HIERARCHICAL AND PARTITIONING ALGORITHM FOR DOCUMENT CLUSTRING: A SURVEY

[1]JITENDRA AGRAWAL, [2]SHIKHA AGRAWAL, [3]SAUMYA AGRAWAL, [4]SANJEEV SHARMA

[1,2,3,4] Rajiv Gandhi Proudyogiki Vishwavidyalaya Bhopal (MP) India
E-mail: [1]jitendra@rgtu.net, [2]shikha@rgtu.net, [3]saumyaagrawal22@gmail.com, [4]sanjeev@rgtu.net

Abstract - Document clustering is the widely researched area because of large amount of rich and dynamic information are available in world wide web. It is the application of cluster analysis to texual documents. There are different applications of document clustering include automatic document organization, data mining , topic extraction and filtering or fast information retrieval. The purpose of this survey is to provide a review of different partitioning and hierarchical techniques used in documentclustering.

Keywords- Document clustering, Hierarchical clustering, Partitioning clustering, k-means.

I. INTRODUCTION

Clustering is the process of organizing data objects into a set of disjoint classes called clusters. Objects within a cluster are similar and they are dissimilar to the objects belonging to other clusters. It is a useful technique for the discovery of data distribution and patterns in the underlying data. The goal of clustering is to discover both the dense and sparse regions in a dataset. . Clustering has ability to handle different types of attributes such as numeric or categorical etc. There are two main approaches to clustering: Partitioning clustering and Hierarchical clustering.
Partitioning algorithm construct partition of a database of N objects into a set of k clusters. The construction involves finding the optimal partition according to an objective function. There are around kN/k ways of partitioning a set of N data points into ksubsets. The partitioning clustering algorithm adopts the iterative optimization paradigm. It starts with an initial partition and uses an iterative control strategy. It tries swapping data points to see if such swapping improves the quality of clustering. When swapping does not return any improvement in clustering, it
finds a locally optimal partition.
There are two categories of partitioning algorithms are:

- K-Means algorithm
- K-Medoid algorithm

K-Means algorithm is developed by MacQueen, it is simplest and best known unsupervised learning algorithm that solves the wellknown clustering problem. It is efficient algorithm in clustering large datasets. This is a top down
clustering algorithm which assigns each document to the cluster whose centroid is nearest. The aim of KMeans
algorithm is to partition a set of objects into kclusters, where k is a user defined constant. For each cluster need to be define k centroids. The centroid of a cluster is formed in such a way that it is nearest to all objects in that cluster.

K-Means algorithm for finding k clusters s
1. Randomly choose 'k' objects as the initial medoids;
2. Repeat,
a. Assign remaining objects to the clusters with the nearest medoid;
b. Select a non-medoid object;
c. Compute the total cost of swapping between old medoid object and newly selected nonmedoid object.
d. If the distance is less than zero, then perform that swap operation to form the new
e. Set of k- medoids.
3. Until no change

K-Medoid algorithm is partitioning algorithm where each cluster is represented by one of the objects of the cluster located near the centre. Here, k data objects are selected randomly as medoids to represent k cluster and remaining all data objects are placed in a cluster according to their nearest distance from any of the medoid. After alloting all data objects, new medoid is found to represent the cluster in better way. In each iteration, medoids change their position step by step. This process is repeated until no change in medoid.

K-Medoid algorithm for finding k clusters
1. Randomly choose 'k' objects as the initial medoids;
2. Repeat,
a. Assign remaining objects tothe clusters with the nearest medoid;
b. Select a non-medoid object;
c. Compute the total cost of swapping between old medoid object and selected nonmedoid object.
d. If the distance is less than zero, then perform that swap operation to form the new set of kmedoids.
3. Until no change.
Hierarchical techniques produce a sequence of partition where each partition is nested into the next sequence of partition. Hierarchical algorithms create a hierarchical decomposition of the database. The algorithms split the database into smaller subsets,

until some termination condition is satisfied. It does not need k as an input parameter, which is an advantage over partitioning algorithm and the disadvantage of hierarchical algorithm is that the termination condition is to be specified. The hierarchical decomposition can be represented as a dendogram.

