DOCUMENT CLUSTERING USING HIERARCHICAL METHODS

1. Dr.R.V.Krishnaiah **2. Katta Sharath Kumar**

3. P.Praveen Kumar

ABSTRACT:

Cluster is a term used regularly in our life is nothing but a group. In the view point of data engineering cluster is a group of objects with similar nature. The grouping mechanism is called as 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

In this paper we mainly focuses on view points and measures in hierarchical clustering. We introduce a novel multi-viewpoint based similarity measure and two related clustering methods , the 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.

Existing work

- 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.

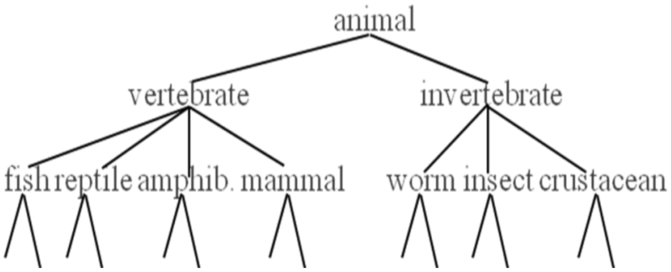
PROPOSED APPROACH

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 this approach can improve the efficiency of clustering and save computing time.

SYSTEM EXPLORATION

Build a tree-based hierarchical taxonomy (Dendogram) from a set of documents.



.

Dendogram

Dendogram: Hierarchical Clustering

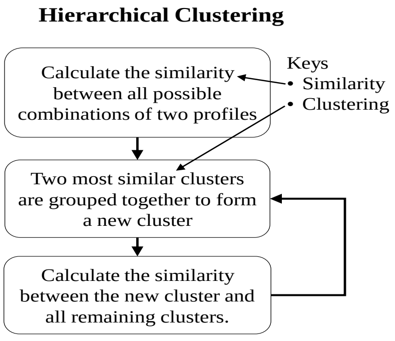
- Clustering obtained by cutting the Dendogram at a desired level: each connected component forms a cluster.

CHALLENGES IN IERARCHICAL 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.

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.


Hierarchical Clustering
Calculate the similarity between all possible combinations of two profiles
Keys
• Similarity
• Clustering
Two most similar clusters are grouped together to form a new cluster
Calculate the similarity between the new cluster and all remaining clusters.

Hierarchical Clustering

HOW THEY WORK?

Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering is this:

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.

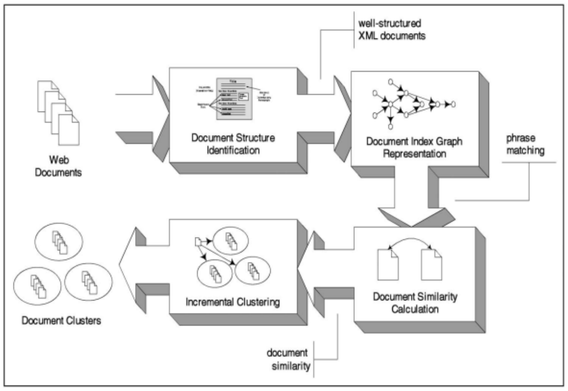
Step 3 can be done in different ways, which is what distinguishes single-linkage from complete-linkage and average-linkage clustering. In single-linkage clustering (also called the connectedness or minimum method), considering the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster.

If the data consist of similarities, consider the similarity between one cluster and another cluster to be equal to the greatest similarity from any member of one cluster to any member of the other cluster. In complete-linkage clustering (also called the diameter or maximum method), consider the distance between one cluster and another cluster to be equal to the greatest distance from any member of one cluster to any member of the other cluster. In average-linkage clustering, consider the distance between one cluster and another cluster to be equal to the average distance. This kind of hierarchical clustering is called agglomerative because it merges clusters iteratively. There is also adivisive hierarchical clustering which does the reverse by starting with all objects in one cluster and subdividing them into smaller pieces. Divisive methods are not generally available, and rarely have been applied.

Of course there is no point in having all the N items grouped in a single cluster but, once the complete hierarchical tree is obtained and need k clusters, k-1 longest links are eliminated.

4.6 ADVANTAGES

- Capable of identifying nested clusters
- They are flexible - cluster shape parameters can be tuned to suit the application at hand.
- They are suitable for automation.
- Can optionally combine the advantages of hierarchical clustering and partitioning around medoids, giving better detection of outliers.
- Reducing effect of initial values of cluster on the clustering results.
- OLR-based clustering algorithm considers more the distribution of data rather than only the distance between data points.
- The method can shorten the computing time and reduce the space complexity, improve the results of clustering.


well-structured XML documents
Web Documents
Document Structure Identification
Document Index Graph Representation
phrase matching
Document Clusters
Incremental Clustering
Document Similarity Calculation
document similarity

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.

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**

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] X. Wu, V. Kumar, J. Ross 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,” *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2007.

[2] I. Guyon, U. von Luxburg, and R. C. Williamson, “Clustering: Science or Art?” *NIPS’09 Workshop on Clustering Theory*, 2009.

[3] I. Dhillon and D. Modha, "Concept decompositions for large sparse text data using clustering," *Mach. Learn.*, vol. 42, no. 1-2, pp. 143–175, Jan 2001.

[4] S. Zhong, "Efficient online spherical K-means clustering," in *IEEE IJCNN*, 2005, pp. 3180–3185.

[5] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, 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," in *Structural, Syntactic, and Statistical Pattern Recognition*, ser. LNCS, vol. 4109, 2006, pp. 871–880.

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

[8] D. Lee and J. Lee, "Dynamic dissimilarity measure for support based clustering," *IEEE Trans. on Knowl. and Data Eng.*, vol. 22, no. 6, pp. 900–905, 2010.

[9] A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *J. Mach. Learn. Res.*, vol. 6, pp. 1345–1382, Sep 2005.

[10] W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization," in *SIGIR*, 2003, pp. 267–273.

[11] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *KDD*, 2003, pp. 89–98.

[12] C. D. Manning, P. Raghavan, and H. Sch ¨ utze, *An Introduction to Information Retrieval*. Press, Cambridge U., 2009.

[13] C. Ding, X. He, H. Zha, M. Gu, and H. Simon, "A min-max cut algorithm for graph partitioning and data clustering," in *IEEE ICDM*, 2001, pp. 107–114.

[14] H. Zha, X. He, C. H. Q. Ding, M. Gu, and H. D. Simon, "Spectral relaxation for k-means clustering," in *NIPS*, 2001, pp. 1057–1064.

[15] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 888–905, 2000.

AUOTHERS PROFILES:

1.Dr.R.V.Krishnaiah
M.Tech(EIE),M.Tech(CSE),
PhD,MIE,MIETE,MISTE
Principal,
DRK INSTITUTE OF SCINCE &
TECHNOLOGY, Hyderabad.
E-mail:- r.v.krishnaiah@gmail.com

2.Katta Sharath Kumar

M.TECH student

Branch: CSE

DRK COLLEGE OF ENGINERING&

TECHNOLOGY, Hyderabad.

E-mail:- sharathkumar.katta@gmail.com

3.

P.Praveen Kumar

M.TECH student

Branch: CS

DRK INSTITUTE OF SCINCE &
TECHNOLOGY, Hyderabad.

mail:praveenkumar.music@gmail.com