IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-1(1) E-ISSN: 2347-2693

Clustering approach based on Efficient Coverage with Minimum Weight for Document Data

D. S. Rajput, R.S. Thakur, G. S. Thakur*

Department of Computer Application, MANIT, Bhopal (MP)

www.ijcaonline.org

Received: 10/July /2013 Revised: 16/July/2013 Accepted: 16/Aug/2013 Published: 30/Sept/203

***Abstract:* -**At present time huge amount of useful data is available on web for access, and this huge amount of data is shared information which can be used by anyone intended to use. The availability of different types and nature of document data has lead to the task of clustering in large dataset. Clustering is one of the very important techniques used for classification of large dataset and widely applicable many areas. High-quality and fast document clustering algorithms play a significant role to successfully navigate, summarize and organize the information. Recent studies have shown that partitional clustering algorithms are suit- able for large datasets. The k-means algorithm [9, 10] is generally used as partitional clustering algorithm because it can be easily implemented and is most efficient in terms of execution time. The major problem with this algorithm is its sensitivity in selection of the initial partition and its convergence to local optima. In this research study we have refined the useful information from document data set using minimum spanning tree for document clustering and good quality of clusters have been generated on several document datasets, and the output show obtained indicates effective improvement in performance.

Keywords: - Minimum Spanning Tree, Document Clustering, World Wide Web, *K*-Means Algorithm

I. INTRODUCTION

In current scenario data mining is one of the most promising tools in computer science. Currently most of the researchers work is focused on [1, 2, 8] Pattern Reorganization [7], Spatial Data Analysis [11], Image Processing [10], Economic Science [6], Biological Data Analysis [10, 20], WWW etc. Document clustering is one of the best known research problems in the field of data mining. The purpose of document clustering is to divide data point into subsets each of which contains homogeneous objects [7]. There is several formulation of clustering algorithm available in literatures. The huge amount of text documents is major problem because the growth of text document is increasing day by day [7, 8]. The need of effectively manage or explore the results of search engine queries, inspires the study of document clustering. The concept behind the document clustering is to find the hidden similarity and discovery of good clusters.

A spanning tree is a connected, undirected graph is sub graph which is a tree connects all the vertices together [5, 16]. Minimum spanning tree is an approach to solve many problems faced by classical document clustering algorithms. In document data we have different document files are available and N is set of data points. In the concept of the minimum spanning tree we are find the distance between the data [2, 11]. Usually the common properties of data are quantitively evaluated by some optimality measures such as minimum intra cluster distance or maximum inter cluster distance [12,19], Therefore clustering analysis has become an essential and valuable tool in various fields.

Corresponding Author: *D. S. Rajput*

To motivate the specification criteria of MST, in this paper we used the MST concept to generate the clusters for document dataset. Although it is usingthe long edge cutting method to determine clustering in dataset but, it cannot directly determine how many clusters there could be in a dataset.

II. RELATED WORK

There are many document clustering techniques based on distance like *k*- mean[9] , *k*-medoid[10], DBSCAN[10], CURE[10] etc. the major drawback of these approach is the restriction in cluster shape, most of the clustering algorithm have specify some input parameters in advance. Initially Zahn[5] proposed MST based clustering algorithm.

Chang J.et. al.[6] proposed in 2010 "A model of classification based on minimum spanning tree for massive data with Map Reduce implementation". In this paper they present a classification model with tries to find an intermediate model between above two extremes aiming at benefiting from their advantages and removing some drawback.

C. Zahn[5] proposed in 1971 "Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters" it was based on Minimum Spanning Tree (MST),andit is widely used.

Y.Xuet. al.[20]proposed in 2001"Minimum Spanning Treesfor geneexpression dataclustering". In this paper author came into result that MST neither assume that data points are gathered around the centre nor separated by regular geometric curve, thus the shape of the cluster boundary has little impact on the performance of the algorithm. They described objective functions and the corresponding cluster algorithm for computing *k*-partition of spanning tree, where$k$> 0. The algorithm simply removes$k$-1 longest edges so that the weight of the sub treesis minimized. The second objective function is defined to minimize the total distance between the centreand each data point in the cluster. The algorithm removes first k-1 edges from the tree, which creates a k-partitions.

