ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering
Vol. 2, Issue 7, July 2013

High Quality Assessment of Similarity by Using Multiple View Points

K.Ramesh[1], C.Vasumurthy[2],Prof.D.Venkatesh[3]

Associate Professor & HOD, Dept. of CSE, GATES College of Engineering, Gooty, Andhra Pradesh, India [1]

PG Student, Dept. of CSE, GATES College of Engineering, Gooty, Andhra Pradesh, India [2]

Dean , Dept. of CSE& IT, GATES College of Engineering, Gooty, Andhra Pradesh, India [3]

ABSTRACT: Clustering is a process of grouping objects based on certain similarity measure. Then groups are known as clusters which can be analyzed and used further for operations like query processing. Clustering algorithms assume certain relationship among objects in the given dataset. The existing clustering algorithms with respect to text mining use single viewpoint similarity measure for partitioned clustering of objects. The main drawback of these algorithms is that the resultant clusters can't make use of fully informative assessment. In this paper we propose a new measure for finding similarity between objects which is multi-viewpoint based. This approach considers multiple viewpoints while comparing objects for clustering. This measure can have more informative assessment of similarity thus making clusters with highest quality. We also proposed two criterion approaches for achieving highest intra-cluster similarity and lowest inter-cluster similarity. The empirical results reveal that the proposed measure is used in making quality clusters.

Keywords: Text mining, single viewpoint, multi-view point, clustering

I. INTRODUCTION

Data mining is the study of the dataset by extracting trends or patterns from data. There are many techniques in the domain of data mining. Clustering is widely used data mining algorithms that help in grouping similar objects. This has many advantages in real world applications such as search engines. Clustering is of two types. They are known as partitioned clustering and hierarchical clustering. This paper focuses on partitioned clustering. K-means is one of the algorithms that are of type clustering algorithms which is widely used in the industry [1].It still appears in the top ten lists of such algorithms. It also has many flavors though they are functionally similar. The K-means algorithm needs dataset and number of clusters as two required arguments. Credit card fraud detection is one of the areas in which K-means is being used. In this application K-Means make a model from the dataset with three different clusters. When new records are added, they are adapted to this model. Based on the pattern in the data, the new transaction is considered normal or probable fraudulent. Thus in the data mining domain, it is simple and effective clustering algorithm [2]. However, it has drawbacks tool as it is sensitive to cluster size, initialization and with low performance comparatively. In spite of these drawbacks, it is still popular and most widely used as it is simple, scalable and intuitive. Its good quality is that it can be used with algorithms in combination to yield good results. The process of clustering with highest quality is an optimization process. With the optimization in place, highest quality clusters can be formed. Therefore, for good quality clusters, it is important to use similarity measure which is suitable. For instance cosine similarity measure is used by K-means. Original K-means also used ED (Euclidean Distance) [3] and [4].

Data grouping, data partitioning and hierarchical clustering are the three types of clustering methods according to Leo Wanner [5]. Clusters with hierarchy are possible with hierarchical clustering while the partitioning clustering focuses on dividing given objects into some groups. The data grouping approach is meant for making a set of overlapping clusters. By investigating the facts provided above, the proposed work is ascertained. Especially the similarity measure is the motivation behind this work. From the review of literature, it is found that the similarity measure being used in clustering has its impact on the resultant clusters. Therefore we thought of making a new similarity measure which is multi-viewpoint based as single viewpoint based measures do not yield highest quality of clusters. After developing the new similarity measure, we focused on making two clustering criteria to achieve highest intra-cluster similarity and lowest inter-cluster similarity. The rest of the paper is structured into sections such as related work, multi-viewpoint based similarity, algorithms, experimental setup and evaluation, results and conclusion.

II. PRIOR WORKS

Document clustering is required in the real world applications such as web search engines. It comes under text mining. It is being used for many years. It is meant for grouping documents into various clusters. These clusters are used by various applications in the real world such as search engines. A document is treated as an object a word in the document is referred as a term. A vector is built to represent each document. The total number of terms in the document is represented by m. Some kind of weighting schemes like Term Frequency – Inverse Document Frequency (TF-IDF) is used to represent document vectors. There are many approaches for document clustering. They include probabilistic based methods [8], non-negative matrix factorization [7] and information theoretic co-clustering [6]. These approaches are not using a particular measure for finding similarity among documents. In this paper, we make use of multi-viewpoint similarity measure for finding similarity. As found it literature, a measure widely used in document clustering is ED (Euclidian Distance).

