Clustering the Web Documents Using the Enhanced Hierarchical Clustering Technique

Konjeti Devi Badari[#1], Kesavarapu Tirumala Reddy[*2]

[#]*PG Scholar, Dept. of CSE, VRS & YRN College of Engg. & Technology, Chirala, AP*

[1] *badarikonjeti@gmail.com*

[*]*Assistant Professor, Dept. of CSE, VRS & YRN College of Engg. & Technology, Chirala, AP*

***Abstract*— Noise gets influenced in a common distance measures in the high dimensional data. In an existing clustering algorithms implemented based on their partitioning, hierarchical, density based and grid based. Assume that some cluster relationship among the data objects that they are applied on in all clustering methods but either explicitly or implicitly it could be the similarity between a pair of objects can be defined in the methods. In clustering method, the major difference between a traditionally is similarity/dissimilarity measures. In the former methods the researchers measure the objects in the cluster by using the single view point. And then later utilizes many different viewpoints, which are objects, assumed to not be in the same cluster with the two objects being measured. To overcome the above issues, in this paper we propose a novel of highly designed multi-viewpoint based similarity measure and two related clustering methods. By using the multi-viewpoint, more informative assessment and effective similarities could be achieved in this method. Theoretical and experimental analysis are conducted for the multi-viewpoint, it could support the above claim. In the clustering methods we proposed the two criterion functions for the document to achieve the similarities. In clustering methods, it consists of several algorithms for similarities measurement but our proposed algorithm get advantages over the old well-known algorithms.**

Keywords— Document clustering, text mining, similarity measure, single viewpoint, multi viewpoint.

I. INTRODUCTION

In the data mining, one of the most significant and interesting topic is clustering. The main aim of clustering is to find intrinsic structures in data, and organize them into meaningful subgroups for further study and analysis. The modified clustering algorithms are published every year for the recent research about the similarities in the data. They proposed for very distinct research in the similarities, and developed using totally different techniques and approaches for the effectiveness. Nevertheless, according to a recent study [1], more than half a century after it were introduced; the simple algorithm k-means still remains as one of the top 10 data mining algorithms nowadays. It is the most frequently used partitional clustering algorithm in practice. Another recent scientific discussion [2] states that k-means is the favourite algorithm that practitioners in the related fields choose to use. Needless to mention, k-means has more than a few basic drawbacks, such as sensitiveness to initialization and to cluster size, and its performance can be worse than other state-of-the-art algorithms in many domains. In spite of that, its simplicity, understand ability and scalability are the reasons for its tremendous popularity.

An algorithm with 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 other methods in larger systems. Our study of similarity of clustering was initially motivated by a research on automated text categorization of foreign language texts, as explained below. As the amount of digital documents has been increasing dramatically over the years as the Internet grows, information management, search, and retrieval, etc., have become practically important problems. Developing methods to organize large amounts of unstructured text documents into a smaller number of meaningful clusters would be very helpful as document clustering is vital to such tasks as indexing, filtering, automated metadata generation, word sense disambiguation, population of hierarchical catalogues of web resources and, in general, any application requiring document organization. Document clustering is also one of the important topics in biomedicine. The clustering deals with the large amounts of data, and its learning approaches are applied through perform Automated Text Clustering (ATC). Given an unlabeled dataset, this ATC system builds clusters of documents that are hopefully similar to clustering (classification, categorization, or labelling) performed by human experts. To identify a suitable tool and algorithm for clustering that produces the best clustering solutions, it becomes necessary to have a method for comparing the results of different clustering algorithms. Though considerable work has been done in designing clustering algorithms, not much research has been done on formulating a measure for the similarity of two different clustering algorithms. Thus, the main goal of this paper is to: First, propose an algorithm for performing similarity analysis among different clustering algorithms;

second, apply the algorithm to calculate similarity of various pairs of clustering methods applied to a Portuguese corpus and the Iris dataset; finally, to cross validate the results of similarity analysis with the Euclidean (centroids) distances and Pearson correlation coefficient, using the same datasets. Possible applications are discussed.

The work in this paper, we propose a novel of highly designed multi-viewpoint based similarity measure and two related clustering methods. By using the multi-viewpoint, more informative assessment and effective similarities could be achieved in this method. Theoretical and experimental analysis are conducted for the multi-viewpoint, it could support the above claim. In the clustering methods we proposed the two criterion functions for the document to achieve the similarities. In clustering methods, it consists of several algorithms for similarities measurement but our proposed algorithm get advantages over the old well-known algorithms. First objective in this method is to find the similarity between the data in sparse and high dimensional domain, particularly text documents. And then formulate new clustering criterion functions and introduce their respective clustering algorithms, which are fast and scalable like k-means, but are also capable of providing high-quality and consistent performance.

