Volume 5, Issue 9, September 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Hierarchical Document Clustering: Review with Comparison

Kavita Nagar
Student of Master of Technology,
Department of Computer science and Engineering
Utter Pradesh Technical University,
Gr. Noida, Utter Pradesh, India

Yatin Agarwal
Associate Professor
Department of Computer science and Engineering
Utter Pradesh Technical University,
Gr. Noida, Utter Pradesh, India

Abstract: Hierarchical Document clustering is automatic organization of documents into clusters so that documents within a cluster have high similarity in comparison to documents in other clusters. It is based on the principle of maximizing intra-similarity and minimizing inter-similarity.It has been studied intensively because of its wide applicability in various areas such as web mining, search engines, and information retrieval. It provides efficient representation and visualization of the documents; thus helps in easy navigation also. In this paper, we have given overview of Hierarchical document clustering with its featureselection process, applications, challenges in document clustering, similarity measures and evaluation of document clustering algorithm.In this paper variousHierarchical document clustering techniques are discussed along with their pros and cons.

Keyword —Clustering,Hierarchical Document Clustering,Similarity measures, frequent item set.

I. INTRODUCTION

Data Mining is one of the important steps for mining orextracting a great deal of information. It is designed toexplore giant amount of information in search of consistentpatterns and to validate the results by the detected patterns tothe new subset of information. Clusters are often thought ofas the foremost necessary unsupervised learning problem,which deals with the problems in data assortment ofunlabeled information[1]. Clustering is the most interesting topics in data mining which aims of findingintrinsic structures in data and find some meaningfulsubgroups for further analysis. It is a common technique forstatistical data analysis, which is used in many fields,including machine learning, data mining, patternrecognition, image analysis and bioinformatics. Thus acluster could also be defined as the "methodology oforganizing objects into groups whose members are similar
in some way."

II. WHY CLUSTERING

Data Clustering is one of the challenging miningtechniques in the knowledge data discovery process.Clustering huge amount of data is a difficult task since thegoal is to find a suitable partition in a unsupervised way (i.e.without any prior knowledge) trying to maximize theintra-cluster similarity and minimize inter-cluster similaritywhich in turn maintains high cluster cohesiveness.Clustering groups data instances into subsets in such amanner that similar instances are grouped together, while Different instances belong to different groups. The instances are thereby organized into an efficient representation thatcharacterizes the population being sampled. Thus the outputof cluster analysis is the number of groups or clusters thatform the structure of partitions, of the data set. In shortclustering is the technique to process the data intomeaningful group for statistical analysis. The exploitation ofData Mining and Knowledge discovery has penetrated to avariety of Machine Learning Systems. A very important areain the field of Machine learning is Text Categorization.Feature selection and Term weighting are two importantsteps that decide the result of any Text Categorizationproblem.[2]

