Bipartite Graph Energy Based Similarity measure for Document Clustering

G. Hannah Grace, Kalyani Desikan

Abstract: Document clustering is a text mining technique wherein a document collection is divided into significant clusters by making use of a suitable distance or similarity measure. Distance measure plays an important role in document clustering. Here similar content is assigned to the same clusters while dissimilar content is assigned to different clusters. This is achieved by minimizing the intra-cluster distance between documents and maximizing the distance between clusters. A variety of distance measures used in document clustering are Euclidean distance, Squared Euclidean distance, Minkowski distance, Chebychev distance, power distance, percent disagreement, Manhattan distance, Bit-vector distance, comparative-clustering distance, Huffman-code distance and Dominance-based distance. The conditions for the distance between two documents to be a distance measure is as follows: If d_1 and d_2 are any two documents in a set and D(d_1, d_2) is the distance between d_1 and d_2, then the following conditions are satisfied.

1. **Non-negativity**: D(d_1, d_2) ≥ 0. i.e., distance between any two documents must be a value greater than or equal to zero.

2. **Identity of indiscernibles**: D(d_1, d_2) = 0 if and only if d_1 = d_2 i.e., the distance between two documents is zero if and only if the two documents are identical.

3. **Symmetry**: D(d_1, d_2) = D(d_2, d_1), i.e., distance between d_1 and d_2 is equal to distance between d_2 and d_1.

A distance which conforms to at least these three conditions is known as a distance measure. A distance measure which also satisfies the triangle inequality is known as a distance metric. In the context of document clustering, the set of terms in the document set is given by X and Y represents the document set. W=| ω_i| represents the frequency of term i in document j.

Index Terms: Bipartite Graph, Document clustering, Similarity measure, Distance measures.

Figure 1 Bipartite Representation of Documents in G

The adjacency matrix of a bipartite weighted graph G(V,E,W) is represented in a block matrix form as follows

\[
\begin{pmatrix}
0 & \mathbf{W} \\
\mathbf{W}^T & 0
\end{pmatrix}
\]

where \(\mathbf{W}^T\) is the transpose of matrix \(\mathbf{W}\). Let \(A = [a_{ij}]\) be the adjacency matrix of a graph G containing n vertices and m edges. The energy of the graph G is defined as \(E(G) = \sum_{i=1}^{n} |\hat{\lambda}_i|\), where the set of eigen values \(\{\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_n\}\) is known as the spectrum of G [12,13].
Bipartite Graph Energy Based Similarity measure for Document Clustering

Energy of a Bipartite Graph:
Consider the bipartite graph representation[10, 12]of the document corpus containing p terms and N documents given in Figure 1. For a bipartite graph G with n vertices and m edges, it [14] has been proved that the energy of G is given by

\[ E(G) \leq \frac{4m}{n} + \sqrt{(n-2)(2m - \frac{8m^2}{n})} \tag{1} \]

We consider only the upper bound of energy of bipartite graph given in equation (1) as a scaling factor in our similarity measure.

II. LITERATURE REVIEW

We present below few of the research work carried out by researchers in the field of similarity measures on document clustering.

Boriah conducted a comparative study on similarity measures for categorical data in the context of outlier detection (Boriah, 2008). He concluded that similarity measures had an influence on the performance of an outlier detection algorithm. Fernando[9] reviewed and compared similarity measures for categorical data. Deshpande [15] conducted an analysis of genetic interaction networks identifying genes with similar profiles using similarity measures. Strehl [16] recognized the impact of similarity measures on web clustering. Zhang [17] used six similarity measures like Euclidean distance, Principal Component Analysis(PCA) & Euclidean distance, Hausdorff distance, Hidden Markov Models (HMM-distance), Longest common Subsequence (LCSS distance), Dynamic time warping (DTW distance) to measure the similarity in trajectories and compared trajectory clustering in outdoor surveillance. Al Khalifa [18] examined twelve similarity measures like Euclidean distance, Average distance, Weighted Euclidean distance, Chord distance, Mahalanobis distance, Cosine measure, Manhattan distance, Mean character Difference, Index of Association, Canberra measure, Czekanowski coefficient, coefficient of Divergence and Pearson coefficient for clustering, and concluded that no single coefficient is appropriate for all methodologies. Inspite of all these studies, there is no empirical analysis and comparison available for continuous data. In all cases, similarity or distance measures are based on vector representation of documents. Our proposed method is based on graph representation.

