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. In this paper we have introduced a new similarity measure namely, Bipartite Graph Energy Based Similarity (BGEBS) based on the energy of a bipartite graph for document clustering. BGEBS helps to cluster the documents by considering the energy of a bipartite graph representation of the document collection. We have compared our measure BGEBS with Euclidean, Jaccard, Cosine, Canberra, Manhattan and Maximum Distance and clustering is carried out using k-means to form clusters. We then compare and analyze our result with a synthetic data set containing 6 documents. we have also evaluated using few benchmark data sets like CLASSIC, WEBKB and BBC. To validate our measure we have used the internal quality measure, sum of squares within (SSW). The values obtained using SSW for the various distance measures when compared to our BGEBS proves to be good.***

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

I. INTRODUCTION

Similarity/distance measures are not only used in clustering, but also in other data mining algorithms. To understand the impact of distance measures on data mining, researchers have performed experimental studies in different fields and have evaluated and compared the results obtained by different distance measures [1]. Though, we cannot say that a similarity measure is best performing, a comparative study would enable us to understand the performance and varied behaviours of the different similarity measures better. Distance measures play an important role in document clustering. The key factor lies in identifying the appropriate distance measure for a given data set. 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**: $D(d_1, d_2) \geq 0$. 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 atleast 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. $W=|\omega_{ij}|$ represents the frequency of term i in document j.

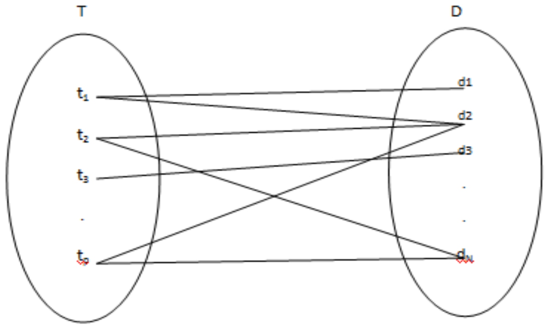


Figure 1 Bipartite Representation of Documents in G

The adjacency matrix of a bipartite weighted graph G(V,E,W) is represented in a block matrix form as follows $\begin{pmatrix} 0 & W \\ W^T & 0 \end{pmatrix}$ where W^T is the transpose of matrix 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 $E(G)=\sum_{i=1}^{n}|\lambda_i|$, where the set of eigen values $\{\lambda_1, \lambda_2, ...\lambda_n\}$ is known as the spectrum of G [12,13].

Revised Manuscript Received on March 20, 2019.

G. Hannah Grace, Division of Mathematics, School of advance science, VIT university, Chennai,India.

Kalyani Desikan, Division of Mathematics, School of advance science, VIT university, Chennai,India.

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

$$E(G) \leq \frac{4m}{n} + \sqrt{(n-2)(2m - \frac{8m^2}{n^2})} \text{-----------------}(1)$$

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 coefficientfor 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 meaure 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 ablge 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:

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

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

F. Maximum Distance

This gives the maximum distance between two documents.

$$D(d_i, d_j) = \max(d_{ij} - d_{jk})$$

IV. BIPARTITE GRAPH ENERGY BASED SIMILARITY MEASURE (BGEBS)

Choosing the correct distance measure for a given document set is important for document clustering. In our proposed method we analyzed the existing distance measures and introduced a new distance measure. The distance measure we have proposed 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

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

where $S(d_i, d_j)$ is used to find the similarity between pairs of documents.

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

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

After Normalization withE, 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,

$$D(d_i, d_j) = 1 - N(d_i, d_j) = 1 - \frac{r^2}{q_i q_j} \quad \forall i, j$$

To prove the first two properties of distance measure i.e.,

$$D(d_i, d_j) \geq 0$$

we equivalently show that

$$1 - N(d_i, d_j) \geq 0$$

(or)

$$N(d_i, d_j) \leq 1$$

sincer is the number of common terms in di, dj, the ratio,

$$\frac{r^2}{q_i q_j} \leq 1$$

Therefore

$$N(d_i, d_j) \leq 1$$

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.

Input: Term Document Matrix

Step 1: Form the Reduced Term Document Matrix

Step 2: Give the Bipartite Representation

Step 3: Calculate the Energy of the graph E

Step 4: Use the BGEBS measure $S(d_i, d_j) = \frac{E * r^2}{q_i q_j}$

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.

Table 1 Reduced Term Document Matrix

Term/Doc	Cluster	Cluto	Distanc	document	Evalu	High	Measur	Use
D1	1	1	0	0	0	1	0	0
D2	1	1	0	0	0	0	0	0
D3	1	0	0	1	0	1	0	0
D4	1	0	1	2	0	0	0	0
D5	1	0	0	1	1	0	2	1
D6	0	0	2	1	1	0	1	1

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.