II. Related Works

In this section, we briefly discuss the works which is similar techniques as our approach but serve for different purposes.

a.k. jain, m.n. murty and p.j. flynn [33] we proposes a paper an overview of pattern clustering methods from a statistical pattern recognition perspective, with a goal of providing useful advice and references to fundamental concepts accessible to the broad community of clustering practitioners. We present a taxonomy of clustering techniques, and identify cross-cutting themes and recent advances. We also describe some important applications of clustering algorithms such as image segmentation, object recognition, and information retrieval.

Yaminee S. Patil and M.B.Vaidya [34] In data mining functionalities, clustering analysis is the most significant tool for distribution of data. Clustering is dynamic field of research in data mining concept. It is related to unsupervised learning in machine learning. On the basis of similarity measures cluster formation process is initiated. With the help of different notations used in clustering algorithms unique clusters are formed with the same data set. In this paper several clustering methods are discussed with their particular algorithms. Clustering methods are drastically affecting the shapes of cluster, quality of cluster, scalability of clusters. In this paper we have discussed integrated clustering algorithm that is multiphase clustering algorithms which improves scalability and efficiency of clusters. Different algorithms perform different task to make cluster more dynamic and effective. Several clustering methods and their corresponding algorithms are described below which helps to further analysis.

ravi Bhusan Yadav and m. Madhu Babu[35] In this paper we are going to present two Clustering methods and a multiview point based similarity measure. In Multipoint Based Similarity Measure we use many different viewpoints that are objects and are assumed to not be in same cluster with two objects being measured, this is the main distinctness of our concept with a traditional dissimilarity/similarity measure is that the aforementioned dissimilarity/similarity exercises only a single view point for which it is the base.We can implement countless descriptive evaluation by utilizing multiple viewpoints and to support this we proposed two functions to document clustering.

Guadalupe J. Torres, Ram B. Basnet, Andrew H. Sung, Srinivas Mukkamala, Bernardete M. Ribeiro[36] This paper introduces a measure of similarity between two clusterings of the same dataset produced by two different algorithms, or even the same algorithm (K-means, for instance, with different initializations usually produce different results in clustering the same dataset). We then apply the measure to calculate the similarity between pairs of clusterings, with special interest directed at comparing the similarity between various machine clusterings and human clustering of datasets. The similarity measure thus can be used to identify the best (in terms of most similar to human) clustering algorithm for a specific problem at hand. Experimental results pertaining to the text categorization problem of a Portuguese corpus (wherein a translation-into-English approach is used) are presented, as well as results on the well-known benchmark IRIS dataset. The significance and other potential applications of the proposed measure are discussed.

Duc Thang Nguyen,Lihui Chen, Chee Keong Chan[37] The aim of this work is to produce fast, easy-to-apply but effective algorithms for clustering large text collections. In this paper, we propose a novel concept of similarity measure among objects and its related clustering algorithms. The similarity between two objects within a cluster is measured from the view of all other objects outside that cluster. As a result, two optimality criteria are formulated as the objective functions for the clustering problem. We analyze and compare the proposed clustering approaches with the popular algorithms for document clustering in the literature. Extensive empirical experiments are carried out on various benchmark datasets and evaluated by different metrics. The results show that our proposed criterion functions consistently outperform the other well-known clustering criteria, and give the best overall performance with the same computational efficiency.

First of all, Table 1 summarizes the basic notations that will be used extensively throughout this paper to represent documents and related concepts. Each document in a corpus corresponds to an m-dimensional vector d, where m is the total number of terms that the document corpus has. Document vectors are often subjected to some weighting schemes, such as the standard Term Frequency-Inverse

Document Frequency (TF-IDF), and normalized to have unit length.

$$Dist(d_i, d_j) = \|d_i - d_j\|. \qquad (1)$$

The principle definition of clustering is to arrange data objects into separate clusters such that the intracluster similarity as well as the intercluster dissimilarity is maximized. The problem formulation itself implies that some forms of measurement are needed to determine such similarity or dissimilarity. There are many state-of-the-art clustering approaches that do not employ any specific form of measurement, for instance, probabilistic model-based method [9], nonnegative matrix factorization [10], information theoretic coclustering [11] and so on. In this paper, though, we primarily focus on methods that indeed do utilize a specific measure. In the literature, euclidean distance is one of the most popular measures. It is used in the traditional k-means algorithm. The objective of k-means is to minimize the euclidean distance between objects of a cluster and that cluster's
centroids

