International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 3 Issue 3, March-2014, pp: (430-437), Impact Factor: 1.252, Available online at: www.erpublications.com

A new approach for finding appropriate number of Clusters using SVD along with determining Best Initial Centroids for K-means Algorithm

M. Ramakrishna Murthy[1], J.V.R. Murthy[2], P.V.G.D. Prasad Reddy[3], Suresh C. Sapathy[4]
[1]Dept of CSE, GMRIT, Rajam, A.P, India
[2]Dept of CSE, JNTU Kakinada, A.P, India
[3]Dept of CSE, Andhra University, A.P, India
[4]Dept of CSE, ANITS, Vizag, A.P, India

ABSTRACT: Everyday terabytes of data is generated in the real world and most of the mare stored in electronic devices, thus offering great potential for data analysis. Data is not only growing in volume, but also expanding in its varieties like text, commercial or business, medical, images, multimedia from various sources, internet being one among them. Most of these data patterns are complex and unstructured. Data analysis on unstructured data is difficult and inefficient until it is transformed into a proper structure. Clustering, in data analysis, is a vital procedure. It involves division of data objects into meaningful groups using unsupervised learning approach. Each group is called as a cluster which contains similar kind of objects and dissimilar objects in other groups. By clustering, we can identify dense and sparse regions and thereby discover all the distribution patterns and interesting correlations among data attributes. In the clustering literature, one of the most popular and simple clustering algorithms is K-means and is widely used in many applications. K-means has many challenges despite its popularity. In this paper, two significant challenges of K-means algorithm are addressed. The first challenge is toselect K value, which is number of clusters to be given by the user. The second challenge is selection of initial centroids. Methods for the above challenges are proposed and implemented.

Keywords: Clustering, K-means ,singular value decomposition, centroids.

1. INTRODUCTION

The data in the real world is remarkably growing primarily because of computerization, electronic facilities and internet. In the real world, most of the data patterns are unstructured and complex, though available in the digitalized form. The data is growing not only in its volume, but also in data types like text, commercial or business, medical, images, multimedia. Data analysis on unstructured data is difficult and inefficient until it is changed into a proper structure. Data mining is a process of discovering valid, novel, potentially useful knowledge from large data repositories. It also provides capabilities to predict the outcome of future patterns through the methods which are used in the mining process. The main methods in the data mining process are association, classification, clustering and the outlier analysis.

2. CLUSTERING

Data clustering was first used in the title of a1954 article saying about anthropological data [jst, 2009]. Data clustering is also called as Q-analysis, typology, clumping, and taxonomy and can also be used by any other name depending on the field where it is applied [Anil. K. Jain, 2009]. 1960's and 1970's was the period wherein cluster analysis became a major area for research. During this period 'Monograph principles of numerical taxonomy' by [Sokal and Sneath, 1963] was initiated, it has accelerated world-wide research on clustering methods. Historical perspective clustering is routed in mathematics, statistics and numerical analysis. Clustering is an important human activity. Early in childhood, one learns how to distinguish between a cat and a dog or between animals and plants by continuously improving subconscious clustering schemes in the human minds. Cluster analysis has been broadly used in several applications including pattern recognition, data analysis, image processing, and market research.
Fundamentally clustering is a division of data objects into meaningful groups using unsupervised learning approach. Each group is called a cluster which contains similar kind of objects and dissimilar objects in other groups. The basic goal of the clustering achieves high similarity within the cluster, high dissimilarity with other clusters. Clustering enables identification

of dense and sparse regions, and therefore overall distribution patterns and interesting correlations among data attributes can be discovered.

Data clustering is under vigorous development. Contributing areas of research include data mining, statistics, machine learning, spatial database technology, biology, and marketing. Clustering has recently become a highly significant topic in data mining research owing to the huge amounts of data collected in databases and data repositories,. The results of clustering methods helps people to retrieve the required information quickly from large data repositories. The similarity and relationship on data objects can be found thereby providing effective decision making on many applications.

Cluster analysis is a challenging and demanding field. There are typical requirements of clustering[Han]: ability to deal with type of attributes, scalability to large data sets, ability to work with high dimensional data, deal with irregular shapes of data, handling outliers, prior domain knowledge to determine some input parameters, user defined constraints, noisy data, etc.