M. Laszlo et. al.[14] proposed in 2005"MST-based tree partitioning algorithms for micro aggregation",and put a constraint on the minimum cluster size rather than the number of cluster. This algorithm was developed for the micro aggregation problem, where the number of clusters in the data set can be figured out by the constraints of the problem itself.

Vathy-Fogarassyet. al. [18] suggests three new pruning criteria for the MST-based clustering. Their goal is to decrease the number of heuristically defined parameters of existing algorithms. This will decrease the influence the user on the clustering results.

In this paper we used minimum spanning tree concept in document dataset and find their clusters. We used Euclidean distance concept to find the distance between two data points. This technique has no specific requirement of prior knowledge of parameter, like number of clusters required and the dimensionality of the data set.

The rest of the paper is organized as follow. In section 3 we define the basic definitions. In section 4pre-processing of datais presented. Further minimum spanning tree based clustering is discussed and in section 5,paper is concluded.

III. BASIC DEFINITIONS

Definition 3.1: Minimum Spanning Tree

Given a connected, undirected graph G = (V, E) , where V is the set of nodes (or vertices), E is the set of edges between pairs of nodes. For a certain pair (V_i, V_j) we may build a direct link between V_i and V_j for a certain cost C(V_i, V_j)>0. Minimum spanning tree play a role one of the most basic formulations. Here suppose we are given a set U of m objects, labelled A_1, A_2, ...$A_m$for each pair $A_i$and $A_j$we have a numerical distance $d(A_i, A_j)$. We require $d(A_i, A_j)=0$ that $d(Ai,Aj)>0$ for district A_i and A_j and that distance are symmetric $d(A_i, A_j)= d(A_j, A_i)$,

A spanning tree of a graph G is a sub graph of G that is a tree and contains all the vertices. The cost of constructing a minimum spanning tree is O (m log n), where m is the number of edges in the graph and n is the number of vertices [16].

Definition 3.2: Euclidean Distance

The dissimilarity (or similarity) between the object described by interval. Scaled variables are typically computed based on distance between each pair of object[3,19]. The most popular distance measure is Euclidean distance. It is the most commonly used due to its simplicity.

Let A, B be two Datapoints $A= (A_1,A_2, A_3...A_n)$, $B= (B_1,B_2,B_3...B_n)$. The Euclideandistance $d_E(A,B)$ in general for an n-dimensional spaceis given by

$$d_E(A, B)=\sqrt{\sum_{k=1}^{n}(A^k - B^k)^2} \quad ....(1)$$

The position of a point in a Euclidean n-space is a Euclidean vector. So, A and B are Euclidean vectors, starting from the origin of the space, and their tips indicate two points. There are following mathematics requirement of a distance functions like if $d_E(A,B)\geq 0$ then distance is a nonnegative and we know that distance is always positive number. And if $d_E(A,B)=0$ then that distance is an object to itself. The befits of choosing Euclidean distance are: It is reportedly faster than most other means of determining correlation and It compares the relationship between actual ratings. This means that the Euclidean distance is a fair measure of how similar ratings are for specific preferences or items.

Definition 3.3: Binary matrix model [BMM], After selection process we have limited terms in each document. Suppose we have n documents The binary matrix M is represented as [8] .

$$M\left[d_i * t_j\right]=\begin{cases}1 & \text{if } t_j \text{is present in } d_i \\ 0 & \text{otherwise}\end{cases} \quad ....(2)$$

Where$i$=1,2,3.............n j= 1,2,3...........m

In binary matrix model each row represents a vector. This means that each document can be represented as a vector. In given model document

$$D_1 \rightarrow [1,\ 1,\ 0,\ 0,\ 1,\ 1,\ 1].$$

Definition 3.4: Document Set

A document set, denoted $D= \{d_1, d_2,\ldots, d_i,\ldots, d_n\}$, also called a document collection, is a set of documents, where n is the total number of documents in D.

Definition 3.5: Term Set

The term set of a document set $D=\{d_1, d_2,\ldots, d_i,\ldots, d_n\}$, denoted $T_D=\{t_1, t_2,\ldots, t_m\}$, is the set of terms appeared in D, where m is the total number of terms.

IV. PROPOSED METHOD

In this section new framework is proposed for document clustering. In this framework consists of three Modules which is shown in figure 1.