The two basic approaches to generating a hierarchical clustering are:

- Agglomerative (Bottom-up approach)
- Divisive (Top-down approach)

In Agglomerative clustering algorithm each object is placed in a unique cluster and for every pair of clusters, value of dissimilarity or distance is computed. The distance may be minimum distance of all pairs of points from the two clusters; the clusters with the minimum distances are merged at every step. One can set the termination criteria by fixing the critical distance Dmin between the clusters. In Divisive clustering algorithm all objects are places in a single cluster. At each step split a cluster until only singleton clusters of individual points remain. At each step we need to decide which cluster to split and how to perform the split. Document clustering is grouping the documents in to various clusters where the similar types of documents are placed in same cluster and documents in different clusters are dissimilar. With the advancement of technologies, large amounts of rich and dynamic information's are available in World Wide Web. A user can quickly browse and locate the documents with web search engines. Search engines returns many documents, many of which are relevant to the topic and some may contain irrelevant documents. Clustering plays an important role in organizing such monolithic amount of documents returned by search engines into meaningful clusters. Clustering algorithm can be classified either as flat and hierarchical or hard and soft. Flat clustering makes a flat set of clusters without any explicit structure that would relate clusters to each other. It partitions the document space into different cluster. Hierarchical clustering creates a hierarchy of cluster. In hierarchical clustering document of lower level cluster is also a member of corresponding higher level cluster. In hard clustering each document is a member of exactly one cluster, it computes hard assignment whereas the assignment in soft clustering algorithm is soft i.e. the document has fractional membership in several clusters.

II. LITERATURE REVIEW

Xiaohui et al. [2] used Particle Swarm Optimization (PSO) for document clustering. KMeans algorithm is most commonly used partitioning algorithm for clustering large datasets but it produces local optimal solution. In contrast to localized searching property of K-Means, PSO performs globalized search using entire solution space. Authors used PSO, K-Means and hybrid PSO clustering algorithm on four document datasets which are derived from Text Retrieval Conference (TREC) and contains 414, 313, 204, 878 documents respectively. In hybrid PSO two modules are used the PSO module and the K-Means module. In each experiment PSO and K-Means run 100 iterations while in hybrid PSO approach, PSO algorithm is executed for 90 iterations and then K-Means is executed for 10 iterations.

For similarity metrics, Euclidian distance and cosine correlation measure are used. Cluster quality is measured by average distance between document and cluster (ADDC) and smaller ADDC value indicates good clustering solution. Performance comparison shows that hybrid PSO algorithm performs better clustering than using either K-Means or PSO alone.

Zhao and Karypis [7] mainly focused on hierarchical document clustering algorithms that make hierarchy of clusters. Hierarchical clustering solutions are used for a number of application domains (phylogenetic trees, biological taxonomies, etc.). This paper demonstrates two approach for hierarchical document clustering firstly they compare the partitional and agglomerative algorithms by using nine agglomerative and six partitional methods in twelve datasets. Experimental results show that partitional clustering methods perform better than agglomerative methods for hierarchical clustering solution. Secondly author proposed new clustering algorithms called constrained agglomerative algorithms in which features of both agglomerative and partitional algorithms are combined. In this algorithm for each partitional cluster agglomerative algorithm is used to make a hierarchical subtree and then merge these clusters to make final hierarchical tree. Experimental results show that these methods give better solution than agglomerative and partitional method alone. CuiandPotok [9] used hybrid Particle Swarm Optimization (PSO) +K-Means for document clustering. PSO is an optimization algorithm and provide glolized search but require more number of iterations and computational time while K-Means is faster than PSO but it is sensitive to initial solution and can be trapped into local optima. So the author combined both, PSO is for initial stage to find the initial seed and then K-Means is used for refining stage. Experimental results on datasets illustrate that hybrid PSO performs better than PSO and K-Means alone. Author also demonstrates various hybridization of PSO with K-Means which are: PSO followed by K-Means, K-Means followed by PSO & K-Means followed by PSO which is further followed by KMeans. From the experimental result reported it is concluded that PSO followed by K-Means performs better than all the other cases. In 2006 Sahoo et al. [8] proposed an algorithm for Incremental hierarchical text document clustering. In many applications where documents need to be processed as soon as they arrived, incremental or online clustering algorithm is

used. Incremental hierarchical clustering algorithms Cobweb and Classitwith normal distribution do not work for text document data. Here variant of Cobweb algorithm which used Katz's distribution is used for text documents. Original Classit algorithm and proposed algorithms are tested on Newswire articles from Reuters-RCV1 dataset and Ohsumeddataset. Results show that Katz distribution improves the performance of existing algorithm.