III. DISTANCE MEASURES IN DOCUMENT CLUSTERING

The following are the different distance measures that are used in our comparative study.

A. Euclidean Distance Measure
Euclidean distance [2,3] is a standard measure for solving geometrical problems. Euclidean distance is widely used in clustering problems for clustering documents, measuring distance between sets. For a N X p matrix the Euclidean distance \( D(d_i, d_j) \) with \( R^p \) dimensions and documents \( d_i \) and \( d_j \) is defined as

\[ D(d_i, d_j) = \sqrt{\sum_{k=1}^{p} (d_{ik} - d_{jk})^2} \]

where \( d_i \) and \( d_j \) represent the documents with \( i = 1...N \) and \( j = 1...N \).

B. Jaccard distance

The Jaccard distance is also known as Tanimoto coefficient[4]. This measures similarity as the intersection divided by the union of the documents. The Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms. The range of the similarity measure for the Jaccard coefficient is between 0 and 1. The extended Jaccard measure presented by Strehl and Ghosh in the year 2000 [4] is the extension of the original Jaccard measure extended to continuous or discrete non-negative features and is given by

\[ D(d_i, d_j) = \frac{d_i^T d_j}{\|d_i\|^2 + \|d_j\|^2 - d_i^T d_j} \]

where \( d_i \) and \( d_j \) are document vectors.

C. Cosine similarity

The Cosine similarity is calculated by measuring the cosine of the abgle between two document vectors. When documents are represented as term vectors, the similarity between two document sets corresponds to the association between the vectors [2]. This is the cosine of the angle between vectors. Cosine similarity is one of the most popular similarity measures applied to sets in information retrieval applications [5] and clustering [6]. The normalized inner product is an appropriate similarity measure given by,

\[ D(d_i, d_j) = \frac{d_i^T d_j}{\|d_i\| \|d_j\|} \]

where \( d_i \) and \( d_j \) are document vectors over the term set \{t1, t2,... tp\} each dimension presents a term with its weight in the document which is always non negative. The \( d_i^T \) is the transpose of the \( i^{th} \) document. The \( \|\|| \) (norm) in the formula helps to scale the results.

D. Canberra distance

The Canberra distance [4] is a measure that is often used for data scattered around an origin. The absolute distance between the variables of the two documents is divided by the sum of the absolute variables prior to summing. The generalized form is given as

\[ D(d_i, d_j) = \sum_{k=1}^{p} \frac{|d_{ik} - d_{jk}|}{|d_{ik}| + |d_{jk}|} \]

E. Manhattan distance

The distance between two documents measured along axes at right angles is the Manhattan distance. In other words it is the sum of differences in each variable:
produced a new distance measure. The distance measure we introduced is called Bipartite Graph energy Based similarity measure(BGES) and it is based on the bipartite representation of documents. The similarity between documents is determined by a distance measure based on graph based representation of documents.

Motivated by the laws of physics for energy and coulombs law, we have introduced a new similarity measure given by

$$S(d_i, d_j) = \frac{E \ast r^2}{q_i q_j}$$

where $E$ is the energy obtained from equation (1), $q_i$ and $q_j$ are the number of terms in documents $i$ and $j$, $r$ is the intersection of words between both the documents $d_i, d_j$.

A. BGEBS as a distance measure
We prove that the BGEBS measure satisfies the distance measure properties: BGEBS is given by

$$S(d_i, d_j) = \frac{E \ast r^2}{q_i q_j}$$

After Normalization with $E$, the total energy of the bipartite graph, we get

$$N(d_i, d_j) = \frac{S(d_i, d_j)}{E} = \frac{r^2}{q_i q_j}$$

Expressed in terms of distance, we have $N(d_i, d_j) ≤ 1$.