$$\text{Dist}\,(d_i,d_j) = ||d_i - d_j|| \qquad (1)$$

K-Means is most widely used clustering algorithm due to its ease of use and simplicity. ED is the measure used in K-Means algorithm to measure the distance between objects to make them into clusters. In this case the cluster centroid is computed as:

$$\text{Min} \sum_{r=1}^{k} \sum_{di \in S_r} ||d_i - C_r||^2 \qquad (2)$$

Another similarity measure being used for text mining is cosine similarity measure. It is best used in hi-dimensional documents [9]. This measure is also being used in Spherical K-Means which is a variant of K-Means. The difference between the two flavors of K-Means that use cosine similarity measure and ED measure respectively is that the former focuses on vector directions while the latter focuses on vector magnitudes. Graph partitioning is yet another approach which is popular. It considers the document corpus as graph and uses min-max cut algorithm which represents centriod as:

$$\text{Min} \sum_{r=1}^{k} \frac{D^t_r\, D}{||Dr||^2} \qquad (3)$$

Average Weight [11] and Normalized Cut [10] are other graph partitioning methods meant for using clustering documents. For aching this they used cosine similarity and pair wise score respectively. Criterion functions are used in [12] for document analysis.

There is a software package by name CLUTO [13] which is meant for document clustering. It makes use of graph partitioning approach. Based on the nearest neighbor graph it builds, it documents are clustered. It is based on the Jacquard coefficient which is computed as:

$$\text{Sim}_{eJacc}(u_i,u_j) = \frac{u^t_i u_j}{||ui||^2 + ||uj||^2 - utiuj} \qquad (4)$$

Jacquard coefficients use both magnitude and direction which is not the case with ED and cosine similarity. However, it is similarity to cosine similarity when the documents are represented as unit vectors. In [14] there is comparison between two techniques namely Jacquard and Pearson correlation. It also concludes that both of them are best used in clustering web documents. For document clustering other approaches can be used which are phrase based and concept based. In [18] phrase based approach is found while in [17] tree-similarity based approach is found. The common algorithm used by both of them is "Hierarchical agglomerative Clustering". The drawback of these approaches is that their computational cost is very high. For clustering XML documents also there are measures. One such measure is named "Structural Similarity" [19] which differs from text document clustering. This paper focuses on a new multi-viewpoint based similarity measure.

III. MULTI-VIEW POINT BASED SIMILARITY MEASURE

Multi-viewpoint based similarity is the approach followed in this paper for clustering documents. It does meant that it uses more than one view point while finding similarity between objects and clustering them into various groups. We calculate the similarity between two documents as:

$$\mathrm{Sim}\,(d_i,d_j) = 1/n\text{-}n_r \sum_{d_t,d_j \in S_r d_h \in S \backslash S_r} \mathrm{Sim}\,(d_i\text{-}d_h,\ d_j\text{-}d_h) \qquad (5)$$

The approach is described here. When di and dj are the two points in cluster Sr, dh is considered the similarity between them which is equal to cosine angle of ED of those points. The assumption used here "dh is not the same cluster as dj and di". When there is similar distance, the likely chances are that dh is in the same cluster. The multi-viewpoint similarity measure being employed in this paper may rarely provide negative results where there are very few documents. However, it can be ignored provided the documents are more while clustering.

IV. ALGORITHMS PROPOSED

Many algorithms have been proposed to work on multi-viewpoint similarity measure. The procedure for similarity matrix is as shown in Listing 1.

```
Procedure BUILDMVSMATRIX(A)
    For r ← 1 : c do
        $D_{s/sr} \leftarrow \Sigma_{di \notin Sr}\ di$
        Ns/sr ← |S\Sr|
    End for
    For r ← 1 : n do
        R ← class of di
        For j ← 1 : n do
            If dj ϵ Sr then
                $aij \leftarrow d^{t}_{j}dj - d^{t\ Ds/Sr}_{i}{}_{nS/Sr} - d^{t\ Ds/sr}_{j}{}_{nS/Sr} + 1$
            else
                $aij \leftarrow d^{t}_{j}dj - d^{t\ Ds/Sr}_{i}{}_{nS/Sr} - d^{t\ Ds/sr - Dj}_{j}{}_{nS/Sr -1} + 1$
            end if
        end for
    end for
    return $A=\{a_{ij}\}$ mxn
end procedure
```

