IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661 Volume 4, Issue 6 (Sep-Oct. 2012), PP 37-42
www.iosrjournals.org

Clustering Algorithm with a Novel Similarity Measure

Gaddam Saidi Reddy[1], Dr.R.V.Krishnaiah[2]
[1,2]*Department of CSE, DRK institute of science and technology, Hyderabad, India.*

Abstract-*Clustering is one of the data mining and text mining techniques used to analyze datasets by dividing it into meaningful groups. The objects in the dataset can have certain relationships among them. All clustering algorithms assume this before they are applied to datasets. The existing algorithms for text mining make use of a single viewpoint for measuring similarity between objects. Their drawback is that the clusters can't exhibit the complete set of relationships among objects. To overcome this drawback, we propose a new similarity measure known as multi-viewpoint based similarity measure to ensure the clusters show all relationships among objects. We also proposed two clustering methods. The empirical study revealed that the hypothesis "multi-viewpoint similarity can bring about more informative relationships among objects and thus more meaningful clusters are formed" is proved to be correct and it can be used in the real time applications where text documents are to be searched or processed frequently.*

Index Terms– *Data mining, text mining, similarity measure, multi-viewpoint similarity measure, clustering methods.*

I. Introduction

Data mining is a process of analyzing data in order to bring about trends or patterns from the data. Many techniques are part of data mining. Other mining such as text mining and web mining also exists. Clustering is one of the important data mining or text mining algorithm that is used to group similar objects together. In other words, it is used to organize given objects into some meaningful sub groups that make further analysis on data easier. Clustered groups make search mechanisms easy and reduce the bulk of operations and computational cost. Many clustering algorithms have been around since the inception of data mining domain. They are used based on the kind of application. One such clustering algorithm being used widelyby the IT industry is k-means . It still remains in the top list of widely used clustering algorithms in the world. It has many variants as well. Basically its functionality is similar. It takes two arguments and forms clusters. The first argument is data set or objects to be clustered while the second argument is the number of clusters to be formed. It has wide range of applications. One such application is credit card fraud detection. In such application, it generates clusters offline and makes a model. And then new transactions are simply added to the model which has clusters indicating high, low and medium range transactions. When a new transaction takes place, it can compare with the general buying patterns of customer and can detect abnormality. Any abnormality is suspected to be a fraudulent transaction. According to also k-means is the most favorite clustering algorithms in the data mining domain. Nevertheless, it has its own drawbacks that are well known to the world. They are sensitiveness to cluster size, sensitiveness to initialization; its performance is lesser than many other clustering techniques used in the data mining domain. Provided these drawbacks, it is still considered popular due to its simplicity, scalability and understandability. As it is less complex with adequate performance, it is widely used in the industry overlooking its known limitations. Another important quality of k-means algorithm is that it can be easily combined with other algorithms for best results. Generally the problem of clustering can be thought as optimization process. By optimizing similarity measures the optimal clusters can be formed thus performance is improved. Therefore the soundness of clustering algorithms depends on their similarity measure adopted. To meet various requirements k-means has many variants. For instance spherical k-means (uses cosine similarity) is used to cluster text documents while original k-means can be used to clustering using Euclidean distance [3].

According to Leo Wanner, clustering methods are classified into hierarchical clustering, data partitioning, data grouping. The hierarchical clustering is used to establish cluster taxonomy. Data partitioning is used to build a set of flat partitions. They are also known as non-overlapping clusters. Data group is used to build a set of flat or overlapping clusters. The proposed work in this paper is motivated by the facts ascertained by investigation of the above. Especially similarity measures are considered. From research findings it is understood that the nature of similarity measured used in any clustering technique has profound impact on the results. The aim of the paper is to develop a new method that is used to cluster text documents that have sparse and high dimensional data objects. Afterwards we formulate new clustering criterion functions and corresponding clustering algorithms respectively. Like k-means the proposed algorithms work faster and provide consistent, high quality performance in the process of clustering text documents. The proposed similarity measure is based on multi-viewpoint which is elaborated in the later sections.

II. Related Work

Document clustering is one of the text mining techniques. It has been around since the inception of text mining domain. It is s process of grouping objects into some categories or groups in such a way that there is maximization of intra-cluster object similarity and inter-cluster dissimilarity. Here an object does mean a document and term refers to a word in the document. Each document considered for clustering is represented as an ***m*** – dimensional vector ***d***. The ***m***represents the total number of terms present in the given document. Document vectors are the result of some sort of weighting schemes like TF-IDF (Term Frequency –Inverse Document Frequency). Many approaches came into existence for document clustering. They include information theoretic co-clustering [4], non – negative matrix factorization, probabilistic model based method [2] and so on. However, these approaches did not use specific measure in finding document similarity. In this paper we consider methods that specifically use certain measurement. From the literature it is found that one of the popular measures is Eucludian distance.

