International Journal of Scientific Research in Computer Science and Engineering
Vol.5, Issue.4, pp.9-15, Aug (2017)

Research Paper

E-ISSN: 2320-7639

Community Structure Detection in Social Networking Data Using Text Mining Approach

S. Arora[1*], P. Shukla[2], N. Karankar[3]

[1*] Dept. of Computer Science and Engineering, IET-DAVV, Indore, India
[2] Dept. of Computer Science and Engineering, IET-DAVV, Indore, India
[3] Dept. of Computer Science and Engineering, IET-DAVV, Indore, India

[*]*Corresponding Author: shrutiaro26@gmail.com, Tel.: 9039520955*

Available online at: www.isroset.org

Received 14th Jun 2017, Revised 26th Jun 2017, Accepted 20th Jul 2017, Online 30th Aug 2017

***Abstract*—**In data mining techniques some of the problems are resolved using the visualization techniques. Among them some of techniques are derived from the graph theory and transparent data modelling. The data structures such as decision trees and semantic graph representation is one of the key implementation of the graph based solution development. Among these technique one of the mathematical model termed as the community detection is a part of data mining solution discovery technique. Data mining techniques are used for finding the application centric patterns recovery from the raw set of data. Additionally the community detection technique is a visual technique for performing the unsupervised learning. During community detection the data objects are keeping connected to represent the bounding among them. Therefore in order to perform categorization task in automatic manner this technique can be employed in different nature of data. In this presented work the social media text is used for community detection. Communities are the group of objects that are highly similar in their properties. Therefore an algorithm is proposed in this work, that first refines the text content, then the text features are computed form raw text. In next the data is evaluated to find the number of possible communities in the data and finally the data is grouped in the communities and their visualization is performed. The proposed algorithm not only used to find the community structure from the data that also provides the relationship among two different communities. The experimental results in terms of precision, recall, f-measures demonstrate the proposed model is efficient and accurate as compared to traditional clustering algorithms namely the k-means clustering.

Keyword Clustering, Community Detection, Complex Network, Tweeter, Data Mining

I. INTRODUCTION

Basically the community is kind of network which demonstrate the connect graph. The nodes of this graph represent the objects of the network or data and the edges of the graph define the similar characteristics among the nodes or data objects. In this context the community detection is a problem to create sub-graph form a given graph based on their structure or the properties. In a broad term the community detection is a unsupervised learning approach or the clustering approach to create multiple sub-graphs from a given graph. This technique is also used in various data mining approaches and representations i.e. decision trees, semantic graphs of text representation.

The proposed work is intended to find the method of community structure detection in text data. Basically the text mining is an approach where the techniques of data mining are employed on text data for recovering the data patterns from it. As initially discussed, the community structure detection is clustering problem thus the proposed algorithm is aimed to prepare an automatic clustering with self-identification of the amount of clusters or communities in data. In addition to that it is also required to design an approach for finding optimal set of communities along with the establishment of relationship between the communities. The effort includes the association of the different phases of real world data processing techniques, analysis techniques and the traditionally available techniques for achieving the desired target

Section I contains the introduction of the presented work , Section II contain the detail explanation of proposed work Section III contain the result analysis and its comparison with the traditional K-mean algorithm Section IV conclude the research work with future directions.

II. PROPOSED WORK

This chapter provides the understanding about the proposed work and the formulation of the proposed community structure detection algorithm. Thus the chapter includes the system overview, the methodology of the work and algorithm steps.

A. System Overview

The social media text is mined in order to discover the various patterns such as the trending topic, user sentiments, topic tracking, spamming and others. In these applications for the analysis of data the text mining techniques are used. The text mining is little bit different from the traditional data mining technique, because the text is not found in a

structured manner or pre-labelled formats. Therefore the mining of such kind of data is a complicated task. Therefore the additional processes in text mining are included to reform the data and make it acceptable for the data mining algorithms. The text mining technique supports both the techniques supervised and unsupervised learning. But in this work the main focus is placed on unsupervised learning technique of text analysis.

The proposed work is intended to mine the text data using the community structure detection approach. In this technique the data simulated as data elements that are placed as the node of the graph and the relation among two data instances are represented using the edge of connected graph. In this approach the community is defined as the cluster of the nodes or sub-graph from the entire developed data graph. In order to formulate the desired concept a new algorithm is proposed for implementation and design. That first process the raw unstructured data and recover the data instances in terms of text features ,data instances are further correlated on the basis of data features similarity and their sub-graphs are developed based on the clustering concept. For performing the same it require, the different optimization cycles that help to improve the accuracy and connectivity of the developed community structures. Finally on the basis of their similarity matrix the relationship among community structures is also defined. This section provides the overview of the proposed system and in the next section the formulation of the model in terms of methodology is defined.

