International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 14
October 2016

It Is Combined With the Similarity Measure Based Multi Viewpoint

[1] Yedla Saikiran, [2] Suvarna.
1.PG Scholar, Department of CSE, RAJAMAHENDRA ENGINEERING COLLEGE, Hyderabad
2.M-tech, Associate professor, Department of CSE, RAJAMAHENDRA ENGINEERING COLLEGE, Hyderabad.

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-viewpoint 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.

1.INTRODUCTION

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), 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.

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.

2.LITERATURE SURVEY

2.1 TYPES OF CLUSTERING

Data clustering algorithms can be hierarchical. Hierarchical algorithms find successive clusters using previously established clusters. Hierarchical algorithms can be agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

Partitional algorithms typically determine all clusters at once, but can also be used as divisive algorithms in the hierarchical clustering.

Two-way clustering, co-clustering or biclustering are clustering methods where not only the objects are clustered but also the features of the objects, i.e., if the data is represented in a data matrix, the rows and columns are clustered simultaneously.

2.2 DISTANCE MEASURE

An important step in any clustering is to select a distance measure, which will determine how the similarity of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and further away according to another. For example, in a 2-dimensional space, the distance between the point (x=1, y=0) and the origin (x=0, y=0) is always 1 according to the usual norms, but the distance between the point (x=1, y=1) and the origin can be 2, $\sqrt{2}$ or 1 if you take respectively the 1-norm, 2-norm or infinity-norm distance.

Common distance functions:

- The Euclidean distance (also called distance as the crow flies or 2-norm distance). A review of cluster analysis in health psychology research found that the most common distance measure in published studies in that research area is the Euclidean distance or the squared Euclidean distance.
- The Manhattan distance (also called taxicab norm or 1-norm)
- The maximum norm
- The Mahalanobis distance corrects data for different scales and correlations in the variables
- The angle between two vectors can be used as a distance measure when clustering high dimensional data. See Inner product space.
- The Hamming distance (sometimes edit distance) measures the minimum number of substitutions required to change one member into another.

2.3 HIERARCHICAL CLUSTERING

Creating clusters

Hierarchical clustering builds (agglomerative), or breaks up (divisive), a hierarchy of clusters. The traditional representation of this hierarchy is a tree (called a dendrogram), with individual elements at one end and a single cluster containing every element at the other. Agglomerative algorithms begin at the leaves of the tree, whereas divisive algorithms begin at the root.

Cutting the tree at a given height will give a clustering at a selected precision. In the following example, cutting after the second row will yield clusters {a} {b c} {d e} {f}. Cutting after the third row will yield clusters {a} {b c} {d e f}, which is a coarser clustering, with a smaller number of larger clusters.

2.4 AGGLOMERATIVE HIERARCHICAL CLUSTERING

For example, suppose this data is to be clustered, and the euclidean distance is the distance metric.

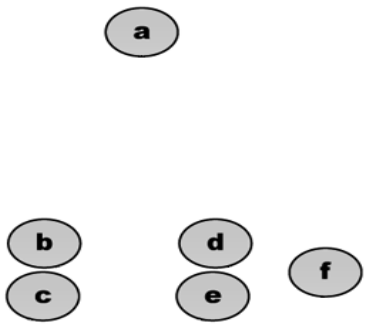


Fig 2.1 Raw data

The hierarchical clustering Dendogram would be as such:

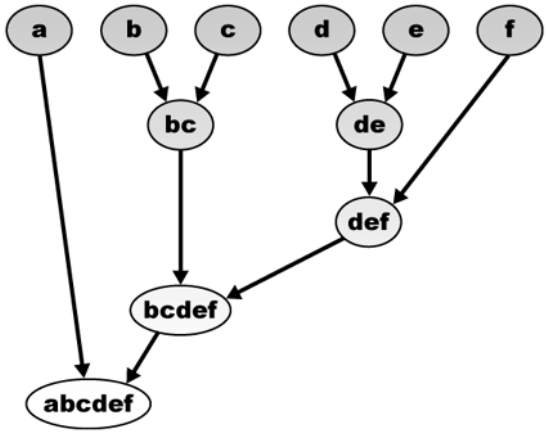


Fig 2.2 Traditional representation

This method builds the hierarchy from the individual elements by progressively merging clusters. In the example, six elements {a} {b} {c} {d} {e} and {f} are represented. The first step is to determine which elements to merge in a cluster. Usually, the major focus is to take the two closest elements, according to the chosen distance.

Optionally, one can also construct a distance matrix at this stage, where the number in the i-th row j-th column is the distance between the i-th and j-th elements. Then, as clustering progresses, rows and columns are merged as the clusters are merged and the distances updated. This is a common way to implement this type of clustering, and has the benefit of caching distances between clusters. A simple agglomerative clustering algorithm is described in the single-linkage clustering page; it can easily be adapted to different types of linkage.

Usually the distance between two clusters $\mathcal{A}$ and $\mathcal{B}$ is one of the following:

- The maximum distance between elements of each cluster (also called complete linkage clustering):

$$\max\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B}\,\}$$

- The minimum distance between elements of each cluster (also called single-linkage clustering):

$$\min\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B}\,\}$$

- The mean distance between elements of each cluster (also called average linkage clustering, used e.g. in UPGMA):

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x,y)$$

- The sum of all intra-cluster variance
- The increase in variance for the cluster being merged (Ward's criterion)
- The probability that candidate clusters spawn from the same distribution function (V-linkage)

Each agglomeration occurs at a greater distance between clusters than the previous agglomeration, and one can decide

to stop clustering either when the clusters are too far apart to be merged (distance criterion) or when there is a sufficiently small number of clusters (number criterion).

2.5 CONCEPT CLUSTERING

Another variation of the agglomerative clustering approach is conceptual clustering.

PARTITIONAL CLUSTERING

K-means clustering

The K-means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster — that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.

The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments. It minimizes intra-cluster variance, but does not ensure that the result has a global minimum of variance.

Fuzzy c-means clustering

In fuzzy clustering, each point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center of cluster. For each point x we have a coefficient giving the degree of being in the kth cluster $u_k(x)$. Usually, the sum of those coefficients is defined to be 1:

$$\forall x \quad \sum_{k=1}^{\text{num. clusters}} u_k(x) = 1.$$

With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$\text{center}_k = \frac{\sum_x u_k(x)^m x}{\sum_x u_k(x)^m}.$$

The degree of belonging is related to the inverse of the distance to the cluster center:

$$u_k(x) = \frac{1}{d(\text{center}_k, x)},$$

then the coefficients are normalized and fuzzyfied with a real parameter m > 1 so that their sum is 1. So

$$u_k(x) = \frac{1}{\sum_j \left(\frac{d(\text{center}_k, x)}{d(\text{center}_j, x)} \right)^{2/(m-1)}}.$$

For m equal to 2, this is equivalent to normalising the coefficient linearly to make their sum 1. When m is close to 1, then cluster center closest to the point is given much more weight than the others, and the algorithm is similar to k-means.

The fuzzy c-means algorithm is very similar to the k-means algorithm:

- Choose a number of clusters.
- Assign randomly to each point coefficients for being in the clusters.
- Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than ε, the given sensitivity threshold) :
  - Compute the centroid for each cluster, using the formula above.
  - For each point, compute its coefficients of being in the clusters, using the formula.

The algorithm minimizes intra-cluster variance as well, but has the same problems as k-means, the minimum is a local minimum, and the results depend on the initial choice of weights. The Expectation-maximization algorithm is a more statistically formalized method which includes some of these ideas: partial membership in classes. It has better convergence properties and is in general preferred to fuzzy-c-means.

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.

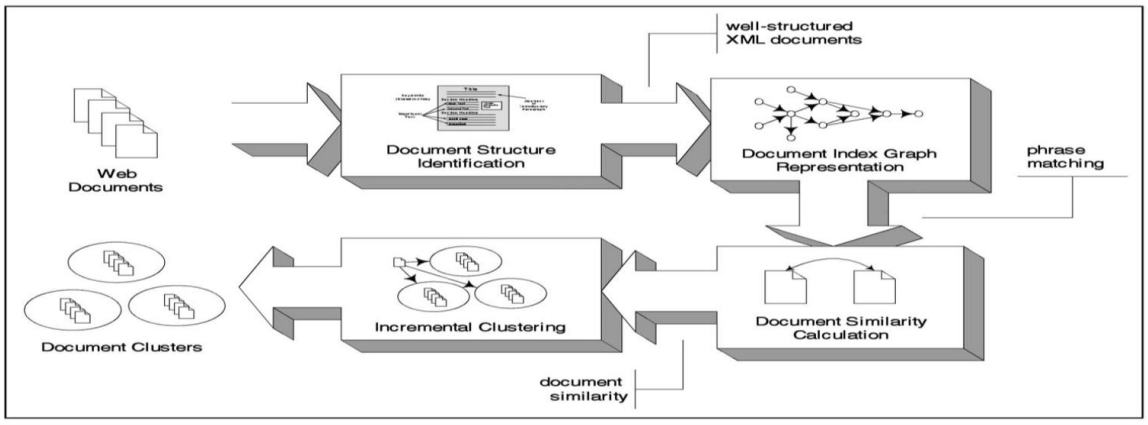
EXISTING SYSTEMS

- 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.

- Existing Systems greedily picks the next frequent item set which represent the next cluster to minimize the overlapping between the documents that contain both the item set and some remaining item sets.
- In other words, the clustering result depends on the order of picking up the item sets, which in turns depends on the greedy heuristic. This method does not follow a sequential order of selecting clusters. Instead, we assign documents to the best cluster.

PROPOSED SYSTEM

- The main work is to develop a novel hierarchal algorithm for document clustering which provides maximum efficiency and performance.
- It is particularly focused in studying and making use of cluster overlapping phenomenon to design cluster merging criteria. Proposing a new way to compute the overlap rate in order to improve time efficiency and "the veracity" is mainly concentrated. Based on the Hierarchical Clustering Method, the usage of Expectation-Maximization (EM) algorithm in the Gaussian Mixture Model to count the parameters and make the two sub-clusters combined when their overlap is the largest is narrated.
- Experiments in both public data and document clustering data show that this approach can improve the efficiency of clustering and save computing time.



Given a data set satisfying the distribution of a mixture of Gaussians, the degree of overlap between components affects the number of clusters "perceived" by a human

operator or detected by a clustering algorithm. In other words, there may be a significant difference between intuitively defined clusters and the true clusters corresponding to the components in the mixture.

MODULES

- HTML PARSER
- CUMMULATIVE DOCUMENT
- DOCUMENT SIMILARITY
- CLUSTERING

MODULE DESCRIPTION:

HTML Parser

- Parsing is the first step done when the document enters the process state.
- Parsing is defined as the separation or identification of meta tags in a HTML document.
- Here, the raw HTML file is read and it is parsed through all the nodes in the tree structure.

Cumulative Document

- The cumulative document is the sum of all the documents, containing meta-tags from all the documents.
- We find the references (to other pages) in the input base document and read other documents and then find references in them and so on.
- Thus in all the documents their meta-tags are identified, starting from the base document.

Document Similarity

- The similarity between two documents is found by the cosine-similarity measure technique.
- The weights in the cosine-similarity are found from the TF-IDF measure between the phrases (meta-tags) of the two documents.
- This is done by computing the term weights involved.
- TF = C / T
- IDF = D / DF.

D → quotient of the total number of documents

DF → number of times each word is found in the entire corpus

C → quotient of no of times a word appears in each document

T → total number of words in the document

- **TFIDF = TF * IDF**

Clustering

- Clustering is a division of data into groups of similar objects.
- Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification.

The similar documents are grouped together in a cluster, if their cosine similarity measure is less than a specified threshold

CONCLUSION

Given a data set, the ideal scenario would be to have a given set of criteria to choose a proper clustering algorithm to apply. Choosing a clustering algorithm, however, can be a difficult task. Even ending just the most relevant approaches for a given data set is hard. Most of the algorithms generally assume some implicit structure in the data set. One of the most important elements is the nature of the data and the nature of the desired cluster. Another issue to keep in mind is the kind of input and tools that the algorithm requires. This report has a proposal of a new hierarchical clustering algorithm based on the overlap rate for cluster merging. The experience in general data sets and a document set indicates that 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. The hierarchical document clustering algorithm provides a natural way of distinguishing clusters and implementing the basic requirement of clustering as high within-cluster similarity and between-cluster dissimilarity.

FUTURE WORKS

In the proposed model, selecting different dimensional space and frequency levels leads to different accuracy rate in the clustering results. How to extract the features reasonably will be investigated in the future work.

There are a number of future research directions to extend and improve this work. One direction that this work might continue on is to improve on the accuracy of similarity calculation between documents by employing different similarity calculation strategies. Although the current scheme proved more accurate than traditional methods, there are still rooms for improvement.

REFERENCES

1) Cole, A. J. & Wishart, D. (1970). An improved algorithm for the Jardine-Sibson method of generating overlapping clusters. The Computer Journal 13(2):156-163.
2) D'andrade,R. 1978, "U-Statistic Hierarchical Clustering" Psychometrika, 4:58-67.
3) Johnson,S.C. 1967, "Hierarchical Clustering Schemes" Psychometrika, 2:241-254.
4) Shengrui Wang and Haojun Sun. Measuring overlap-Rate for Cluster Merging in a Hierarchical Approach to Color Image Segmentation. International Journal of Fuzzy Systems,Vol.6,No.3,September 2004.
5) Jeff A. Bilmes. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. ICSI TR-97-021, U.C. Berkeley, 1998.
6) E.M. Voorhees. Implementing agglomerative hierarchical clustering algorithms for use in document retrieval. Information Processing and Management, 22(6):465–476, 1986.
7) Sun Da-fei,Chen Guo-li,Liu Wen-ju. The discussion of maximum likehood parameter estimation based on EM algorithm. Journal of HeNan University. 2002,32(4):35~41
8) Khaled M. Hammouda, Mohamed S. Kamel , efficient phrase-based document indexing for web document clustering , IEEE transactions on knowledge and data engineering, October 2004
9) Haojun sun, zhihui liu, lingjun kong, A Document Clustering Method Based On Hierarchical Algorithm With Model Clustering, 22nd international conference on advanced information networking and applications,
10) Shi zhong, joydeep ghosh, Generative Model-Based Document Clustering: A Comparative Study, The University Of Texas.