1. DocumentPre-processingModule (Remove stop ward, Stemming and Term Selectionin documents)
2. Document Clustering Module (using MST approach)
3. Result phase (visualization of results)

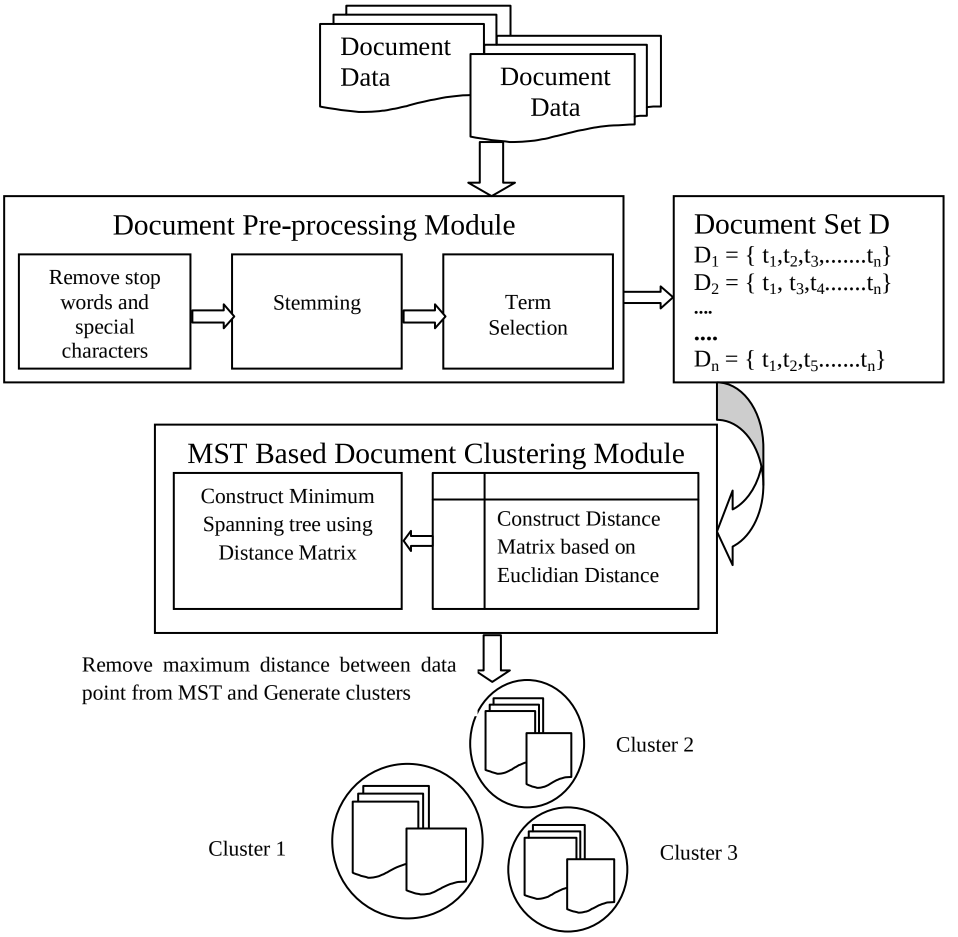

Figure1: Framework for Document Clustering based on MST

Algorithm 1: MST_ Document Clustering
Algorithm to obtain the selected representation of document
Input: 1. A document dataset $D=(D_1, D_2, ...D_n)$,
2. Stop word list,
3. Stemming word list,
4. Threshold value,
5. int*x*.
Output: Number of Clusters $C= \{C_1, C_2......C_n\}$, Mean, Standard Deviation

1. For (all d_i in *D*) do
2. Remove all stop word from document dataset *D*.
3. Apply stemming operation.
4. End For
5. Calculate frequency of each term t_i in all document *D*.
6. For (all d_i in *D*) do
7. For (1 to *j*) do
   a. Count total occurrence of t_{ij} in document d_i
   b. Assign the total occurrence of t_{ij} in *N*
   c. If (*N*<*Threshold*)
   d. Remove t_{ji} from the document d_i
8. End for
9. End for
10. For each $d_i\,\mathcal{E}$ *D*, count the frequency f_{ij} of t_i in d_i. The final representation $d_i=(t_1, f_{ij})(t_2,f_{ij}).....(t_n,f_{ij})$
11. Calculate distance between all data point and create Distance Matrix
12. Construct Minimum spanning tree using distance matrix