B. Methodology

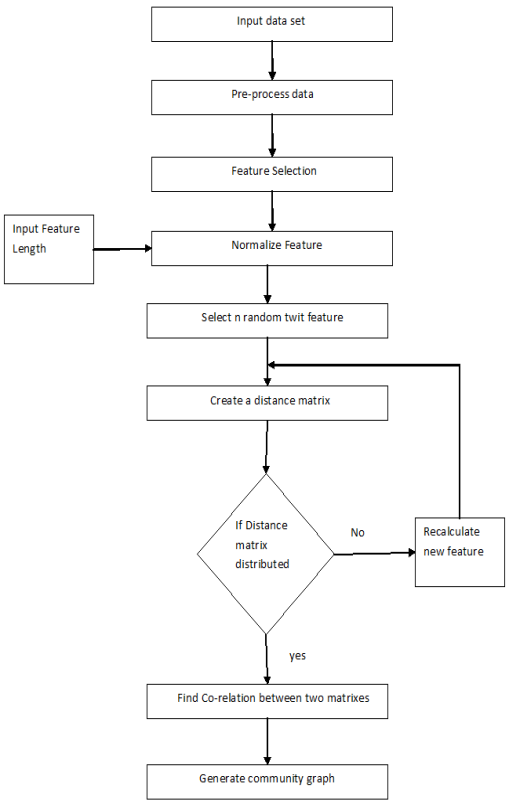

Figure 2.1 proposed system

After reviewing different recently developed techniques and contributions in the field of community detection. It is concluded that, the community detection technique is an unsupervised technique of categorization of the involved elements by their internal similarity among different objects. By considering the previously suggested facts a community detection model for unstructured text data is presented in this work.

The overview of the proposed concept can be understood by the figure 2.1. This section provides the component level details of the proposed system.

The twits are collected on the different domains that are Sports, Movies, Polities and business. Basically the available data in twitter dataset is not labelled with these class labels and additionally it is unstructured kind of data.

Input dataset: the system requires an initial input for which the communities are needed to be approximate. In this presented work the social media based text is targeted for extraction of communities. In order to perform such task the twitter data set is used. First of all the twitter dataset downloaded and the available twits in dataset is sub-divided according to their subjects. In this context the following four areas of twits are collected the collected twits in different subjects are needed to be refined and clean first.

Pre-processing:in data mining and its applications, pre-processing is an essential part of information processing. The main aim of pre-processing is to reduce the noise from the data and remove the unwanted information from the available data. In order to perform this task data can be transpose, transformed or mapped for improving the quality of data. In this work the data is pre-processed for finding the valuable data which can able to identify the subject of twits. Therefore two phase pre-processing is used in this work as:

1. **Removing stop words:** first a list of stop words are created such as (is, am, are, this, that,…). This list is used with the find and replaces function to reduce the target contents from the given twits.
2. **Removing characters:**after that the different characters which are used in text are removed using find and replace function such as (%, &, #,…). Finally these elements are also reduced from the given contents.

Feature selection:after pre-processing of the data the significant amount of text is reduced from all the twits. Now the tokenization process is taken place. Therefore all the twits are converted into a set of words or tokens. Now for each token the word frequency is computed. For computing the word frequency the following formula is used.

$$T_f = \frac{W_c}{W_t}$$

Where, T_f= the term frequency of a given token, W_c is the total count of the word or token and W_t is the total amount of tokens in document or domain.

After computing the word frequency each word is associated with their frequency values also.

Input feature length:during the frequency count it is observed, the amount of tokens in all the twits is different. Therefore an additional input by the user is required which is used to decide the length of features in the twits. That input is a real number produced by the user who experimented with the system.

Normalize features:in this phase the two inputs are accepted first the list of features which are associated with the twits. Secondly the length of features is also provided in this phase. Using both the input values the features list are regulated to achieve the similar length of tokens from all the twits. Therefore if the length of tokens is larger than the given length of features then the top frequent words is selected and remaining are removed from the feature list. Additionally if the amount of tokens is less than the given length then the null strings are added to prepare the similar length of features. After normalizing the features the now the next process is to identify the similar contents as the community.

