A NOVEL K-VARIANT ALGORITHM FOR DOCUMENT CLUSTERING

K. Venkata Ratnam[1], H. Devaraju[2], Y. Ramesh Kumar[3]

[1] Final M.Tech Student ,Dept of Computer Science and Engineering, Avanthi Institute of Engineering and Technology (JNTUK) Cherukupally,Vizianagaram Dist ,Andhra pradesh, India.
[2] Assistant professor, ,Dept of Computer Science and Engineering, Avanthi Institute of Engineering and Technology (JNTUK) Cherukupally,Vizianagaram Dist ,Andhra pradesh, India.
[3]Head of the department, ,Dept of Computer Science and Engineering, Avanthi Institute of Engineering and Technology (JNTUK) Cherukupally,Vizianagaram Dist ,Andhra pradesh, India.

Abstract

Now a days most of the traditional clustering mechanisms based on linear space. Relation exists between the pair data objects either implicitly or explicitly. In the traditional mechanism uses a single view point, In this paper we proposes a novel mechanism for multiview point (i.e. n –dimensional space) with different similarity measure. Using the multiple viewpoints, more informative assessment of similarity can be achieved. Different mechanisms used for efficient clustering mechanisms.

Keywords: *N-Dimensional Space, Document clustering, Similarity Measure.*

---***---

1. INTRODUCTION

Clustering is one of the most interesting and important topics in data mining. The aim of clustering is to find intrinsic structures in data, and organize them into meaningful subgroups for further study and analysis. There have been many clustering algorithms published every year. They can be proposed for very distinct research fields, and developed using totally different techniques and approaches. Nevertheless, according to a recent study, more than half a century after it was introduced, the simple algorithm k-means still remains as one of the top 10 data mining algorithms nowadays. It is the most frequently used partition clustering algorithm in practice. Another recent scientific discussion states that k-means is the favourite algorithm that practitioners in the related fields choose to use. Needless t mention, k-means has more than a few basic drawbacks, such as sensitiveness to initialization and to cluster size, and its performance can be worse than other state-of-the –art algorithms in many domains. In spite of that, its simplicity, understandability and scalability are the reasons for its tremendous popularity. An algorithm with adequate performance and usability in most of application scenarios could be preferable to one with better performance in some cases but limited usage due to high complexity. While offering reasonable results, k-means is fast and easy to combine with other methods in larger systems.

A common approach to the clustering problem is to treat it as an optimization process. An optimal partition is found by optimizing a particular function of similarity among data. Basically, there is an implicit assumption that the true intrinsic structure of data could be correctly described by the similarity formula defined and embedded in the clustering criterion function. Hence, effectiveness of clustering algorithms under this approach depends on the appropriateness of the similarity measure to be data at hand. For instance, the original k-means has sum-of –squared –error objective function that uses Euclidean distance. In a vey sparse and high dimensional domain like text documents, spherical k-means, which uses cosine similarity instead of Euclidean distance as the measure, is deemed to be more suitable.

In [5], Banerjee et al. showed that Euclidean distance was indeed one particular form of a class of distance measures called Bregman divergences. They proposed Bregman hard-clustering algorithm, in which any kind of the Bregman divergences could be applied. Kullback- Leibler divergence was a special case of Bregman divergences that was said to give good clustering results on document datasets. Kullback-Leibler divergence is a good example of non-symmetric measure. Also on the topic of capturing dissimilarity in data, Pakalska et al.[6] found that the discriminative power of some distance measures could increase when their non-Euclidean and non-metric attributes were increased. They concluded that non-Euclidean and non-metric measures could be informative for statistical learning of data. In [7], Pelillo even argued that the symmetry and non-negativity assumption of similarity measures was actually a limitation of current state-of-the-art clustering approaches.

Simultaneously, clustering still requires more robust dissimilarity or similarity measures; recent works such as [8] illustrate this need.