To prove the first two properties of distance measure i.e., $D(d_i, d_j) ≥ 0$ we equivalently show that

$$1 - N(d_i, d_j) ≥ 0$$

(or)

$$N(d_i, d_j) ≤ 1$$

since $r$ is the number of common terms in $d_i, d_j$, the ratio,

$$\frac{r^2}{q_i q_j} ≤ 1$$

Therefore

$$N(d_i, d_j) ≤ 1$$

Hence our distance satisfies the first two properties of a distance measure.

To prove the third property

$$D(d_i, d_j) = D(d_j, d_i)$$

is equivalent to proving

$$N(d_i, d_j) = N(d_j, d_i)$$

$$N(d_i, d_j) = 1 - \frac{r^2}{q_i q_j} = 1 - \frac{r^2}{q_j q_i} = N(d_j, d_i)$$

The property of symmetry is satisfied. Hence our proposed method, BGEBS proves to be a distance measure.

B. BGEBS algorithm
The algorithm for computing Bipartite Graph Energy Based Similarity is as follows.

1. **Input:** Term Document Matrix
2. **Step 1:** Form the Reduced Term Document Matrix
3. **Step 2:** Give the Bipartite Representation
4. **Step 3:** Calculate the Energy of the graph $E$
5. **Step 4:** Use the BGEBS measure $S(d_i, d_j) = \frac{E \ast r^2}{q_i q_j}$

Output: Similarity matrix

V. EXPERIMENTAL ANALYSIS AND RESULTS

In this work we provide a new graph based document clustering technique known as Bipartite Graph Energy Based Similarity. This was inspired by the current advances in the area of graph based document clustering. We considered a document collection, identified the unique terms in the collection after preprocessing and gave a graph-based representation for the documents and unique terms/words. Depending on the number of occurrences (frequency) of the terms, the sparsity of the term document matrix was reduced and we obtained a reduced term document matrix. We then gave a bipartite graph representation for the document collection based on the reduced set of words. We then found the similarity between documents using our novel method based on the energy of a bipartite graph.

A. Analysis for Synthetic document set
For illustration, we have limited our analysis to a document set consisting of 6 documents. In our paper [9] we have given the description of this 6 documents, the preprocessing steps, their graphical representation as an undirected graph, adjacency matrix and the reduced term document matrix. The reduced term document matrix is as given in Table 1.

$$D(d_i, d_j) = \sum_{k=1}^{p} |d_{ik} - d_{jk}|$$
Bipartite Graph Energy Based Similarity measure for Document Clustering

Table 1 Reduced Term Document Matrix

<table>
<thead>
<tr>
<th>Term</th>
<th>Cluster</th>
<th>Cluster</th>
<th>Distant</th>
<th>Document</th>
<th>Evalu</th>
<th>High</th>
<th>Measure</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>D6</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The above reduced matrix contains eight frequently occurring terms. Fig. 1 shows the bipartite representation for this document set and the eight most frequently occurring terms. The edge weights indicate the frequency of a particular word in the corresponding document.

![Bipartite representation of 6 documents](image)

Fig. 1. Bipartite representation of 6 documents

We have employed our distance measure in k-means algorithm and clustered the documents. For sample six document set, we have used 2 clusters and for other large data sets we have considered 4 clusters.After reducing the term document matrix, we have used the similarity measure given in equation 1 and found the similarity between the documents. Table 2 gives the values of similarity matrix. From Table 2 it can be noted that the distance between documents 1 and 6 and documents 2 and 6 are zero as there are no terms in common between these two pairs of documents. It can also be noted that the distance between documents 1 and 2, 0.6666 is the maximum.