Centroid selection:it is the process of clustering which is used to find the similar valued data from the available set of data. These are the points or feature objects which are randomly selected as the centre point of the community and the similarity or distance from other points are decided which point are concerned with which community.

Table 2.1 centroid selection

Input : list of features F
Output: possible centroid C

Process:
1. $Temp = null$
2. Centroid =0
3. $for(i = 1; i \leq F.length; i++)$
   a. $x = F[i]$
   b. $for(j = i + 1; j < F.length; j++)$
      i. $y = F[j]$
      ii. $d = \sqrt{x^2 - y^2}$
      iii. $if\ (d \leq 0.5)$
         1. $Temp.Add(y)$
         2. Remove(y, F)
      iv. End if
   c. End for
   d. centroid = centroid+1
4. End for

Distance matrix:now the each point or twit is compared with the selected centroids. In order to perform this task rapidly the distance matrix is computed. The example distance matrix is given using table 2.2.

Table 2.2 distance matrix

	C_1	C_2	C_3
$Twit_1$			
$twit_2$			

In this example distance matrix the centroids are provided as the columns of the matrix and the rows defines the twits and their isolability with the given centroids.

Matrix distribution: is the matrix contains the similar amount of twits in all the selected centroids then the matrix is uniformly distributed. The distribution of twits is not achievable by a single evaluation therefore the new centroid selection and their distance matrix computation needs more than one cycles to find most appropriate or uniform matrix distribution.

Table 2.3 matrix distribution

Input: number of domains N_d, number of twit per-domain D_t, centroid wise twit count C_t
Output: Boolean true / false

Process:
1. $for(i = 1; i \leq C_t.length; i++)$
   a. $T_h^i = \frac{D_t}{N_d} * C_t^i$
   b. $if\left(\left(C_t^i * 0.9\right) \leq T_h^i\right)$
      i. $return\ false$
   c. Else
      i. Return true
   d. End if
2. $end\ for$

New centroid selection:if the previous computed distance matrix is not uniformly distributed than the new centroids are computed and again the entire process is performed.

Correlating two distance matrix: matrix if the algorithm reaches its stoping criteria or the uniformly distributed is obtained then the correlation between two centroids are computed. The correlations between two centroids are their relativity between two communities.

Community generation: finally all the twits are considered as the community nodes and the distance matrix values are used to join edges of the nodes. Additionally for providing the relation between two communities the correlation values are used to join the two community centroids.

C. Proposed algorithm

The previous section provides the different definitions and the algorithms for demonstrating the proposed model. This section provides the steps of combining the entire components in functional steps using table 2.4:

Table 2.4 proposed algorithm

Input: text data set D
Output: communities C

Process:
1. $T_r = ReadDataset(D)$
2. $P_r = preProcessData(T_r)$
3. $F = ComputeFeatures(P_r)$
4. $Cen = ComputCentroid(F)$
5. $Dmat = ComputeDistanceMatrix(Cen, F)$
6. $C_{distribute} = Dmat.computeDistribution(Cen, F)$
7. $if(C_{distribute} == regular)$
   a. Stop optimization
8. Else
   a. Generate new centroid

b. Go to step 5
9. End if
10. $Cr = Correlation(Dmat)$
11. C= Generate Community
12. Return C

III. RESULTS AND DISCUSSION

The given chapter provides the detailed understanding about the evaluated results of the proposed *Community Detection for Social Networks*. Therefore this chapter includes the different performance parameters and their description on which the proposed system is evaluated using different runs of project.

A. Precision

Precision measure is the ratio of the number of correct positive results and number of all positive results. It measures the exactness of the clustering. The higher the precision means that less false positives (FP), whereas the lower precision means that more the false positives are. Here, we are showing precision rate formula.

$$\text{Precision Rate} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

-TP is the number of true positives

-FN is the number of false positives

The evaluation of the clustering is demonstrated using figure 3.1 and table 3.1. In this figure, X-axis show the number of different runs during algorithm process and Y-axis depict the precision rate of proposed and traditional clustering. Blue line show proposed approach using clustering algorithm, orange line show the traditional k-mean clustering performance. In this we have compared and analysis their comparison on the basis of the produced output. By this demonstration, the community structure node in different domain of the proposed approach performance is most efficient and accurate in terms of the precision rate other than evaluated method.

