Document Clustering using Multi-View Points

A.Susheel Kumar
M.Tech Student, CVSR College of Engineering, Department of Computer Science, A.P. India

S.Kalyani
Associate Professor of Computer Science and Engineering, Anurag group of Institutions, A.P. India

Abstract - All clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. In this paper, we introduce a novel multi view point-based similarity measure and two related clustering methods. The major difference between a traditional dissimilarity/similarity measure and ours is that the former uses only a single viewpoint, which is the origin, while the latter utilizes many different viewpoints, which are objects assumed to not be in the same cluster with the two objects being measured. Using multiple viewpoints, more informative assessment of similarity could be achieved. Theoretical analysis and empirical study are conducted to support this claim. Two criterion functions for document clustering are proposed based on this new measure. We compare them with several well-known clustering algorithms that use other popular similarity measures on various document collections to verify the advantages of our proposal.

Index Terms— Document clustering, text mining, similarity measure.

I. INTRODUCTION

CLUSTERING is one of the most interesting and important topics in data mining. The aim of clustering is to find intrinsic structures in data, and organize them into meaningful subgroups for further study and analysis. There have been many clustering algorithms published every year. They can be proposed for very distinct research fields, and developed using totally different techniques and approaches. Nevertheless, according to a recent study [1], more than half a century after it was 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 zmention, 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. 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. A common approach to the clustering problem is to treat it as an optimization process. An optimal partition is found by optimizing a particular function of similarity (or distance) among data.

II. RELATED WORK

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. The principle definition of clustering is to arrange data objects into separate clusters such that the intra cluster similarity as well as the inter cluster 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], and nonnegative matrix factorization [10], information theoretic co clustering [11] and so on. In this paper, though, we primarily focus on methods that indeed do utilize a specific measure.

III. PROPOSED ALGORITHM

A. QT clustering algorithm

QT (quality threshold) clustering is an alternative method of partitioning data, invented for gene clustering. It requires more computing power than k-means, but does not require specifying the number of clusters a priori, and always returns the same result when run several times.The user chooses a maximum diameter for clusters. Build a candidate cluster for each point by including the closest point, the next closest, and so on, until the diameter of the cluster surpasses the threshold. Save the candidate cluster with the most points as the first true cluster, and remove all points in the cluster from further consideration. Must clarify what happens if more than 1 cluster has the maximum number of points? Recurse with the reduced set of points.

B. Comparisons between data clustering's

There have been several suggestions for a measure of similarity between two clustering's. Such a measure can be used to compare how well different data clustering algorithms perform on a set of data. Many of these measures are derived from the matching matrix (aka confusion matrix), e.g., the Rand measure and the Fowlkes-Mallows B_k measures. Several different clustering systems based on mutual information have been proposed. One is Marina Meila's 'Variation of Information' metric and another provides hierarchical clustering.

C. Hierarchical Document Clustering Using Frequent Item sets

Document clustering has been studied intensively because of its wide applicability in areas such as web mining, search engines, information retrieval, and topological analysis. Unlike in document classification, in document clustering no labeled documents are provided. Although standard clustering techniques such as k-means can be applied to document clustering, they usually do not satisfy the special requirements for clustering documents: high dimensionality, high volume of data, ease for browsing, and meaningful cluster labels. In addition, many existing document clustering algorithms require the user to specify the number of clusters as an input parameter and are not robust enough to handle different types of document sets in a real-world environment. For example, in some document sets the cluster size varies from few to thousands of documents. This variation tremendously reduces the clustering accuracy for some of the state-of-the art algorithms. *Frequent Itemset-based Hierarchical Clustering (FIHC)*, for document clustering based on the idea of *frequent item sets* proposed by Agrawalet. al. The intuition of our clustering criterion is that there are some frequent item sets for each cluster (topic) in the document set, and different clusters share few frequent item sets. A frequent item set is a set of words that occur together in some minimum fraction of documents in a cluster. Therefore, a frequent item set describes something common to many documents in a cluster. In this technique use frequent item sets to construct clusters and to organize clusters into a topic hierarchy. Here are the features of this approach.

*Reduced dimensionality***.** This approach uses only the frequent items that occur in some minimum fraction of documents in document vectors, which drastically reduces the dimensionality of the document set. Experiments show that clustering with reduced dimensionality is significantly more efficient and scalable. This decision is consistent with the study from linguistics (Longman Lancaster Corpus) that only 3000 words are required to cover 80% of the written text in English and the result is coherent with the Zipf's law and the findings in Mladenic et al. and Yang et al.

High clustering accuracy. Experimental results show that the proposed approach FIHC outperforms best documents clustering algorithms in terms of accuracy. It is robust even when applied to large and complicated document sets.

*Number of clusters as an optional input parameter***.** Many existing clustering algorithms require the user to specify

the desired number of clusters as an input parameter. FIHC treats it only as an optional input parameter. Close to optimal clustering quality can be achieved even when this value is unknown.