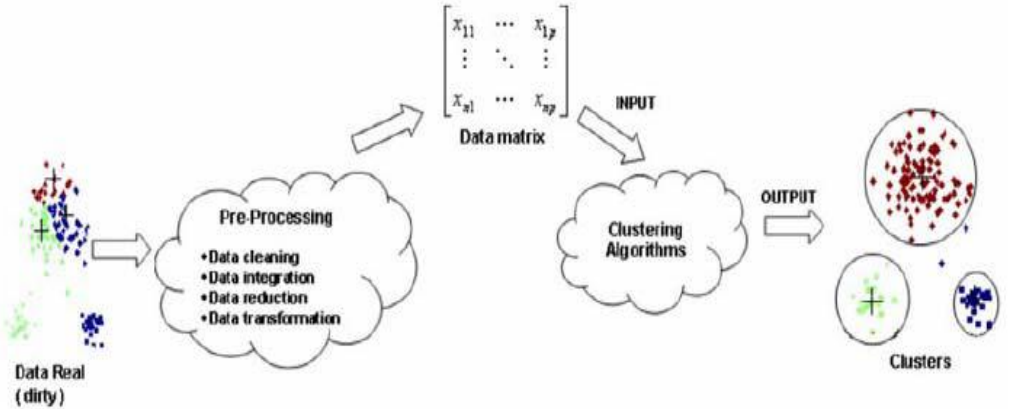

Figure 1.Clustering Process

© 2015, IJARCSSE All Rights Reserved

III. TYPES OF CLUSTERING

Document clustering is define as the grouping of similartext document into clusters such as that the documentswithin the clusters have high similarity in comparison to oneanother but are dissimilar to documents in other clusters. Asthousands of electronic documents have been added on theWorld Wide Web it becomes very important to browse orsearch the relevant data effectively. To identify suitablealgorithms for clustering that produces the best clustering solutions, it becomes necessary to have a method forcomparing the results of different clustering algorithms.Many different clustering techniques have been defined inorder to solve the problem from different perspective.

According to [3], document clustering is divided into two major subcategories, hard clustering and soft clustering. Soft clustering also known as overlapping clustering is again divided into partitioning, hierarchical, and frequent item set based clustering.

- Hard (Disjoint): Hard clustering compute the hard assignment of a document to a cluster i.e., each document is assigned to exactly one cluster; giving a set of disjoint clusters.
- Soft (Overlapping): Soft clustering compute the soft assignment i.e., each document is allowed to appear in multiple clusters; thus, generates a set of overlapping clusters. For instance, a document discussing "Natural language and Information Retrieval" will be assigned to "Natural language" and "Information Retrieval" clusters.
- Partitioning: Partitioning clustering allocates documents into a fixed number of non-empty clusters. The most well-known partitioning methods are the K-means and its variants [3]. The basic K-means method initially allocates a set of objects to a number of clusters randomly. In each iteration, the mean of each cluster is calculated and each object is reassigned to the nearest mean. It stops when there is no change for any of the clusters between successive iterations.
- Hierarchical: Hierarchical document clustering is to build dendrogram, a hierarchical tree of clusters, whose leaf nodes represent the subset of a document collection. Hierarchical Agglomerative Clustering (HAC) and Unweighted Pair Group Method with Arithmetic Mean (UPGMA) fall in this category [3].Hierarchical methods are classified into agglomerative methods and divisive methods. In an agglomerative method, each object forms a cluster. Two most similar clusters are combined iteratively until some termination criterion is satisfied. Thus, it follows bottom up approach. Whereas, in a divisive method top-down approach is there; i.e., from a cluster consisting of all the objects, one cluster is selected and split into smaller clusters recursively until some termination criterion issatisfied [4]. The major decision criteria, at each step, are to find which cluster to split and how to perform the split.The bisecting K-means, a variant of K-means, is a divisive hierarchical clustering algorithm. The algorithm recursively selects the largest cluster and uses the basic K-means algorithm to divide it into two sub-clusters until the desired number of clusters is formed [4]. Out of agglomerative and divisive, agglomerative techniques are more common [5].In [5], UPGMA is proved to be the best among three agglomerative clustering algorithms, IST (Intra-Cluster Similarity Technique), CST (Centroid Similarity Technique), and UPGMA through experiments.Hierarchical clustering gives better quality clustering, but is limited because of its quadratic time complexity. Whereas, partitioning methods like K-means and its variants have a linear time complexity, making it more suitable for clustering large datasets, but are thought to produce inferior clusters [5]. Also, the major problem with K-means is that it is sensitive to the selection of the initial partition and may converge to local optima [6].
- Frequent itemset-based: These methods use frequent itemsets generated by the association rule mining to cluster the documents. Also, these methods reduce the dimensionality of term features efficiently for very large datasets, thus improves the accuracy and scalability of the clustering algorithms. Another advantage of frequent itemset based clustering method is that each cluster can be labeled by the obtained frequent itemsets shared by the documents in the same cluster [3]. These methods include Hierarchical Frequent Term-based Clustering (HFTC) [7], Hierarchical Document Clustering Using Frequent Itemsets (FIHC) [8], and Fuzzy Frequent Item set based Document Clustering (F2IDC) [10].HFTC method minimizes the overlap of clusters in terms of shared documents. But the experiments of Fung et al. showed that HFTC is not scalable. For a large datasets in [8] FIHC algorithm is given where frequent itemsets derived from the association rule mining are used to construct a hierarchical topic tree for clusters. FIHC uses only the global frequent items in document vectors, which drastically reduces the dimensionality of the document set. Thus, FIHC is not only scalable, but also accurate [9]. In F2IDC [10] fuzzy association rule mining is combined with WordNet. A term hierarchy generated from WordNet is applied to discover generalized frequent itemsets as candidate cluster labels for grouping documents. The generated clusters with conceptual labels are easier to understand than clusters annotated by isolated terms for identifying the content of individual clusters.

