A SURVEY PAPER ON DIFFERENT TECHNIQUES OF DOCUMENT CLUSTERING

[1]Mamta Mahilane, [2]Mr. K. L. Sinha
[1]M.Tech Scholar, [2]Sr. Assistant Professor
Department of Comp. Sci. & Engg.
Chhatrapati Shivaji Institute of Technology, Durg
Chhattisgarh, India
Email: [1]mamta.csit@gmail.com , [2]khomlalsinha@csitdurg.in

***Abstract*—Now these days, there is great increase in the amount of information available on Internet. Also the amount of data kept in computer files and databases is rising at an amazing rate. Document clustering has been investigated for use in a number of different areas of text mining and information retrieval. In this paper various document clustering techniques are discussed along with their pros and cons. Clustering is currently one of the most crucial techniques for dealing with massive amount of information on the web and system. The principle of clustering depends on the concept of dividing a set of objects into a specified number of clusters on the basis of characteristics found in the actual data. Finally, we proposed the future techniques and methodologies which promise to enhance the ability of computer system in the field of text mining and information retrieval.**

***Keywords*—clustering, text mining, data mining, information retrieval.**

I. Introduction

Data mining and knowledge discovery in databases have been attracting a significant amount of research, industry, and media attention. Historically the notion of finding useful patterns in data has been given a variety of names including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing [11]. Data mining or knowledge discovery in database is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. Data mining is the application of specific algorithms for extracting patterns from data. This includes a number of technical approaches, such as clustering, data summarisation, classification, finding dependency networks, analysing changes and detecting anomalies [12]. The term *data mining* has been mostly used by statisticians, data analysts, and the management information systems (MIS) communities.

Data mining applications can be broadly divided into two categories namely, business and e-commerce data mining, scientific, engineering and health care data mining.Other application areas include crime detection, mortgage loan delinquency prediction, portfolio management, risk analysis etc.

II. Problem Identification

Every day, people encounter a large amount of information and for further analysis and management; they store or represent it as data. One of the vital means in dealing with these data is to classify or group them into a set of categories called clusters. Cluster can be defined as a set of like elements. Elements from various clusters must not alike [5]. In order to understand a new fact or to experience a new object, people always try to search the features that can characterize it, and then compare it with other known facts or objects, based on the similarity or dissimilarity, according to some certain standards or rules. Actually, classification is very simple activity that plays an important and vital role in the long history of human development. Basically, classification systems are either supervised or unsupervised, depending on whether they assign new inputs to one of a finite number of discrete supervised classes or unsupervised categories, respectively [6], [7], [8]. In supervised classification, the mapping from a set of input data vectors to a finite set of discrete class labels, is modelled in terms of some mathematical function, $y=y(X,W)$ where W is a vector of adjustable parameters. The values of these parameters are determined by an inductive learning algorithm, which is aimed to minimize an experimental risk functional on a finite data set of input–output; for example $(X_i,Y_i),i=1.....N$, where N is the finite cardinality of the available representative data set [6], When the inducer algorithm terminates or a convergence point occurs, an induced classifier is generated.

In unsupervised classification, no labeled data are available [9]. Clustering is an unsupervised learning data mining technique. The principle of clustering depends on the concept of dividing a set of objects into a specified number of clusters on the basis of characteristics found in the actual data. Clustering can be done without the knowledge of the category structure of class or pre-assumptions. They use the technique to define similarity or dissimilarity among the objects [4]. The keynote of clustering is to make individual clusters, i.e. partition the data into groups so that most similar objects are to be found within the group. The aim of clustering is to maximize the intra similarity values and to minimize inter similarity values. Document clustering has been investigated for use in a number of different areas of text mining and information retrieval. Initially, document clustering was investigated for improving the extraction in information retrieval systems. It was only an efficient way of finding the nearest neighbors of a document. More recently, clustering has been proposed for use in browsing a collection of documents or in organizing the results returned by a search engine in response to a user's query [1].

Clustering has been used in many application domains, including biology, medicine, anthropology, marketing and economics. Clustering applications include plant and animal classification, disease classification, image classification, pattern recognition, and document retrieval. The objectives of clustering is to uncover natural clustering, to make guess about the data and to find consistent and valid organization of data.