13. Compute the mean value Mean= $\frac{1}{n-1}\sum_{1}^{n-1}\sigma(e)$
14. Compute The standard deviation = $\frac{1}{n-1}\sum_{1}^{n-1}(\sigma(e)-Mean)^2$
15. For(i=1 to$x$) do
16. Select the maximum edge
17. Remove from the MST and construct the clusters

Suppose we have data set $D= (D_1, D_2, D_3, D_4, D_5, D_6, D_7)$ First, apply pre-processing technique in raw document dataset [9,10], we find the pre proposed data set. In this paper Table 1 shows the BMM table of any example dataset. Here we have sevendocumentsin dataset than after pre-processing we get seven term set those are present in their documents.

Module 1: Document Pre-processing

The most important procedure in the pre-processing stage of documents is to convert the word forms into meaning combination. The objective of this module is to optimize the efforts of the next phase. Each document dataset have numerous stop words, special marks, punctuation marks and spaces. This process includes various sub processes like stop word elimination, stemming, Term selection etc.

a) Stop Words Elimination:-

First we remove all stop words and special symbols. Stop words are the words which don't have meaning with respect to the classification. So these words are removed when the term matrix is created for the classification purpose. In short the words are removed from the documents which are not necessary for the next stage. Stop words are 'of, 'it', 'the', 'was', 'were' etc., along with all removed prepositions, conjunction and articles from the data set D.

b) Stemming:-

After stop words elimination, the stemming process will be applied. The stemming process is elimination of prefixes and suffixes. The objective is to remove the variation that arises from the amount of different grammatical forms of the similar word. The stemming process helps to decrease the size of the data dictionary file.

c) Feature Term Selection:-

In text classification applications, selection is a critical task for the classifier performance. With increasing number of documents, the number of features also increases. To reduce the size of the dictionary, the threshold term selection method is used. In this method, the upper and lower thresholds are decided according to the number of words in the dictionary. After that the term which exceeds the upper threshold and the terms below lower threshold are extracted from the document. This helps to reduce the size of the dictionary.

The weighting scheme TF-IDF (Term Frequency, Inverse Document Frequency)[9]is used to assign higher weights to distinguish terms in a document, and it is the most widely used weighting scheme which is defined as.

Once text pre-processing is applied over the raw document datasets, it will be converted into form of binary matrix. To convert all documents in the form of binary matrix [7] we have used BMM Tablewhich shown in Table 1.

S. No.	Document set	1	2	3	4	5	6	7
1	D_1	1	1	0	0	1	1	1
2	D_2	0	1	0	1	0	0	1
3	D_3	0	0	0	1	1	0	0
4	D_4	0	1	1	0	0	1	1
5	D_5	0	0	0	0	1	1	0
6	D_6	0	1	1	1	0	0	1
7	D_7	1	1	0	0	0	1	1

Table 1: BMM after Document Preprocessing

Module 2: MST Based Document Clustering Module

After performing the document pre-processing module we find the distance matrix using Euclidean distance concept in table. We find the distance between all data points, and construct the distance matrix. Table 2 show the distance matrix between different data points.

Calculate distance between two data point using Euclidian distance formula which show in equation 1.

