 ISBN:978-93-5254-640-4

Review: Hierarchical Document clustering

Harsha Patil

Research Scholar, Maulana Azad Institute of Technology, Bhopal, Madhya Pradesh, India.

Dr. Ramjeevan Singh Takur

Maulana Azad Institute of Technology, Bhopal, Madhya Pradesh, India

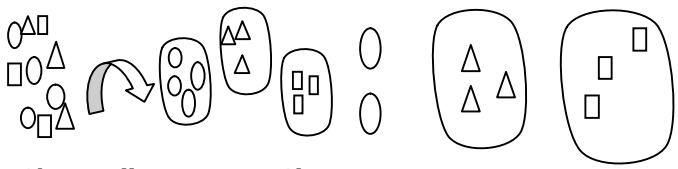
Abstract—***As we know use of Internet flourishes with its full velocity and in all dimensions. Enormous availability of Text documents in digital form like (email, webpages, blog post, news articals, ebooks and other text files) on internet challenges technology to appropriate retrieval of document as a response for any search query. As a result there has been an eruption of interest in people to mine these vast resources and classify them properly. It invigorates researchers and developers to work on numerous approaches of document clustering. Clustering is the process of subseting data objects. Subsets are based on characteristics of the objects.Objects with similar characteristics are come together and make a set.So, objects from one set have high similarity while objects from other set. Hierarchical document clustering organizes clusters into a tree or a hierarchy that facilitates browsing. This paper review several special challenges in hierarchical document clustering: high dimensionality, high volume of data, ease of browsing, and meaningful cluster labels. Hierarchical document clustering techniques has quality contribution in improvement of document clustering results.Nowadays many researchers are working on Hierarchical document clustering techniques and methods. This paper gives light on contribution of authors in this era.State-of-the-art document clustering algorithms are reviewed: agglomerative,Hierarchical Frequent Term-based Clustering (HFTC), Frequent Item set-based Hierarchical Clustering(FIHC).***

Keywords— **Hierarchical Document clustering, Frequent Item sets, agglomerative c clustering.**

I. INTRODUCTION

Clustering is the process of subseting data objects. Subsets are based on characteristics of the objects.Objects with similar characteristics are come together and make a set.So, objects from one set have high similarity while objects from other set.

Example: Suppose we have collection of 10 objects.Here we consider shape characteristic of objects .So objects of same shape are come together in one set.This set is known as cluster and process of dividing similar objects in setwise is known as clustering.After partitioning we get three clusters of objects.



Object collection Clusters

Figure.1 Clustering

In clustering problem, defining concepts of optimization criteria is very important.

In above example partitioning objects on the basis of their shapes is very easy task,but this is not the always case.In reality , finding the properties of object by which we can make cluster is very difficult,basically its very important that objects which belongs to different cluster somehow must be more dissimilar to each other.Suppose we have collection of n objects .lets o1,o2,03.....on. Lets s= {o1,o2,o3.....on} and suppose number of clusters are k.so we need to partition n objects into k clusters.That means we need to make k partitions.

II. BACKGROUND

Text data is ubiquitous in nature.In age of Internet generation of digital documents are very fast. As the volume of text data increases, management and analysis of text data becomes unprecedentedly important. Text mining is an emerging technology for handling the increasing text data.document clustering is one of the widely used application of text mining.Text clustering is to divide a collection of text documents into different category groups so that documents in the same category group describe the same topic, such as movie discription or sports details.Document clustering technique needs to handle large and high dimensional data and should be able to handle sparsity and semantics.Due to the sparsity nature of doccuments its become imppossible to finalized general technique of document clustering which is suitable for all kinds of text data.The main objective of document clustering technique is to minimize intra cluster distance between documents and increase inter cluster distance.

Document Clustering can be categorised in two broad category on the basis of their assignment pattern:

1. Hard document clustering
2. Soft document clustering (Fuzzy clustering)

1. Hard document clustering:Each document is exactly belong to only one cluster.
2. Soft document clustering:Each document has probabilistic membership to each cluster.

A. Challenges in Document Clustering

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

1. Selecting appropriate features of the documents that should be used for clustering.
2. Selecting an appropriate similarity measure between documents.

