Volume 6, Issue 2, February 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Overview on Efficient Document Clustering Techinques

[1]Subha S S*, [2]Kalaiselvi R
[1] PG Student & Kumaraguru College of Technology, India
[2] Assistant Professor & Kumaraguru College of Technology, India

Abstract Clustering is an technique of unsupervised learning that groups the objects such that objects are similar within the same group but different from other groups. As different clustering techniques involved and used appropriately in different fields, hierarchical clustering is an efficient clustering method in document searching. Document searching is a cumbersome process where cost increases as the number of documents increases. Document clustering process speeds up the search process such that cost is reduced. The main goal of clustering is to measure the two data points are similar or dissimilar. As many types of clustering techniques like hierarchical, partition, grid and density clustering are available, hierarchical clustering employs best on document searching. This paper explains about various document clustering techniques.

Keywords— Clustering, document search, performance cost

I. INTRODUCTION

Clustering is a data mining techniques mainly in data analysis used for grouping similar kind of data into groups (clusters), where high similarity data are placed in one cluster and dissimilar data in another cluster. Similarity and dissimilarity of data values are measured based on attribute values describing by data. Clustering mainly used to organize and categorize data. Cluster's center (centroid) is identified for all data points Cluster analysis can be used as a standalone data mining tool. Cluster center (centroid) measuring a similarity between input vector and all data point and determining which data point is nearest or most similar to centroid

Document clustering mainly used to divide a set of documents into clusters.The documents within each group or cluster should have large similarity and the similarity among different clusters should be reduced .Basically document clustering techniques divided into "Hierarchical" and "Partitioning .Hierarchical Clustering is like a tree structure called a dendrogram(tree) which displays cluster and sub clusters tree relationships [3,4]. Partition Clustering method used for partition a group of documents into a set of non-overlapping clusters. Hierarchical Clustering method is a better quality clustering approach than other methods [1]. Agglomerative clustering initially each data point as a singleton cluster and merges clusters until all points have been merged into a single cluster . some of the agglomerative clustering algorithms are CURE, ROCK, HAMELEON, BIRCH[2]. Divisive clustering initially consider all the data points in one cluster and then split desired number of clusters is obtained. some of the divisive clustering algorithm are bisecting k-means method, k means[2].

II. LITERATURE SURVEY

An efficient clustering algorithm used for large databases using CURE [5] agglomerative hierarchical clustering algorithm. CURE mainly partitions the random sample and partially data points cluster in each partition. After eliminating outliers, the pre clustered data in each partition is clustered and generate the final clusters. CURE algorithm features are: (1) arbitrarily shaped clusters (e.g., ellipsoidal), (2) robust to the presence of outliers (3) the algorithm uses space that is linear time complexity. (4) It appropriate for handling large data sets.

BIRCH is an agglomerative hierarchical clustering mainly used for very large databases and design for minimize the number of I/O operations .BIRCH mainly used to find a good clustering with a single scan of the data and improve the quality. BIRCH clustering algorithm proposed mainly handle the "noise" effectively. BIRCH algorthim groups the data set into compact sub clusters (called Clustering Features (CF). CF's are computed and updated as the sub clusters are being constructed. [6]. BIRCH can achieve a computational complexity of $O(n)$. BIRCH two generalizations are BUBBLE and BUBBLE-FM algorithms . BIRCH algorithm can find approximate solution to combinatorial problems with very large data sets.

ROCK (robust hierarchical clustering) is an clustering based on the notion of links suitable for handling large data sets using agglomerative hierarchical [7].ROCK algorthim used to combines k nearest neighbour, relocation and hierarchical agglomerative methods. ROCK calcuates cluster similarity is based on the number of points from different clusters that have neighbours in common in this algorthim [8]. A robust hierarchical clustering algorithm (ROCK) was develop that employs links and not distances for merging clusters [9].A quick version of the ROCK algorithm for clustering of categorical data is proposed it is called QROCK. And complexity of ROCKO(n2). QROCK is quicker than ROCK because of improve performance .

CHAMELEON as a hierarchical agglomerative clustering algorithm can find dynamic modeling. It is based on two phases: at first partitions the data points into sub-clusters, using a graph partitioning, then repeatedly merging sub-clusters, com from previous stage to obtain final clusters. The algorithm is proven to find clusters of diverse shapes, densities, and sizes in two-dimensional space [10]. Chameleon is an efficient algorithm that uses a dynamic model to obtain clusters of arbitrary shapes and arbitrary densities .

Particle Swarm Optimization (PSO) algorithm is stochastic optimization technique . PSO mainly used to find an optimal or near optimal solution. PSO algorithm used to generate initial cluster centroid for the K-means. In [11], a hybrid PSO+K-means document clustering algorithm which performs fast clustering and only local optimal solution is proposed. PSO+K-means is compared with PSO(Particle Swarm Optimization).PSO+K-means algorithm can generate the most accurate clustering results.

A simple and efficient techniques of K-means, bisecting K-means, where centroid are updated periodically are represented in[12]. Produced better performance than other regular K-means. Bisecting K-means has a linear time complexity. Agglomerative three hierarchical techniques are Intra-Cluster Similarity Technique (IST), Centroid Similarity Technique (CST) and UPGMA are compared with K means. UPGMA is the best hierarchical technique compared with K-means and bisecting K-means. Bisecting k-means is better than UPGMA and regular k-means. bisecting K-means is better performance because of production uniform size clusters.

