High Quality Assessment of Similarity by Using Multiple View Points

C.S Mahaboobbee*1, and M.Venkatesh Naik#2

*Student, Dept of CSE, CRIT, Affiliated to JNTUA University, ANANTAPURAMU, AP, India
# Asst Professor, Dept of CSE, CRIT, Affiliated to JNTUA University, ANANTAPURAMU, AP, India

1sadiyaghouse786@gmail.com
2venkateshnaikm0@gmail.com

Abstract—Data grouping, data partitioning and hierarchical clustering are the three types of well known clustering methods. The data grouping approach is meant for making a set of overlapping clusters. The K-means algorithm, a kind of partitioned clustering 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. Even though it is simple and effective algorithm, it suffers from the drawbacks of low performance, initialization and sensitive to cluster size. Hence always, The process of clustering with highest quality is an optimization process. In order to attain highest quality clusters, we go for Similarity measure approach.

In this paper we propose a multi-view point based similarity measure which 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.

Keywords: Similarity measure, text mining, document clustering

I. INTRODUCTION

Clustering data described by categorical attributes [3] is a challenging task in data mining applications. Unlike numerical attributes, it is difficult to define a distance between pairs of values of the same categorical attribute, since they are not ordered. Clustering is a popular data mining technique that enables to partition data into groups (clusters) in such a way that objects inside a group are similar, and objects belonging to different groups are dissimilar. When objects are described by numerical (real, integer) features, there is a wide range of possible choices. Objects can be considered as vectors in a n-dimensional space, where n is the number of features. Then, many distance metrics can be used in n-dimensional spaces. Clearly, these distance metrics do not distinguish between the different values taken by the attribute, since they only measure the equality between pair of values. This is a strong limitation for a clustering algorithm, since it prevents to capture similarities that are clearly identified by human experts.

Clustering is "unsupervised classification" or "unsupervised segmentation". The aim is to assign instances to classes that are not defined a priori and that are supposed to somehow reflect the underlying structure of the entities that the data represents. Most of the problems encountered with the clustering algorithms involve dealing with the large number of dimensions and a large number of objects becoming prohibitive due to time complexity, the effectiveness of an algorithm depends upon the similarity measure.

II. EXISTING SYSTEM

Document clustering is a form of text mining meant for grouping documents into various clusters. 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 existing document clustering algorithms include probabilistic based methods, nonnegative matrix factorization and information theoretic co-clustering.

The most widely used clustering algorithm [1] is ED-Euclidean distance which is measured as

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

Based on the ease of use and the simplicity, K-Means is most widely used clustering algorithm. ED is the measure used in K-Means algorithm to measure the distance between objects to make them into clusters. The cluster centroid is computed as

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

Cosine similarity measure is another algorithm used in high-dimensional documents. 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 in which it considers the document corpus as a graph and uses min-max cut algorithm which represents centroid as:

$$\text{Min } \sum_{i=1}^{k} \frac{D_i^2 D}{|D_i|^2}$$

Another graph partitioning approach CLUTO documents are clustered based on the nearest neighbor graph

$$\text{Sim}_{\text{nearest}}(u_i, u_j) = \frac{u_i \cdot u_j}{||u_i||^2 + ||u_j||^2 - u_i \cdot u_j}$$

For document clustering other approaches can be used which are phrase based and concept based. 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” which differs from text document clustering.

III. PROPOSED SYSTEM

The implemented work in this paper is based on Multi-viewpoint based similarity measure. It does mean that it uses more than one view point while finding similarity between objects and clustering them into various groups. The similarity between the two documents can be given as

$$\text{Sim} \left( d_i, d_j \right) = \frac{1}{n} \sum_{d_r, d_s \in S_r, d_t, d_u \in S_t} \text{Sim} \left( d_r - d_s, d_t - d_u \right)$$

where 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 procedure for similarity matrix is as given below:

Algorithm 1: Procedure for Similarity matrix

The validity is calculated as an average of all the rows. If validation score is higher, it reflects that the similarity is higher and thus eligible for clustering.

From the above figure it can be clearly observed that the performance of Multi-viewpoint based similarity is higher when compared to Cosine Similarity.

