International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol. 3, Issue. 5, Sep - Oct. 2013 pp-2823-2826 ISSN: 2249-6645

A Novel Clustering Method for Similarity Measuring in Text Documents

Preethi Priyanka Thella[1], G. Sridevi[2]
[1]M.Tech, Nimra College of Engineering & Technology, Vijayawada, A.P., India.
[2]Assoc.Professor, Dept.of CSE, Nimra College of Engineering & Technology, Vijayawada, A.P., India.

ABSTRACT: *Clustering is the process of grouping data into subsets in such a manner that identical instances are collected together, while different instances belong to different groups. The instances are thereby arranged into an efficient depiction that characterizes the populace that is being sampled. A general move towards the clustering process is to treat it as an optimization process. A best partition is found by optimizing an exacting function of similarity, or distance, among data. Basically, there is a hidden assumption that the true inherent structure of data could be correctly describe by using the similarity formula defined and fixed in the clustering decisive factor. In this paper, we introduce clustering with multi- view points based on different similarity measures. The multi- view point approach to learning is one in which we have 'views' of the data (sometimes in a rather abstract sense) and the goal is to use the relationship between these views to alleviate the difficulty of a learning problem of interest.*

Keywords: *Clustering, Text mining, Similarity measure, View point.*

I. INTRODUCTION

Clustering[1] or cluster analysis is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of explorative data mining techniques, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Cluster analysis itself is not one specific algorithm or procedure, but the general task to be solved. It can be achieved by using various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with low distances among the cluster members, intervals or particular statistical distributions, dense areas of the data space. Clustering can therefore be formulated as a Multi- objective optimization process.

The appropriate clustering algorithm and parameter settings, including values such as the distance function to use, a density threshold or the number of expected clusters, depend on the individual data set and intended use of the results. Clustering as such is not an automatic task, but an iterative process of Knowledge discovery or interactive multi- objective optimization that involves trial and failure. It will often be necessary to modify parameters and preprocessing until the result achieves the desired properties. Cluster analysis can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering process could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects or items which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Figure 1 shows clustering process.

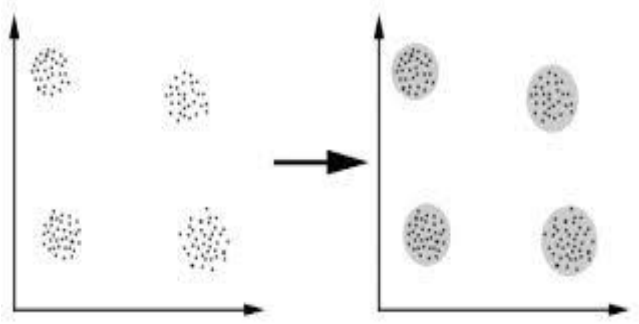

Figure 1: Clustering Process

In this case we easily identify the four clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called as distance based clustering. Another kind of clustering is called conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to the simple similarity measures. The multi-view point approach to learning is one in which we have 'views' of the data (sometimes in a rather abstract sense) and the goal is to use the relationship between these views to alleviate the difficulty of a learning problem of interest.

II. RELATED WORK

Text clustering is required in the real world applications such as web search engines. It comes under text mining process. It is meant for grouping text documents into various clusters. These clusters are used by various applications in the real world, for example, search engines. A text document is treated as an object a word in the document is referred as a term. A vector is built to represent each text document. The total number of terms in the text 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 text document clustering. They include probabilistic based methods [2], nonnegative matrix factorization [3] and information theoretic co-clustering [4]. These approaches are not using a particular measure for finding similarity among text documents. In this paper, we make use of multi- view point similarity measure for finding the similarity. As found it literature, a measure widely used in text document clustering is ED (Euclidian Distance).

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

K-Means algorithm is most widely used clustering algorithm due to its ease of use and simplicity. Euclidian distance 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 follows:

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

Another similarity measure being used for text document mining is cosine similarity measure. It is best useful in hi-dimensional documents [5]. This measure is also being used in Spherical K-Means which is a variant of K-Means algorithm. The difference between the two flavors of K-Means algorithm 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 very popular. It considers the text document corpus as graph and uses min-max cut algorithm which represents centriod as follows:

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

There is a software package called CLUTO [6] which is meant for document clustering. It makes use of the graph partitioning approach. Based on the nearest neighbor graph it builds, it text documents are clustered. It is based on the Jacquard coefficient which is computed as follows:

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

Jacquard coefficients use both magnitude and direction which is not the case with Euclidian distance and cosine similarity. However, it is similarity to cosine similarity when the documents are represented as unit vectors. In [7] there is comparison between the two techniques namely Jacquard and Pearson correlation. It also concludes that both of them are best used in clustering process of web documents. For tsxt document clustering other approaches can be used which are phrase based and concept based. In phrase based approach is found while in [8] tree similarity based approach is found. The common procedure used by both of them is "Hierarchical agglomerative Clustering". The drawback of these approaches is that their computational cost is too high. For clustering XML documents also there are some measures. One such measure is called "Structural Similarity" which differs from text document clustering. This paper focuses on a new multi-view point based similarity measure for text clustering.

III. PROPOSED WORK

In proposed work, 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 text document clustering. As per our approach the similarity between two documents is calculated as follows:

$$\underset{d_i, d_j \in S_r}{sim(d_i, d_j)} = \frac{1}{n - n_r} \sum_{d_h \in S \setminus S_r} sim(d_i - d_h, d_j - d_h)$$

Consider two point "di" and "dj" in the 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 the cosine angle between those points with respect to Euclidean 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 very small, then the chances are higher that the "dh" is in the same cluster. Though various viewpoints are useful in increasing the accuracy of the similarity measure there is a possibility of having that give negative result. However the possibility of such a drawback can be ignored provided plenty of documents to be clustered.

Now we have to carry out the validity test for the cosine similarity and multi view based similarity as follows. For each type of the similarity measure, a similarity matrix called A = {aij}n×n is created. For CS, this is very simple, as aij = dti dj . The algorithm for building Multi view Similarity (MVS) matrix is described in Algorithm 1.

ALGORITHM 1: BUILDMVSMATRIX(*A*)

Step 1: for r ← 1 : c do

Step 2: DS \Sr ← $\sum_{d_i \notin S_r} d_i$

Step 3: nS \Sr←|S \ Sr|

Step 4: end for

Step 5: for i ← 1 : n do

Step 6: r ← class of di

Step 7: for j ← 1 : n do

Step 8: if dj ∈ Sr then

Step 9: $$a_{ij} \leftarrow d_i^t d_j - d_i^t \frac{D_{S\setminus S_r}}{n_{S\setminus S_r}} - d_j^t \frac{D_{S\setminus S_r}}{n_{S\setminus S_r}} + 1$$

Step 10: else

Step 11: $$a_{ij} \leftarrow d_i^t d_j - d_i^t \frac{D_{S\setminus S_r} - d_j}{n_{S\setminus S_r} - 1} - d_j^t \frac{D_{S\setminus S_r} - d_j}{n_{S\setminus S_r} - 1} + 1$$

Step 12: end if

Step 13: end for

Step 14: end for

Step 15: return A = {aij}n×n

First, the outer composite with respect to each class is determined. Then, for each row **a**i of "A", i = 1, . . . , n, if the pair of text documents di and dj, j = 1, . . . , n are in the same class, aij is calculated as in line 9. Otherwise, dj is assumed to be in di's class, and aij is calculated as shown in line 11.

After matrix "A" is formed, the code in Algorithm 2 is used to get its validity score:

ALGORITHM 2: GETVALIDITY(*validity,A, percentage*)

Step 1: for r ← 1 : c do

Step 2: qr ← floor(percentage × nr)

Step 3: if qr = 0 then

Step 4: qr ← 1

Step 5: end if

Step 6: end for

Step 7: for i ← 1 : n do

Step 8: {aiv[1], . . . , aiv[n] } ←Sort {ai1, . . . , ain}

Step 9: s.t. aiv[1] ≥ aiv[2] ≥ . . . ≥ aiv[n]

{v[1], . . . , v[n]} ← permute {1, . . . , n}

Step 10: r ← class of di

Step 11: $$validity(d_i) \leftarrow \frac{|\{d_{v[1]}, \ldots, d_{v[q_r]}\} \cap S_r|}{q_r}$$

Step 12: end for

Step 13: $$validity \leftarrow \frac{\sum_{i\leftarrow 1}^{n} validity(d_i)}{n}$$

Step 14: return validity

For each document "di" corresponding to row "**a**i" of matrix A, we select "qr" documents closest to point "di". The value of "qr" is chosen relatively as the percentage of the size of the class r that contains "di", where percentage ∈ (0, 1]. Then, validity with respect to "di" is calculated by the fraction of these "qr" documents having the same class label with "di", as shown in line 11. The final validity is determined by averaging the over all the rows of matrix A, as shown in line 13. It is clear that the validity score is bounded within values 0 and 1. The higher validity score a similarity measure has, the more suitable it should be useful for the clustering process.