Algorithm 1 –Procedure for making similarity matrix

As per the procedure in Algorithm 1, it is known that dl and di are closer and the dl is also considered closer to di as per the multi-viewpoint simairlity measure. The Algorithm 2 showsn the validation procedure.

```
Procedure GETVALIDITY(validity,A,percentage)
    For r ← 1 : c do
        qr ← [percentage x $n_r$
        if qr=0 then     percentage too small
            qr ← 1
        end if
    end for
    For i ← 1 : n do
        $\{a_{iv[1]}, a_{iv[N]}\} \leftarrow$ Sort $\{a_{i1}, a_{in}\}$
        s.t. $a_{iv[1]} > a_{iv[2]} >\text{---}> a_{iv[n]}$
        {v[1] , v[n]} ← permute {1,..n}
        r ← class of di
        validity $(d_i) \leftarrow |\{d_{v[1]}, d_{v[qr]}\} \cap Sr|$
    end for
    validity ← $\Sigma^{n}_{i=1}$ validity (di)
    return validity
end procedure
```

Algorithm 2 –Validation Procedure

By averaging overall rows, the final validity is calculated. It is as given in line 14. It is known that when validation score is higher, it reflects that the similarity is higher and thus eligible for clustering. Fig. 1 shows the validity scores of multi-viewpoint simialirty and cosine similarity.

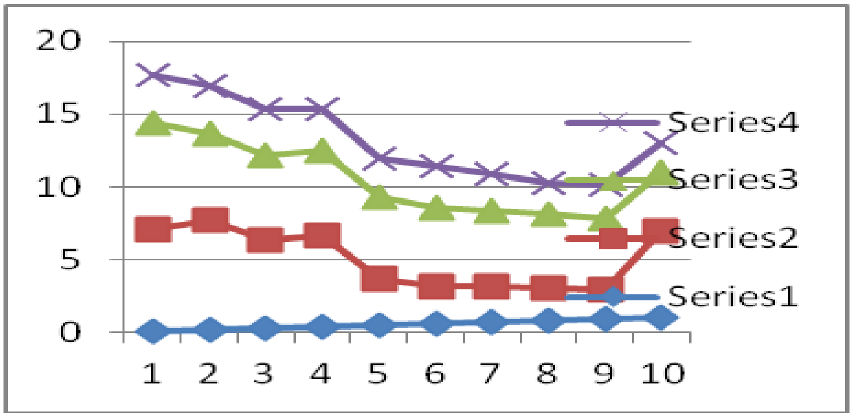


Fig. 1 – Validity of Cosine Similarity and Multi-Viewpoint Based Similarity

As can be seen in fig. 1, Series 4 is related to klb-MVS, series 3 corresponds to klb-CS, series 2 corresponds to reutors-7 while series 1 corresponds to reutors -7 CS. As shown in fig. 1 performance of MVS is higher when compared to that of CS.

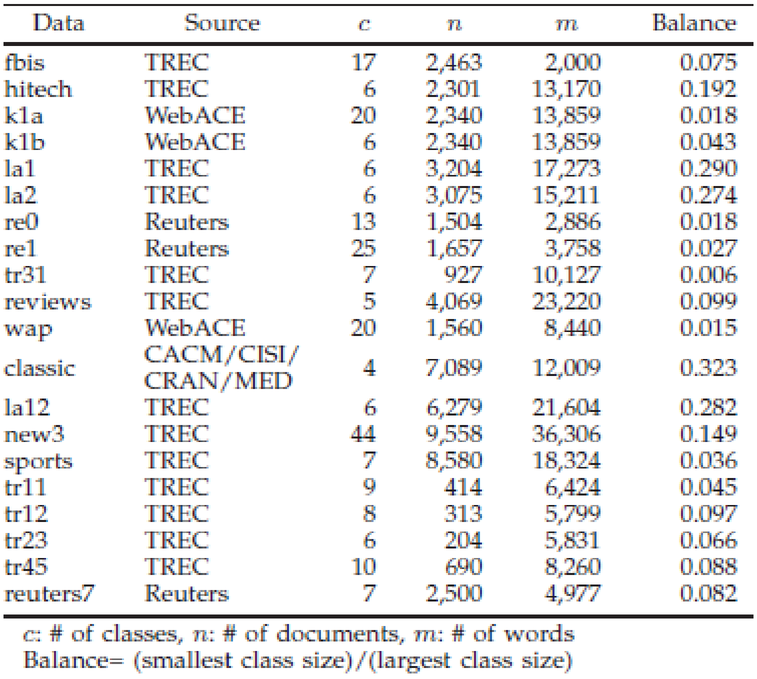
```
1.  Select k seeds S1…………..,Sk randomly
2.  Cluster[di] ← p=argmaxr{strdi}, ʉi=1,…..,n
3.  Dr ← Σ $_{di∉Sr}$ di, nr ← |Sr|, ʉr=1,….,k
4.  End procedure
5.  Procedure REFEINEMENT
6.  Repeat
7.  {v[1 : n]} ← random permutation of {1,….,n}
8.  For j ← 1: n do
9.  I ←v[j]
10. P ← cluster[di]
11. ΔIp ← I(np-1,Dp-di) – I(np,Dp)
12. q ← arg max r,r=p {I(nr+1, Dr+di)-I(nr,Dr)}
13. ΔIp ← I(nq+1, Dq+di) – I(nq,Dq)
14. If ΔIp + ΔIq > 0 then
15. Move di to cluster q: cluster[di] ← q
16. Update Dp,np,Dq,nq
17. End if
18. End for
19. until No move for all n documents
20. end procedure
```

