 10 and the RMSE is calculated for different combinations of Height of Generalization (h) and Threshold (α).

Tab 4.6 results for Reuter's top 10 dataset

Number of Clusters	Height of Generalization (h)	Threshold (α)	RMSE
10	0	10	45.9
10	0	13	47.5
10	2	9	42.2
10	2	11	41.6
10	3	13	42.8
10	4	10	43.4
10	4	12	46.8

Synthetic Dataset: The *Table 4.7* shows RMSE for the synthetic dataset when number of clusters are 2 and for varying Height of Generalization (h) and Threshold (α).

Tab 4.7 RMSE results when no. of clusters 2

Number of Clusters	Height of Generalization (h)	Threshold (α)	RMSE
2	0	25	46.2
2	3	25	43.2
2	4	25	41.5
2	2	20	40.6
2	3	20	41.7

The *Table 4.8* shows RMSE for the synthetic dataset when number of clusters are 3 and for varying Height of Generalization (h) and Threshold (α).

Tab 4.8 RMSE results when no. of clusters 3

Number of Clusters	Height of Generalization (h)	Threshold (α)	RMSE
3	2	20	41.8
3	3	20	39.5
3	4	20	38.9
3	4	25	38.4

The *Table 4.9* shows RMSE for the synthetic dataset When number of clusters are 4 and for varying Height of Generalization (h) and Threshold (α).

Tab 4.9 RMSE results when no. of clusters 4

Number of Clusters	Height of Generalization (h)	Threshold (α)	RMSE
4	3	20	37.6
4	3	25	36.4
4	4	20	34.9
4	4	25	34.1

The *Table 4.10* shows RMSE for the synthetic dataset when number of clusters are 5 and for varying Height of Generalization (h) and Threshold (α).

Tab 4.10 RMSE results when no. of clusters 5

Number of Clusters	Height of Generalization (h)	Threshold (α)	RMSE
5	3	20	38.2
5	4	20	39.5
5	4	25	36.2
5	3	20	37.5

Let us consider in detail how the Bisecting K-means is able to reduce the RMSE. Suppose consider the situation for number of clusters is 4, height of generalization (h) is 4 and Threshold (α) is 25.

Tab 4.11 reduce RMSE by bisecting k-means

Iteration	Number of Clusters	RMSE
1	2	68.7
2	3	53.4
3	4	36.6

In the first iteration of Bisecting K-means, the Document Vectors are clustered into two clusters. Its RMSE is 68.4. Then during the second iteration, the cluster with more number of Vectors is considered for further clustering and by the end of second iteration three clusters are obtained with RMSE as 53.4.During the third iteration, four clusters are obtained with RMSE 36.6.Thus we can conclude from the results that Bisecting K-means better clusters with mostly uniform in size.

C. Comparing Results of K-means and Bisecting K-means Algorithms

Let us compare the results of both K-means and Bisecting K-means Algorithms together.

Table4. 12 Comparison of k-means and bisecting k-means

Number of Clusters	Height of Generalization (h)	Threshold (α)	RMSE of K-means Algorithm	RMSE of Bisecting K-means Algorithm
2	2	20	68.3	40.6
3	4	25	68.2	38.4
4	4	25	70.6	34.1
5	4	25	70.4	36.2

From the results of Table 4.12, it is clear that Bisecting K-means is much efficient than K-means Algorithm. This is because K-means may not be efficient for high dimensional dataset.

V. CONCLUSION

The datasets Reuters and Synthetic have given better results when clustered with Bisecting K-means algorithm compared to K-means algorithm. Both these algorithms have O (n) complexity.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Bisecting K-means and K-means has a time complexity which is linear in the number of documents

Bisecting K-means tends to produce clusters of relatively uniform size whereas K-means produce clusters of non-uniform size. For bisecting K-means, there is not much change in the results from one run to another, whereas for regular K-means, the results vary quite a bit from one run to another. If the number of clusters is large, then bisecting K-means is more efficient than the regular K-means algorithm. Hence, Bisecting K-means gives better results for larger data sets. The RMSE values of Bisecting K-means have significant improvement over that of K-means based clustering.