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

K-means is one of the popular clustering algorithms in the world. It is in the list of top 10. Due to its simplicity and ease of use it is still being used in the mining domain. Euclidian distance measure is used in *k*-means algorithm. The main purpose of the k-means algorithm is to minimize the distance, as per Euclidian measurement, between objects in clusters. The centroid of such clusters is represented as:

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

In text mining domain, cosine similarity measure is also widely used measurement for finding document similarity, especially for hi-dimensional and sparse document clustering . The cosine similarity measure is also used in one of the variants of ***k***-means known as spherical ***k***-means. It is mainly used to maximize the cosine similairity between cluster's centroid and the documents in the cluster. The difference between k-means that uses Euclidian distance and the k-means that make use of cosine similarity is that the former focuses on vector magnitudes while the latter focuses on vector directions. Another popular approach is known as graph partitioning approach. In this approach the document corpus is considered as a graph. Min – max cut algorithm is the one that makes use of this approach and it focuses on minimizing centroid function.

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

Other graph partitioning methods include Normalized Cut and Average Weight are used for document clustering purposes successfully. They used pairwise and cosine simialarity score for document clustering. For document clustering analysis of criterin functions is made.

CLUTO [1] software package where another method of document clustering based on graph partitioning is implemented. It builds nearest neighbor graph first and then makes clusters. In this approach for given non-unit vectors of document the extend Jaccard coefficient is:

$$Sim_{ejacc}\ (u_i,u_j) = \frac{ui\ uj}{||ui||2 + ||uj||2 - uti\ uj} \qquad (4)$$

Both direction and magnitude are considered in Jaccard coefficients when compared with cosine similarity and Euclidean distance.When the documents in clusters are represented as unit vectors, the approach is very much similar to cosine similarity. All measures such as cosine, Euclidean, Jaccard, and Pearson correlation are compared . The conclusion made here is that Ecludean and Jaccard are best for web document clustering. In [1] and research has been made on categorical data. They both selected related attributes for given subject and calculated distance between two values. Document similarities can also be found using approaches that are concept and phrase based. In [1] tree-milarity measure is used conceptually while proposed phrase-based approach. Both of them used an algorithm known as Hierarchical Agglomerative Clustering in order to perform clustering. Their computational complexity is very high that is the drawback of these approaches. For XML documents also measures are found to know structural similarity [5]. However, they are different from normal text document clustering.

III. Multi-View Point Based Similarity

Our approach in finding similarity between documents or objects while performing clustering is multi-view based similarity. It makes use of more than one point of reference as opposed to existing algorithms used for clustering text documents. As per our approach the similarity between two documents is calculated as:

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

Here is the description of this approach. Consider two point ***di*** and ***dj*** in cluster ***Sr***. The similarity between those two points is viewed from a point ***dh*** which is outside the cluster. Such similarity is equal to the product of cosine angle between those points with respect to Eucludean distance between the points. An assumption on which this definition is based on is "***dh*** is not the same cluster as ***di*** and ***dj***. When distances are smaller the chances are higher that the ***dh*** is in the same cluster. Though various viewpoints are useful in increasing the accuracy of similarity measure there is a possibility of having that give negative result. However the possibility of such drawback can be ignored provided plenty of documents to be clustered.

IV. Algorithms Proposed

A series of algorithms are proposed to achieve MVS (Multi-View point Similarity). Listing 1 give a procedure for building similarity matrix of MVS.

```
1: procedure BUILDMVSMATRIX(A)
2: for r ← 1 : c do
3: DSISrI←_di/∈∑Sr di
4: nSISr ← |S ISr|
5: end for
6: for i ← 1 : n do
7: r ← class of di
8: for j ← 1 : n do
9: if dj ∈ Sr then
10: aij ← dti dj – dti DSISr nSISr – dt j DS\Sr nSISr + 1
11: else
12: aij←dti dj−dti DSISr –dj nSISr −1–dt j DSISr –dj nSISr −1
13: end if
14: end for
15: end for
16: return A = {aij}n×n
17: end procedure
```