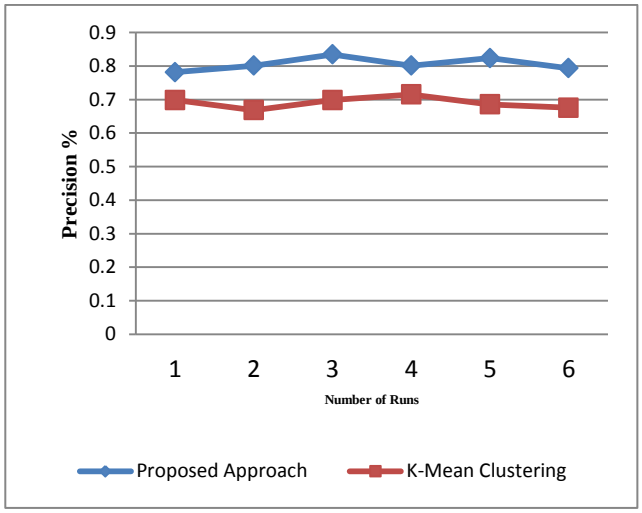

Figure 3.1 Precision Rate

Table 3.1 Tabular Values of Precision Rate

Number of Runs	Proposed Approach	K-Mean Approach
1	0.78161	0.69841
2	0.80137	0.66825
3	0.83471	0.69841
4	0.80115	0.71538
5	0.82349	0.68561
6	0.79345	0.67548

B.Recall

Recall is the ratio of the number of correct positive results and number of positive results that should have been returned. Community recall, which provide information about how much the nodes of a given community tend to be in the same ground truth community. Higher the recall means that small false negatives (FN), whereas lower the recall is more false negatives it leads to. In this, following recall rate formula used to find community for social network is:

$$\text{Recall Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

-TP is the number of true positives

-FN is the number of false negatives

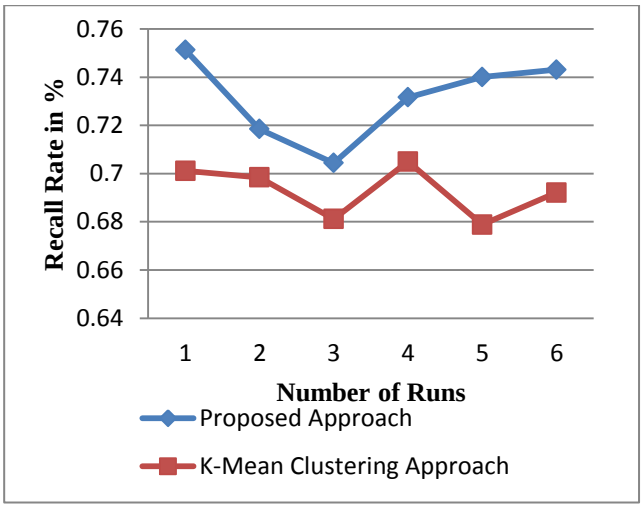

Figure 3.2 Recall Rate

The figure 3.2 and table 3.2 shows the comparative recall rate of implemented clustering algorithm. In order to show the performance of the system the X-axis contains the different execution of the project for analysis and the Y axis shows the performance in terms of recall rate percentage. The recall rate of the traditional k-mean is given using the orange line and the performance of the proposed clustering approach is given using the blue line. The performance of the proposed clustering is effective and efficient during different execution and reducing with the amount of data increases; thus, the presented community detection of more efficient and accurate than the evaluated traditional k-mean clustering.

Table 3.2 Tabular Values of Recall Rate

Number of Runs	Proposed Approach	K-Mean Approach
1	0.75134	0.70115
2	0.71845	0.69841
3	0.70443	0.68114
4	0.73164	0.70501
5	0.74005	0.67887
6	0.74314	0.69215

C. F-Measure

We can combine precision and recall into their structure indicates to obtaining the F1-measure, a concise quality score for the individual pairing. F-Measure or F1 Score is the harmonic mean of the Precision and Recall often used as a weighted average for balancing quality vs. quantity of true positives selection of an algorithm given by:

$$\text{F} - \text{Measure} = 2 * \frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}}$$