III. Clustering Algorithm

Most document clustering algorithms can be classified into two groups namely "partitioning" and "Hierarchal" algorithms. Hierarchical techniques produce a nested sequence of partition, with a single, all-inclusive cluster at the top and single clusters of individual points at the bottom. The result of a hierarchical clustering algorithm can be graphically displayed as tree, called a dendogram. Dendrogram graphically displays the merging process and the intermediate clusters.

There are two basic approaches to generating a hierarchical clustering:

A. *Agglomerative*

Start with the points as individual clusters and, step-by-step, merge the most similar or closest pair of clusters. This requires a definition of cluster similarity or distance.

B. *Divisive*

These techniques take the opposite approach from agglomerative technique. In this case, we need to decide, at each step, which cluster to split and how to perform the split.

The partitional clustering method partitions the database into predefined number of clusters. Partitional algorithm construct partitions of a database of N objects into a set of k clusters. The cluster construction is performed by determining the optimal partition with respect to an objective function. It usually adopts the iterative optimization paradigm. K-means and k-medoid are the two main categories of partitioning algorithm. In K-means algorithm, each cluster is represented by the center of gravity of the cluster, while in k-medoid algorithm the cluster is represented by one of the object of the cluster located near the center [13].

The remaining part of this section lists the various clustering algorithms those are in use now these days, and there is a short introductory information about some of these clustering algorithms in table(1).

Clustering Algorithms

1. *Hierarchical*
   - Agglomerative

   Single linkage, complete linkage, average linkage, balanced iterative reducing and clustering using hierarchies (BIRCH), clustering using representatives (CURE), robust clustering using links (ROCK), centroid linkage, median linkage.
   - Divisive

   divisive analysis (DIANA), monothetic analysis (MONA), Bisecting k-means (BKMS).

2. *Squared Error-Based (Vector Quantization)*
   K-Means, iterative self-organizing data analysis technique (ISODATA), genetic -means algorithm (GKA), partitioning around medoids (PAM).
3. *Mixture Densities based*
   Gaussian mixture density decomposition (GMDD).
4. *Graph Theory-Based*
   Delaunay triangulation graph (DTG), highly connected subgraphs (HCS), Chameleon,
5. *Combinatorial Search Techniques-Based*
   Genetically guided algorithm (GGA), TS cluster affinity search technique (CAST), clustering identification via connectivity kernels (CLICK) clustering, SA clustering.
6. *Fuzzy*
   Fuzzy C-means (FCM), mountain method (MM), possibilistic C-means clustering algorithm (PCM), fuzzy-shells (FCS).
7. *Kernel-Based*
   Kernel K-means, support vector clustering (SVC).
8. *Large-Scale Data Sets*
   Clustering Large Applications (CLARA), Clustering Large Applications based upon RANdomized Search (CLARANS), CURE, BIRCH, Density-based spatial clustering of applications with noise (DBSCAN), Density-based Clustering (DENCLUE), Wave Cluster, fuzzy clustering (FC).

[Table (1): Document Clustering Techniques]

Clustering Algorithm	**Compatibility of handling high Dimensional data**	**Complexity**	**Type**	**Advantage**	**Disadvantage**
DBSCAN	NO	$O(N^2)$ *[TIME]* $O(N^2)$ *[SPACE]*	Hierarchical	• DBSCAN is resistant to noise and can handle clusters of various shapes and sizes • DBSCAN has a notion of noise, and is robust to outliers. • DBSCAN can find arbitrary shape cluster.	• DBSCAN does not work well when dealing with clusters of varying densities or with high dimensional data not entirely deterministic • DBSCAN depends on the distance measure • Multi density dataset are not complete by DBSCAN. • Run time complexity is high.
CLARA	NO	$O(K(40+K)^2+K(N-K))^4$ *[TIME]*	Partitional	• Good clustering performance for unbiased sampling.	• If the sampling is biased we cannot have a good clustering • Trade off-of efficiency
CLARANS	NO	Quadratic in total performance	Partitional	• It is more effective than both PAM and CLARA • Handles outliers	• Iterative • The clustering quality depends on the sampling method
Average link	YES	$O(N^2)$ *[TIME]* $O(N^2d)$ *[SPACE]*	Hierarchical	• Quite Simple	• Less efficient than single link • sensitive to the shape and size of clusters
k-means	NO	O(NKd) [TIME] O(N) [SPACE]	Partitional	• If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls. • K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.	• Difficult to predict K-Value. • With global cluster, it didn't work well. • Different initial partitions can result in different final clusters. • It does not work well with clusters (in the original data) of Different size and Different density
BIRCH	NO	O(N) [TIME] O(N) [SPACE]	Hierarchical	• Scales linearly • It is also an incremental method that does not require the whole data set in advance. • It makes full use of available memory to derive the finest possible sub-clusters while minimizing I/O costs.	• It cannot handle all kinds of distance functions or dissimilarity function.