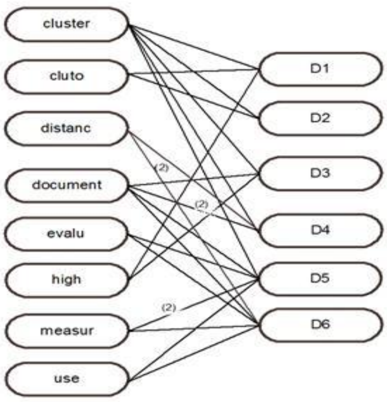


Fig. 1. Bipartite representation of 6 documents

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

	D1	D2	D3	D4	D5	D6
D1	1	0.66666	0.44444	0.08333	0.05555	0
D2	0.66666	1	0.16666	0.125	0.08333	0
D3	0.44444	0.167	1	0.33333	0.22222	0.05555
D4	0.08333	0.125	0.33333	1	0.16666	0.16666
D5	0.05555	0.08333	0.22222	0.16666	1	0.44444
D6	0	0	0.05555	0.16666	0.44444	1

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

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

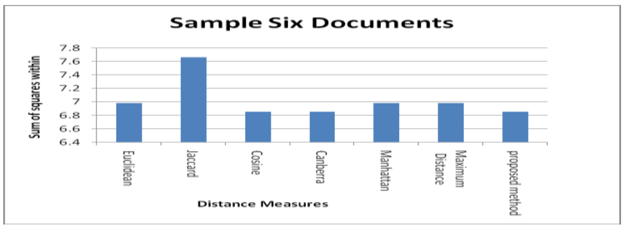
where k is the number of clusters,
C_i is the i^{th} cluster,
c_j is the mean of j^{th} cluster,
d_j is the document contained in the j^{th} cluster,
N is the total number of documents
which is minimized over all k-partitions.

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

Table 3: SSW for Sample six documents

Sample six Documents	
Distance measure	Sum of squares within
Euclidean	6.981565
Jaccard	7.66248
Cosine	6.854124
Canberra	6.854124
Manhattan	6.981565
Maximum Distance	6.981565
Proposed method	6.854124

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 bench mark 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 bench mark 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

***Classic* Dataset**	
Distance measure	Sum of squares within
Euclidean	22977.16
Jaccard	22721.85
Cosine	22719.96
Canberra	22976.3
Manhattan	22719.96
Maximum Distance	22719.96
Proposed method	**22719.62**

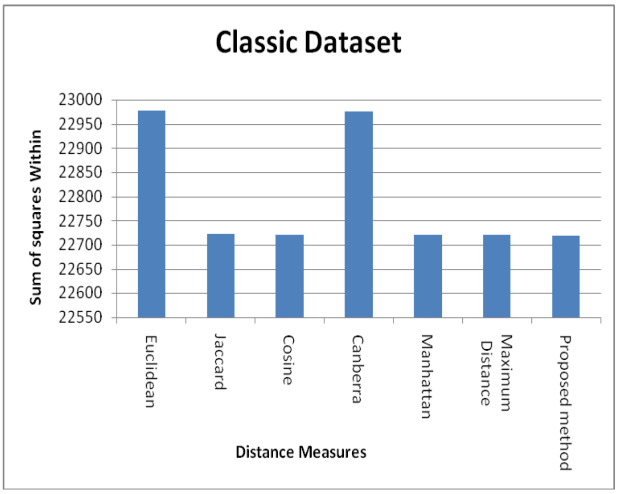


Figure 3: Comparison for *Classic* dataset

Figure 3 shows the comparison of SSW for *Classic* 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 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 internal quality, sum of squares within.

Table 5: SSW for *Webkb* data set

***Webkb* Dataset**	
Distance measure	Sum of squares within
Euclidean	3147.653
Jaccard	3067.615
Cosine	3067.615
Canberra	3055.211
Manhattan	3133.241
Maximum Distance	3181.935
Proposed method	**3049.574**

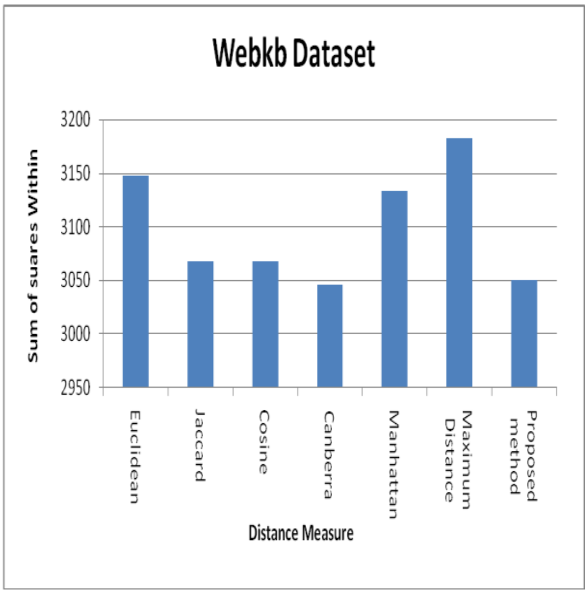


Figure 4: Comparison for *Webkb* dataset

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 bench mark data set BBC to validate our result. Table 6 shows the different distance measure use along with the sum of squares within, internal quality measure.