Clustering algorithms can be categorized mainly into two groups: partitioning and hierarchical. The partitioning method divides the given n number of objects into specific number of groups. Each group represents a cluster; each object belongs to any one of the groups. Each group may be represented by a centroid or cluster representative. The representative object or centroid is a summary description of all the objects contained in a cluster. For example, a representative point where real-valued data is available, the arithmetic mean of the attribute considers representative point of the cluster. In case of non-numeric objects, the representative point or centroid may be determined in other ways. For example, a cluster of documents can be represented by a list of keywords.

K-means is the oldest and most popular algorithm. It is popular because of its simplicity, easier implementation and higher efficiency. But it has some challenges which are addressed in this paper.

3. K-MEANS ALGORITHM

A standard algorithm was first proposed by Stuart Lloyd in the year 1957. The term K-means was first used by James Macqueen in 1967. K-means is partitioning based clustering algorithm, it groups the objects in continuous n-dimensional space, which uses centroid as mean of the group of objects.

Let us take P= {Pi}, i=1,…..n be the set of data points, to be clustered into a set of K number of clusters without any prior knowledge of the input objects. Predefined number of groups is indicated with K, where K is provided as an input parameter. Assigning of each object to a cluster is based on the objects' proximity to the mean of the cluster. Then the mean of the cluster is in turn recomputed and the process of assigning objects to cluster resumes.

The algorithm works in the following manner.

Step 1: Choose randomly K input objects as initial cluster centroids.
Step 2: Each object in the input data set is assigned to a close cluster, based on the similarity or distance between input object and cluster centre.
Step3: Each cluster centre is recomputed as the average of the points in the cluster.
Step 4: Step2 &3 are repeated until the objects do not change from one cluster to another.

A proximity measure is needed to assign data objects to the closest centroid. In the K-means algorithm, many different distance measures are possible, the most widely used one is Euclidean distance. This is defined as

$$d(x,y) = \sqrt{(y1 - x1)^2 + (y2 - x2)^2 + \cdots + (yn - xn)^2} = \sqrt{\sum_{i=1}^{n}(yi - xi)^2} \quad \text{--- (1)}$$

Here x and y are two representative points contain n number of values.
There are other distance measures: Manhattan, Cosine similarity and Jaccard.Cosine and Jaccard measures are appropriated for document clustering. Cosine similarity is mostly appropriate for document similarity. The cosine similarity is measured in the following manner.

$$\text{Cosine}(d1, d2) = \frac{d1.d2}{|d1|.|d2|} = \frac{\sum_{i=1}^{n} d1i \, Xd2\, i}{\sqrt{\sum_{i=1}^{n}(d1)^2} \, X \sqrt{\sum_{i=1}^{n}(d2)^2}} \text{--} \qquad (2)$$

Where d1, d2 are document1 and document2, |d1|, |d2| are the lengths of the documents i.e., $\sqrt{tf1^2 + tf2^2 + \cdots + tfn^2}$. Here tf is the term frequency of the document.

3.1 Challenges of K-means:

K-means is a simple and easy algorithm, but it has manyproblems to be solved. A few of them are:

1. Finding the correct number of clusters i.e., K-value is one of the challenges in this algorithm.
2. The selection of initial centroid data points is very vital to avoid local minima. Basic algorithm selects initial centroid data points randomly.
3. K-means algorithm is found in local minima and might return a bad clusters.

Users specify three important input parameters for K-means algorithm.

i) Number of clusters "K value".
ii) Initial centroids.
iii) Distance metric or proximity measures.

The quality of the K-means clustering is based on the sum of the squared error(SSE). SSE is a measure to calculate the summation of the distance among the centroids and their objects of each cluster.

[Tan,Steinbach,vipin]. Sum of squared error is defined in the following equation:

$$SSE = \sum_{i=1}^{K} \sum_{x \in Ci} dist(ci - x)^2 \text{ --- (3)}$$

Here K is the number of clusters,Ci is the i^{th} cluster, x is a point in Ci is the mean of the i^{th}cluster. The object of the K-means algorithm is to minimize the sum of the squared error.

The above challenges are addressed in this work, with text data.