Data Points	Distance Calculation using Euclidean	Results
D_1D_2	$\sqrt{(1-0)^2+(1-1)^2+(0-0)^2+(0-1)^2+(1-0)^2+(1-0)^2+(1-1)^2}$	$\sqrt{4}$ =2
D_1D_3	$\sqrt{(1-0)^2+(1-0)^2+(0-0)^2+(0-1)^2+(1-1)^2+(1-0)^2+(1-0)^2}$	$\sqrt{5}$ =2.236
D_1D_4	$\sqrt{(1-0)^2+(1-1)^2+(0-1)^2+(0-0)^2+(1-0)^2+(1-1)^2+(1-1)^2}$	$\sqrt{3}$ =1.732
D_1D_5	$\sqrt{(1-0)^2+(1-0)^2+(0-0)^2++(0-0)^2+(1-1)^2+(1-1)^2+(1-0)^2}$	$\sqrt{3}$ =1.732
D_1D_6	$\sqrt{(1-0)^2+(1-1)^2+(0-1)^2+(0-1)^2+(1-0)^2+(1-0)^2+(1-1)^2}$	$\sqrt{5}$ =2.236
D_1D_7	$\sqrt{(1-1)^2+(1-1)^2+(0-0)^2+(0-0)^2+(1-0)^2+(1-1)^2+(1-1)^2}$	$\sqrt{1}$ =1
D_2D_3	$\sqrt{(0-0)^2+(1-0)^2+(0-0)^2+(1-1)^2+(0-1)^2+(0-0)^2+(1-0)^2}$	$\sqrt{3}$ =1.732
D_2D_4	$\sqrt{(0-0)^2+(1-1)^2+(0-1)^2+(1-0)^2+(0-0)^2+(0-1)^2+(1-1)^2}$	$\sqrt{3}$ =1.732
D_2D_5	$\sqrt{(0-0)^2+(1-0)^2+(0-0)^2+(1-0)^2+(0-1)^2+(0-1)^2+(1-0)^2}$	$\sqrt{5}$ =2.236
D_2D_6	$\sqrt{(0-0)^2+(1-1)^2+(0-1)^2+(1-1)^2+(0-0)^2+(0-0)^2+(1-1)^2}$	$\sqrt{1}$=1
D_2D_7	$\sqrt{(0-1)^2+(1-1)^2+(0-0)^2+(1-0)^2+(0-0)^2+(0-1)^2+(1-1)^2}$	$\sqrt{3}$ =1.732
D_3D_4	$\sqrt{(0-0)^2+(0-1)^2+(0-1)^2+(1-0)^2+(1-0)^2+(0-1)^2+(0-1)^2}$	$\sqrt{6}$ =2.450
D_3D_5	$\sqrt{(0-0)^2+(0-0)^2+(0-0)^2+(1-0)^2+(1-1)^2+(0-1)^2+(0-0)^2}$	$\sqrt{2}$ =1.414
D_3D_6	$\sqrt{(0-0)^2+(0-1)^2+(0-1)^2+(1-1)^2+(1-0)^2+(0-0)^2+(0-1)^2}$	$\sqrt{4}$ =2
D_3D_7	$\sqrt{(0-1)^2+(0-1)^2+(0-0)^2+(1-0)^2+(1-0)^2+(0-1)^2+(0-1)^2}$	$\sqrt{6}$ =2.450
D_4D_5	$\sqrt{(0-0)^2+(1-1)^2+(1-1)^2+(0-1)^2+(0-0)^2+(1-0)^2+(1-1)^2}$	$\sqrt{4}$ =2
D_4D_6	$\sqrt{(0-0)^2+(1-1)^2+(1-1)^2+(0-1)^2+(0-0)^2+(1-0)^2+(1-1)^2}$	$\sqrt{2}$ =1.414
D_4D_7	$\sqrt{(0-1)^2+(1-1)^2+(1-0)^2+(0-0)^2+(0-0)^2+(1-1)^2+(1-1)^2}$	$\sqrt{2}$ =1.414
D_5D_6	$\sqrt{(0-0)^2+(0-1)^2+(0-1)^2+(0-1)^2+(1-0)^2+(1-0)^2+(0-1)^2}$	$\sqrt{6}$ =2.450
D_5D_7	$\sqrt{(0-1)^2+(0-1)^2+(0-0)^2+(0-0)^2+(1-0)^2+(1-1)^2+(0-1)^2}$	$\sqrt{4}$ =2
D_6D_7	$\sqrt{(0-1)^2+(1-1)^2+(1-0)^2+(1-0)^2+(0-0)^2+(0-1)^2+(1-1)^2}$	$\sqrt{4}$ =2

Table 2: Distance Calculation between all Data Points

To find the all possible distance between the documents we construct the distance matrix. Table 3 shows the distance matrix based on Euclidean distance.

D_1	0						
D_2	2	0					
D_3	2.236	1.732	0				
D_4	1.732	1.732	2	0			
D_5	1.732	2.236	1.414	2	0		
D_6	2.236	1	2	1.414	2.450	0	
D_7	1	1.732	2.450	1.414	2	2	0
	D_1	D_2	D_3	D_4	D_5	D_6	D_7

Table 3: Distance Matrix based on Euclidean Distance

Figure 2(a) shows the document representation in *n* dimensional space. Every document shown a data points. Now apply the algorithm and construct minimum spanning tree based on all properties of MST. Figure 2(b) show the minimum spanning tree of data set which we have considered. Now we apply the cut property in MST and divide the objects *U* into *k* groups. For a given parameter *k*.