Table 6: SSW for *BBC* data set

***BBC* Dataset**	
Distance measure	Sum of squares within
Euclidean	26664.93
Jaccard	26642.6
Cosine	26725.71

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Canberra	26711.26
Manhattan	26713.24
Maximum Distance	26666.24
Proposed method	**26636.72**

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.

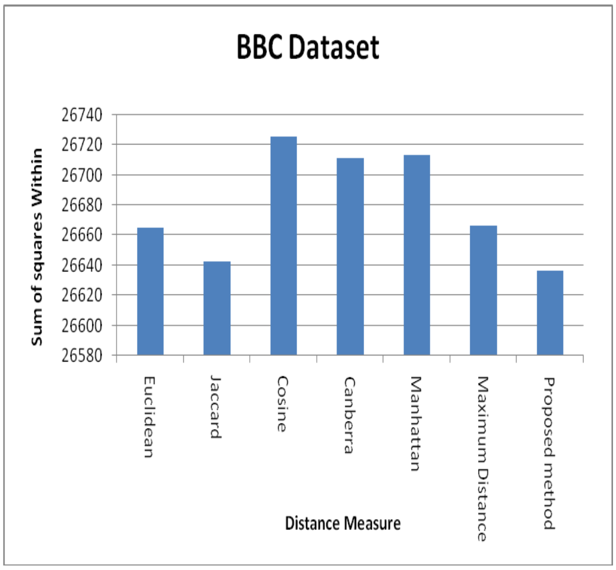


Figure 5: Comparison for BBC dataset

Table 2, 3, 4 and 5 clearly present the comparison of different distance measures like Euclidean, Jaccard, Cosine, Canberra, Manhattan, Maximum distance with our proposed method. We have used SSW for different datasets like *classic* dataset, *BBC* dataset and *webkb* dataset. The above results show that our proposed distance measure gives good clustering solutions comparable with some of the popular distance measures. In some cases it even gives better clustering solutions than some of the well known distance measures..

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

1. Z. Chen, "Graph-based clustering and its application in coreference resolution," in *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*, ACL 2010, 2010, pp. 1–9.
2. A. Huang, "Similarity measures for text document clustering," in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)*, Christchurch, New Zealand, 2008, pp. 49–56.
3. A. Vimal, S. R. Valluri, and K. Karlapalem, "An experiment with distance measures for clustering." in *COMAD*, 2008, pp. 241–244.
4. A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering," in *Workshop on artificial intelligence for web search* (AAAI 2000), Vol. 58, 2000, p. 64.
5. R.-N. Baeza-Yates, "Ricardo baeza-yates, berthier ribeiro-neto: *Modern information retrieval*. chapter 3," 1999.
6. B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering," in Proceedings of the fifth ACM SIGKDD *international conference on Knowledge discovery and data mining*. ACM, 1999, pp. 16–22.
7. A. S. Shirkhorshidi, S. Aghabozorgi, and T. Y. Wah, "A comparison study on similarity and dissimilarity measures in clustering continuous data," *PloS one*, Vol. 10, no. 12, 2015, p. e0144059.
8. G. H. Grace and K. Desikan, "Document clustering using a new similarity measure based on energy of a bipartite graph," *Indian Journal of Science and Technology*, Vol. 9, no. 40, 2016.
9. Lourenco F, Lobo V, Bacao F. Binary-based similarity measures for categorical data and their application in Self-Organizing Maps. 2004; 1–18.
10. 10. D.Chakrabarti. Tools for large graph miners. Thesis, School of Computer Science, Carnegie Mellon University, CMUCALD
11. -05-107, Center for Automated Learning and Discovery, 2005, 1-117.
12. 11. DB, West. Introduction to Graph Theory. Prentice Hall, 2001.
13. 12. Jack H, Koolen K. "Maximal Energy graphs." Advances in Applied Mathematics 26, no. 1 (2001): 47-52.
14. 13. R, Balakrishnan. "The Energy of a graph." Linear Algebra and its Applications 387 (2004): 287-95.
15. 14. Jack H, Koolen K. "Maximal Energy graphs." Advances in Applied Mathematics 26, no. 1 (2001): 47-52.
16. 15. Deshpande R, VanderSluis B, Myers CL. Comparison of Profile Similarity Measures for Genetic Interaction Networks. PLoS One. 2013; vol8.
17. 16. Strehl, A., Ghosh, J., Mooney, R.:" Impact of similarity measures on web-page Clustering". In Proc. AAAIWorkshop on AI forWeb Search, pp 58–64, 2000.
18. 17. Zhang Z, Huang K, Tan T. Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. Proceedings—International Conference on Pattern Recognition. IEEE; 2006. pp. 1135–1138
19. 18. Khalifa A Al, Haranczyk M, Holliday J. Comparison of Nonbinary Similarity Coefficients for Similarity Searching, Clustering and Compound Selection. J Chem Inf Model. 2009; 49: 1193–1201.

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication

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.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication