Volume 3, Issue 11, November 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Document Clustering with Similarity Measures

Ashish Moon*
Department of M.Tech CSE
R.T.M. Nagpur University, India

T. Raju
Asst. Prof. Department of M.Tech CSE
R.T.M. Nagpur University, India

Abstract— Clustering is one of the most important data mining or text mining algorithm that is used to group similar objects together. It aims to find essential structures in data, and arrange them into meaningful subgroups. The main difference of the novel method from the existing one is that it uses only single view point for clustering and where as in Multi-Viewpoint based similarity measure uses many different viewpoints. In Multi-Viewpoint the two objects are measured are assumed to not be in the same cluster. In this paper, we introduce Hierarchical Clustering with Multiple view points based on different similarity measures. We use two measures for intercluster and intracluster relation between objects. The former clustering process focuses on partitional clustering of multi viewpoint documents, which are not focused on sparse and high dimensional data. Using Hierarchical Multiview point, we can achieve more informative evaluation of similarity.

Keywords— Document Clustering, Similarity Measure, Text Mining, K-Mean Clustering Algorithm, Hierarchical Methods, High Dimensional Data

I. OVERVIEW

Clustering is a process of grouping a set of objects into classes of similar objects and is the most interesting concept of data mining. Purpose of Clustering is to group fundamental structures in data and classify them into meaningful subgroups. There have been many clustering algorithms published every year. K-means is one of the top most data mining algorithms. Even though it is a top most algorithm, it has a few basic disadvantages such as sensitiveness to initialization and to cluster size. Its performance can be worse than other algorithms in many areas. In spite of that, its simplicity, understandability, and scalability are the reasons for its incredible popularity. While offering reasonable results, K-means is fast and easy to combine with other methods in larger systems. The efficiency of clustering algorithms depends on the accuracy of the similarity measure to the data. Similarity is measure that reflects the strong point of relationship between two objects or two features.

SIMILARITY MEASURE

Some of the similarity measures explained based on single view point.

A. Euclidean Distance

It is common distance between two points and can be without difficulty measured with a ruler in two or three dimensional space. Euclidean distance is one of the most popular measures:

$$\text{Dist}(di, dj) = \|di - dj\|$$

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 centroid:

$$min \sum_{r=1}^{K} \sum_{di \in Sr} ||di - cr||^2$$

Particularly, similarity of two documents vector di and dj, Sim (di, dj), is defined as the cosine of angle between them. For unit vectors, this equals to their inner product:

$$\text{sim}(di, dj) = \cos(di, dj) = di^t\ dj$$

B. Cosine Similarity

When documents are represented in terms vectors, the similarity of two documents corresponds to the correlation between the vectors. In a sparse and high dimensional space, cosine similarity is widely used. It is also a popular similarity score in text mining and information retrieval.

$$\max \sum_{r=1}^{k} \sum_{di \in Sr} \frac{di^t\ Cr}{||Cr||}$$

C. Jaccard Coefficient

It sometimes referred to as the Tanimoto coefficient, measures similarity as the intersection divided by the union of the objects. For text 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.

$$\mathrm{Sim}_{\mathrm{eJacc}}[\mathrm{Ui}, \mathrm{Uj}] = \frac{\mathrm{Ui}^{\mathrm{t}}\ \mathrm{Uj}}{\|\mathrm{Ui}\|^2 + \|\mathrm{Uj}\|^2 - \mathrm{Ui}^{\mathrm{t}}\ \mathrm{Uj}}$$

D. Pearson Correlation Measure

It provides a method for clustering a set of objects into the set of objects into the best possible number of clusters, without specifying that number in proceed. The normalized Pearson correlation defined as:

$$S(x_i, x_j) = \frac{(X_i - \overline{x_I})^T\ (x_j - \overline{X_j})}{\|X_i - \overline{x_I}\| x_j - X_I\|}$$

Where $\overline{x_I}$ denotes the average feature value of x over all dimensions.

II. DOCUMENT CLUSTERING

It is automatic grouping of text documents into clusters so that documents within a cluster have high similarity in comparison to one another, but dissimilar documents in other clusters. To accomplish more correct document clustering, more informative features including phrases and their weights are particularly important in such scenarios. Document vectors are the result of some sort of weighting schemes like TF-IDF (Term Frequency –Inverse Document Frequency). Many approaches came into existence for document clustering. They include information theoretic co-clustering, non–negative matrix factorization, and probabilistic model based method and so on. However, these approaches did not use specific measure in finding document similarity.

Document clustering is particularly useful in many applications such as automatic classification of documents, grouping search engine results, building classification of documents, and others. For this Hierarchical Clustering method provides a better improvement in achieving the result. Hierarchical document clustering categorizes clusters into a tree or a hierarchy that facilitates browsing. Most popular document clustering methods are:

1. K-means Clustering
2. Hierarchical Agglomerative Clustering

K-means clustering is a partitioning method. It is one of the popular clustering algorithms in the world. Due to its simplicity and ease of use it is top 10 in data mining domain. Euclidean distance measure is used in k-means algorithm. The main purpose of the k-means algorithm is to minimize the distance, as per Euclidian measurement, between objects in clusters. K-means clustering is a method of cluster analysis. It aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

Basic K-means Algorithm for finding *K clusters.*

1. Select *K* points as the initial centroids.
2. Assign all points to the closest centroid.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids don't change (or change very little).

Hierarchical clustering methods are classified into agglomerative (bottom-up) and divisive (top-down). An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more most suitable clusters.

Steps in Hierarchical Agglomerative Clustering

1. Start with N clusters, each containing a single entity, and an N × N symmetric matrix of distances (or similarities).Let dij = distance between item i and item j.
2. Search the distance matrix for the nearest pair clusters (i.e., the two clusters that are separated by the smallest distance). Denote the distance between these most similar clusters A and B by dAB.
3. Merge clusters A and B into a new cluster, labelled S. Update the entries in the distance matrix by
a. Deleting the rows and columns corresponding to clusters A and B, and
b. Adding a row and column giving the distances between the new cluster S and all the remaining clusters.
4. Repeat steps (2.) and (3.) a total of N-1 times.

HIGH DIMENSIONAL DATA

The rapid growth in various new application domains, like bioinformatics and e-commerce, reflects the need for studying high dimensional data. Thus mining high dimensional data is an urgent problem of great realistic importance. In

a gene expression microarray data set, there could be tens or hundreds of dimensions, each of which corresponds to an experimental condition. In a customer purchase behaviour data set, there may be up to hundreds of thousands of merchandizes, each of which is mapped to a dimension. Researchers and practitioners are very eager in analysing these data sets. Various data mining models have been proven to be very successful for examining very large data sets. In data mining, the objects can have hundreds of attributes or dimensions. Clustering in such high dimensional data spaces presents a tremendous difficulty, much more so than in predictive learning. However, there are some unique challenges for mining high dimensional data including

1) The clustering tendency will lose when the dataset contains irrelevant attributes. Searching for clusters is a hopeless enterprise where there are no relevant attributes for finding clusters. Attribute selection is the best approach to address the problem of selecting irrelevant attributes.
2) Dimensionality curse is another problem in high dimensional data. As the number of attributes or dimensions increases in a dataset, the distance measures will become increasingly meaningless. The resultant clusters with high dimensions; they are equidistant from each other.

III. EXISTING APPROACH

- The aim of clustering is to find essential structures in data, and arrange them into meaningful subgroups for further study and analysis. There have been many clustering algorithms published every year.
- Existing Systems greedily selects the next frequent item set which represent the next cluster to reduce the overlapping between the documents that contain both the item set and some remaining item sets.
- The clustering result depends on the order of selection of the item sets, which in turns depends on the greedy heuristic. This method does not follow a chronological order of selecting clusters. Instead, we allocate documents to the best cluster.

IV. PROPOSED APPROACH

- In our proposed method, we are using correlation similarity and cosine similarity to measure the similarity between objects in the same cluster and dissimilarity between objects in the different cluster groups.
- The main work is to develop multiviewpoint based algorithm for document clustering which provides maximum efficiency and performance.
- It is particularly focused in making the use of cluster overlapping phenomenon to design cluster merging criteria. So suggest a new way to calculate the overlap rate in order to improve time efficiency of clustering and save computing time.

V. CONCLUSION

Choosing a clustering algorithm however can be a difficult task. For that we propose Hierarchical clustering method based on Multiview point similarity measuring method. Hierarchical MVS is suitable for sparse and high dimensional data compared with partitional MVS clusters. The new method can decrease the time cost, reduce the space complexity and improve the accuracy of clustering. Specially, in the document clustering, the newly proposed algorithm measuring result show great advantages.

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

[3] E. Mooi and M. Sarstedt, "Cluster Analysis", DOI10.1007/978-3-642-12541-6_9, ©Springer-Verlag Berlin Heidelberg 2011.

[4] A.K. JAIN, M.N. MURTY and P.J. FLYNN, "Data Clustering: A Review", ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[5] 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.

[6] Anna Huang, "Similarity Measures for Text Document Clustering", NZCSRSC 2008, April 2008, Christchurch, New Zealand.

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

[8] 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.

[9] D. 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.

[10] 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.