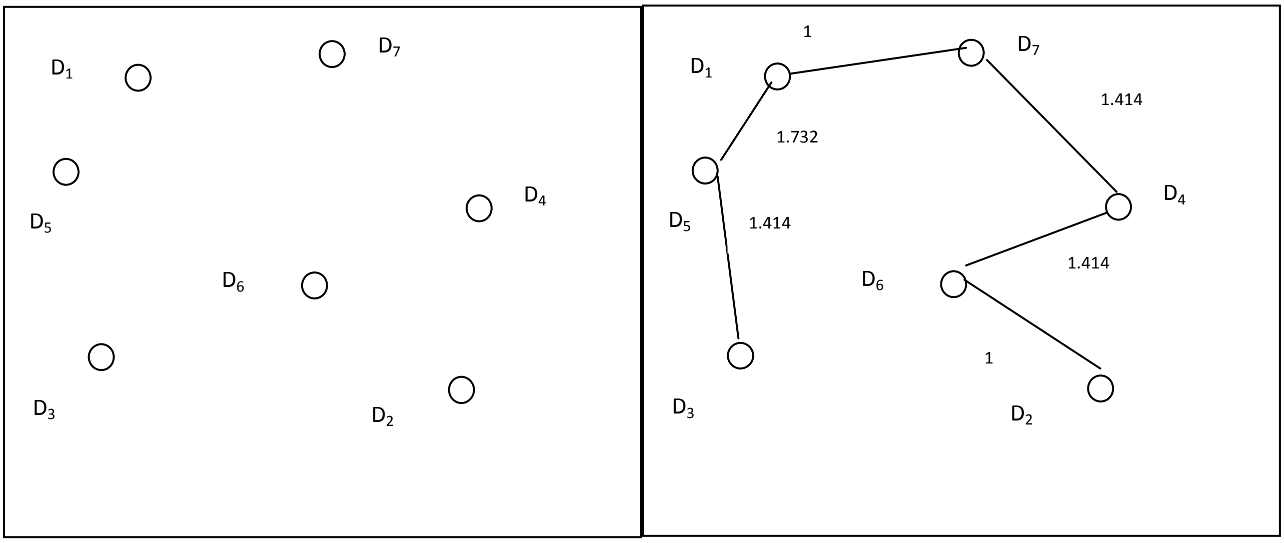

Figure 2 (a): Data Point in *n*-dimensional space (b) Minimum Spanning Tree based on Distance Matrix

In this MST the maximum distance of two data point is $d_E(D_1,D_5)$=1.732. Now apply cut property and divide the MST into two parts. Now we have two tree available. In the first we have (D_3, D_5) and other is (D_2,D_3, D_4,D_6,D_7).