$$\min \sum_{r=1}^{k} \sum_{d_i \in S_r} \|d_i - C_r\|^2. \qquad (2)$$

However, for data in a sparse and high-dimensional space, such as that in document clustering, cosine similarity is more widely used. It is also a popular similarity score in text mining and information retrieval [12]. Particularly, similarity of two document vectors di and dj, Sim di; dj Þ, is defined as the cosine of the angle between them. For unit vectors, this equals to their inner product

$$Sim(d_i, d_j) = \cos(d_i, d_j) = d_i^t d_j. \qquad (3)$$

Cosine measure is used in a variant of k-means called spherical k-means [3]. While k-means aims to minimize euclidean distance, spherical k-means intends to maximize the cosine similarity between documents in a cluster and that cluster's centroids

$$\max \sum_{r=1}^{k} \sum_{d_i \in S_r} \frac{d_i^t C_r}{\|C_r\|}. \qquad (4)$$

The major difference between euclidean distance and cosine similarity, and therefore between k-means and spherical kmeans, is that the former focuses on vector magnitudes, while the latter emphasizes on vector directions. Besides direct application in spherical k-means, cosine of document vectors is also widely used in many other document clustering methods as a core similarity measurement. The min-max cut graph-based spectral method is an example [13]. In graph partitioning approach, document corpus is consider as a graph G ¼ ðV ;EÞ, where each document is a vertex in V and each edge in E has a weight equal to the similarity between a pair of vertices. Min-max cut algorithm tries to minimize the criterion function

$$\min \sum_{r=1}^{k} \frac{Sim(S_r, S \setminus S_r)}{Sim(S_r, S_r)}$$
$$\text{where } Sim(S_q, S_r) \underset{1 \le q, r \le k}{=} \sum_{d_i \in S_q, d_j \in S_r} Sim(d_i, d_j), \qquad (5)$$

and when the cosine as in (3) is used, minimizing the criterion in (5) is equivalent to

$$\min \sum_{r=1}^{k} \frac{D_r^t D}{\|D_r\|^2}. \qquad (6)$$

There are many other graph partitioning methods with different cutting strategies and criterion functions, such as Average Weight [14] and Normalized Cut [15], all of which have been successfully applied for document clustering using cosine as the pairwise similarity score [16], [17]. In [18], an empirical study was conducted to compare a variety of criterion functions for document clustering. Another popular graph-based clustering technique is implemented in a software package called CLUTO [19]. This method first models the documents with a nearest neighbor graph, and then splits the graph into clusters using a min-cut algorithm. Besides cosine measure, the extended Jaccard coefficient can also be used in this method to represent similarity between nearest documents. Given nonunit document vectors ui, uj ðdi ¼ ui=kuik; dj ¼ uj=kujkÞ, their extended Jaccard coefficient i

$$Sim_{eJacc}(u_i, u_j) = \frac{u_i^t u_j}{\|u_i\|^2 + \|u_j\|^2 - u_i^t u_j}. \qquad (7)$$

Compared with euclidean distance and cosine similarity, the extended Jaccard coefficient takes into account both the magnitude and the direction of the document vectors. If the documents are instead represented by their corresponding unit vectors, this measure has the same effect as cosine similarity. In [20], Strehl et al. compared four measures: euclidean, cosine, Pearson correlation, and extended Jaccard, and concluded that cosine and extended Jaccard are the best ones on web documents. In nearest neighbor graph clustering methods, such as the CLUTO's graph method above, the concept of similarity is somewhat different from the previously discussed methods. Two documents mayhave a certain value of cosine similarity, but if neither of them is in the other one's neighborhood, they have no connection between them. In such a case, some context-based knowledge or relativeness property is already taken into account when considering similarity. Recently, Ahmadand Dey [21] proposed a method to compute distance between two categorical values of an attribute based on their relationship with all other attributes. Subsequently, Ienco et al. [22] introduced a similar context-based distance learning method for categorical data. However, for a given attribute, they only selected a relevant subset of attributes from the whole attribute set to use as the context for calculating distance between its two values. More related to text data, there are phrase-based and concept-based document similarities. Lakkaraju et al. [23] employed a conceptual tree-similarity measure to identify similar documents. This method requires representing documents as concept trees with the help of a classifier. For clustering, Chim and Deng [24] proposed a phrasebased document similarity by combining suffix tree model and vector space model. They then used Hierarchical Agglomerative Clustering algorithm to perform the clustering task. However, a drawback of this approach is the high computational complexity due to the needs of building the suffix tree and calculating pairwise similarities explicitly before clustering. There are also measures designed specifically for capturing structural similarity among XML documents [25]. They are essentially different from the document-content measures that are discussed in this paper. In general, cosine similarity still remains as the most popular measure because of its simple interpretation and easy computation, though its effectiveness is yet fairly limited. In the following sections, we propose a novel way to evaluate similarity between documents, and consequently formulate new criterion functions for document clustering.