Algorithm 3 –Algorithm for incremental clustering

Algorithm 3 is shows algorithm with two phases. They are refinement and initialization. Selecting k documents as seeds are known as initialization which making initial positions while the refinement makes much iteration to form best clusters. Each iteration in refinement phase visits n number of documents in random fashion. After that the process of verification is done for each document. If the document is considered similar, it is moved to the cluster. When no documents are there the iterations come to an end.

V. PERFORMANCE EVALUATION

For performance evaluation for two criterion functions such as Ir, and Iv with respect to multi-viewpoint similarity measure. Bench mark datasets have been used to test the efficiency of our approach. The results are tabulated in table 1.



Data	Source	c	n	m	Balance
fbis	TREC	17	2,463	2,000	0.075
hitech	TREC	6	2,301	13,170	0.192
k1a	WebACE	20	2,340	13,859	0.018
k1b	WebACE	6	2,340	13,859	0.043
la1	TREC	6	3,204	17,273	0.290
la2	TREC	6	3,075	15,211	0.274
re0	Reuters	13	1,504	2,886	0.018
re1	Reuters	25	1,657	3,758	0.027
tr31	TREC	7	927	10,127	0.006
reviews	TREC	5	4,069	23,220	0.099
wap	WebACE	20	1,560	8,440	0.015
classic	CACM/CISI/ CRAN/MED	4	7,089	12,009	0.323
la12	TREC	6	6,279	21,604	0.282
new3	TREC	44	9,558	36,306	0.149
sports	TREC	7	8,580	18,324	0.036
tr11	TREC	9	414	6,424	0.045
tr12	TREC	8	313	5,799	0.097
tr23	TREC	6	204	5,831	0.066
tr45	TREC	10	690	8,260	0.088
reuters7	Reuters	7	2,500	4,977	0.082

c: # of classes, n: # of documents, m: # of words
Balance= (smallest class size)/(largest class size)

Table 1 –Benchmark documents datasets

VI. EXPERIMENTAL SETUP AND EVALUATION

Our algorithm is compared with other algorithms for evaluation. They include M-means, Min Max Cut Algorithm, graph EJ which is nothing but CLUTO's graph with extended Jacquard, graphCS which is nothing but CLUTO's graph with Cosine Similarity, SpkMeans which is nothing but Spherical K-Means with Cosine Similarity, MVSC Iv which is nothing but the proposed approach with Iv criterion function and MVSC Ir which is nothing but the proposed approach with Ir criterion. The results are presented in the next section.

VII. RESULTS

The results of experiments are presented in fig. 2 and 3. It shows graphically the results of all clustering algorithms for 20 benchmark datasets. The results are presented into two different graphs. Each graph shows the experimental results of 10 datasets.