Listing 1 –Procedure for building MVS similarity matrix

From the consition it is understood that when ***di*** is considered closer to ***dl***, the ***dl*** can still be consideredbeing closer to ***di*** as per MVS. For validation purpose listing 2 is used.

```
Require: 0 < percentage ≤ 1
1: procedure GETVALIDITY(validity,A, percentage)
2: for r ← 1 : c do
3: qr ← _percentage × nr
4: if qr = 0 then _ percentage too small
5: qr ← 1
6: end if
7: end for
8: for i ← 1 : n do
9: {aiv[1], . . . , aiv[n] } ←Sort {ai1, . . . , ain}
10: s.t. aiv[1] ≥ aiv[2] ≥ . . . ≥ aiv[n] {v[1], . . . , v[n]} ← permute {1, . . . , n}
11: r ← class of di
12: validity(di) ← |{dv[1], . . . , dv[qr] } ∩ Sr| qr
13: end for
14: validity ← _ni←1 validity(di)n
15: return validity
16: end procedure
```

Listing 2 –Procedure for get validity score

The final validity is calculated by averaging overall the rows of A as given in line 14. When the validity score is higher, the suitability is more for clustering. The validity scores of Cosine Simialirty (CS) and MVS are presented in fig. 1.

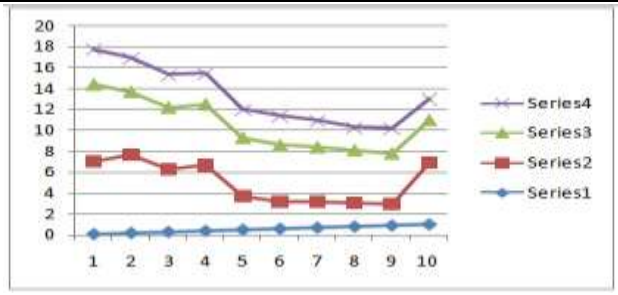

Fig. 1 – Validity test of CS and MVS

Here series 1 corresponds to reutors – 7 CS; series 2 corresponds to reutors-7 MVS; series 3 corresponds to klb-CS; and series 4 corresponds to klb-MVS. The validitity scores of CS and MVS are shown in fig. 1. In the validity test as per the results shown in fig. 1, MVS is better than that of CS.

```
1: procedure INITIALIZATION
2: Select k seeds s1, . . . , sk randomly
3: cluster[di] ← p = argmaxr{strdi}, ∀i = 1, . . . , n
4: Dr ← _di∈Srdi, nr ← |Sr|, ∀r = 1, . . . , k
5: end procedure
6: procedure REFINEMENT
7: repeat
8: {v[1 : n]} ← random permutation of {1, . . ., n}
9: for j ← 1 : n do
10: i ← v[j]
11: p ← cluster[di]
12: ΔIp ← I(np − 1,Dp − di) − I(np,Dp)
13: q ← arg max r,r_=p{I(nr+1,Dr+di)−I(nr,Dr)}
14: ΔIq ← I(nq + 1,Dq + di) − I(nq,Dq)
15: if ΔIp +ΔIq >0 then
16: Move di to cluster q: cluster[di] ← q
17: Update Dp, np,Dq, nq
18: end if
19: end for
20: until No move for all n documents
21: end procedure
```

Listing 3 –Algorithm for incremental clustering

The algorithm provided in listing 3 has two phases known as initialization and refinenement. Initialization is the process of selecting ***k*** documents as seeds and forming initial positions while refinement has number of iterations.In each iteration***n*** number of documents are randomly visited. A verification is done for each document to find whether moving it to a cluster increases objective function. If improvement is not estimated, the object is not moved to the cluster else it is moved to the cluster that provides highest improvement. This process is terminated when iteration finds no document to be moved to new clusters.

V. Performance Evaluation Of Mvs

As part of the performance evaluation, the comparison is made between MVSC Ir, MVSC Iv with existing algorithms. The document database, data corpora, has benchmark datasets for clustering purposes. These benchmark datasets details are given 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

To demonstrate MVSCs we compared them with 5 other clustering algorithms. All the clustering algorithms used in evaluation are:

- MVSC Ir : MVSC with criterion function Ir
- MVSC Iv : MVSC with criterion function Iv
- K-means : conventional k-means with Ecludean distance
- Spkmeans: Spherical k-means with CS
- graphCS : CLUTO's graph method with CS
- graphEJ: CLUTO's graph with extended Jaccard
- MMC: Min Max Cut algorithm

VII. Results

The experimental results are shown in fig. 2 and fig. 3 for all clustering algorithms using 20 bench mark document databases. As the results are not fit into one graph they are split into two graphs and eash graph shows results with 10 datasets.