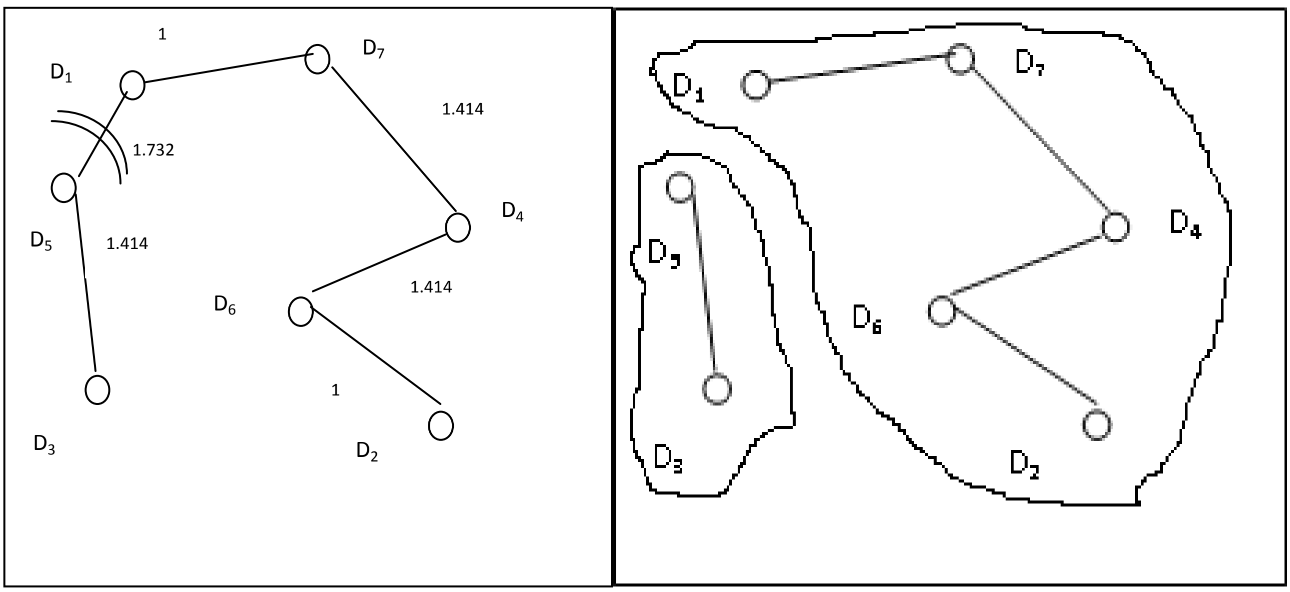

Figure 3 (a): Remove Maximumdistance edge From MST (b) Two cluster after Remove the edge

After the second iteration now again choose the highest distance between two nodes in remaining trees. So highest distance is 1.414 which have which have (D_3,D_5), (D_4, D_6) and (D_4,D_7). If we remove this edge between them then.

(D_1, D_7),(D_2, D_6), D_3, D_4, D_5

Module 3: Results Visualization

Mean=1.329

Standard deviation= 7.359

Finally the Cluster formed are

Number of Clusters	Documents
1	(D_1, D_7)
2	(D_2, D_6)
3	D_3
4	D_4
5	D_5

Table 4: Clustering Results

We have used “20 newsgroup” dataset and “Reuter’s” dataset, which are widely used in many publications. A summary description of these data sets is given in Table 5. The experiments were performed on an Intel core 2 Duo, 2.94 GHz system running Windows 7 professional with 2 GB of RAM.

Data Set	Documents	Classes
20 newsgroup dataset	20000	20
Reuter’sText Categorization Collection	8654	52

Table 5: Dataset Description

The final comparison of these algorithms is shown in Table 6.

Parameter Name	MST Based Clustering	K-Means
Dataset	20 newsgroup, Reuter’s	20 newsgroup, Reuter’s
Stop word Removal	Yes	Yes
Stemming	Yes	Yes
Length of smallest term(threshold)	5(Five)	5(Five)
Cluster Count *k*	Depend on cuts	Depend on Value of *k*
Overlapping	No	Yes
Work with High Dimensional data	Yes	No
Scalability	Yes	No

Table 6: Parameters list for our Approach and the other Approaches

V. CONCLUSION

In this paper, we discussed MST based document clustering and find the cluster in document datasets. We have also demonstrated the effectiveness of our approachwhich works more reliable. In first step of the framework we convert the unstructured data into structured format, then in second step produce the distance matrix and finally find the clusters. In the future we will explore and test our proposed document clustering algorithm in various domains and also reduce the complexity. For calculate of distance between two points. Ourmethod does not have the problem of dimensnality and no need togenerate initial seed.

VI. ACKNOWLEDGMENT

This work is supported by research grant from MANIT, Bhopal, India under Grants in Aid Scheme 2010-11, No. Dean(R&C)/2010/63 dated 31/08/2010.

REFERENCES

[1] A. Vathy-Fogarassy, A. Kiss, and J. Abonyi , “Hybrid Minimal Spanning Tree and Mixture of Gaussians Based Clustering Algorithms”, Proceeding. IEEE InternationalConferance Tools with Artificial Intelligence, pp 73-81, 2006.

[2] Andreas C. Muller, S. Nowozin, christoph H. Lampert, “Information theoretic clustering using

minimum spanning tree"Pattern Recognition, pp. 205-215, 2012.