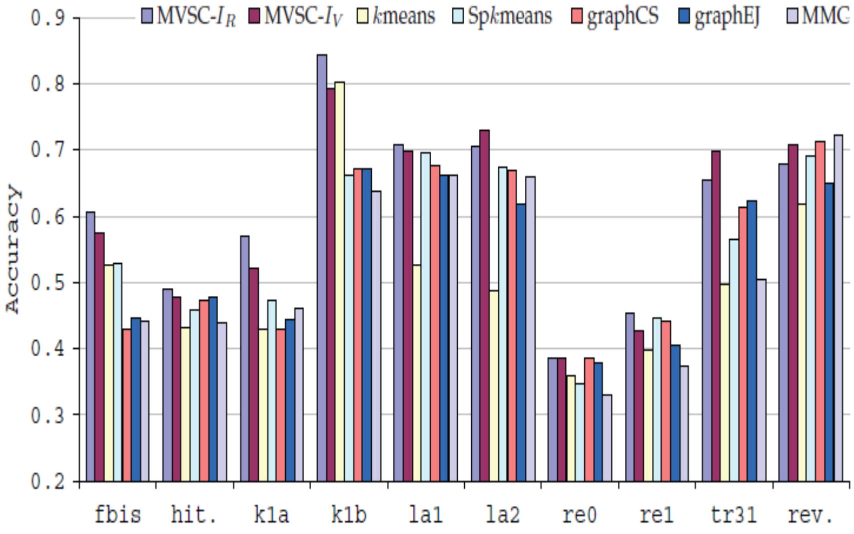


Fig. 2 (a) Results of Algorithms for first 10 datasets

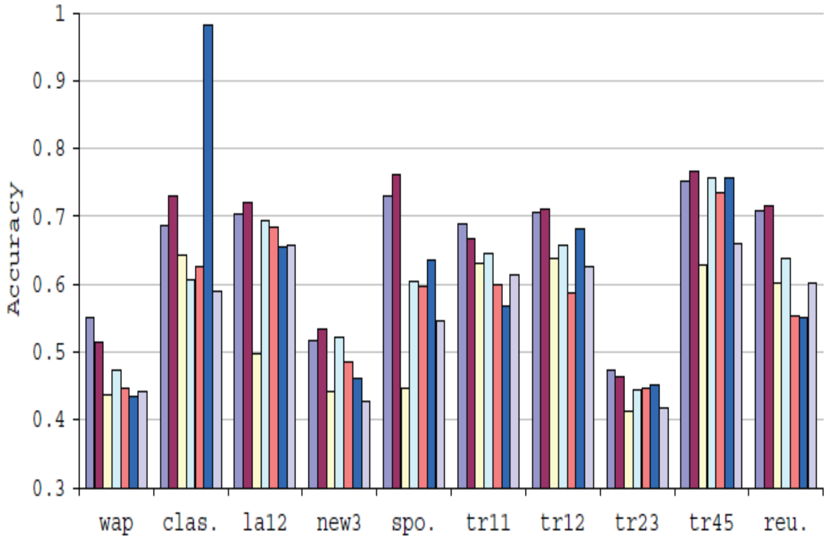


Fig. 2 (b):Results of Algorithms for next 10 datasets

As shown in fig. 2(a) and (b), the proposed approach performs better when compared with other algorithms. GraphEJ performed better only in some cases. Overall performance of MVSC Kr, and MVSC Iv is far better than others. The effect of α is also presented on the performance of MVSC Ir.

VIII. THE EFFECT OF α ON THE PERFORMANCE OF MVSC IR

It is understood that the cluster size and balance have their impact on the methods of partitional clustering based on criterion functions. In terms of NMI, FScore, and accuracy the assessment is done and the results are presented in fig. 3.

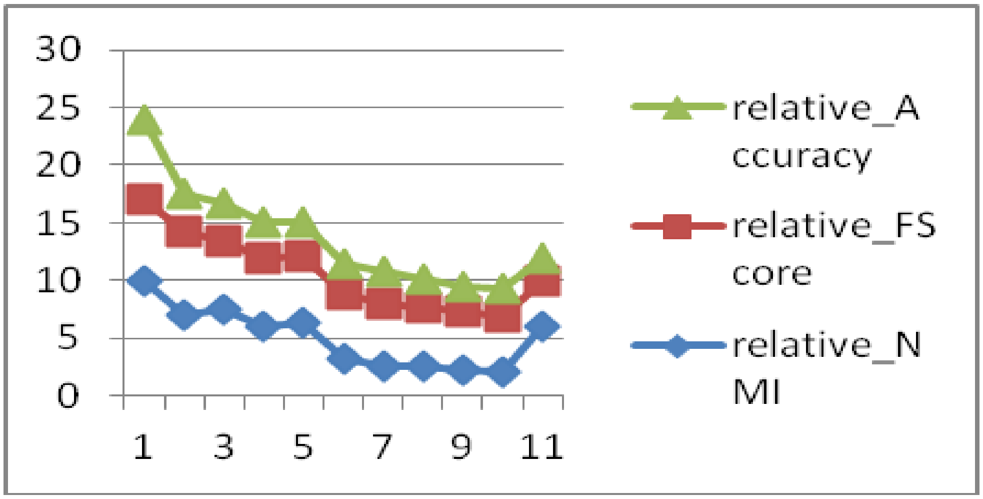


Fig. 3 - Performance of MVSC Ir with respect to α

As shown in fig. 3, MVSR Ir's performance is within 5% with regard to any evaluation metric for best case.