4. Procedure to Find The Correct Number Of Clusters (K)

Large collections of documents are rapidly increasing day by day. Document clustering has been used intensively because of its wide applicability in areas such as web mining, search engines, information retrieval, and topological analysis. In information retrieval system, document clustering has been used for improving the precision of retrieval results. The objective of document clustering is categorizing or grouping similar documents together and assigning them to the same inherent theme or topic. The first point for applying clustering algorithms to unstructured text data is to create a vector space model or bag-of–words. The process of vector space model construction involves the application of basic preprocessing procedure like elimination ofstop words, stemming words to extract, unique content bearing words or keywords from the set of documents. All these keywords are used to construct vector space model, which is conceptually represented by a vector of keywords extracted from the documents.

Vector space model contains word-by-document matrix, whichhas keyword frequency values as its rows anddocuments as the columns.Thus for large document collections, both row and column dimensions of the matrix are quite large as well as sparse. According to the size of the Vector space model,the most critical problem for text clustering is the high dimensionality. This high dimensionality is a challenge for document clustering.There are methods to reduce the dimensionality like Singular Value Decomposition, principal component analysis, independent principal component analysis etc. Singular Value Decomposition (SVD) method is used to reduce the high-dimensionality of the vector space model or a term-document matrix. In this method,the term-document matrix is decomposed into three smaller matrices for dimensionality reduction Here we proposed Singular Value Decomposition method used for dimensionality reduction along with to find the appropriate number of clusters for K-means algorithm rather than randomly selecting K-value.

5. SINGULAR VALUE DECOMPOSTION (SVD)

Singular Value Decomposition is a mathematical method it uses for dimensionality reduction. The basic idea behind singular value decomposition taking a high dimensional, highly variable set of data points and reducing it to a lower dimension space into that exposes the substructure of the original data more clearly and orders it from most variation to the

least [KirkBaker]. The rank reduced singular value decomposition is performed on the matrix to determine patterns in the relationship between the term and concepts contained in text.
In this method, the document-term matrix$A_{m \times n}$is split into three smaller rank matrices called U, V and S in the following way.

Calculation of U matrix:

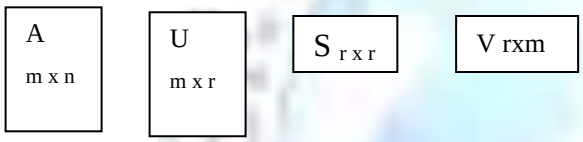
Step1: Calculate $A.A^T$
Step2: Eigen values of AA^T are calculated.
Step3: Eigen vectors for corresponding Eigen values are calculated.
Step4: Those Eigen vectors are placed in a matrix. U is obtained.

Calculation of V matrix:

Step1: Calculate A^TA
Step2: Eigen values of A^TA are calculated.
Step3: Eigen vectors for corresponding Eigen values are calculated.
Step4: Those Eigen vectors are placed in a matrix. V is obtained..

Calculation of S matrix:

Step1: Calculate square root of Eigen values of AA^T or A^TA.
Step2: Place these values as diagonal in the decreasing order. Put the remaining values as zero. S is obtained.



After applying the Singular Value Decomposition method on the term-document matrix A, it splits into three smaller rank matrices USV^T where $U^TU=I$, $V^TV=I$; the columns of U are orthonormal Eigen vectors of AA^T, the columns of V are orthonormal eigenvectors of A^TA, and S is a diagonal matrix containing the square roots of Eigen values from U or V arranged in descending order. Later on, split the term-document matrix A where U is term matrix of size m x m, S is a diagonal matrix of size m x n, and V^T is a transposed matrix of V of size n x n having rows of the document vector.

After applying singular value decomposition, it makes similar items even more similar, and dissimilar items become more dissimilar. Clustering is to be applied after the singular value decomposition on the low ranked matrix; it gives the advantage to reduce the computational burden. Our main proposal is to address the challenges of K-means algorithm. The first challenge as mentioned above is to find the correct number of clusters i.e., K-value based on the following proposal. Diagonal matrix S contains the square roots of the singular values ordered from the highest to the least along its diagonal. These values indicate the variance of the linearly independent components along each dimension. The S matrix is the concept matrix in the SVD decomposition procedure. These normalized number of Eigen values are considered for the K-value (number of clusters) rather than selecting K-value randomly. This proposal is better than the selection of K-value randomly. It is used for both text and image data.