Table 2. Similarity matrix for 6 sample documents

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>D5</th>
<th>D6</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>0.66666</td>
<td>0.44444</td>
<td>0.08333</td>
<td>0.05555</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td>0.66666</td>
<td>1</td>
<td>0.16666</td>
<td>0.125</td>
<td>0.08333</td>
<td>0</td>
</tr>
<tr>
<td>D3</td>
<td>0.44444</td>
<td>0.167</td>
<td>1</td>
<td>0.33333</td>
<td>0.22222</td>
<td>0.05555</td>
</tr>
<tr>
<td>D4</td>
<td>0.08333</td>
<td>0.125</td>
<td>0.33333</td>
<td>1</td>
<td>0.16666</td>
<td>0.16666</td>
</tr>
<tr>
<td>D5</td>
<td>0.05555</td>
<td>0.08333</td>
<td>0.22222</td>
<td>0.16666</td>
<td>1</td>
<td>0.44444</td>
</tr>
<tr>
<td>D6</td>
<td>0</td>
<td>0</td>
<td>0.05555</td>
<td>0.16666</td>
<td>0.44444</td>
<td>1</td>
</tr>
</tbody>
</table>

We use our proposed distance measure, BGEBS in k-means clustering algorithm to cluster the documents. We have used R software to implement the algorithm. We have also used other distance measures such as Euclidean, Jaccard, Cosine, Canberra, Manhattan and Maximum distance in k-means algorithm to form clusters and compare and analyze the impact of our distance measure. To validate our measure we have used the sum of squares within cluster quality measure. The Sum of Squares Within (SSW) is an internal quality index which measures the goodness of a clustering solution without any external information. Unlike the external validation measures, which use external information, internal validation measures rely only on the information in the data. Sum of Squares within is useful in comparing two clustering solutions or two clusters. The formula for SSW is given by

$$SSW = \frac{1}{N} \sum_{i=1}^{k} \sum_{j \in C_i} ||d_j - c_j||$$

where k is the number of clusters, $C_i$ is the $i^{th}$ cluster, $c_j$ is the mean of $j^{th}$ cluster, $d_j$ is the document contained in the $j^{th}$ cluster, N is the total number of documents which is minimized over all k-partitions.

We now compare the clustering results used with our proposed distance measure and other existing methods.

Table 3: SSW for Sample six documents

<table>
<thead>
<tr>
<th>Sample six Documents</th>
<th>Sum of squares within</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>6.981565</td>
</tr>
<tr>
<td>Jaccard</td>
<td>7.66248</td>
</tr>
<tr>
<td>Cosine</td>
<td>6.854124</td>
</tr>
<tr>
<td>Canberra</td>
<td>6.854124</td>
</tr>
<tr>
<td>Manhattan</td>
<td>6.981565</td>
</tr>
<tr>
<td>Maximum Distance</td>
<td>6.981565</td>
</tr>
<tr>
<td>Proposed method</td>
<td>6.854124</td>
</tr>
</tbody>
</table>

Table 3 clearly shows that the SSW value for our proposed method is the lowest and it matches with those of Canberra distance and cosine measure. The graphical representation of Table 3 is given in Figure 2.

Figure 2: Comparison for 6 document dataset
B. Illustration and Comparison for Classic data set, BBC and webkb datasets:

We have used some well known benchmark data sets like Classic dataset, BBC and Webkb and applied the different distance measures to study the efficacy of our proposed measure. The table below shows the description about the benchmark data sets used. After reducing the term document matrix, we have used our distance measure BGEBS in the well known k-means algorithm and obtained the clustering solution. To validate our result we have used the sum of squares within (SSW) cluster quality measure. The below tables and figures give the value of SSW for all the clustering solutions obtained using k means clustering algorithm.

Table 4: SSW for Classic dataset

<table>
<thead>
<tr>
<th>Classic Dataset</th>
<th>Distance measure</th>
<th>Sum of squares within</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Euclidean</td>
<td>22977.16</td>
</tr>
<tr>
<td></td>
<td>Jaccard</td>
<td>22721.85</td>
</tr>
<tr>
<td></td>
<td>Cosine</td>
<td>22719.96</td>
</tr>
<tr>
<td></td>
<td>Canberra</td>
<td>22976.3</td>
</tr>
<tr>
<td></td>
<td>Manhattan</td>
<td>22719.96</td>
</tr>
<tr>
<td></td>
<td>Maximum Distance</td>
<td>22719.96</td>
</tr>
<tr>
<td></td>
<td>Proposed method</td>
<td>22719.62</td>
</tr>
</tbody>
</table>