IX. CONCLUSION

This paper presents a novel similarity approach known as multi-viewpoint based similarity measure. The similaritymeasure is capable of providing informative assessment and bestows high quality clusters. The proposed approach achieves highest similarity between objects of same cluster and lowest similarity between the objects of different clusters. Two criterion functions were implemented with MVS. The proposed similarity measure is tested with bench mark datasets. The proposed clustering algorithms in this paper are compared with five other clustering algorithms used for document clustering. The results revealed that the multi-viewpoint based similarity measure outperforms them.

REFERENCES

[1] A. Ahmad and L. Dey, "A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set," Pattern Recognit. Lett., vol. 28, no. 1, pp. 110 – 118, 2007.
[2] A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," J. Mach. Learn. Res., vol. 6, pp. 1345–1382, Sep 2005.
[3] I. Dhillon and D. Modha, "Concept decompositions for large sparse text data using clustering," *Mach. Learn.*, vol. 42, no. 1-2, pp. 143–175, Jan 2001.
[4] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in KDD, 2003, pp. 89–98.
[5] S. Flesca, G. Manco, E. Masciari, L. Pontieri, and A. Pugliese, "Fast detection of xml structural similarity," IEEE Trans. on Knowl. And Data Eng., vol. 17, no. 2, pp. 160–175, 2005.
[6] I. Guyon, U. von Luxburg, and R. C. Williamson, "Clustering: Science or Art?" *NIPS'09 Workshop on Clustering Theory*, 2009.
[7] D. Ienco, R. G. Pensa, and R. Meo, "Context-based distance learning for categorical data clustering," in Proc. of the 8th Int. Symp. IDA, 2009, pp. 83–94.
[8] Leo Wanner (2004). "Introduction to Clustering Techniques". Available online at: http://www.iula.upf.edu/materials/040701wanner.pdf [viewed: 16 August 2012]
[9] C. D. Manning, P. Raghavan, and H. Sch ¨ utze, An Introduction to Information Retrieval. Press, Cambridge U., 2009.
[10] on web-page clustering," in Proc. of the 17th National Conf. on Artif. Intell.: Workshop of Artif. Intell.for Web Search. AAAI, Jul. 2000, pp. 58–64.
[11] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, pp. 888–905, 2000.
[12] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures.
[13] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl.Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2007.
[14] W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization," in SIGIR, 2003, pp. 267–273.
[15] H. Zha, X. He, C. H. Q. Ding, M. Gu, and H. D. Simon, "Spectral relaxation for k-means clustering," in NIPS, 2001, pp. 1057–1064.
[16] Y. Zhao and G. Karypis, "Empirical and theoretical comparisons of selected criterion functions for document clustering," Mach. Learn., vol. 55, no. 3, pp. 311–331, Jun 2004.
[17] S. Zhong, "Efficient online spherical K-means clustering," in *IEEE IJCNN*, 2005, pp. 3180–3185.

BIOGRAPHY

K.Ramesh graduated from JNTU Anantapur in Bachelor of Technology (Computer Science& Engineering) in the the year 2002, M.Tech specialized in Computer Science from JNTU Anantapur in the year 2010 and currently working as Asso.Prof & Head of the Department of Computer Science in GATES Institute of Technology .His area of interest includes Data Ware Housing Mining, Formal Languages and Automata Theorem

C.Vasumurthy graduated in Master of Computer Applications from Dravidian University Kuppam in the year 2010, and currently perusing M.Tech specialized in computer Science from GATES Institute of Technology, Gooty (Affiliated to JNTU Anantapur).His area of interest includes Data Ware Housing Mining, Computer Networks

D.Venkatesh graduated in Master of Computer Applications, and received his M.Tech from Satyabhama University and currently pursuing Ph.D from Rayalaseema University. He is associated with GATES Institute of Technology as Dean of CSE &IT departments since 2010. His area of interest includes Design and Analysis of Algorithms, Computer Networks