VI. ACKNOWLEDGMENTS

It gives me immense pleasure to express deep sense of gratitude to my guide Mr. B. S. Vamsi Krishna, Sr. Assistant Professor, Department of Computer Science & Engineering, for his Whole-hearted and invaluable guidance throughout the work. Without his sustained and sincere effort, this work would not have taken his shape. He encouraged and helped me to overcome various difficulties that I have faced at various stages of paper.

REFERENCES

[1]. A. Hotho and S. Staab A. Maedche (2001), ―Ontology-based Text Clustering‖, In proceedings of the IJCAI-2001 workshop Text Learning Beyond Supervision
[2]. Andreas Hotho and Alexander Maedche and Steffen Staab (2002), ―Ontology based Text-Document Clustering‖, C, Kunstliche Intelligez, 2002, Vol. 16, No. 4. (2002), pp. 48-54.
[3]. Hai-Tao Zheng, Charles Borchert , Hong-Gee Kim(2003), ―GOClonto: An ontological clustering approach for conceptualizing PubMed abstracts‖, Journal of Biomedical Informatics 43 (2010) 31–40
[4]. Julian Sedding, ―WordNet-based Text Document Clustering‖, Department of Computer Science, University of York Heslington, York YO10 5DD, United Kingdom
[5]. Michael Steinbach, George Karypis and Vipin Kumar(2001), ―A Comparison of Document Clustering Techniques‖, Department of Computer Science and Egineering,University of Minnesota, Technical Report 00-034
[6]. Rekha Baghel and Dr. Renu Dhir (2010), ―A Frequent Concepts Based Document Clustering Algorithm‖, International journal of Computer Applications (0975-8887),Volume 4-No.5,July 2010.
[7]. S C Punitha, K Mugunthadevi and M Punithavalli (2011), ―Impact of Ontology based Approach on Document Clustering‖ International Journal of Computer Applications 22(2):22–26, May 2011. Published by Foundation of Computer Science.
[8]. Samah Fodeh · Bill Punch · Pang-Ning Tan (2011), ―On ontology-driven document clustering using core semantic features‖, Received: 10 December 2009 / Revised: 6 September 2010 / Accepted: 26 November 2010, Springer-Verlag London Limited 2011
[9]. Fellbaum, Christiane (2005), ―WordNet and wordnets‖, In Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670
[10]. Chihli Hung (2004), ―Neural Network Based Document Clustering Using WordNet Ontologies‖, Journal International Journal of Hybrid Intelligent Systems Volume 1 Issue 3,4, December 2004
[11]. Brian S. Everitt, Sabine Landau, and Morven Leese. Cluster Analysis. Oxford University Press, fourth edition, 2001.
[12]. Sam Scott, Stan Matwin(1997), ―Text Classification Using WordNet Hypernyms‖, Computer Science Dept., University of Ottawa, Ottawa, Canada.
[13]. DataMining,Wikipedia, http://en.wikipedia.org/wiki/Data_mining‖.
[14]. DocumentClustering,Wikipedia, http://en.wikipedia.org/wiki/Document_clustering‖.
[15]. Document-termmatrix,Wikipedia, http://en.wikipedia.org/wiki/Document-term_matrix”
[16]. WordSenseDisambiguation,Wikipedia, http://en.wikipedia.org/wiki/Word-ense_disambiguation.
[17]. WordNet, http://wordnet.princeton.edu/.

B.S.Vamsi Krishna received M.Tech in computer Science and engineering in 2008 from Andhra University; he has seven years of teaching experience. He is currently employed as Sr.Assistant Professor in CSE department, MVGR College of Engineering. He has more than eight papers in journals.

P.Satheesh received M.Tech in computer Science and Technology in 2006 from Andhra University; he has ten years of teaching experience. He is currently employed as a an Associate professor in CSE department, MVGR College of Engineering. He has more than ten papers in journals.

Suneel Kumar R currently pursuing M.Tech in computer Science and engineering from Jawaharlal Nehru Technological University Kakinada; he has two years of teaching experience. He has two papers in journals.