The following example is for easy understood of our proposal.
Case1:

4	0	4	4	0	4
1	0	3	1	0	3
2	0	1	2	0	1
0	4	0	0	4	0
0	2	0	0	2	0
1	0	3	1	0	3
1	0	3	1	0	3
2	0	3	2	0	3
0	0	0	0	0	0

Table1: Term document matrix

12.2856	0	0	0	0	0
0	6.3246	0	0	0	0
0	0	3.0105	0	0	0
0	0	0	0.0000	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

Table 2: Singular values order from greatest to least diagonally.

In this case, the term-document matrix contains six columns i.e., six different documents, but the values of the first three columns are the same as the rest of the three columns. Table 2 shows four singular values of the input matrix.Here this proposal is for clustering the documents by using K-means which uses the singular value as the K value instead of selecting K value randomly.

Case 2:

2	0	8	6	0
1	6	0	1	7
5	0	7	4	0
7	0	8	5	0
0	10	0	0	7

Table 3: term-documents matrix

17.9184	0	0	0	0
0	15.1714	0	0	0
0	0	3.5640	0	0
0	0	0	1.9842	0
0	0	0	0	0.3496

Table 4: Singular Values order from greatest to least

In this case, table 3 shows the term document matrix containing five columns which have values a little variant to the values of case 1. Table 4 shows five singular values but it does not mean that the K value is five. The variations of singular values are normalized and then the K value is selected. In this example the first singular value is 17.9184, the second value is 15.1714, and the variation is 2.74. So they can be considered similar. But the variation between 15.1714 and 3.5640 are 11.6074 which show more variance. Similarly the variance between 3.5640 and 1.9842 is 1.5798which is very less. According to this case as shown in the table 4,the singular values in the first two columnsare related and the next three columns values are related. So when the above term-document matrix (table 3)is clustered using K-means, the chosen K value is 2 rather than random selection of K value.

6. Selection of initial centroids for the best cluster formation

The drawback with the K-means algorithm is that the final formation of clusters and convergence time are completely dependent on the selection of initial centroids. As per the algorithm of K-means, the random selection initial centroids produce non-deterministic results. In general, the algorithm has to be executed several times with all possible different set of initial centroids. Each run produces asquared error value. The least squared error value of all the attempts is considered as initial centroids. But this method is time consuming. Some initial seeds produce non-representative clusters. Three methods have been proposed to select initial centroids that provide the best clusters. The measurement of the best clusters is based on SSE values. The formula to find the SSE value is given in the equation (3). When finding a single cluster, the initial centroid value does not matter because all data points are assigned to the same cluster. If there is more than one cluster, the following three methods are proposed to find initial centroids.

Method1:

The first method is based on the indexes of the given integer data set. The data set contains N number of objects which will be grouped into K number of clusters. In this approach, initial centroids are selected based on the following steps :

Step 1: Finding the M value, which is index value of the given data set, using the following formula:

$$M = \frac{N}{K} \text{---} (4)$$

Step2: The indexes are taken as product M. where I=1, 2…K.

Step 3: The product values in the corresponding index are taken as initialcentroids

Method2:

In the second method, the initial centroids selection is based on the following procedure.

Step1: Finding the highest (H) and smallest (L) value from the data set.

Step2: The following formula is used todetermine the range value (x) which is used to determine the initial centroids.

$$x = \frac{(H-L)}{K} \text{---} (5)$$

Step3:The initial centroids are taken as the mean of the values in the range of L to L+x, L+x to L+2x ….L+(k-1)x to H

Method 3:

In the third method, the initial centroids selection is in the following manner.

Step1: Sorting the dataset either in ascending or descending order.

Step2: Finding the number of objects in each partition (D) using the following formula.

$$D = \frac{N}{K} \text{--} (6)$$

In each partition D number of values are assigned continuously, but in some cases k-1 values ,which do not belong to any of the partitions may be left out.

Step 3: The middle value of the each partition is taken as initial centroid if the number of values in the partition odd: otherwise i.e., even number of value, the average of the two middle value consider an centroid.

The above three proposed methods help in finding the best initial centroids for K-means clustering algorithm to provide best cluster results. The above three proposed methods provide better results than selecting the initial centroid randomly as mentioned in the K-means algorithm.