[Table (1): Document Clustering Techniques (Continue)]

Clustering Algorithm	**Compatibility of handling high Dimensional data**	**Complexity**	**Type**	**Advantage**	**Disadvantage**
Single Link	YES	O(N2) [TIME] O(N2d) [SPACE]	Hierarchical	• Quite Simple	• Low efficiency • Create clusters with long chains • sensitive to the presence of outliers and the difficulty in dealing with severe • differences in the density of clusters • displays total insensibility to shape and size of clusters
Complete link	YES	O(N2) [TIME] O(N2d) [SPACE]	Hierarchical	• Quite Simple	• Expensive • Complete-linkage is not strongly affected by outliers, but can break large clusters, and has trouble with convex shapes
MST	YES when using Dunn's Index value [16].	O(N2) [TIME] O(N2) [SPACE]	Partitional/ hierarchical	• Solves the problems of single link	• Quite complicated
PAM	NO	O(dK(N-K)2) [TIME] O(N2) [SPACE]	Partitional	• more robust against outliers • can deal with any dissimilarity measure • easy to find representative objects per cluster • more robust than k-Means in the presence of noise and outliers	• It is more costly than the k-Means method, • Like k-means, it requires the user to specify k

IV. Conclusion

In this paper we studied and summarized the different kind of clustering techniques in table form. We also discussed about the advantages and disadvantages of clustering techniques. A good clustering of text requires effective feature selection and

a proper choice of the algorithm for the task at hand. So this paper provides a quick review of the different clustering techniques in data mining.

References

[1] Michael Steinbach, George Karypis, Vipin Kumar, "A Comparison of Document Clustering Techniques",

[2] Xiaohui Cui, Thomas E. Potok, "Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm", Journal of Computer Sciences (Special Issue): 27-33, 2005.

[3] Rui Xu, Donald Wunsch, “Survey of Clustering Algorithms”, IEEE transactions on neural networks, vol. 16 no. 3, may 2005

[4] K. Rajendra Prasad, or. P.govinda rajulu, “A Survey on Clustering Technique for Datasets using Efficient Graph Structures”, International Journal of Engineering Science and Technology Vol. 2 (7), 2010, 2707-2714.

[5] Margaret H. Dunham, Data Mining : Introductory and Advanced topic, Pearson education, 2011.

[6] L. Bobrowski and J. Bezdek, “C-Means Clustering with the l and l norms”,” IEEE Trans. Syst., Man, Cybern., vol. 21, no. 3, pp. 545–554, May-Jun. 1991.

[7] V. Cherkassky and F. Mulier, Learning From Data: Concepts, Theory, and Methods”, New York: Wiley, 1998.

[8] R. Duda, P. Hart, and D. Stork, “Pattern Classification”, 2nd ed. New York: Wiley, 2001.

[9] B. Everitt, S. Landau, and M. Leese, “Cluster Analysis”, London:Arnold, 2001.

[10] Jain, R. Duin, and J. Mao, “Statistical Pattern Recognition: A Review,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 1, pp. 4–37, 2000.

[11] Jain, A.K., M.N. Murty and P.J. Flynn, “Data Clustering: A Review” ACM Computing Survey, 31:264-323, 1999.

[12] M. Steinbach, G. Karypis, V. Kumar, “A Survey of Document Clustering Techniques”, Text Mining Workshop, KDD, 2000.

[13] Arun K.Pujari, Data Mining Techniques, University publication, 2001

[14] Deepti Sisodia, Lokesh Singh, Sheetal Sisodia, Khushboo saxena, “Clustering Techniques: A Brief Survey of Different Clustering Algorithms”, International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol. 1 Issue 3 September 2012.

[15] Yogita Rani, Dr. Harish Rohil, “A Study of Hierarchical Clustering Algorithm”, International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 11 (2013), pp. 1225-1232.

[16] Dr. B. Eswara Reddy1 and K. Rajendra Prasad, “Reducing Runtime Values in Minimum Spanning Tree based Clustering by Visual Access Tendancy”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.3, May 2012.