III. PROPOSED WORK

In clustering method, the major difference between a traditionally is similarity/dissimilarity measures. In the former methods the researchers measure the objects in the cluster by using the single view point. And then later utilizes many different viewpoints, which are objects, assumed to not be in the same cluster with the two objects being measured. To overcome the above issues, in this paper we propose a novel of highly designed multi-viewpoint based similarity measure and two related clustering methods. By using the multi-viewpoint, more informative assessment and effective similarities could be achieved in this method. Theoretical and experimental analysis are conducted for the multi-viewpoint, it could support the above claim. In the clustering methods we proposed the two criterion functions for the document to achieve the similarities.

First objective in this method is to find the similarity between the data in sparse and high dimensional domain, particularly text documents. And then formulate new clustering criterion functions and introduce their respective clustering algorithms, which are fast and scalable like k-means, but are also capable of providing high-quality and consistent performance.

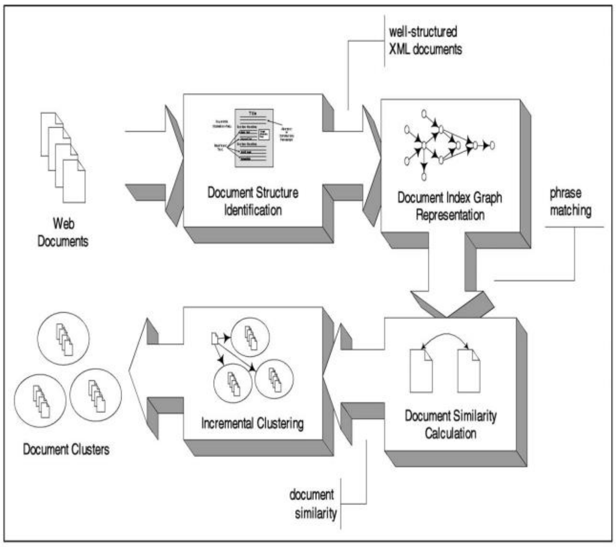

well-structured XML documents
Web Documents
Document Structure Identification
Document Index Graph Representation
phrase matching
Document Clusters
Incremental Clustering
Document Similarity Calculation
document similarity

Figure 1: Proposed Architecture

Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering is this:

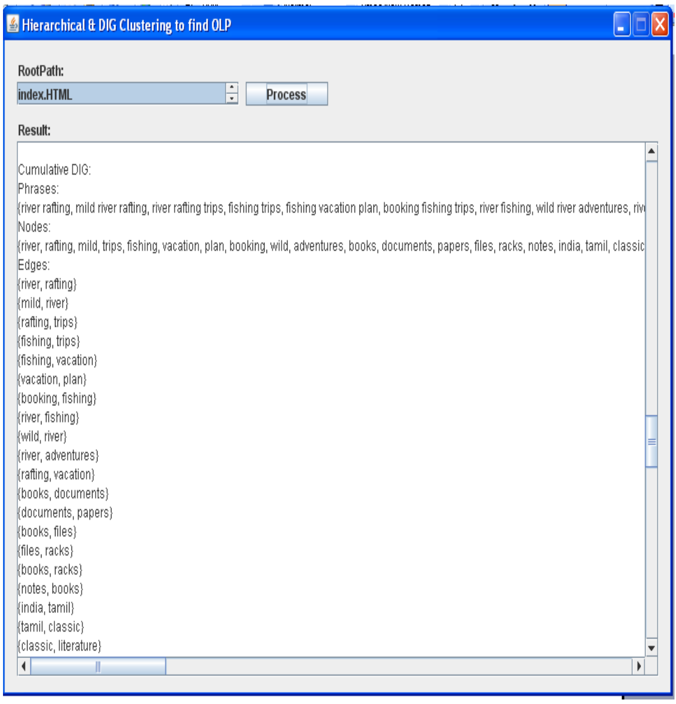
STEP1 - Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.

STEP2 - Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less with the help oh tf - itf.

STEP3 - Compute distances (similarities) between the new cluster and each of the old clusters.

STEP4 - Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Step 3 can be done in different ways, which is what distinguishes single-linkage from complete-linkage and average-linkage clustering. In single-linkage clustering (also called the connectedness or minimum method), considering the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster.



```
Cumulative DIG:
Phrases:
{river rafting, mild river rafting, river rafting trips, fishing trips, fishing vacation plan, booking fishing trips, river fishing, wild river adventures, riv
Nodes:
{river, rafting, mild, trips, fishing, vacation, plan, booking, wild, adventures, books, documents, papers, files, racks, notes, india, tamil, classic
Edges:
{river, rafting}
{mild, river}
{rafting, trips}
{fishing, trips}
{fishing, vacation}
{vacation, plan}
{booking, fishing}
{river, fishing}
{wild, river}
{river, adventures}
{rafting, vacation}
{books, documents}
{documents, papers}
{books, files}
{files, racks}
{books, racks}
{notes, books}
{india, tamil}
{tamil, classic}
{classic, literature}
```

Figure 2: Node and Edge Identification

This involves the document similarity analysis and thereby finding the Overlapping Rate (OLP Rate).