Modified K-Means algorithm with Jaccard distance measure was proposed by Shameenand Ferdous [3], which computes most dissimilar k documents as initial centroid for k clusters. Clustering algorithms uses iterative approach for clustering which traps into local minima. Iterative techniques are very sensitive to initial partition and the more distance between cluster centroids gives better performance. In simple K-Means algorithm initiallyk centroids are randomly choosen hence it does not provide most dissimilar documents as centroid. Here Jaccard distance measure is used for finding the most dissimilar k documents as centroid. Jaccard coefficient finds similarity between datasets by dividing the size of the intersection to the size of union of the sample sets. Experiments are performed on Reuter's collection dataset andresults illustrate that the use of jaccard distance for computing the initial centroid improves cluster quality of the simple K-Means.

Singh et al. [6] used flat clustering algorithms like K-Means, Heuristic K-Means and Fuzzy C-means in clustering for text documents. K-Means is a hard flat clustering and C-means is a soft flat clustering algorithm. In their experiment authors used different representation such as term frequency (tf), term frequency Inverse document frequency (tf.idf) and Boolean. Different selection schemes (with or without stop word removal &with or without stemming) are also used. Stop words are common words like 'the', 'am', 'is', 'are', 'who' etc. which do not provide any information about the representation of the topic. Different form of terms like 'computer', 'computes', 'computational', 'computing' are represented by its root word 'computer', this process is called stemming. Different experiments are performed using K-Means, heuristic K-Means and Fuzzy C-means and results illustrate that ztf.idf performs better than both tf and Boolean representation while tf performs better than only Boolean. Performance of Fuzzy C-means is better than K-Means and Heuristic K-Means both. The results of Stemming alone produce better clustering than stop word removal and stemming &stop word removal together.

In 2012 Forsati et al. [1] presented Harmony Search (HS) optimization method for document clustering. Authors first proposed pure HS based clustering for finding near optimal solution which is called HSCLUST. Then HS is integrated with K-Means which combines explorative power of K-Means with refining power of HS. In contrast localized searching property of existing K-Means HS performs globalized search and it is less dependent on the initial partition. Authors combine Harmony Search with K-Means in different ways. The Sequential hybridization, in which optimum region is found by HSCLUST and then optimum centroid is found using K-Means. In Interleaved Hybridization, after every iteration of harmony search K-Means is used. And Hybridization K-Means as one step of HSCLUST is used in which HSCLUST and K-Means are combined for every iteration. In this paper HS is applied with K-Means and Genetic Algorithm (GA) based clustering algorithm on five different document sets such as Politics dataset, TREC, DMOZ collection, 20 NEWSGROUP, WebACE project(WAP). Quality of clusters is compared based on Entropy, F-measure, Purity, and Average Distance of Documents to Cluster Centroid

(ADDC).Experimental results yields that the proposed algorithms generate better clusters.

Akter and Chung [5] proposed an evolutionary approach for document clustering based on genetic algorithm. In this paper genetic algorithm is not applied on the whole dataset directly. Authors propose two phase genetic algorithm approach in which dataset is partitioned into some groups and genetic algorithm is applied into each separate partition and another phase of genetic algorithm is applied on the result. This avoids the problem of local minima. Another advantage of this approach is that it does not need to specify the total number of clusters in advance. Authors compare the performance of K-Means, Genetic algorithm and proposed algorithm using benchmark database REUTERS-21578 which include 1000 texts from topics such as acq, crude, trade, grain and money-fx. Performances are compared using F-measure metric and latent semantic indexing (LSI) is also applied on dataset. Results show that proposed algorithm performs better than K-Means and Genetic algorithm. In 2013 [10] Changchun and Wang proposed a query specific density clustering in IR in order to improve the effectiveness of clustering. Here relationships of documents that are relevant to specific query are taken into consideration. Proposed model has been evaluated using mant TREC collections based on density clusters. The result reported verifies the superiority of the proposed methodology over other algorithms compared

In 2013[12] Minjuan proposed a Semantic Optimization Clustering Method for XML documents. In this paper, for XML element clustering Latent Semantic Indexing Model is used to find semantic relationship between terms and evolution function for K-medioid clustering algorithm is performed to automatic generate the optimal cluster number. Evolution function for clustering is based on compaction and resolution. Compaction is the intra-

cluster distance and Resolution is the inter-cluster distance. In this research IEEE CS data collection is used and to compare the performance of cluster quality information gain criteria is used. The information gain is clustering optimization is compared with non-optimization (fixed the cluster number in advance) and results indicate that clustering with optimization provides better clustering quality

