ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Hierarchical Clustering with Multiviewpoint Based Similarity Measure for document Clustering

V. Dhanalakshmi [*1], M. Sabrabeebe [2]

[*1,2] M.E Computer Science Engineering, IFET College of Engineering,Villupuram, Tamil Nadu, India
dhana12ucev@gmail.com

Abstract

A cluster is a group of similar objects placed together and are dissimilar to other cluster objects. In this paper, we introduce Hierarchical Clustering with Multiple view points based on different similarity measures. The major difference between a traditional dissimilarity and similarity measure 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. The main objective is to cluster web documents. Using Hierarchical Multiview point, we can achieve more informative assessment of similarity. We compare our approach with former model on various document collections to verify the advantages of our proposed method.

Keywords: Hierarchical clustering, document clustering, MVP similarity measure.

Introduction

Cluster Analysis

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, biology 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.

About the project

Document clustering techniques mostly rely on single term analysis of the document data set, such as the Vector Space Model. To achieve more accurate document clustering, more informative features including phrases and their weights are particularly important in such scenarios. Document clustering is particularly useful in many applications such as automatic categorization of documents, grouping search engine results, building taxonomy of documents, and others. For this Hierarchical Clustering method provides a better improvement in achieving the result. Our project presents two key parts of successful Hierarchical document clustering. The first part is a document index model, the Document Index Graph, which allows for incremental construction of the index of the document set with an emphasis on efficiency, rather than relying on single-term indexes only. It provides efficient phrase matching that is used to judge the similarity between documents. This model is flexible in that it could revert to a compact representation of the vector space model if we choose not to index phrases. The second part is an incremental document clustering algorithm based on maximizing the tightness of clusters by carefully watching the pair-wise document similarity distribution inside clusters.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Both the phases are based upon two algorithmic models called Gaussian Mixture Model and Expectation Maximization. The combination of these two components creates an underlying model for robust and accurate document similarity calculation that leads to much improved results in Web document clustering over traditional methods.

Related works

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.

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.

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.

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

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

(distance criterion) or when there is a sufficiently small number of clusters (number criterion).

Comparisons between data clusterings

There have been several suggestions for a measure of similarity between two clusterings. 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.

Hierarchical Document Clustering Using Frequent Itemsets

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. Here are the features of this approach.

- *Reduced dimensionality.* This approach use 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.

Challenges In Hierarchical Document Clustering

- High dimensionality: Each distinct word in the document set constitutes a dimension. So there may be 15~20 thousands dimensions. This type of high dimensionality greatly affects the scalability and efficiency of many existing clustering algorithms. This is been cleared described in the following paragraphs.
- High volume of data: In text mining, processing of data about 10 thousands to 100 thousands documents are involved
- Consistently high accuracy: Some existing algorithms only work fine for certain type of document sets, but may not perform well in some others.
- Meaningful cluster description: This is important for the end user. The resulting hierarchy should facilitate browsing.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Existing approach

TABLE 1
Notations

Notation	Description
n	No. of Documents
m	No. of terms
c	No. of classes
k	No. of clusters
d	document vector.||d||=1
S={d1,...dn}	set of all documents
Sr	set of documents in cluster r
D =∑di ε S di	composite vector of all documents
Dr =∑ di ε Sr di	composite vector of cluster r
C =D/n	centroid vector of all the documents
Cr =Dr/nr	centroid vector of cluster r,nr =|Sr|

Euclidean Distance

Euclidean distance is a regular metric for geometrical problems. It is the common distance between two points and can be without difficulty measured with a ruler in two- or three dimensional space. It is also the default distance measure used with the K-means algorithm. Euclidean distance is one of the most popular measures: 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.

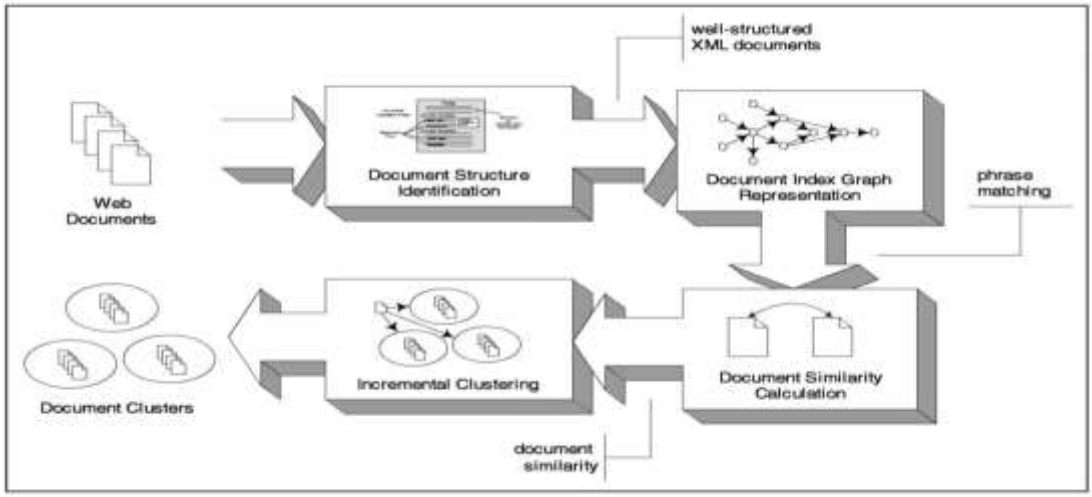
$$\min\sum_{r=1}^{k}\sum_{d_i\epsilon S_r}||d_i - c_r||^2$$