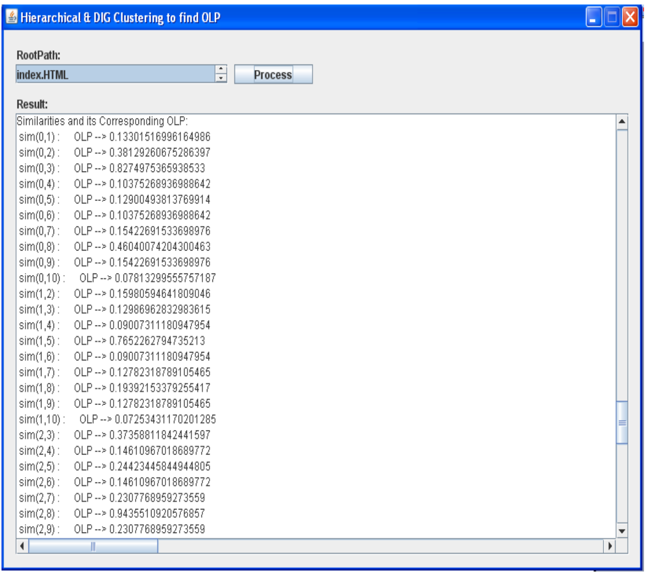
By taking into account these two factors — term frequency (TF) and inverse document frequency (IDF) — it is possible to assign “weights” to search results and therefore ordering them statistically. Put another way, a search result’s score (”ranking”) is the product of TF and IDF:

TFIDF = TF * IDF where:

TF = C / T where C = number of times a given word appears in a document and T = total number of words in a document

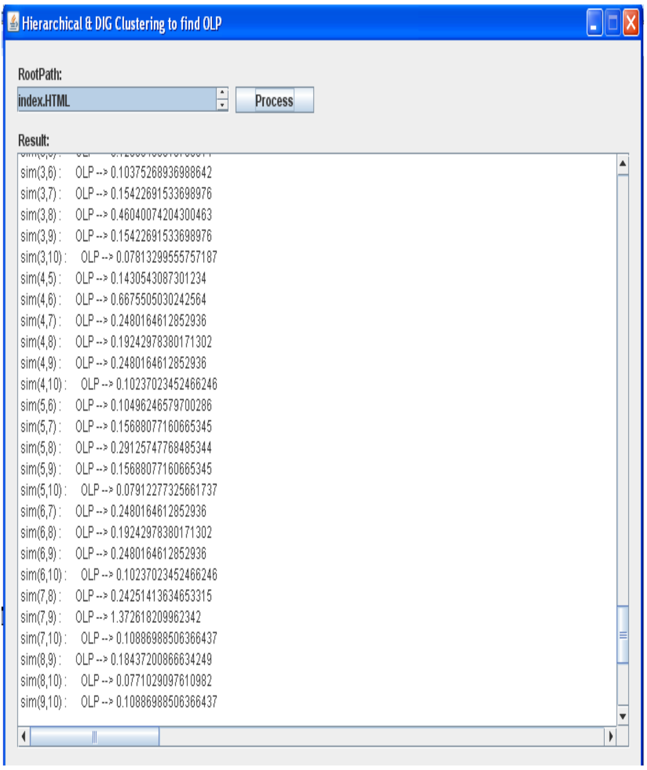
IDF = D / DF where D = total number of documents in a corpus, and DF = total number of documents containing a given word