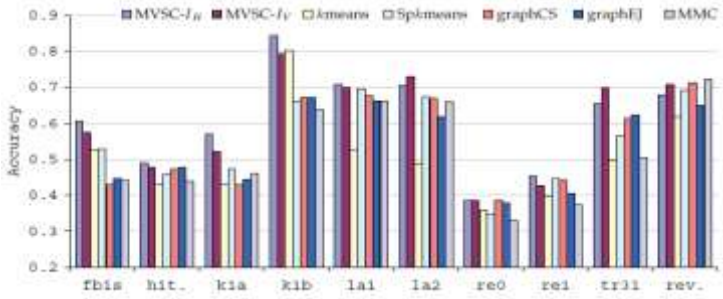

Fig. 2 (a) Experimental Results for first 10 datasets

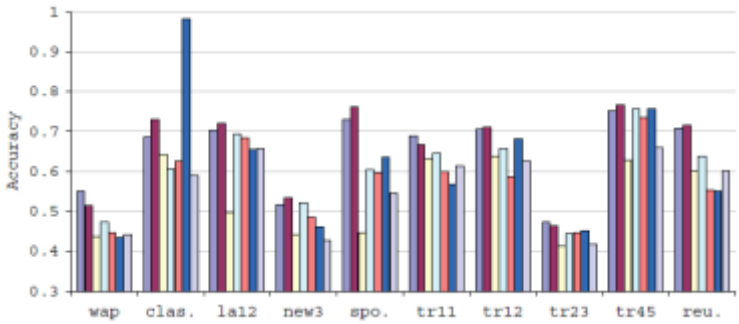

Fig. 2 (b) : Experimental results for next 10 datasets

As can be seen in fig. 2 (a) and fig. 2 (b), it is evident that with respect to many data sets MVSC is performing better. In some cases only other algorithms like graphEJ performed well. Both MVSC Ir and MVSC Iv outperform many other existing algorithms in most of the cases. As part of experiments we also present the effect of α on the performance of MVSC Ir.

The Effect Of α On The Performance Of MVSC Ir

Cluster size and balance have impact on the partitional clustering methods that are based on criterion functions. Based on the clustering results in Accuracy, FScore and NMI, this assessment is done. The results are as shown in fig. 3.

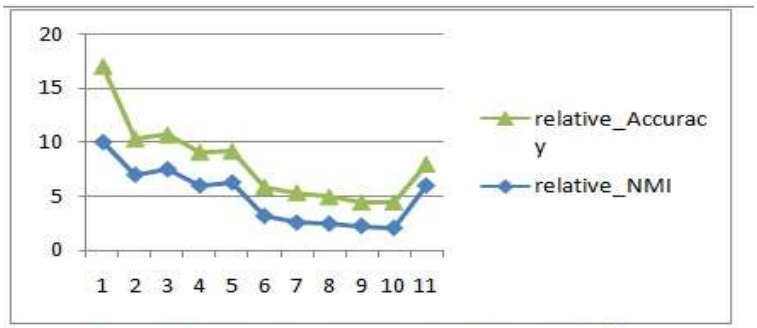

Fig. 3 - Performance of MVSC Ir with respect to α

As can be seen in fig. 3, MVSR Ir's performance worst at 0 and 1 while it has significant performance improvement in the middle. MVSR Ir performs within 5% of the best case with respect to any type of evaluation metrics.

VIII. Conclusion

In this paper we proposed a new similarity measure known as MVS (Multi-Viewpoint based Similarity). When it is compared with cosine similarity, MVS is more useful for finding the similarity of text documents. The empirical results and analysis revealed that the proposed scheme for similarity measure is efficient and it can be used in the real time applications in the text mining domain. IR and IV are the two criterion functions proposed based on MVS. Their respective clustering algorithms are also introduced. The proposed scheme is tested with large datasets with various evolution metrics. The results reveal that the clustering algorithm provides performance that is better than many state – of – the - art clustering algorithms. Similarity measure from multiple viewpoints is the main contrition of this paper. The paper also provides partitioned clustering that can be applied on documents. The future work is that the proposed algorithms can be altered and applied to hierarchical clustering. Our novel approach to measure document similarity is described in the following sections.

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.

Gaddam Saidi Reddy(M.Tech) is student of DRK institute of science and technology, Hyderabad, AP, INDIA. He has received B.Tech Degree computer science and engineering&M.Tech Degree in computer science and engineering. His main research interest includes data mining. Cloud computing.

Dr.R.V.Krishnaiah(Ph.D) is working as Principal at DRK INSTITUTE OF SCINCE & TECHNOLOGY, Hyderabad, AP, INDIA. He has received M.TechDegree(EIE&CSE). His main research interest includes Data Mining, Software Engineering.