3. Selecting an appropriate clustering method utilising the above similarity measure.
4. Implementing the clustering algorithm in an efficient way that makes it feasible interms of required memory and CPU resources.
5. Finding ways of assessing the quality of the performed clustering.

Document Clustering Techniques

A. Distance based

Distance-based clustering algorithms are based on finding similarity between the text objects. As defined earlier cosine similarity function is used commonly in the text domain. Computation of text similarity is a fundamental problem in information retrieval. Although most of the work in information retrieval has focused on how to assess the similarity of a keyword query and a text document, rather than the similarity between two documents, many weighting heuristics and similarity functions can also be applied to optimize the similarity function for clustering. Effective information retrieval models generally capture three heuristics, i.e., TF weighting, IDF weighting, and document length normalization (Brown et al.,1992)

B. Partitioning based

Partitioning algorithms are widely used techniques for document clustering. The two most widely used distance-based partitioning algorithms (Chakrabarti et al.,2000) are as follows:
k-medoid clustering algorithms: In k-medoid clustering algorithms, we use a set of points from the original data as the anchors (or medoids) around which the clusters are built. The key aim of the algorithm is to determine an optimal set of representative documents from the original corpus around which the clusters are built. Each document is assigned to its closest representative from the collection. This creates a running set of clusters from the corpus which are successively improved by a randomized process.

The algorithm works with an iterative approach in which the set of k representatives are successively improved with the use of randomized inter-changes. Specifically, we use the average similarity of each document in the corpus to its closest representative as the objective function which needs to be improved during this interchange process. In each iteration, we replace a randomly picked representative in the current set of medoids with a randomly picked representative from the collection, if it improves the clustering objective function. This approach is applied until convergence is achieved.

C. Hierarchical Based

In hierarchical clustering algorithm the observation vectors are group together on the basis of their mutual distance. Agglomerative hierarchicalClustering is popular approach for document clustering. agglomerative hierarchical document clustering approach operates by successive merger of cases. Suppose we have n number of documents, we put all documnets in differnt clusters.So if we have n number of documnets, we need to start with n number of clusters.At each stage we merge two most similar groups to form a new Cluster thus reducing the number of clusters by one.This keeps going unitill all subgroups of documents are fused togetheer to form one single cluster.

The hierarchy can also be built top-down which is known as the divisive approach. It starts with all the data objects in the same cluster anditeratively splits a cluster into smaller clusters until a certain termination condition is fulfilled.Methods in this category usually suffer from their inability to perform adjustment once a merge or split has been performed.

D. Word and Phrased based

Since text documents are drawn from an inherently high-dimensional domain, it can be useful to view the problem in a dual way, in which important clusters of words may be found and utilized for finding clusters of documents. In a corpus containing d terms and n documents, one may view a term-document matrix as an n × d matrix, in which the (i, j)th entry is the frequency of the jth term in the ith document.

We note that this matrix is extremely sparse since a given document contains an extremely small fraction of the universe of words. We note that the problem of clustering rows in this matrix is that of clustering documents, whereas that of clustering columns in this matrix is that of clustering words. In reality, the two problems are closely related, as good clusters of words may be leveraged in order to find good clusters of documents and vice-versa. For example, the work in determines frequent itemsets of words in the document collection, and uses them to determine compact clusters of documents. This is somewhat analogous to the use of clusters of words for determining clusters of documents. The most general technique for simultaneous word and document clustering is referred to as co-clustering . This approach simultaneous clusters the rows and columns of the term-document matrix, in order to create such clusters. This can also be considered to be equivalent to the problem of re-ordering the rows and columns of the term-document matrix so as to create dense rectangular blocks of non-zero entries in this matrix. In some cases, the ordering information among words may be used in order to determine good clusters. The work in determines the frequent phrases in the collection and leverages them in order to determine document clusters. It is important to understand that the problem of word clusters and document clusters are essentially dual problems which are closely related to one another. The former is related to dimensionality reduction, whereas the latter is related to traditional clustering. The boundary between the two problems is quite fluid, because good word clusters provide hints for finding good document clusters and vice-versa. For example, a more general probabilistic framework which determines word clusters and document clusters simultaneously is referred to as topic modeling (Croft,1997). Topic modeling is a more general framework than either clustering or dimensionality reduction.