-

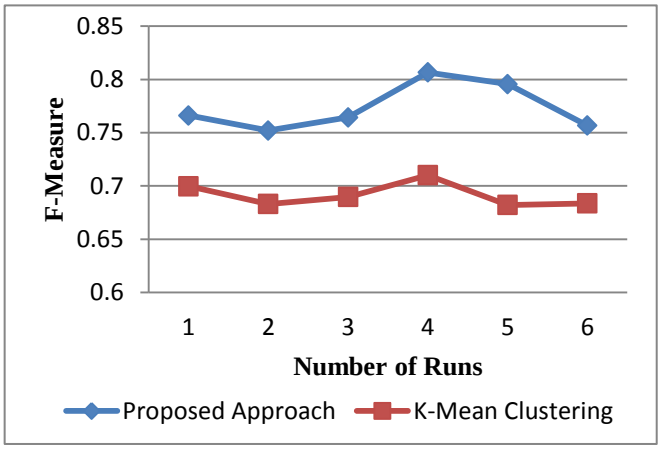

Figure 3.3 F-Measure

The figure 3.3 and the table 3.3 show the performance of proposed clustering algorithm in terms of f-measures. To demonstrate the performance of the system the X axis shows the different runs for data execution and the Y axis shows the obtained performance in terms of f-measures. According to the obtained results the performance of the proposed system is much stable and enhancing approach of community structure for a complex network based on centroid. In addition of that the results are in more progressive manner as the amount of data base is increases. Thus the obtained results are adoptable and efficient for community detection over the tweeter data set.

Table 3.3 Tabular Values of F-Measure

Number of Runs	Proposed Approach	K-Mean Approach
1	0.76617	0.69977
2	0.75203	0.68299
3	0.76432	0.68966
4	0.80663	0.71015
5	0.79564	0.68222
6	0.75666	0.68371

D. Time Consumption

The amount of time required to make cluster node which are belonging different community whereas community containing level of data is known as the time consumption. That can be computed using the following formula:

$$Time\ Consumed = End\ Time - Start\ Time$$

The time consumption of the proposed algorithm is given using figure 3.4 and table 3.4. In this diagram the X axis contains the number of experiments and the Y axis contains time consumed for finding the centriod node in community in terms of milliseconds. The depiction is comparison of the proposed clustering and k-mean clustering. According to the comparative results analysis the performance of the proposed technique minimize the time consumption. But the amount of time is increases in similar manner as the amount of data for analysis is increases.

Figure 3.4 Time Consumption

Table 3.4 Tabular Values of Time Consumption

Number of Runs	Proposed Approach	K-Mean Approach
1	100	115
2	82	125
3	85	130
4	90	131
5	101	121
6	105	128

E. Memory Consumption

Memory consumption of the system also termed as the space complexity in terms of algorithm performance. That can be calculated using the following formula:

$$\text{Memory Consumption} = \text{Total Memory} - \text{Free Memory}$$

The amount of memory consumption depends on the amount of data reside in the main memory, therefore that affect the computational cost of an algorithm execution. The performance of the implemented method for community detection is given using figure 3.5 and table 3.5. For reporting the performance the X- axis contains the number of execution by executing algorithms and the Y

axis shows the respective memory consumption during runs in terms of kilobytes (KB). According to the obtained results the performance of algorithm demonstrates similar behaviour with different system performance and variation of the memory space according the structure of community.

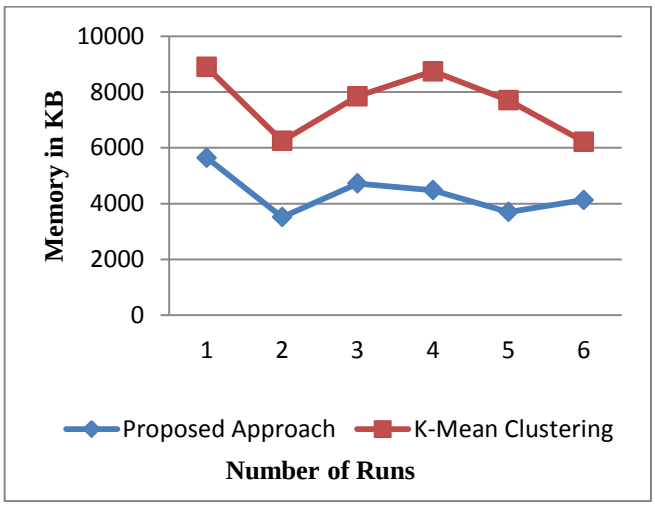

Figure 3.5 Memory Consumption

Table 3.5 Tabular form of Memory Consumption

Number of Runs	Proposed Approach	K-Mean Approach
1	5647	8895
2	3514	6246
3	4721	7845
4	4481	8741
5	3701	7701
6	4129	6215

IV. CONCLUSION AND FUTURE SCOPE

This chapter includes the summary of the proposed work and the performed efforts. The conclusion is made on the basis of the efforts and experimentation analysis. Additionally the future extension is provided on the basis of the possible future scope of the work.