The incremental clustering algorithm for clustering the documents has been implemented in two phases: Refinement and Initialization.

Initialization involves selecting k documents as seeds for making the initial positions. The refinement phase makes each iteration to form best clusters. Each iteration in refinement phase visits n number of documents in random fashion. Once the verification process is done for each document, it is moved to the cluster if the document is considered to be similar. When no documents are there the iterations come to an end.

The following benchmark datasets have been used to test the efficiency of our approach:
The evaluation results are best compared to M-means, Min Max Cut Algorithm, graph EJ 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 Ir the proposed algorithm with Ir iteration.

Table 1: Benchmark documents datasets

<table>
<thead>
<tr>
<th>Data</th>
<th>Source</th>
<th>c</th>
<th>n</th>
<th>m</th>
<th>Balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>fbis</td>
<td>TREC</td>
<td>17</td>
<td>2,463</td>
<td>3,000</td>
<td>0.075</td>
</tr>
<tr>
<td>hitex</td>
<td>TREC</td>
<td>6</td>
<td>2,301</td>
<td>13,170</td>
<td>0.192</td>
</tr>
<tr>
<td>k1a</td>
<td>WebACE</td>
<td>20</td>
<td>2,340</td>
<td>13,859</td>
<td>0.018</td>
</tr>
<tr>
<td>k1b</td>
<td>WebACE</td>
<td>6</td>
<td>3,204</td>
<td>17,273</td>
<td>0.043</td>
</tr>
<tr>
<td>la1</td>
<td>TREC</td>
<td>6</td>
<td>3,075</td>
<td>15,211</td>
<td>0.274</td>
</tr>
<tr>
<td>la2</td>
<td>TREC</td>
<td>7</td>
<td>927</td>
<td>10,127</td>
<td>0.006</td>
</tr>
<tr>
<td>re0</td>
<td>Reuters</td>
<td>13</td>
<td>1,504</td>
<td>2,886</td>
<td>0.018</td>
</tr>
<tr>
<td>re1</td>
<td>Reuters</td>
<td>25</td>
<td>1,657</td>
<td>3,758</td>
<td>0.027</td>
</tr>
<tr>
<td>re2</td>
<td>TREC</td>
<td>6</td>
<td>3,075</td>
<td>15,211</td>
<td>0.274</td>
</tr>
<tr>
<td>reviews</td>
<td>WEBACE</td>
<td>10</td>
<td>2,004</td>
<td>8,440</td>
<td>0.015</td>
</tr>
<tr>
<td>classic</td>
<td>CACM/CS/</td>
<td>4</td>
<td>7,089</td>
<td>12,009</td>
<td>0.323</td>
</tr>
<tr>
<td>la2</td>
<td>TREC</td>
<td>6</td>
<td>1,679</td>
<td>21,604</td>
<td>0.282</td>
</tr>
<tr>
<td>new3</td>
<td>TREC</td>
<td>44</td>
<td>9,558</td>
<td>36,306</td>
<td>0.149</td>
</tr>
<tr>
<td>sports</td>
<td>TREC</td>
<td>7</td>
<td>8,580</td>
<td>18,524</td>
<td>0.256</td>
</tr>
<tr>
<td>tr11</td>
<td>TREC</td>
<td>9</td>
<td>414</td>
<td>6,424</td>
<td>0.045</td>
</tr>
<tr>
<td>tr12</td>
<td>TREC</td>
<td>7</td>
<td>313</td>
<td>5,799</td>
<td>0.067</td>
</tr>
<tr>
<td>tr25</td>
<td>TREC</td>
<td>6</td>
<td>204</td>
<td>9,831</td>
<td>0.067</td>
</tr>
<tr>
<td>tr45</td>
<td>TREC</td>
<td>10</td>
<td>690</td>
<td>8,260</td>
<td>0.080</td>
</tr>
<tr>
<td>reuters7</td>
<td>Reuters</td>
<td>7</td>
<td>2,500</td>
<td>4,977</td>
<td>0.082</td>
</tr>
</tbody>
</table>

The similarity measure 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. The data for experimental evaluation considers benchmark datasets.

IV. CONCLUSION

REFERENCES