E. Frequent Word Based

Frequent pattern mining is a technique which has been widely usedin the data mining literature in order to determine the most relevant patterns in transactional data. The clustering approach in is designed on the basis of such frequent pattern mining algorithms. A frequent itemset in the context of text data is also referred to as a frequent termset, because we are dealing with

documents rather than transactions. The main idea of the approach is to not cluster the high dimensional document data set, but consider the low dimensional frequent term sets as cluster candidates. This essentially means that a frequent terms set is a description of a cluster which corresponds to all the documents containing that frequent term set. Since a frequent term set can be considered a description of a cluster, a set of carefully chosen frequent terms sets can be considered a clustering. The appropriate choice of this set of frequent term sets is defined on the basis of the overlaps between the supporting documents of the different frequent term sets.

III. HIERARCHICAL CONCEPTS

The Hierarchical Frequent Term-based Clustering (HFTC)method proposed by (Beil, Ester, & Xu, 2002) introduced document clustering using frequent itemsets.HFTC method select itemset one by one which are frequent, on the basis of which new clusters are created.So the clustering result depends on the order of selected itemsets.but HFTC is not scalable (Fung, Wang, & Ester, 2003).

The Frequent Itemset Hierarchical Cluster (FIHC) proposed by (Fung, Wang, & Ester, 2003),is very scalable method. First step of the method is to find all global frequentitemsetsfrom document set. Then assign each global frequent itemset in separate cluster.FIHC assigns documents to the best cluster from among all available clusters. Initial clusters are overlapping because one document may contain multiple global frequent itemsets. For removing this overlapping the next step of the algorithm finds the best cluster for the document.In which for each document, the "best" initial cluster is identified and the document is assigned only to the best matching initial cluster. Any cluster Ci is best cluster for docj if it achieve the minimum score for cluster membership.Themininimumscore can be measured with the help of score function based on cluster frequent item.

IV. LITERATURE REVIEW

Document clustering has been an interesting topic of study since a long time. There are many method which are discussed earlier are used by researcher for clustering as per their requirements. Many algorithms are developed my researchers in this era. For agglomerative clustering [3] UPAGMA (Unweighted Pair Group Method with ArithmeticMean) [2] work very well. BisectingK-means has best results as compare to K-means for partition clustering [3].Many researcher work with frequent itemsets for document clustering. Hierarchical Frequent Term based Clustering (HFTC) [4] has been great contribution in this regards.In which author used Apriori calculate frequent .But HFTC result depends on the order of selected itemsets,which in turn depends on the greedy heuristic used also HFTC was not scalable. Fung, et al came up with HierarchicalDocument Clustering using Frequentitemsets (FIHC) [5] which provides light on limitation of HFTC. According to researchers FIHC is "cluster –centered". It means FIHC measures cohesiveness of cluster directly using frequent itemsets.Someof the drawbacks of FIHC include (i)using all the frequent itemsets to get theclustering (number of frequent itemsets may be very large and redundant) (ii)Not comparable with previous methods like UPGMA and Bisecting K-meansin terms of clustering quality. [6] (iii) Use hard clustering (each document canbelong to at most one cluster), etc. Then Yu, et al came up with a much moreefficient algorithm using closed frequent itemsets for clustering(TDC) [7].

Theyalso provide a method for estimating the support correctly. But they use closeditemsets which also may be redundant.Recently Hasan H Malik, et al. proposed Hierarchical Clustering using ClosedInteresting Itemsets, (which we refer to as HCCI) [6] which is the current stateof the art in clustering using frequent itemsets.