[3] BhaskarAdepu, K.K. bejjanki, "A Novel Approach for Minimum Spanning Tree based Clustering Algorithm"

[4] B. Eswara Reddy, K. Rajendra Prasad, "reducing runtime values in minimum spanning tree based clustering by visual access tendency" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.3, pp 11-22, May 2012.

[5] C. Zahn. "Graph-theoretical methods for detecting and describing gestalt clusters". IEEE Transactions on Computers, C-20:pp. 68-86, 1971.

[6] Chang, J., Luo, J., Huang, J.Z., Feng, S., Fan, J.: Minimum spanning tree based classification model for massive data with mapreduce implementation. In: Fan, W., Hsu, W., Webb, G.I., Liu, B., Zhang, C., Gunopulos, D., Wu, X. (eds.) ICDM Workshops,. IEEE Computer Society pp. 129–137, 2010.

[7] CongnanLuoa, Yanjun Lib, Soon M. Chungc, "Text document clustering based on neighbours" Data & Knowledge EngineeringVolume 68, Issue 11, Pages 1271–1288, November 2009.

[8] D.S Rajput, R.S. Thakur, G.S. Thakur "Rule Generation from Textual Data by using Graph Based Approach", International Journal of Computer Application (IJCA) 0975 – 8887, New york USA, ISBN: 978-93-80865-11-8, Vol. 31–No.9,pp. 36-43 , October 2011.

[9] D. S. Rajput, R. S. Thakur, G. S. Thakur ,NeerajSahu, " Analysis of Social Networking Sites Using K- Mean Clustering Algorithm", International Journal of Computer & Communication Technology (IJCCT) ISSN (ONLINE): 2231 - 0371 ISSN (PRINT): 0975 – 7449 Vol-3, Iss-3, pp. 88-92, 2012.

[10] Han I and Kamber M, "Data Mining concepts and Techniques," M. K. Publishers, pp.335–389, 2000.

[11] Jiaxiang Lin, Dongyi Ye, Chongcheng Chen, MiaoxianGao, "Minimum Spanning Tree Based Spatial Outlier Mining and Its Applications", Third International Conference, RSKT 2008, Chengdu, China, May 17-19,. pp 508-515, 2008.

[12] J. Zhang and N. Wang, "Detecting outlying subspaces for high-dimensional data: the new task, Algorithms and Performance", Knowledge and Information Systems, 10(3):pp. 333-555, 2006.

[13] Lijuan Zhou , Linshuang Wang ; XuebinGe ; Qian Shi , "A clustering-Based KNN improved algorithm CLKNN for text classification", Informatics in Control, Automation and Robotics (CAR), 2nd International Asia Conference onVol.-3
pp: 212 – 215, 2010.

[14] M. Laszlo and S. Mukherjee, "Minimum Spanning Tree Partitioning Algorithm for Micro aggregation", IEEE Transaction, Knowledge and Data Engineering, Vol. 17, no 7, pp 902-911, July 2005.

[15] O. Grygorash, Y. Zhou, Z. Jorgensen, "Minimum spanning tree based clustering algorithm", in Proceeding of the 18th International Conference on Tools with Artificial Intelligence, pp. 73–81, 2006.

[16] PiotrJuszczak, David M.J. Taxa, ElżbietaPe¸kalskab, Robert P.W. Duina, "Minimum spanning tree based one-class classifier "Advances in Machine Learning and Computational Intelligence, Volume 72, Issues 7–9, , pp. 1859–1869, March 2009.

[17] P.Sampurnima, J Srinivas&Harikrishna, "Performance of Improved Minimum Spanning Tree Based on Clustering Technique" Global Journal of Computer Science and Technology Software & Data Engineering, ISSN: 0975-4172 Volume 12 Issue 13 pp 16-22, 2012.

[18] Vathy-Fogarassy ,A.Kiss, J.Abnoyi,"Hybrid Minimal Spanning tree based clustering and mixture of Gaussians based clustering algorithm", Foundations of Information and Knowledge systems, Springer, pp 313-330, 2006.

[19] William B. March, Parikshit Ram, Alexander G. Gray "Fast Euclidean minimum spanning tree: algorithm, analysis, and applications" In proceeding of: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010.

[20] Y.Xu, V.Olman and D.Xu. "Minimum spanning trees for gene expression data clustering". Genome Informatics, 12:pp24-33, 2001.

