Evaluation of Similarities Measure in Document Clustering

Hemalatha Immandhi

P.G Student, B.V.C. Engineering College, Odalarevu

Abstract: ***Clustering is a technique of collecting data into subsets in such a manner that identical instances are collected together, at the same time as different instances belong to different groups. The occurrences are thereby organized into an efficient depiction that characterizes the populace being sectioned. Clustering of entities is as earliest as the human need for describing the salient characteristics of mean and objects and identifying them with a style. Consequently, it squeezes a choice of scientific regulations from mathematics and statistics to biology and genetics, the entire of which uses different terms to describe the topologies formed using this analysis. As of biological "taxonomies" to medical "syndromes" and genetic "genotypes" to manufacturing "group technology"-the problem is same forming groups i.e. cluster text documents that have sparse and high dimensional data objects. Subsequently we originate new clustering criterion functions and corresponding clustering algorithms respectively. Divisive algorithms initiated with just only one cluster that contains all sample data. After that, the single cluster splits into two or more clusters that have higher dissimilarity between them until the number of clusters becomes number of samples or as specified by the user. The most important work is to build up a novel hierarchical algorithm for document clustering which provides maximum efficiency and performance. It is mainly spotlighted in studying and making use of cluster overlapping phenomenon to design cluster merging criteria. Recommending a new method to compute the overlap rate in order to improve time efficiency and "the veracity" is mainly concentrated. Multi-view learning algorithms characteristically assume a complete bipartite mapping between the different views in order to exchange information during the learning process. The remaining of this paper is ordered.***

Keywords: Technology, clustering, Algorithm, data, analysis.

1. Introduction

We are facing an ever increasing volume of text documents. The abundant texts flowing over the Internet, huge collections of documents in digital libraries and repositories, and digitized personal information such as blog articles and emails are piling up quickly every day. These have brought challenges for the effective and efficient organization of text documents.

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.

Text document clustering groups similar documents that to form a coherent cluster, while documents that are different have separated apart into different clusters. However, the definition of a pair of documents being similar or different is not always clear and normally varies with the actual problem setting. For example, when clustering research papers, two documents are regarded as similar if they share similar thematic topics. When clustering is employed on web sites, we are usually more interested in clustering the component pages according to the type of information that is presented in the page. For instance, when dealing with universities' web sites, we may want to separate professors' home pages from students' home pages, and pages for courses from pages for research projects. This kind of clustering can benefit further analysis and utilize of the dataset such as information retrieval and information extraction,

2. Literature Survey

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.

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.

3. Proposed System

The work in this paper is motivated by investigations from the above and similar research findings. It appears to us that the nature of similarity measure plays a very important role in the success or failure of a clustering method. Our first objective is to derive a novel method for measuring similarity between data objects in sparse and high-dimensional domain, particularly text documents. From the proposed similarity measure, we 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

4. 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 itemsets* proposed by Agrawal et. al. The intuition of our clustering criterion is that there are some frequent itemsets for each cluster (topic) in the document set, and different clusters share few frequent itemsets. A frequent itemset is a set of words that occur together in some minimum fraction of documents in a cluster. Therefore, a frequent itemset describes something common to many documents in a cluster. In this technique use frequent itemsets 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.

5. Modules

- **Select File**
  HTML root file is selected from the list of files displayed in the window
- **Process**
  By processing the root file, we can get the child files which are linked to root file.
- **Histogram**
  Histogram displays the no of documents by showing the similarity range between 0 to 1.
- **Clusters**
  Clusters formed by considering similarity of the documents.
- **Similarity**
  Similarity is calculated between the keyword tags between two files.
- **Result**
  Result is displayed as a bar chart which axis has similarity between file to file.

6. 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].

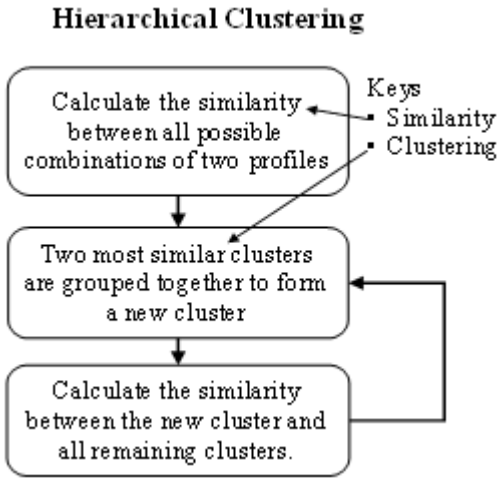


Figure 1: Hierarchical Clustering

Step 1- 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.

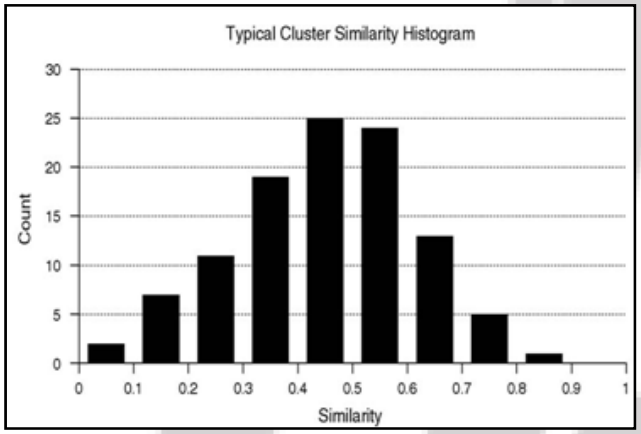
STEP 2- Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less with the help oh tf - itf.
STEP 3- Compute distances (similarities) between the new cluster and each of the old clusters.
STEP 4- Repeat steps 2 and 3 until all items are 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 a*verage-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.

7. TFIDF Analysis

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



8. 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.

9. 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.