IV. HIERIRCHICAL DOCUMENT CLUSTERING PROCEDURE

It is important to emphasize that getting from a collection of documents to a clustering of the collection, is not merely a single operation. It involves multiple stages; which generally comprise three main phases: feature extraction and selection, document representation, and clustering.

Feature extraction begins with the parsing of each document to produce a set of features and exclude a list of pre-specified stop words which are irrelevant from semantic perspective. Then representative features are selected from the set of extracted features [15]. Feature selection is an essential pre-processing method to remove noisy features. It reduces the high dimensionality of the feature space and provides better data understanding, which in turn improves the clustering result, efficiency and performance. It is widely used in supervised learning, such as text classification [16]. Thus, it is important for improving clustering efficiency and effectiveness. Commonly employed feature selection metrics are term frequency (TF), inverse document frequency (TF · IDF), and their hybrids. These are discussed further in same section. Also some improvements in traditional methods is discussed.In the document representation phase, each document is

© 2015, IJARCSSE All Rights Reserved

represented by k features with the highest selection metric scores according to top-k selection methods. Document representation methods include binary (presence or absence of a feature in a document), TF (i.e., within-document term frequency), and TF.IDF.

In the final phase of document clustering, the target documents are grouped into distinct clusters on the basis of the selected features and their respective values in each document by applying clustering algorithms [15].

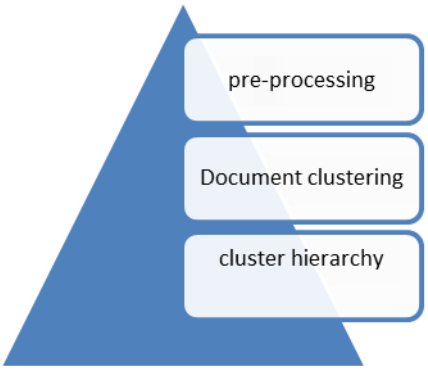

Figure 2:Basic architecture of Hierarchical document clustering

V. CHALLENGES IN HIERARCHICAL DOCUMENT CLUSTERING

Document clustering is being studied from many decades but still it is far from a trivial and solved problem. The challenges are [12]:

- Selection of appropriate features of the documents.
- Selection of appropriate similarity measure
- Selection of appropriate clustering method
- Assessment ofthe quality of the clusters.
- Implementation of the clustering algorithm in an efficient way by making optimal use of available memory and CPU resources.
- Associate meaningful label to each final cluster [11].
- To consider semantic relationship betweenwords like synonyms [11].

In the context of hierarchical document clustering, some other major challenges are given in [13]:

- Very high dimensionality of the data: With medium to large document collections (10,000+ documents), the number of term-document relations is millions+, and the computational complexity of the algorithm applied is thus a central factor. If the vector model is applied, the dimensionality of the resulting vector space will likewise be 10,000+ [12]. The computational complexity should be linear with respect to the number of dimensions (terms).
- The algorithms must be efficient and scalable to large data sets.
- Overlapping between document clusters should be allowed.
- The algorithms must be able to update the hierarchy when new documents arrive (or are removed).
- The clustering algorithm should be able to find number of clusters on its own.

VI. EVALUATION OF HIERARCHICAL DOCUMENT CLUSTERING ALGORITHM

One of the most important issues in clusters analysis is the evaluation of the clustering results. Evaluation is the analysis of the output to understand how well it reproduces the original structure of the data [12].