A New Approach of Document Clustering using K Mean clustering[13] . In this approach document representation as document term matrix .In matrix represents the row and column in that rows represents documents and columns represent number of terms. Term number are arranged first and calculate weight and arranged weight(frequency in document) in decreasing order document are represented in vector space model or matrix representation in this paper using k-means clustering algorithm provide less time complexity.

A novel document partitioning method based on the non-negative factorization (NMF) of the term-document matrix is presented in [14]. The method differs from the latent semantic indexing method based on the singular vector decomposition (SVD) and the related spectral clustering methods. This is because semantic space derived by NMF does not need to be orthogonal and each document takes only non-negative values in all the latent semantic directions. An important benefit that each axis in the space derived by the NMF has a much more straightforward with each document cluster. Experimental result shows that the proposed document clustering method SVD and the eigen decomposition clustering methods in the easy and reliable clustering results and clustering accuracy.

A survey on k mean clustering and particle swarm optimization [15] proposed pso(particle swarm optimization) mainly a technique pulse code modulation.k mean clustering algorithm is a unsupervised techniques according to that k object as inital cluster and then calculate distance between each cluster from center and object and allocate to nearest cluster process continue until average of centroid distance equally .

III. CONCLUSION

Many clustering techniques are discussed in this paper .It has been proved that k mean clustering is the best among all the clustering techniques. Document clustering is mainly used for cluster the large no of document and mainly used among to retrieve the more relevant document into some cluster in that more relevant document are retrieved. document clustering mainly used to reduce the time complexity of the retrieve document .Future wok retrieve the document with low time period using some clustering algorthim.

REFERENCES

[1] Inderjit Dhillon Yuqiang Guan and Brian Kulis,"*A Fast Kernel-based Multilevel Algorithm for Graph Clustering*", International Journal of Business Intelligence and Data Mining ,Vol. 1, No. 1, July 2005.

[2] Bhagyashree Umale and Nilav M, "*Overview of K-means and Expectation Maximization Algorithm for Document Clustering",* International Journal of Computer Applications (0975 – 8887) International Conference on Quality Up-gradation in Engineering, Science and Technology (ICQUEST-2014).

[3] Michael Steinbach , George Karypis, and Vipin Kumar, "*A comparison of document clustering techniques*," In KDD Workshop on Text Mining, Vol. 58, pp. 236-244, May 2002 .

[4] 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, Aug 2002.

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

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

[7] William-Chandra Tjhi and Lihui Chen, "*A heuristic-based fuzzy co-clustering algorithm for categorization of high-dimensional data,"* Journal of Fuzzy Sets and Systems, vol. 159, issue 4, pp. 371-389, Feb 2008 .

[8] Malay K. Pakhira, "*A Modified k-means Algorithm to Avoid Empty",* International Journal of Recent Trends in Engineering, Vol. 1, No. 1, pp. 6220-226, March 2009 .

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

[10] Wei Xu, Xin Liu, and Yihong Gong, "*Document Clustering Based On Non-negative Matrix Factorization*," In Proc. of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 267-273,January 2010.

[11] Li Taoying, Chne Yan, Qu Lili and Mu Xiangwei, "*Incremental clustering for categorical data using clustering ensemble",* 29th Chinese Control Conference (CCC), pp. 2519-2524,May 2010.

[12] Shin-Jye Lee and Xiao-Jun Zeng, *"A three-part input-output clustering-based approach to fuzzy system identification",* 2010 10th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 55-60, June 2010 .

[13] Guo-Yan Huang, Da-Peng Liang, Chang-Zhen Hu and Jia-Dong Ren, *"An algorithm for clustering heterogeneous data streams with uncertainty"*, 2010 International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 4, pp. 2059-2064,August 2010.

[14] Xiaodi Huang, Xiaodong Zheng, Wei Yuan, Fei Wang, and Shanfeng Zhu, "*Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization,*" an International Journal on Information Sciences, vol. 181, issue 11, pp. 2293-2302, June 2011 .

[15] Yuan-chao Liu, Chong Wu, and Ming Liu, "*Research of fast SOM clustering for text information*," An International Journal Expert Systems with Applications, vol. 38, issue 8, pp. 9325-9333, August 2011 .

[16] Abdolreza Hatamloua, Salwani Abdullah, and Hossein Nezamabadi-pour, *"A combined approach for clustering based on K-means and gravitational search algorithms,"* Swarm and Evolutionary Computation, Available online 12 March 2012 .

[17] Neha Soni1, Dr. Amit Ganatra2.," *Comparative study of several Clustering Algorithms*", International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)Volume-2 Number-4 Issue-6 December-2012.

[18] S. Guha, R. Rastogi and K. Shim, CURE: An efficient clustering algorithm for large databases,Information Systems, 26 (2001), 35-58

[19] G. Karypis, E.H. Han and V. Kumar, CHAMELEON: Hierarchical clustering using dynamic modeling, IEEE Computer, 32 (1999), 68-75.

[20] S. Guha, R. Rastogi and K. Shim, ROCK: A robust clustering algorithm for categorical attributes ,IEEE Information Systems, 25 (2000), 345-36