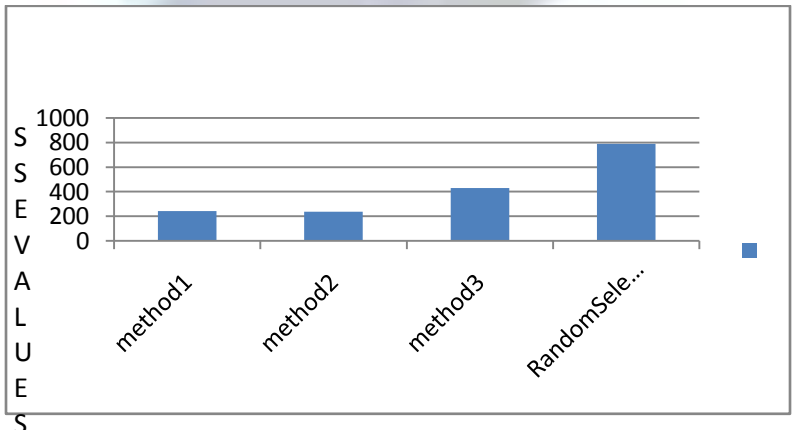


Fig 1: Comparison of three proposed methods with random selection of centroids for 3 numbers of clusters.

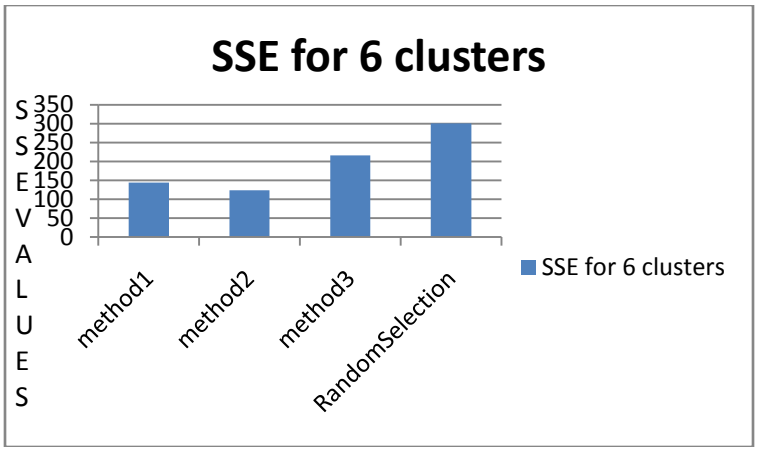


Fig 2: SSE Value for 6 clusters with three proposed methods comparison with random selection method.

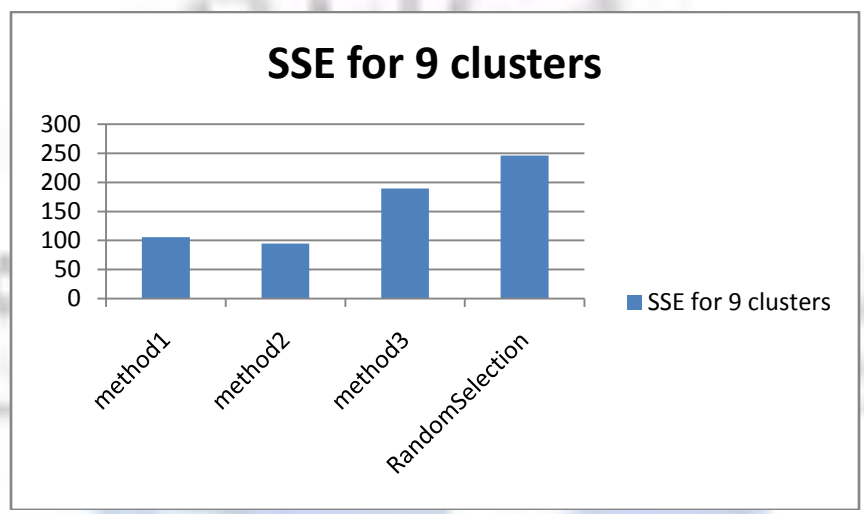


Fig 3: SSE Value for 9 clusters with threeproposed methods comparison with random selection of k value.