The ways of evaluation are divided in two parts:

- **Internal quality measure**: Here, the overall similarity measure is used based on the pair wise similarity of documents and no external knowledge is used. The cohesiveness of clusters can be used as a measure of cluster similarity. One method for computing the cluster cohesiveness is the usage of the weighted similarity of the internal cluster similarity [12].
- **External Quality measure**: Some external knowledge for the data is required. One external measure is entropy. It provides a measure of goodness for un-nested clusters or for the clusters at one level of a hierarchical clustering. Another external measure is the F-measure which measures the effectiveness of a hierarchical clustering [5].

Shannon's Entropy: Entropy is used as a measure of quality of the clusters [5]. For each cluster, the category distribution of data is calculated first i.e. let p_{ij} be the probability that a member of cluster j belongs to category i. Then the entropy of each cluster j is calculated as [12]:

$$E_j = - \Sigma\, p_{ij} \log (p_{ij})$$

The total entropy is calculated by adding the entropies of each cluster weighted by the size of each cluster:

$$E_{en} = \sum_{j=1}^{m} ((n_j * E_j) / n)$$

Where m is the total number of clusters, n_j is the size of j^{th} cluster and n is the total number of documents.

F-measure: This is an aggregation of precision and recall concept of information retrieval. Precision is the ratio of the number of relevant documents to the total number of documents retrieved for a query. Recall is the ratio of the number of

© 2015, IJARCSSE All Rights Reserved

relevant documents retrieved for a query to the total number of relevant documents in the entire collection [12]. For cluster j and class i

Recall (i, j) = n_{ij} / n_i

Precision (i, j) = n_{ij} / n_j

Where n_{ij} is the number of members of class i in cluster j, n_j is the number of members of cluster j and n_i is the number of members of class i.

The F-measure of cluster j and class i is calculated from precision and recall as

F(i, j) = (2 * Recall(i, j) * Precision(i, j)) / ((Precision(i, j) + Recall(i, j))

For an entire hierarchical clustering the F-measure of any class is the maximum value it attains at any node in the tree and an overall value for the F-measure is computed by taking the weighted average of all values for the F-measure as given by the following.

$F = \Sigma\ n_i / n \max \{F (i, j)\}$

Where the max is taken over all clusters at all levels, and n is the number of documents [5]. Higher value of F-measure indicates better clustering [12].

In [13], authors have shown that F-measure has bias towards hierarchical clustering algorithms so F_{norm} which is the normalized version of the F-measure is proposed to solve the cluster validation problem for hierarchical clustering. Experimental results show that F_{norm} is more suitable than the unnormalized F-measure in evaluating the hierarchical clustering results acrossdatasets with different datadistribution.

VII. STUDY TABLE FOR COMPARISON OF HIERARCHICAL CLUSTERING TECHNIQUES

Nameof Algorithm	**Algorithm Key- Idea**	**Data type**	**Advantage**	**Disadvantage**
Single link	Closest pair of points	_	it does not need to specify no.of clusters	Terminati-on condition needs to be satisfied. Sensitive to outliers
Average link	Centroids of Clusters	_	It considers all members in cluster rather than single point	It produces clusters with same variance
Complete link	Farthest pair of Points	_	Not strongly affected by Outliers	It has problem withconvex shape clusters
ROCK	Notion of links	Numerical	Robust Appropriate for large dataset	space complexity depends on,, initialization of local heaps
BRICH	Multi –dimensional	Numerical	-suitable for large databases -scales linearly	-Handles only numeric data -sensitive to data records
CURE	Partition samples		-Robust to outliers -Appropriate for handling large Dataset	Ignores information about inter-connectivity of objects
APRIORI	Frequent Item-sets, Apriori property, Join Operation	Boolean, categrical, Quantative	-easy to implement -easily parallised -use large item set property	-slow -Bottleneck in candidate generation -requires many database scan
FP-growth	Discovery of frequent item set without candidate itemset generation	Boolean, categrical, Quantative	-Scan only twice -small execution time	-May not fit into memory -Expensive

VIII. CONCLUSION

Hierarchical Document clustering is a fundamental operation used in unsupervised document organization, automatic topic extraction, and information retrieval. In this paper we presented the Hierarchical Document clustering algorithms. In this paper, we have explained the document clustering procedure with feature selection, various improvements for it.We concluded the discussion on Hierarchical clustering algorithms by a comparative study with pros and cons of each category. We have also discussed the concept of Similarity measures which proves to be the most important criteria for

© 2015, IJARCSSE All Rights Reserved

document clustering.We have tried to provide detailed and exhaustive overview on Hierarchicaldocument clustering method.We feel this survey paper will be very useful in research area of Hierarchical document clustering.