DISCUSSION

In this survey paper various hierarchical and partitioning clustering algorithms are discussed for document clustering. Hierarchical Agglomerative clustering algorithm provide solution by firstly assigning each document to their own cluster and then selecting and merging pairs of clusters repeatedly to obtain a single cluster.
Partitioning clustering algorithm partition the whole document
into a set of k clusters at once. K-Means and K-Medoid algorithm of partitioning clustering algorithm are discussed in this paper. Hierarchical clustering algorithm provide better clustering but it has quadratic time complexity while K-Means partitioning algorithm has linear time complexity but it produce inferior cluster. Previous studies have shown that among various algorithm K-Means algorithm is more suitable for clustering large datasets but it only produces a local optimal solution. To find global optimal solution various optimization algorithms such as Genetic Algorithm, Particle Swarm Optimization, Hybrid PSO, Harmony Search are applied for document clustering.

These optimization algorithms improved the quality of clustering but require their own algorithm specific control parameters with common controlling parameters like population size and number of generations. The algorithm specific parameter is a crucial factor which arise difficulty with modification and hybridization. In future we can use any optimization algorithm that does not require algorithm specific parameter for document clustering.

REFERENCES

[1]. R. Forsati, M. Mahdavi, M. Shamsfard and M.R. Meybodi, "Efficient stochastic algorithms for document clustering", Information Sciences, Vol 220, pp. 269-291, 2012.

[2]. C. Xiaohui, E. P. Thomas and P. Paul, "Document Clustering using Particle Swarm Optimization", Swarm Intelligence Symposium, pp. 185-191, Pasadena, CA, USA, 2005.

[3]. M.S. Hameem and R. Ferdous, "An efficient K-Means Algorithm integrated with Jaccard Distance Measure for Document Clustering", First Asian Himalayas International Conference, pp. 1-6, Kathmandu, 2009.

[4]. S.T. Deokar, "Text Documents clustering using K Means Algorithm", International Journal of Technology and Engineering Science, Vol 1, pp. 282-286, 2013.

[5]. R. Akter and Y. Chung, "An Evolutionary Approach for Document Clustering", International Conference on Electronic Engineering and Computer Science, pp. 370-375, 2013.

[6]. V.K. Singh, N. Tiwari and S. Garg, "Document Clustering using K-Means, Heuristic K-Means and Fuzzy C-means", IEEE International Conference on Computational Intelligence and Communication Systems, pp. 297-301, 2011.

[7]. Y. Zhao, G. Karypis and U. Fayyad, "Hierarchical Clustering Algorithms for Document Datasets", Vol 10, pp. 141-168, 2005.

[8]. N. Sahoo, J. Callan, R. Krishnan, G. Duncan and R. Padman, "Incremental Hierarchical Clustering of Text Documents", ACM international conference on Information and knowledge management, pp. 357-366, New York, USA, 2006.

[9]. C. Xiaohui and E. Thomas Potok, "Document Clustering Analysis Based on Hybrid PSO+K-Means Algorithm", Journal of Computer Sciences, pp. 27-33, 2005.

[10]. C. Li and J.Y. Wang, "A Clustering Approach to Improving Pseudo-Relevance Feedback", Information Science and Engineering, pp. 35-38, 2012.

[11]. S. P. Kasivishwanathan and G. Kong G, " Novel document detection for massive data streams using distributed dictionary learning", IBM Journal of Research and Development, Vol 5, pp. 1-15, 2013.

[12]. Z. Minjuan, "An Effective Search Results Semantic Optimization Clustering Method for XML Fragments", Computer Science and Applications, pp. 479-482, 2013.

[13]. T. Kanungo, D. M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman and A.Y. Wu, "An Efficient K-Means Clustering Algorithm : Analysis and Implementation", Pattern Analysis and Machine Intelligence", Vol 24, pp. 881-892, 2002.

[14]. A.K. Jain, "Data Clustering: 50 years beyond K-Means", 19th International Conference on Pattern Recognition, pp. 651-666. Tampa, Florida, 2010.

[15]. S.S. Singh and N.C. Chauhan, "K-Means v/s K-Medoids: A Comparative Study", National Conference on Recent Trends in Engineering & Technology, pp.1-5. Gujarat, India, 2011.

★★★