IV. INCREMENTAL CLUSTERING ALGORITHM

The main goal of this algorithm is to perform text document clustering by optimizing I_R and I_V as shown below:

$$I_R = \sum_{r=1}^{k} \frac{1}{n_r^{1-\alpha}} \left[\frac{n+n_r}{n-n_r} \|D_r\|^2 - \left(\frac{n+n_r}{n-n_r} - 1 \right) D_r^t D \right]$$

$$I_V = \sum_{r=1}^{k} \left[\frac{n + \|D_r\|}{n - n_r} \|D_r\| - \left(\frac{n + \|D_r\|}{n - n_r} - 1 \right) \frac{D_r^t D}{\|D_r\|} \right]$$

With this general form, the incremental optimization algorithm, which has two major steps Initialization and Refinement, is shown in Algorithm 3 and Algorithm 4.

ALGORITHM 3: INITIALIZATION

Step 1: Select k seeds s1, . . . , sk randomly

Step 2: $cluster[d_i] \leftarrow p = \arg\max_r \{s_r^t d_i\}, \forall i = 1, \ldots, n$

Step 3: $D_r \leftarrow \sum_{d_i \in S_r} d_i,\ n_r \leftarrow |S_r|, \forall r = 1, \ldots, k$

Step 4: end

ALGORITHM 4: REFINEMENT

Step 1: repeat

Step 2: {v[1 : n]} ← random permutation of {1, . . ., n}

Step 3: for j ← 1 : n do

Step 4: i ← v[j]

Step 5: p ← cluster[di]

Step 6: $\Delta I_p \leftarrow I(n_p - 1, D_p - d_i) - I(n_p, D_p)$

Step 7: $q \leftarrow \arg\max_{r, r \neq p} \{I(n_r + 1, D_r + d_i) - I(n_r, D_r)\}$

Step 8: $\Delta I_q \leftarrow I(n_q + 1, D_q + d_i) - I(n_q, D_q)$

Step 9: if $\Delta I_p + \Delta I_q > 0$ then

Step 10: Move di to cluster q: cluster[di] ← q

Step 11: Update Dp, np,Dq, nq

Step 12: end if

Step 13: end for

Step 14: until No move for all n documents

Step 15: end

At Initialization, "k" arbitrary documents are selected to be the seeds from which initial partitions are formed. Refinement is a process that consists of a number of iterations. During each iteration, the "n" text documents are visited one by one in a totally random order. Each text document is checked if its move to another cluster results in improvement of the objective function. If yes, then the text document is moved to the cluster that leads to the highest improvement. If no clusters are better than the current cluster, the text document is not moved. The clustering process terminates when iteration completes without any text documents being moved to new clusters.

V. CONCLUSION

In the view point of data engineering, a cluster is a group of objects with similar nature. The grouping mechanism is called as clustering process. The similar text documents are grouped together in a cluster, if their cosine similarity measure is less than a specified threshold. In this paper we mainly focuses on view points and we introduce a novel multi-viewpoint based similarity measure for text mining. The nature of similarity measure plays a very important role in the success or failure of the clustering method. From the proposed similarity measure, we then formulate new clustering criterion functions and introduce their respective clustering algorithms, which are fast and scalable like k-means algorithm, but are also capable of providing high quality and consistent performance.

REFERENCES

[1] I. Guyon, U. von Luxburg, and R. C. Williamson, "Clustering: Science or Art?" , || NIPS'09 Workshop on Clustering Theory, 2009.
[2] Leo Wanner (2004). "Introduction to Clustering Techniques". Available online at: http://www.iula.upf.edu/materials/040701wanner.pdf [viewed: 16 August 2012]
[3] 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.
[4] I. Guyon, U. von Luxburg, and R. C. Williamson, "Clustering: Science or Art?" NIPS'09 Workshop on Clustering Theory, 2009.
[5] C. D. Manning, P. Raghavan, and H. Sch ¨ utze, An Introduction to Information Retrieval. Press, Cambridge U., 2009.
[6] 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.
[7] W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization," in SIGIR, 2003, pp. 267–273.
[8] S. Zhong, "Efficient online spherical K-means clustering," in IEEE IJCNN, 2005, pp. 3180–3185.