ACKNOWLEDGEMENT

I would like to express great pleasure and gratitude to Prof. Rajesh Pathak and Associate Prof. Yatin Agrwal for their invaluable guidance and constant encouragement for my work. I would like express my gratitude to all my friends in Department of Computer Science & Engineering of GNIOT GR. Noida, UTTERPRADESH

REFERENCES

[1] Prof. NehaSoni, Dr.AmitGanatra. "Comparative Study of Several Clustering Algorithms", International Journal of Advanced ComputerResearch, Volume-2, Number-4, Issue-6 December 2012.

[2] (2013) The Wikipedia website. [Online]. Available: http.

[3] Chun-Ling Chen, Frank S.C. Tseng, and Tyne Liang, "An integration of WordNet and fuzzyassociation rule mining for multi-label document clustering," Data and Knowledge Engineering, vol. 69, issue 11, pp. 1208-1226, Nov. 2010

[4] Yong Wang and Julia Hodges, "Document Clustering with Semantic Analysis," In Proc. of the 39th Annual Hawaii International Conference on System Sciences, HICSS 2006, vol. 03, pp. 54.3

[5] Michael Steinbach , George Karypis, and Vipin Kumar, "A comparison of document clustering techniques," In KDD Workshop on Text Mining, 2002

[6] Xiaohui Cui and Thomas E. Potok, "Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm," Special Issue, 2005

[7] F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering," Proc. of Int'l Conf. on knowledge Discovery and Data Mining (KDD'02), pp. 436–442, 2002.

[8] Benjamin C.M. Fung, Ke Wang, and Martin Ester, "Hierarchical Document Clustering Using Frequent Itemsets," In Proc. Siam International Conference On Data Mining 2003, SDM 2003

[9] Chun-Ling Chen, Frank S. C. Tseng, and Tyne Liang, "Mining fuzzy frequent item sets for hierarchical document clustering," Published in an Int'l Journal of Information Processing and Management, vol. 46, issue 2, pp. 193-211, Mar. 2010

[10] C.L. Chen, F.S.C. Tseng, T. Liang, An integration of fuzzy association rules and Word Net for document clustering, Proc. of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-09), 2009, pp. 147–159.

[11] Rekha Baghel and Dr. Renu Dhir, "A Frequent Concepts Based Document Clustering Algorithm," International Journal of Computer Applications, vol. 4, No.5, pp. 0975 – 8887, Jul. 2010 [2] A. Huang, "Similarity measures for text document

[12] Pankaj Jajoo, "Document Clustering," Masters' Thesis, IIT Kharagpur, 2008

[13] Reynaldo Gil-García and Aurora Pons-Porrata, "Dynamic hierarchical algorithms for document clustering," Pattern Recognition Letters 31, pp. 469–477, 2010

[14] Junjie Wu, HuiXiong, and Jian Chen, "Towards understanding hierarchical clustering: A data distribution perspective," Neurocomputing 72, pp. 2319–2330, 2009

[15] Chih-Ping Wei, Chin-Sheng Yang, Han-Wei Hsiao, and Tsang-Hsiang Cheng, "Combining preference- and content-based approaches for improving document clustering effectiveness," Published in Int'l Journal of Information Processing and Management, vol. 42, issue 2, pp. 350-372, Mar. 2006

[16] MS. K.Mugunthadevi, MRS. S.C. Punitha, and Dr. M. Punithavalli, "Survey on Feature Selection in Document Clustering," Int'l Journal on Computer Science and Engineering (IJCSE), vol. 3, No. 3, pp. 1240-1244, Mar

© 2015, IJARCSSE All Rights Reserved