International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol.3, Issue.4, Jul - Aug. 2013 pp-1906-1909 ISSN: 2249-6645

A Novel Multi- Viewpoint based Similarity Measure for Document Clustering

S. Neelima[1], A. Veerabhadra Rao[2]
[1]*M. Tech(CSE), Sri Sai Madhavi Institute of Science & Technology, A.P., India.*
[2]*Asst. Professor, Dept.of CSE, Sri Sai Madhavi Institute of Science & Technology, A.P., India.*

Abstract: *Data mining is a process of analyzing data in order to bring about patterns or trends from the data. Many techniques are part of data mining techniques. Other mining techniques such as text mining and web mining also exists. Clustering is one of the most important data mining or text mining algorithm that is used to group similar objects together. In other words, it is used to organize the 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 the computational cost. Clustering methods are classified into data partitioning, hierarchical clustering, data grouping. The aim of this paper is to develop a new method that is used to cluster text documents that have sparse and high dimensional data objects. Like k-means algorithm, the proposed algorithm work faster and provide consistent, high quality performance in the process of clustering text documents. The proposed similarity measure is based on the multi-viewpoint.*

Keywords: *Clustering, Euclidean distance, HTML, Similarity measure.*

I. INTRODUCTION

Clustering is a process of grouping a set of objects into classes of similar objects and is the most interesting concept of data mining in which it is defined as a collection of data objects that are similar to one another. Purpose of Clustering is to group fundamental structures in data and classify them into a meaningful subgroup for additional analysis. Many clustering algorithms have been published every year and can be proposed for developing several techniques and approaches. The k-means algorithm has been one of the top most data mining algorithms that is presently used. Even though it is a top most algorithm, it has a few basic drawbacks when clusters are of various sizes. Irrespective of the drawbacks is understandability, simplicity, and scalability is the main reasons that made the algorithm popular. K-means is fast and easy to combine with the other methods in larger systems.

A common approach to the clustering problem is to treat it as the optimization process. An optimal partition is found by optimizing the particular function of similarity among data. Basically, there is an implicit assumption that the true intrinsic structure of the data could be correctly described by the similarity formula defined and embedded in the clustering criterion function. An algorithm with an 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 the other methods in larger systems. The original k-means has sum-of-squared-error objective function that uses the Euclidean distance. In a very sparse and high-dimensional domain like text documents, spherical k- means, which uses cosine similarity (CS) instead of the Euclidean distance as the measure, is deemed to be more suitable [1], [2]. The nature of similarity measure plays a very important role in the success or failure of the clustering methods. Our objective is to derive a novel method for multi viewpoint similarity between data objects in the sparse and high-dimensional domain, particularly text documents. From the proposed method, we then formulate new clustering criterion functions and introduce their respective clustering algorithms, which are fast and scalable like k-means.

II. RELATED WORK

Document clustering is one of the important text mining techniques. It has been around since the inception of the text mining domain. It is the 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 the 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 the weighting schemes like TF-IDF (Term Frequency –Inverse Document Frequency). Many approaches came into existence for the document clustering. They include the information theoretic co-clustering [3], non – negative matrix factorization, probabilistic model based method [4] and so on. However, these approaches did not use specific measure in finding the document similarity. In this paper we consider methods that specifically use the certain measurement. From the literature it is found that one of the popular measures is the Eucludian distance:

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

K-means is one of the important clustering algorithms in the world. It is in the list of top 10 clustering algorithms. Due to its simplicity and ease of use it is still being used in the data 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 the Euclidian measurement, between objects in clusters. The centroid of such clusters is represented as follows:

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

In text mining domain, the 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 the spherical ***k***-means. It is mainly used to maximize the cosine similairity between the cluster's centroid and the documents in the cluster. The difference between k-means that uses the 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 the graph partitioning approach. In this approach the document corpus is considered as the graph. Min – max cut algorithm is the one that makes use of this approach and it focuses on minimizing the centroid function:

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

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

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

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

III. PROPOSED WORK

The main work is to develop a novel multi viewpoint based algorithm for document clustering which provides maximum efficiency and performance. It is particularly focused in studying and making the use of cluster overlapping phenomenon to design cluster merging criteria. Proposing a new way to compute the overlap rate in order to improve the time efficiency and ―the veracity is mainly concentrated. Based on the Hierarchical Clustering Method, the usage of the Expectation-Maximization (EM) algorithm in the Gaussian Mixture Model to count the parameters and make the two sub-clusters combined when their overlap is the largest is narrated. In the simplest case, an optimization problem consists of maximizing or minimizing a real function by systematically choosing the input values from within an allowed set and computing the value of the function. The generalization of optimization theory and the techniques to other formulations comprises a large area of applied mathematics.

The cosine similarity can be expressed be expressed as follows:

$$Sim(d_i, d_j) = \cos(d_i - 0, d_j - 0) = (d_i - 0)^t (d_j - 0) \quad \text{---- (5)}$$

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 the two documents is defined as follows :

$$sim(d_i, d_j) = \frac{1}{n - n_r} \sum_{d_i, d_j \in S_r} sim(d_i - d_h, d_j - d_h) \quad \text{---- (6)}$$

The multi view based similarity in equ. (6) depends on particular formulation of the individual similarities within the sum. If the relative similarity is defined by the dot product of the difference vectors, we have:

$$\mathrm{MVS}(d_i, d_j \mid d_i, d_j \in S_r)$$

$$= \frac{1}{n - n_r} \sum_{d_h} \cos(\ d_i - \ d_h, d_j - d_h) \| d_i - \ d_h \| \ \| d_j - d_h \| \qquad \text{---- (7)}$$

The similarity between the two points di and dj inside cluster Sr, viewed from a point dh outside this cluster, is equal to the product of the cosine of the angle between di and dj looking from dh and the Euclidean distances from dh to these two points.

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

procedure 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

Firstly, 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 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 in line 11.

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

procedure 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 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 $\in$ (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 in line 11. The final validity is determined by averaging the over all the rows of A, as in line 13. It is clear that the validity score is bounded within 0 and 1. The higher validity score a similarity measure has, the more suitable it should be for the clustering process.

IV. CONCLUSION

Clustering is one of the data mining and text mining techniques used to analyze datasets by dividing it into various meaningful groups. The objects in the given dataset can have certain relationships among them. All the clustering algorithms assume this before they are applied to datasets. The existing algorithms for the text mining make use of a single viewpoint for measuring similarity between objects. Their drawback is that the clusters cannot exhibit the complete set of relationships among objects. To overcome this drawback, we propose a new similarity measure known as the multi-viewpoint based similarity measure to ensure the clusters show all relationships among objects. This approach makes use of different viewpoints from different objects of the multiple clusters and more useful assessment of similarity could be achieved.

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 Recognition Letters, vol. 28, no. 1, pp. 110-118, 2007

[2] S. Zhong, "Efficient Online Spherical K-means Clustering," Proc. IEEE Int'l Joint Conf. Neural Networks (IJCNN), pp. 3180-3185, 2005

[3] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in KDD, 2003, pp. 89–98.

[4] 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.

[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.

.