IV. MULTI-VIEWPOINTBASED SIMILARITY

The cosine similarity in Eq. (3) can be expressed in the following form without changing its meaning:

$$Sim(di, dj) = \cos(di-0, dj-0) = (di-0)t\,(dj-0)$$

Where 0 is vector 0 that represents the origin point. According to this formula, the measure takes 0 as one and only reference point. The similarity between two documents *di* and *dj* is determined w.r.t. the angle between the two points when looking from the origin. To construct a new concept of similarity, it is possible to use more than just one point of reference. We may have a more accurate assessment of how close or distant a pair of points are, if we look at them from many different viewpoints [2][3]. From a third point *dh*, the directions and distances to *di* and *dj* are indicated respectively by the difference vectors (*di* − *dh*) and (*dj* − *dh*). By standing at various reference points *dh* to view *di*, *dj* and working on their difference vectors, we define similarity between the two documents as:

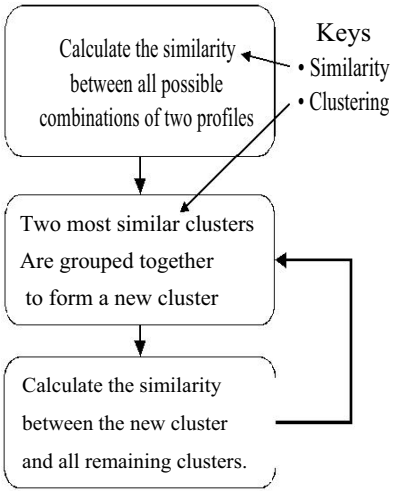
A. Analysis and practical examples of MVS

In this section, we present analytical study to show that the proposed MVS could be a very effective similarity measure for data clustering. In order to demonstrate its advantages, MVS is compared with cosine similarity(CS) on how well they reflect the true group structure in document collections Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters)[6], so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. The computational task of classifying the data set into k clusters is often referred to as k-clustering. Besides the term data clustering (or just clustering), there are a number of terms with similar meanings, including cluster analysis, automatic classification, numerical taxonomy, botryology and typological analysis [6].Document clustering aims to group, in an unsupervised way, a given document set into clusters such that documents within each cluster are more similar between each other than those in different clusters. It is an enabling technique for a wide range of information retrieval tasks such as efficient organization, browsing and summarization of large volumes of text documents. Cluster analysis aims to organize a collection of patterns into clusters based on similarity. Clustering has its root in many fields, such as mathematics, computer science, statistics, biology, and economics. In different application domains, a variety of clustering techniques have been developed, depending on the methods used to represent data, the measures of similarity between data objects, and the techniques for grouping data objects into clusters.

V. HIERARCHICAL ANALYSIS MODEL

A hierarchical clustering algorithm creates a hierarchical decomposition of the given set of data objects. Depending on the decomposition approach, hierarchical algorithms are classified as agglomerative (merging) or divisive (splitting). The agglomerative approach starts with each data point in a separate cluster or with a certain large number of clusters. Each step of this approach merges the two clusters that are the most similar. Thus after each step, the total number of clusters decreases. This is repeated until the desired number of clusters is obtained or only one cluster remains. By contrast, the divisive approach starts with all data objects in the same cluster. In each step, one cluster is split into smaller clusters, until a termination condition holds. Agglomerative algorithms are more widely used in practice. Thus the similarities between clusters are more researched [9].

Hierarchical Clustering



VI. Experimental Setup and Evaluation

To demonstrate how well MVSCs can perform, we compare them with five other clustering methods on the 20 data sets. In summary, the seven clustering algorithms are . MVSC-IR: MVSC using criterion function IR . MVSC-IV : MVSC using criterion function IV . k-means: standard k-means with euclidean distance . Spkmeans: spherical k-means with CS . graphCS: CLUTO's graph method with CS . graphEJ: CLUTO's graph with extended Jaccard . MMC: Spectral Min-Max Cut algorithm [13].Our MVSC-IR and MVSC-IV programs are implemented in Java. The regulating factor _ in IR is always set at 0.3 during the experiments. We observed that this is one of the most appropriate values. A study on MVSC-IR's performance relative to different _ values is presented in a later section. The other algorithms are provided by the C library interface which is available freely with the CLUTO toolkit [9]. For each data set, cluster number is predefined equal to the number of true class, i.e., k ¼ c.None of the above algorithms are guaranteed to find global optimum, and all of them are initialization dependent. Hence, for each method, we performed clustering a few times with randomly initialized values, and chose the best trial in terms of the corresponding objective function value. In all the experiments, each test run consisted of 10 trials. Moreover, the result reported here on each data set by a particular clustering method is the average of 10 test runs. After a test run, clustering solution is evaluated by comparing the documents' assigned labels with their true labels provided by the corpus. Three types of external evaluation metric are used to assess clustering performance. They are the FScore, Normalized Mutual Information (NMI),and accuracy.

VII. FUTURE WORK

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.

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 Algorithms in Data Mining," Knowledge Information Systems, vol. 14, no. 1, pp. 1-37, 2007.

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

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

[6] E. 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] A. 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. Schutze, An Introduction to Information Retrieval. Cambridge Univ. Press, 2009.

[13] C. 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.