Research is also being done about improving the clustering quality by usingan ontology to enhance the document representation. Some of the most commonlyavailable ontologies include WordNet, MESH, Wikipedia etc. Several works [8–10]have been done to include these ontologies to enhance document representationby replacing the words with their synonyms or the concepts related to them.But all these methods have a very limited coverage. It can also happen thataddition of new words could bring in noise into the document or while replacingthe original content, there might be some information loss. Existing knowledgerepositories like Wikipedia and ODP(open directory project) can be used asbackground knowledge. Gabrilovich and Markovitch [11, 12] propose a method to improve text classification performance by enriching document representationwith Wikipedia concepts. Wang P, Hu J, Zeng H-J, Chen Z [14] has great contribution in semantic analysis of text. They proposed method in which they automatically construct a thesaurus of concepts from Wikipedia. They also introduce framework to expand the BOW representation with semantic relation (synonymy, hyponymy and associative relations) Xiaoke Su, Yang Lan, Renxia Wan and Yuming Qin[15] suggested a fast incremental hierarchical clustering algorithm which was feasible and effective. Theoretical analysis and experiments results shows that it not only overcome the inadequate impact of memory while clustering large data set but also reflect the accurate features of the data set.

V. CONCLUSION

In this paper we investigated many existing algorithms.. We conclude that it is hardly possible to get a general algorithm, which can work the best in clustering all types of datasets. Thus we will try to implement algorithms which can work well in different types of document sets. Finally we would conclude that though many algorithms have been proposed for document clustering but it is still an open problem and looking at the rate at which the web is growing, for any application using web documents, clustering will become an essential part of the application.

REFERENCES

1. Pudi, V., Haritsa, J.R.: Generalized Closed Itemsets for Association Rule Mining,In Proc. of IEEE Conf. on Data Engineering. (2003)
2. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data, An introduction to ClusterAnalysis, John Wiley & Sons, Inc (1990)
3. Zhao, Y., Karypis, G.: Evaluation of Hierarchial Clustering Algorithms for DocumentDatasets, In Proc. of Intl. Conf. on Information and Knowledge Management.(2002)
4. Beil, F., Ester, M., Xu, X.: Frequent Term-based Text Clustering, In Proc. of Intl.Conf. on Knowledge Discovery and Data Mining,. (2002)
5. Fung, B., Wang, K., Ester, M.: Hierarchical Document Clustering using FrequentItemsets, In Proc. of SIAM Intl. Conf. on Data Mining. (2003)
6. Malik, H.H., Kender, J.R.: High Quality, Efficient Hierarchical Document ClusteringUsing Closed Interesting Itemsets, In Proc. of IEEE Intl. Conf. on DataMining. (2006)
7. Yu, H., Searsmith, D., Li, X., Han, J.: Scalable Construction of Topic Directorywith Nonparametric Closed Termset Mining, In Proc. of Fourth IEEE Intl. Conf.on Data Mining. (2004)
8. Sedding J, Kazakov D (2004) WordNet-based text document clustering. In: COLING-2004 workshop on robust methods in analysis of natural language data
9. Hotho, A., Maedche, A., Staab, S.: Text Clustering Based on Good Aggregations,In Proc. of IEEE Intl. Conf. on Data Mining. (2001)
10. Zhang, X., Jing, L., Hu, X., et al: A Comparative Study of Ontology BasedTerm Similarity Measures on Document Clustering,In Proc. of 12th Intl. Conf.on Database Systems for Advanced Applications. (2007)
11. Gabrilovich, E., Markovitch, S.: Overcoming the Brittleness Bottleneck UsingWikipedia: Enhancing Text Categorization with Encyclopedic Knowledge, In Proc.of The 21st National Conf. on Artificial Intelligence. (2006)
12. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness UsingWikipedia-based Explicit Semantic Analysis, In Proc. of The 20th Intl. Joint Conf.on Artificial Intelligence. (2007)
13. Hu, X., Zhang, X., Lu, C., et al.: Exploiting Wikipedia as External Knowledge forDocument Clustering, In Proc. of Knowledge Discovery and Data Mining. (2009)
14. Wang P, Hu J, Zeng H-J, Chen Z (2009) Wikipedia knowledge to improve text classification. KnowlInfSyst 19(3): 265–281
15. XiaokeSu,Yang Lan, Renxia Wan and Yuming Qin," A fast Incremental Clustering Algorithm," proceedings of the 2009 Int. Sumposium on Information processing(ISIP'09), Huangshan, P. R. China, August 21-23, 2009,pp. 175-178.