A. Conclusion

The proposed text mining approach is a graph based approach of clustering of data. That graphical modelling of data is working in an unsupervised manner. Therefore the work is community detection through clustering technique in visual manner. The representation of data in this environment is performed as connected graphs; additionally the clusters are demonstrated as the sub-graphs of the entire graph model. The nodes of this connected graph are prepared using the data instances or data features and the edges are demonstrating the similarity of data. Although the text mining based on cluster analysis is defined by computing the internal object similarity.

Therefore the proposed algorithm includes the three main phases of the system execution. First the pre-processing of the data, which is used to make clean the data set from the different unwanted symbols and words. In next phase the possible amount of centroids is computed and based on the centroids the distance matrix is computed. In the final phase the data instances which are centroids are optimized and the possible community structure is defined. In addition to that correspondence among two communities is computed using the distance matrix.

The implementation of the proposed technique is performed using JAVA environment. The implemented technique is compared with the traditional k-means clustering with the similar dataset. According to obtained performance, the proposed community structure detection technique for text mining is efficient and accurate as compared to the traditional k-means based text mining technique. Thus the proposed model is efficient and accurate for real world application and their usages.

B. Future work

The main aim of the proposed work is to find community structure and obtaining the relation among two communities is accomplished in this work. The proposed work is extended in the following manner in near future.

1. The use of technique can be feasible in developing the information retrieval systems and document categorization techniques
2. The work is also extendable for improving the topic tracking and trending topic detection purposed.

REFERENCES

[1] A.Biswas, B.Biswas, "*Investigating community structure in perspective of ego network*", Expert Systems with Applications, 42 (2015) 6913–6934, 2015 ,Elsevier Ltd
[2] Arnaboldi, Passarella, Conti, Pezzoni "*Analysis of ego network structure in online social network"*,pp. 31-40, 2012
[3] Y.Wang, &L.Gao"*An edge-based clustering algorithm to detect social circles in ego networks*", Journal of Computers, Vol 8,2013
[4] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. "*Measurement and analysis of online social networks*". In IMC '07, pp 29–42, 2007.
[5] G. Palla, I. Der´enyi, I. Farkas, and T. Vicsek. "*Uncovering the overlapping community structure of complex networks in nature and society*". Nature, 435(7043):pp814–818, 2005.
[6] J. Leskovec, L. Adamic, and B. Huberman. "*The dynamics of viral marketing*". TWeb, ACM, 1(1), 2007.
[7] S. L. Feld. "*The focused organization of social ties*". Am. J. of Sociology, 86(5):1015–1035, 1981.
[8] G. Flake, S. Lawrence, and C. Giles." *Efficient identification of web communities*". In KDD '00, pages 150–160, 2000.
[9] M. Girvan and M. Newman. "*Community structure in social and biological networks*". PNAS, 99(12):7821–7826, 2002.
[10] M. S. Granovetter. "The strength of weak ties". Am. J. of Sociology, 78:1360–1380, 1973.
[11] R. Solanki, "*Principle of Data Mining*", McGraw-Hill Publication, **India**, pp. 108-128, 1998.
[12] Liu, H., &juan Ban, X. ."*Clustering by growing incremental self-organizingneural network*". Expert Systems with Applications, 42:4965–4981,2015,ElsevierLtd
[13] R.R.Khorasgani., J.Chen, &O.R.Zaïane, O. R. "*Top leaders community detectionapproach" in information networks*. In International conference on knowledge discovery and data mining,KDD'10: ACM,Citeseer, (pp. 9),2010

Authors Profile

Ms. S. Arora pursed Bachelor of Engineering from Medicaps Institute of science and Technology,RGPV,Indore. and currently she is student of Institute of Engineering & Technology,Davv,indore pursuing Master of Engineering in Computer Engineering.

Dr Pragya Shukla pursed bachelor of Engineering form government Engineering College Bhopal,MP,india, Masters in Technology from School of computers, DAVV,Indore,MP,India &Ph.D from Institute of Engineering and Technology,Davv,indore,MP.India .She has a teaching experience of 18 years and she has published 28 papers in reputated journals.Currently she is working as Professor in Institute of Engineering and Technology,Davv.Indore,MP,India

Mrs Nilima Karankar pursed bachelor of Engineering in Information Technology from mandsore Institute of Technology, Masters in Engineering in information technology with specialization in information security from IET,DAVV,Indore,MP,India.She has a teaching experience of more than 10 years. She is currently working as Assistant professor in IET,DAVV,Indore.MP,India.