IV. SIMULATION WORKS/RESULTS



```
Similarities and its Corresponding OLP:
 sim(0,1) :    OLP --> 0.13301516996164986
 sim(0,2) :    OLP --> 0.38129260675286397
 sim(0,3) :    OLP --> 0.8274975365938533
 sim(0,4) :    OLP --> 0.10375268936988642
 sim(0,5) :    OLP --> 0.12900493813769914
 sim(0,6) :    OLP --> 0.10375268936988642
 sim(0,7) :    OLP --> 0.15422691533698976
 sim(0,8) :    OLP --> 0.46040074204300463
 sim(0,9) :    OLP --> 0.15422691533698976
 sim(0,10) :    OLP --> 0.07813299555757187
 sim(1,2) :    OLP --> 0.15980594641809046
 sim(1,3) :    OLP --> 0.12986962832983615
 sim(1,4) :    OLP --> 0.09007311180947954
 sim(1,5) :    OLP --> 0.7652262794735213
 sim(1,6) :    OLP --> 0.09007311180947954
 sim(1,7) :    OLP --> 0.12782318789105465
 sim(1,8) :    OLP --> 0.19392153379255417
 sim(1,9) :    OLP --> 0.12782318789105465
 sim(1,10) :    OLP --> 0.07253431170201285
 sim(2,3) :    OLP --> 0.37358811842441597
 sim(2,4) :    OLP --> 0.14610967018689772
 sim(2,5) :    OLP --> 0.24423445844944805
 sim(2,6) :    OLP --> 0.14610967018689772
 sim(2,7) :    OLP --> 0.2307768959273559
 sim(2,8) :    OLP --> 0.9435510920576857
 sim(2,9) :    OLP --> 0.2307768959273559
```

Figure 3: Similarity Calcuation



```
sim(3,6) :    OLP --> 0.10375268936988642
sim(3,7) :    OLP --> 0.15422691533698976
sim(3,8) :    OLP --> 0.46040074204300463
sim(3,9) :    OLP --> 0.15422691533698976
sim(3,10) :    OLP --> 0.07813299555757187
sim(4,5) :    OLP --> 0.1430543087301234
sim(4,6) :    OLP --> 0.6675505030242564
sim(4,7) :    OLP --> 0.2480164612852936
sim(4,8) :    OLP --> 0.19242978380171302
sim(4,9) :    OLP --> 0.2480164612852936
sim(4,10) :    OLP --> 0.10237023452466246
sim(5,6) :    OLP --> 0.10496246579700286
sim(5,7) :    OLP --> 0.15688077160665345
sim(5,8) :    OLP --> 0.2912574776848534
sim(5,9) :    OLP --> 0.15688077160665345
sim(5,10) :    OLP --> 0.07912277325661737
sim(6,7) :    OLP --> 0.2480164612852936
sim(6,8) :    OLP --> 0.19242978380171302
sim(6,9) :    OLP --> 0.2480164612852936
sim(6,10) :    OLP --> 0.10237023452466246
sim(7,8) :    OLP --> 0.24251413634653315
sim(7,9) :    OLP --> 1.372618209962342
sim(7,10) :    OLP --> 0.10886988506366437
sim(8,9) :    OLP --> 0.18437200866634249
sim(8,10) :    OLP --> 0.0771029097610982
sim(9,10) :    OLP --> 0.10886988506366437
```

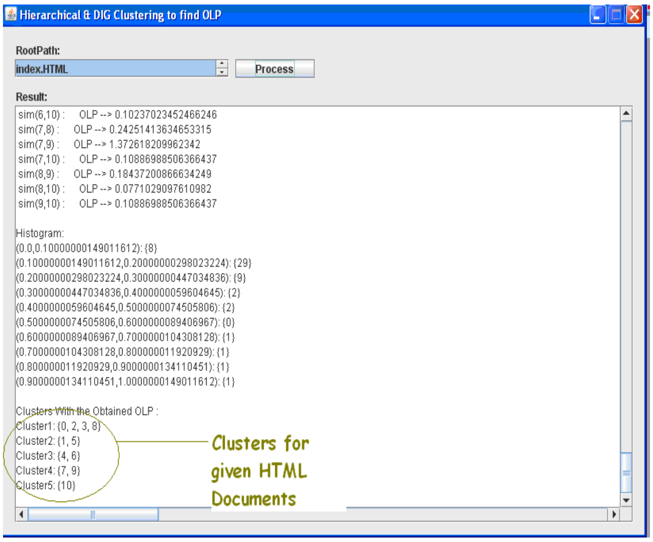
Figure 4: Similarity Calculation with OLP values

HISTOGRAM FORMATION

After finding the similarity and the OLP Rate, Histogram is formed. Histogram is also called as Dendogram.

CLUSTER FORMATION

Then the final step is the formation of clusters. This is shown in the below figure. Thus the Document clustering using Hierarchical Clustering is done and the causes are documented.



```
sim(6,10) :    OLP --> 0.10237023452466246
sim(7,8) :   OLP --> 0.24251413634653315
sim(7,9) :   OLP --> 1.372618209962342
sim(7,10) :    OLP --> 0.10886988506366437
sim(8,9) :   OLP --> 0.18437200866634249
sim(8,10) :    OLP --> 0.0771029097610982
sim(9,10) :    OLP --> 0.10886988506366437

Histogram:
(0.0,0.10000000149011612): {8}
(0.10000000149011612,0.20000000298023224): {29}
(0.20000000298023224,0.30000000447034836): {9}
(0.30000000447034836,0.4000000059604645): {2}
(0.4000000059604645,0.5000000074505806): {2}
(0.5000000074505806,0.6000000089406967): {0}
(0.6000000089406967,0.7000000104308128): {1}
(0.7000000104308128,0.800000011920929): {1}
(0.800000011920929,0.9000000134110451): {1}
(0.9000000134110451,1.0000000149011612): {1}

Clusters With the Obtained OLP :
Cluster1: {0, 2, 3, 8}
Cluster2: {1, 5}
Cluster3: {4, 6}
Cluster4: {7, 9}
Cluster5: {10}
```