Cosine Similarity

When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity. Cosine similarity is one of the most popular similarity measure practical to text documents, such as in various information retrieval applications and clustering too. An important property of the cosine similarity is its independence of document length.

$$\max\sum_{r=1}^{k}\sum_{d_i\epsilon S_r}\frac{d_i^t C_r}{||C_r||}$$

Arcitecture Diagram For Existing System



Proposed approach

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

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

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

Architecture diagram for proposed Approach

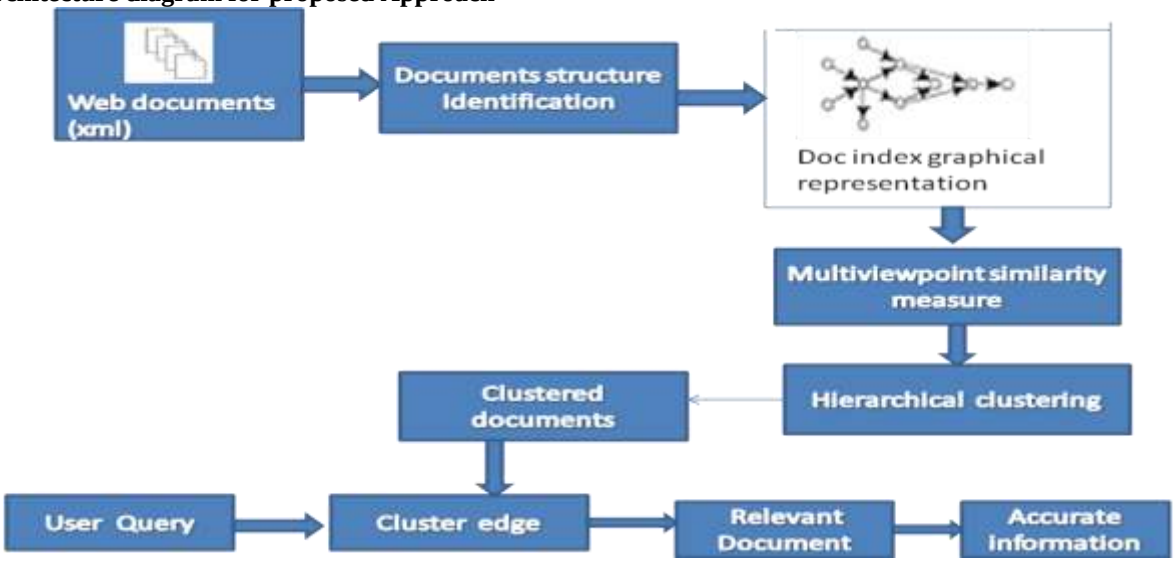


Fig 4.2 System Architecture

Conclusion

In this paper, we propose Hierarchical Multiview point based similarity measuring method. Theoritical analysis and empirical examples show that Hierarchical MVS is suitable for sparse and High dimensional data. Compared with partitional MVS clusters. The key contribution of this paper is the fundamental concept of hirarchical clustering from multiple view points. Future based on the same concept using different alternative measures and use other methods to combine the relative similarities according to the different viewpoints.

References

1. *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.*
2. *S. Zhong, "Efficient Online Spherical K-means Clustering," Proc. IEEE Int'l Joint Conf. Neural Networks (IJCNN), pp. 3180-3185, 2005.*
3. *D. Lee, J. Lee,—Dynamic dissimilarity measure for support based clustering‖, IEEE Trans. on Knowl. and Data Eng., Vol. 22, No. 6, pp. 900–905, 2010.*
4. *R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification. John Wiley & Sons, 2001.*
5. *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.*
6. *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.*
7. *I.S. Dhillon, S. Mallela, and D.S. Modha, "Information-Theoretic Co-Clustering," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge.*
8. *H. Chim and X. Deng, "Efficient Phrase-Based Document Similirity for Clustering,"IEEE Trans. Knowledge and Data Eng., vol.20,no. 9, pp.1217-1229,Sept.2008.*
9. *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*

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Statistical Pattern Recognition, vol. 4109,pp-880,2006.

10. *M. Pelillo, What is a cluster? Perspectives from game theory, in Proc. of the NIPS Workshop on Clustering Theory, 2009.*