2. LITERATURE SURVEY

The principle definition of clustering is to arrange data objects into separate clusters such that the intra-cluster similarity as well as the inter-cluster dissimilarity is maximized. The problem formulation itself implies that some forms of measurement are needed to determine such similarity or dissimilarity. There are many state-of-theart clustering approaches that do not employ any specific form of measurement, for instance, probabilistic model based method [9], non-negative matrix factorization [10], information theoretic co-clustering [11] and so on. In this paper, though, we primarily focus on methods that\ indeed do utilize a specific measure. In the literature, Euclidean distance is one of the most popular measures:

$$Dist\left(d_i, d_j\right) = \|d_i - d_j\|$$

It is used in the traditional k-means algorithm. The objective of k-means is to minimize the Euclidean distance between objects of a cluster and that cluster's centroid:

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

However, for data in a sparse and high-dimensional space, such as that in document clustering, cosine similarity is more widely used. It is also a popular similarity score in text mining and information retrieval [12]. Particularly, similarity of two document vectors di and dj , Sim(di, dj), is defined as the cosine of the angle between them. For unit vectors, this equals to their inner product:

$$Sim\left(d_i, d_j\right) = \cos\left(d_i, d_j\right) = d_i^t d_j \qquad (3)$$

Cosine measure is used in a variant of k-means called spherical k-means [3]. While k-means aims to minimize Euclidean distance, spherical k-means intends to maximize the cosine similarity between documents in a cluster and that cluster's centroid:

$$\max \sum_{r=1}^{k} \sum_{d_i \in S_r} \frac{d_i^t C_r}{\|C_r\|} \qquad (4)$$

3. PROPOSED SYSTEM

In this novel approach Initially we calculate the weights of the documents and the respective multi view point similarity matrix can be constructed and then cosine similarity can calculated for the keywords in the document with the help of weight calculated for respective documents and then incremental clustering mechanism can be applied for the documents.

3.1 Our novel similarity measure:

The cosine similarity in Eq. (3) can be expressed in the following form without changing its meaning: *Sim*(*di*, *dj*) = cos(*di*−0, *dj*−0) = (*di*−0)*t* (*dj*−0) where 0 is vector 0 that represents the origin point. According to this formula, the measure takes 0 as one and only reference point. The similarity between two documents *di* and *dj* is determined w.r.t. the angle between the two points when looking from the origin.

3.2 MVS Similarity matrix:

We present analytical study to show that the proposed MVS could be a very effective similarity measure for data clustering. In order to demonstrate its advantages, MVS is compared with cosine similarity (CS) on how well they reflect the true group structure in document collections.

procedure BUILDMVSMATRIX(A)
  for $r \leftarrow 1 : c$ **do**
    $D_{S \setminus S_r} \leftarrow \sum_{d_i \notin S_r} d_i$
    $n_{S \setminus S_r} \leftarrow |S \setminus S_r|$
  end for
  for $i \leftarrow 1 : n$ **do**
    $r \leftarrow$ class of d_i
    for $j \leftarrow 1 : n$ **do**
      if $d_j \in S_r$ **then**
        $a_{ij} \leftarrow d_i^t d_j - d_i^t \frac{D_{S \setminus S_r}}{n_{S \setminus S_r}} - d_j^t \frac{D_{S \setminus S_r}}{n_{S \setminus S_r}} + 1$
      else
        $a_{ij} \leftarrow d_i^t d_j - d_i^t \frac{D_{S \setminus S_r} - d_j}{n_{S \setminus S_r} - 1} - d_j^t \frac{D_{S \setminus S_r} - d_j}{n_{S \setminus S_r} - 1} + 1$
      end if
    end for
  end for
  return $A = \{a_{ij}\}_{n \times n}$
end procedure

Fig. 1. Procedure: Build MVS similarity matrix.

To further justify the above proposal and analysis, we carried out a validity test for MVS and CS. The purpose of this test is to check how much a similarity measure coincides with the true class labels. It is based on one principle: if a similarity measure is appropriate for the clustering problem, for any of a document in the corpus, the documents that are closest to it based on this measure should be in the same cluster with it.

3.3 A Novel K-Variant Algorithm

Consists of a number of iterations. During each iteration, the *n* documents are visited one by one in a totally random order. Each document is checked if its move to another cluster results in improvement of the objective function. If yes, the document is moved to the cluster that leads to the highest improvement. If no clusters are better than the current cluster, the document is not moved. The clustering process terminates when an iteration completes without any documents being moved to new clusters. Unlike the traditional *k*-means, this algorithm is a stepwise optimal procedure. While *k*means only updates after all *n* documents have been reassigned, the incremental clustering algorithm updates immediately whenever each document is moved to new cluster. Since every move when happens increases the objective function value, convergence to a local optimum is guaranteed. During the optimization procedure, in each iteration, the main sources of computational cost are Searching for optimum clusters to move individual documents to: $O(nz \cdot k)$. • Updating composite vectors as a result of such moves: $O(m \cdot k)$. where *nz* is the total number of non-zero entries in all document vectors. Our clustering approach is partitional and incremental; therefore, computing similarity matrix is absolutely not needed. If τ denotes the number of iterations the algorithm takes, since *nz* is often several tens times larger than *m* for document domain, the computational complexity required for clustering with *IR* and *IV* is $O(nz \cdot k \cdot \tau)$.

3.4. Fitness Function

For Each and every iteration ,Fitness score can be calculated by placing the documents in the clusters, if the next move has the Optimal fitness values than the previous fitness value of the respective cluster up to number of iterations and Process continues until the specified number of iterations or the consecutive fitness values occurred.

procedure INITIALIZATION
  Select k seeds $s_1, \ldots, s_k$ randomly
  $cluster[d_i] \leftarrow p = \arg\max_r \{s_r^t d_i\}, \forall i = 1, \ldots, n$
  $D_r \leftarrow \sum_{d_i \in S_r} d_i,\ n_r \leftarrow |S_r|, \forall r = 1, \ldots, k$
end procedure
procedure REFINEMENT
  repeat
    $\{v[1:n]\} \leftarrow$ random permutation of $\{1, \ldots, n\}$
    for $j \leftarrow 1 : n$ **do**
      $i \leftarrow v[j]$
      $p \leftarrow cluster[d_i]$
      $\Delta I_p \leftarrow I(n_p - 1, D_p - d_i) - I(n_p, D_p)$
      $q \leftarrow \arg\max_{r, r \neq p} \{I(n_r + 1, D_r + d_i) - I(n_r, D_r)\}$
      $\Delta I_q \leftarrow I(n_q + 1, D_q + d_i) - I(n_q, D_q)$
      if $\Delta I_p + \Delta I_q > 0$ **then**
        Move d_i to cluster q: $cluster[d_i] \leftarrow q$
        Update D_p, n_p, D_q, n_q
      end if
    end for
  until No move for all n documents
end procedure

4. EXPERIMENTAL ANALYSIS

It ecperimentally proved that the vectorized document can be withe respect to their localal frequencies,global frequiencies and relative frequencies as follows.

Vectorized Documents

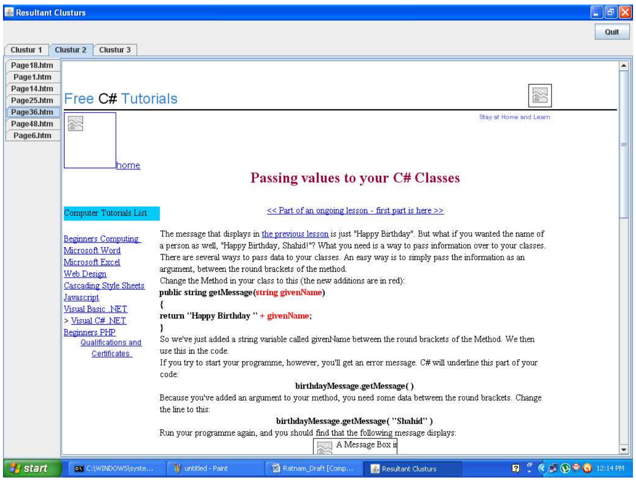
Compute Document Weights

Documents: Page0.htm, Page1.htm, Page10.htm, Page11.htm, Page12.htm, Page13.htm, Page14.htm, Page15.htm, Page16.htm, Page17.htm, Page18.htm, Page19.htm, Page2.htm, Page20.htm, Page21.htm, Page22.htm, Page23.htm, Page24.htm, Page25.htm, Page26.htm, Page27.htm, Page28.htm, Page29.htm, Page3.htm, Page30.htm, Page31.htm, Page32.htm, Page33.htm

Word	Local Freq	Global Freq	Relative Freq
	57	3112	1584.5
<	3	117	60.0
<<	1	24	12.5
>	1	103	52.0
>	1	46	23.5
A	26	771	398.5
ACCESS	1	12	6.5
ADD	20	236	128.0
ADDED	1	19	10.0
ADDING	4	24	14.0
ADDS	1	5	3.0
AFTER	5	71	38.0
ALL	5	118	61.5
ALMOST	1	5	3.0
ALPHABETICALLY	2	2	2.0
ALSO	3	24	13.5
AN	13	150	81.5
AND	17	563	290.0
ANY	2	30	16.0
ANYTHING	1	16	8.5
APPEAR	1	32	16.5
ARE	4	95	49.5
ARRAY	7	119	63.0
ARRAYLIST	19	20	19.5
ARRAYLIST	1	1	1.0
ARRAYLISTS	4	4	4.0
ARRAYLISTS	1	1	1.0
ARRAYPOS	2	2	2.0
ARRAYS	3	41	22.0
AS	3	116	59.5
AT	5	169	87.0

Vectorized Documents

After the clusterization ,documents can be clusterized based on fitness function with incremental algorithm,they are as follows.



Clusters

CONCLUSION

In this proposed mechanism of multi view point clusterization Theoretical analysis and empirical examples show that MVS is potentially more suitable for text documents than the popular cosine similarity. This novel move mechanism on documents with respect to the clusters shows efficient results then the single view point Clustering mechanisms.

REFERENCES

[1] 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.

[2] I. Guyon, U. von Luxburg, and R. C. Williamson, “Clustering: Science or Art?” *NIPS’09 Workshop on Clustering Theory*, 2009.

[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] S. Zhong, “Efficient online spherical K-means clustering,” in *IEEE IJCNN*, 2005, pp. 3180–3185.

[5] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, “Clustering with Bregman divergences,” *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, Oct 2005.

[6] E. Pekalska, A. Harol, R. P. W. Duin, B. Spillmann, and H. Bunke, “Non-Euclidean or non-metric measures can be informative,” in *Structural, Syntactic, and Statistical Pattern Recognition*, ser. LNCS, vol. 4109, 2006, pp. 871–880.

[7] M. Pelillo, “What is a cluster? Perspectives from game theory,” in *Proc. of the NIPS Workshop on Clustering Theory*, 2009.

[8] D. Lee and J. Lee, “Dynamic dissimilarity measure for support based clustering,” *IEEE Trans. on Knowl. and Data Eng.*, vol. 22, no. 6, pp. 900–905, 2010.

[9] 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.

[10] W. Xu, X. Liu, and Y. Gong, “Document clustering based on nonnegative matrix factorization,” in *SIGIR*, 2003, pp. 267–273.

[11] I. S. Dhillon, S. Mallela, and D. S. Modha, “Information-theoretic co-clustering,” in *KDD*, 2003, pp. 89–98.

[12] C. D. Manning, P. Raghavan, and H. Sch ¨ utze, *An Introduction to Information Retrieval*. Press, Cambridge U., 2009.

[13] C. Ding, X. He, H. Zha, M. Gu, and H. Simon, “A min-max cut algorithm for graph partitioning and data clustering,” in *IEEE ICDM*, 2001, pp. 107–114.

[14] 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.

[15] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 888–905, 2000.

[16] I. S. Dhillon, “Co-clustering documents and words using bipartite spectral graph partitioning,” in *KDD*, 2001, pp. 269–274.

[17] Y. Gong and W. Xu, *Machine Learning for Multimedia Content Analysis*. Springer-Verlag New York, Inc., 2007.

BIOGRAFIES

K. Venkata Ratnam was born in **Visakhapatnam**, Andhra Pradesh, India. She received B.Tech in C.S.E from JNTU University, Hyderabad, Andhra Pradesh, India. After having 5 years of Teaching and software programmer she is pursuing M.Tech in C.S.E from Avanthi Institute of Engg and echnology,Vizianagaram, Andhra Pradesh, India. Her Research interest Data Mining and Data Warehousing.

Mr H. Devaraju is MCA and M.Tech(CSE) from Nagarjuna university , Andhra Pradesh, India. He is working as Assistant professor in Computer Science & Engineering department in Avanthi Institute of Engineering and Technology (JNTUK) Cherukupally, Vizianagaram Dist, Andhra Pradesh, India. He has 5 years of experience in teaching Computer Science and Engineering related subjects . He has guided more than 15 students of Bachelor degree, 25 Students of Master degree in Computer Science and Engineering in their major projects. He can be reached at **devaraju.mtech@gmail.com**.

Y. Ramesh Kumar obtained his M.Sc (Computer Science) degree from Andhra University. Later he obtained his M.Tech (CST) degree from Andhra University. Presently he is working as Associate Professor and Head of the Department (CSE) at Avanthi Institute of Engineering and Technology, Cherukupally, Vizianagaram Dist. His Research interest includes **Ontology-based Information Extraction** based on Web search and mining in Data Mining and Data are housing