Figure 5: Efficient Cluster Formation

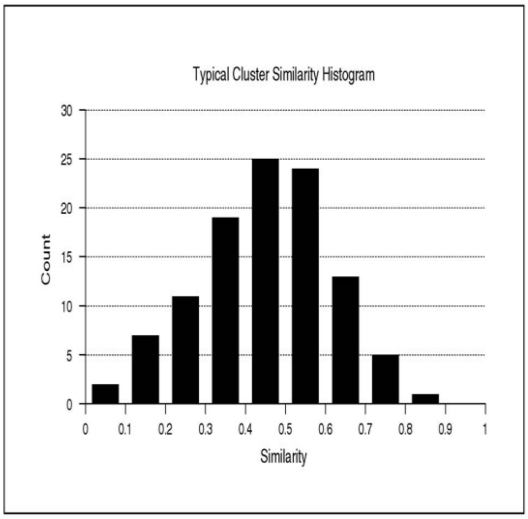


Figure 6: Efficiency Improved Results

V. Conclusion and Future Works

In this paper we proposed a Clustering with Multi-viewpoint-Based Similarity Measure and it also named as MVS. This method is analysed by theoretically and empirically for potential suitable for the text documents other than the popular cosine similarity. Based on the MVS methods we proposed the two criterion functions for the document to achieve the similarities. MVSC-IR and MVSC-IV have been introduced. Compared with other state-of-the-art clustering methods that use different types of similarity measure, on a large number of document data sets and under different evaluation metrics, the proposed algorithms show that they could provide significantly improved clustering performance.

The key contribution of this paper is the fundamental concept of similarity measure from multiple viewpoints. Future methods could make use of the same principle, but define alternative forms for the relative similarity in (10), or do not use average but have other methods to combine the relative similarities according to the different viewpoints. Besides, this paper focuses on partitional clustering of documents. In the future, it would also be possible to apply the proposed criterion functions for hierarchical clustering algorithms. Finally, we have shown the application of MVS and its clustering algorithms for text data. It would be interesting to explore how they work on other types of sparse and high-dimensional data.

VI. References

[1] X. Wu, V. Kumar, J.R. 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,” Knowledge Information Systems, vol. 14, no. 1, pp. 1-37, 2007.

[2] Guyon, U.V. Luxburg, and R.C. Williamson, “Clustering: Science or Art?,” Proc. NIPS Workshop Clustering Theory, 2009.

[3] Dhillon and D. Modha, “Concept Decompositions for Large Sparse Text Data Using Clustering,” Machine Learning, vol. 42, nos. 1/2, pp. 143-175, Jan. 2001.

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

[5] Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, “Clustering with Bregman Divergences,” J. Machine Learning Research, vol. 6, pp. 1705-1749, Oct. 2005.

[6] Pekalska, A. Harol, R.P.W. Duin, B. Spillmann, and H. Bunke, “Non-Euclidean or Non-Metric Measures Can Be Informative,” Structural, Syntactic, and Statistical Pattern Recognition, vol. 4109, pp. 871-880, 2006

[7] M. Pelillo, “What Is a Cluster? Perspectives from Game Theory,” Proc. NIPS Workshop Clustering Theory, 2009.

[8] D. Lee and J. Lee, “Dynamic Dissimilarity Measure for Support Based Clustering,” IEEE Trans. Knowledge and Data Eng., vol. 22, no. 6, pp. 900-905, June 2010.

[9] Banerjee, I. Dhillon, J. Ghosh, and S. Sra, “Clustering on the Unit Hypersphere Using Von Mises-Fisher Distributions,” J. Machine Learning Research, vol. 6, pp. 1345-1382, Sept. 2005.

[10] W. Xu, X. Liu, and Y. Gong, “Document Clustering Based on Non-Negative Matrix Factorization,” Proc. 26th Ann. Int’l ACM SIGIR Conf. Research and Development in Informaion Retrieval, pp. 267-273, 2003.

[11] I.S. Dhillon, S. Mallela, and D.S. Modha, “Information-Theoretic Co-Clustering,” Proc. Ninth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 89-98, 2003.