Table 5: SSW for Webkb data set

<table>
<thead>
<tr>
<th>Webkb Dataset</th>
<th>Distance measure</th>
<th>Sum of squares within</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Euclidean</td>
<td>3147.653</td>
</tr>
<tr>
<td></td>
<td>Jaccard</td>
<td>3067.615</td>
</tr>
<tr>
<td></td>
<td>Cosine</td>
<td>3067.615</td>
</tr>
<tr>
<td></td>
<td>Canberra</td>
<td>3055.211</td>
</tr>
<tr>
<td></td>
<td>Manhattan</td>
<td>3133.241</td>
</tr>
<tr>
<td></td>
<td>Maximum Distance</td>
<td>3181.935</td>
</tr>
<tr>
<td></td>
<td>Proposed method</td>
<td>3049.574</td>
</tr>
</tbody>
</table>

Figure 3: Comparison for Classic dataset

Figure 3 shows the comparison of SSW for Classic dataset with Euclidean distance, Jaccard distance, Cosine measure, Canberra distance, Manhattan distance, Maximum distance along with our proposed method. It is clear from Table 4 that the SSW value for our proposed method is the least at 22719.62. It is marginally better than Cosine, Manhattan and Maximum distance. Hence, our proposed measure gives a relatively better cluster quality. Table 5 gives the description for webkb data set used along with the sum of squares within.

Figure 4: Comparison for Webkb dataset

Figure 4 shows the comparison of SSW for webkb data set with Euclidean distance, Jaccard distance, Cosine measure, Canberra distance, Manhattan distance, Maximum distance along with our proposed method. It is clear from Table 5 that the SSW value for our proposed method is the least at 3049.574. It is marginally better than Cosine, Jaccard and Canberra distance. Hence, our proposed measure gives a relatively better cluster quality. We then use the benchmark data set BBC to validate our result. Table 6 shows the different distance measure used along with the sum of squares within, internal quality measure.

Table 6: SSW for BBC data set

<table>
<thead>
<tr>
<th>BBC Dataset</th>
<th>Distance measure</th>
<th>Sum of squares within</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Euclidean</td>
<td>26664.93</td>
</tr>
<tr>
<td></td>
<td>Jaccard</td>
<td>26642.6</td>
</tr>
<tr>
<td></td>
<td>Cosine</td>
<td>26642.6</td>
</tr>
</tbody>
</table>

Retrieval Number: F2183037619/19©BEIESP
Figure 5 shows the comparison of SSW for webkb data set with Euclidean distance, Jaccard distance, Cosine measure, Canberra distance, Manhattan distance, Maximum distance along with our proposed method. It is clear from Table 6 that the SSW value for our proposed method is the least at 26636.72. It is marginally better than other distance. Hence, our proposed measure gives a relatively better cluster quality.

VI. CONCLUSION

We have proposed a new graph energy based distance method using bipartite representation of the document set. The measure satisfies the criteria for a distance measure and proves to be a valid distance measure. We have illustrated the computation of BGEBS for a sample of 6 documents. We have compared our proposed distance measure with six different distance measures for three different benchmark data sets. To validate our result we have used the internal cluster quality measure, sum of within square. For classic data set our BGEBS proves to be better than Cosine, Manhattan and Maximum distance. For webkb and BBC data sets BGEBS is better when compared to all other measures. This experimental analysis shows that our proposed distance measure is comparably good.

REFERENCES

11. 19©BEIESP.
AUTHORS PROFILE

Dr. G. Hannah Grace is currently an Assistant Professor in the Department of Mathematics, School of Advanced Sciences, VIT Chennai. She has over 15 years of Academic experience. Her area of research is Document Clustering. She has completed her PhD at VIT Chennai under the guidance of Dr. Kalyani Desikan, Professor, VIT Chennai. She has 7 journal publications to her credit and has presented in 4 international conferences.

Dr. Kalyani Desikan is a Professor of Mathematics in the School of Advanced Sciences at VIT Chennai, India. She has over 23 years of experience in teaching and research. Her research interests include Data mining, Clustering, Link analysis, Automata theory, Spectral graph theory and Cosmology. She has guided one Ph.D. student and is currently guiding 5 Ph.D. students. She has close to 25 publications in National and International Journals.