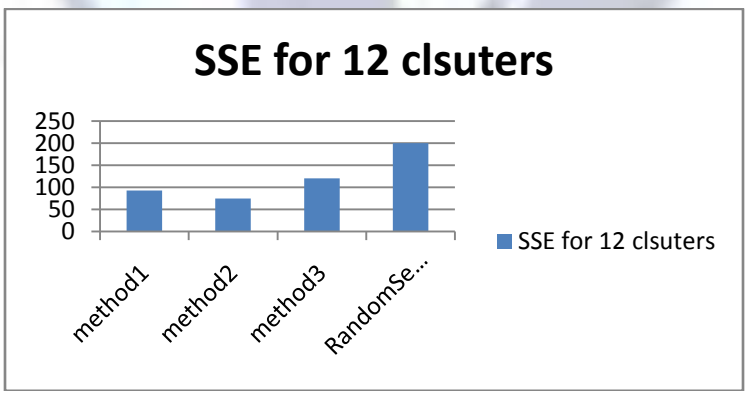


Fig 4: SSE Value for 12 clusters with three proposed methods comparison with random selection of K value.

While analyzing the results of three proposed methods, shown in the above figures provide better results than random selection of initial seeds. In particular the second method provides better results among the three. We also observed if number of clusters is increases, sum of squared error is decreases proportionately.

7. Conclusion

In this paper two important challenges of K-means algorithm have been addressed. One is finding appropriate number of clusters and the second is the selection of initial centroids for the best cluster results. In the first challenge, K value selection is based on dimensionality reduction technique i.e., singular value decomposition. It helps to find the best number of clusters selection based on singular values. The other one is initial seed selection which affects the cluster efficiency. Three methods are proposed to find initial centroids that provide best cluster results. The experimental results show that our proposed methods provide efficient cluster results than normal run of the K-means algorithm.

References

[1]. A.K Jain, 2009. "Data Clustering: 50 years beyond k-means" in pattern recognition letters, Elsevier.
[2]. A.K Jain, M.N.Murty, P.J Flynn "DataClustering: A review, ACM Computing Survey, Vol.31, No.3, 2000.
[3]. P.S Bradley, Usama M.Fayyad "Refining Initial points for K-Means Clustering".
[4]. Congnan Luo, Yanjun Li,Soon M Chung "Text document clustering based on neighbors" in data and knowledge engineering 68,2009,1271-1288.
[5]. Han-Hermann Bock "origins and extensions of the K-means algorithm in cluster analysis" in electronic journal for History of probabilistic and statistics, vol.4 no.2 Dec 2008.
[6]. Han, Kamber, "Data Mining: Concepts and Techniques" Morgan Kaufmann publishers, 2001.
[7]. Jon R.Kettenring "The practice of cluster analysis" in journal of classification, 2006, Vol 23, No 1, page3-30, Springer.
[8]. Kirk Baker " Singular Value Decomposition Tutorial" March 2005, revised 2013.
[9]. P.S.Bredley and U.M Fayyad. Refining Initial Points for K-means clustering. In proc of the 15th intl. Conference on Machine Learning, Madison.WI,July 1998. Morgan Kaufmann Publishers Inc.
[10]. Tapas Kanungo, David M. Mount,Nathan S.Netanyahu, D.Piatko, Ruth Silverman, Angela Y.Wu, "An Efficient K-means clustering algorithm: Analysis and implementation, in IEEE Transactions on pattern analysis and machine intelligence Vol.24, No.7 July 2002.
[11]. Tan,Pang-Ning,Steinbach,Michael, Vipan kumar "Introduction to Data Mining", Pearson Education,2008.
[12]. Yu Luo, Li Yu, Xicng-hua Liu " Improvement Study and Application Based on K-Means Clustering Algorithm " in Fuzzy Information and Engineering Vol 2, Springer 2009.
[13]. Duff, R.G., and Lewis J.(1989). Sparse matrix test problems. ACM Trans Math Soft, page 1-14.
[14]. M.Ramakrishna Murty,JVR Murty,Prasad Reddy"A Dimensionality reduced Text data clustering with prediction of optimal number of clusters" IJARIT Vol.2,Issue 2 Page41-49, 2011.
[15]. Pena J., Lozano J and Larranage P., " An Empirical comparison of four initialization methods for the K-means algorithm", Pattern Recognition Letter 29(1999) 1027-1040.
[16]. M.Ramakrishna Murty,JVR Murty, Prasadreddy " Text Document Classification Based-on Least Square Support Vector Machines with Singular Value Decomposition" International Journal of Computer Applicaions, Aug 2011.