[12] C.D. Manning, P. Raghavan, and H. Schu¨ tze, An Introduction to Information Retrieval. Cambridge Univ. Press, 2009.

[13] Ding, X. He, H. Zha, M. Gu, and H. Simon, “A Min-Max Cut Algorithm for Graph Partitioning and Data Clustering,” Proc. IEEE Int’l Conf. Data Mining (ICDM), pp. 107-114, 2001.

[14] H. Zha, X. He, C. Ding, H. Simon, and M. Gu, “Spectral Relaxation for K-Means Clustering,” Proc. Neural Info. Processing Systems (NIPS), pp. 1057-1064, 2001.

[15] J. Shi and J. Malik, “Normalized Cuts and Image Segmentation,” IEEE Trans. Pattern Analysis Machine Intelligence, vol. 22, no. 8, pp. 888-905, Aug. 2000.

[16] I.S. Dhillon, “Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning,” Proc. Seventh ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 269-274, 2001.

[17] Y. Gong and W. Xu, Machine Learning for Multimedia Content Analysis. Springer-Verlag, 2007.

[18] Y. Zhao and G. Karypis, “Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering,” Machine Learning, vol. 55, no. 3, pp. 311-331, June 2004.

[19] G. Karypis, “CLUTO a Clustering Toolkit,” technical report, Dept. of Computer Science, Univ. of Minnesota, http://glaros.dtc.umn.edu/~gkhome/views/cluto, 2003.

[20] Strehl, J. Ghosh, and R. Mooney, “Impact of Similarity Measures on Web-Page Clustering,” Proc. 17th Nat’l Conf. Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI), pp. 58-64, July 2000.

[21] 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.

[22] Ienco, R.G. Pensa, and R. Meo, “Context-Based Distance Learning for Categorical Data Clustering,” Proc. Eighth Int’l Symp. Intelligent Data Analysis (IDA), pp. 83-94, 2009.

[23] P. Lakkaraju, S. Gauch, and M. Speretta, “Document Similarity Based on Concept Tree Distance,” Proc. 19th ACM Conf. Hypertext and Hypermedia, pp. 127-132, 2008.

[24] H. Chim and X. Deng, “Efficient Phrase-Based Document Similarity for Clustering,” IEEE Trans. Knowledge and Data Eng., vol. 20, no. 9, pp. 1217-1229, Sept. 2008.

[25] Flesca, G. Manco, E. Masciari, L. Pontieri, and A. Pugliese, “Fast Detection of xml Structural Similarity,” IEEE Trans. Knowledge and Data Eng., vol. 17, no. 2, pp. 160-175, Feb. 2005.

[26] E.-H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore, “Webace: A Web Agent for Document Categorization and Exploration,” Proc. Second Int’l Conf. Autonomous Agents (AGENTS ’98), pp. 408-415, 1998.

[27] J. Friedman and J. Meulman, “Clustering Objects on Subsets of Attributes,” J. Royal Statistical Soc. Series B Statistical Methodology, vol. 66, no. 4, pp. 815-839, 2004.

[28] L. Hubert, P. Arabie, and J. Meulman, Combinatorial Data Analysis: Optimization by Dynamic Programming. SIAM, 2001.

[29] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, second ed. John Wiley & Sons, 2001.

[30] S. Zhong and J. Ghosh, “A Comparative Study of Generative Models for Document Clustering,” Proc. SIAM Int’l Conf. Data Mining Workshop Clustering High Dimensional Data and Its Applications, 2003.

[31] Zhao and G. Karypis, “Criterion Functions for Document Clustering: Experiments and Analysis,” technical report, Dept. of Computer Science, Univ. of Minnesota, 2002.

[32] T.M. Mitchell, Machine Learning. McGraw-Hill, 1997.

[33] a.k. jain, m.n. murty and p.j. Flynn, “Data Clustering: A Review” ACM Computing Surveys, Vol. 31, No. 3, September 1999

[34] Yaminee S. Patil and M.B.Vaidya , “A Technical Survey onCluster Analysis inData Mining”- ISSN 2250-2459, Volume 2, Issue 9, September 2012

[35] ravi Bhusan Yadav and m. Madhu Babu, “Measurement Of Multiview Point Similarity Considering Concepts Of Data Clustering”- Vol. 3, Issue 2, March -April 2013, pp.1299-1305

[36] Guadalupe J. Torres, Ram B. Basnet, Andrew H. Sung, Srinivas Mukkamala, Bernardete M. Ribeiro, “Similarity Measure for Clustering and Its Applications”

[37] Duc Thang Nguyen,Lihui Chen, Chee Keong Chan, “Multi-viewpoint Based Similarity Measure and Optimality Criteria for Document Clustering”