A Survey Based on Data Clustering Algorithms

Shreyata Khatri[1] Dr. Kanwal Garg[2]
[1]Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science & Application
[1,2]Kurushetra University, Kurushetra

Abstract— Grouping of document on the bases of clustering is vital task in document categorization. Documents are increasing day by day so, it is very crucial to segregate these documents in clusters. The objective of this paper is to discuss clustering algorithms and issues and challenges concerned with document clustering. The algorithms under consideration are: k-means, hierarchical clustering, density based, grid based and model based algorithm.

Key words: Data Mining, Clustering, Various Types of Clustering Algorithm

I. INTRODUCTION

Data mining refers to extracting or mining information from huge amounts of data. Data mining is uniquely categorized as a way for handling large amount of data sets. The documents may be web pages, blog posts, news articles, or other text files. The text documents are impulsively mounting over the internet, e-mail and web pages and are stored in the electronic database format. For organizing the large anthology of documents, clustering technique is widely used. Clustering is known as an unsupervised classification that means clustering as no predefined classes. Clustering organize the similar types of documents under single clusters. Documents in same cluster are more similar to one another than the documents in another cluster. Clustering techniques are widely used in various fields like machine learning, statistical data analysis, text mining, pattern reorganization, information retrieval, data compression etc.

The process of document categorization with clustering consist of various steps like document preprocessing, term selection, attribute reduction and maintaining the relationship between the important terms using background knowledge, wordnet. Figure1 shows various step of categorization process. First of all the data set is divide into small sets of document using divide and conquer strategy. Preprocesing is performed which take plain text as input and set of tokens are obtained as a result. Preprocessing is performed for removal of special characters, puncatation marks, word with no meaning and low frequency. Further, base form of words is generated with stemming process.

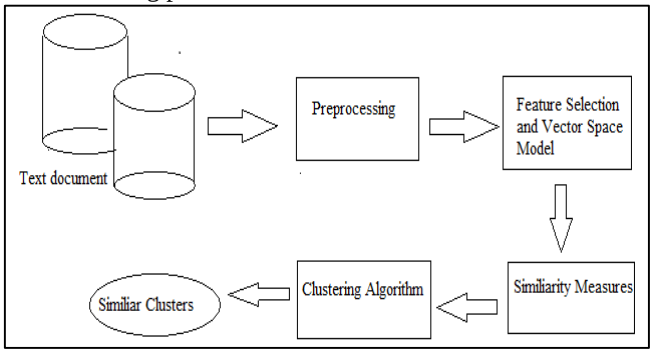

Fig. 1: Document categorization with clustering

Vector space model is a retrieval technique which is widely used in text mining. VSM is performed for vocabulary building which is performed on the basis of term frequency inverse document frequency model. After VSM, distance between various clusters is calculated using various similarity techniques. Similarity measures reflect the degree of closeness or separation between various clusters. Cosine similarity is the most commonly used similiarity measure.

In this paper there is brief overview of the clustering algorithms and issues and challenges concerned with clustering. Section 1 gives the introduction, section 2 explores types of clustering algorithm, section 3 presents the reviews of various researchers, section 4 explores comparison evaluation, section 5 describes various issues and challenges concerned with clustering, section 6 concludes the paper.

II. TYPES OF CLUSTERING ALGORITHM

Clustering of similar data together is performed with various clustering algorithm. Various clustering algorithm are available in literature.

- Partition based algorithm
- Hierarchical algorithm
- Density based algorithm
- Grid based algorithm
- Model based algorithm

A. Partition Based Algorithm

K-means algorithm comes under partition based algorithm which partitions a given set of observation(x1, x2....an) in d-dimensional vector into k-clusters where data in one cluster is similar to each other in such a way that the total deviation of each object from is cluster Centre. It is efficient in processing large data sets and the clusters formed have spherical shapes. This algorithm is sensitive to noise.

B. Hierarchical Algorithm

Hierarchical based algorithm is also known as connectivity based algorithm which create a hierarchical decomposition of the set of data. It differs by the way distances are computed. No a-priori information about the number of Clusters is required but the algorithm can never undo what was done previously.

C. Density Based Algorithm

Density based clustering algorithm try to find clusters based on density of data points in s region. It does not require a-priori specification and able to identify noisy data while clustering. It fails in case of neck type of data set and it does not work well in case of high dimensionality data. It needs only one scan of the input dataset and density parameters to be initialized

All rights reserved by *www.ijsrd.com*

D. Grid Based Algorithm

Grid based clustering, partition the space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed[4]. Its fast processing time, which is typically independent of the number of data objects, yet dependent on only the number of cells in each dimension in the quantized space.

E. Model Based Clustering

A model is hypothesized for each of the clusters and produces a classification scheme for a set of unlabelled objects.

III. RELATED WORKS

Various researchers had worked on various clustering algorithms, some improved them, some implemented new ones, and some studied and make a distinction among them. Following are some of the previous work performed by researchers.

The paper analyzed [2]k-mean and hierarchical algorithm by applying validation measures like entropy, coefficient of variance, f-measure and time. The experimental results show that k-mean algorithm shows better results as compared to hierarchical algorithm and takes less time for execution.[1] compared various algorithm based on size of dataset, numbers of clusters, type of data set, and type of software. The results shows that performance of k-means and EM algorithm is better than hierarchical algorithm and SOM algorithm shows more accuracy in classifying the object into similar clusters than other algorithm.

K-means clustering comes under partition algorithm which partition n observation into k clusters. Several other clustering algorithm are proposed for dealing with document clustering task including Novel algorithm[5] for automatic clustering suggested how clustering is done automatically, Improved partition K-means algorithm[6] presented new method for initializing centroids. Ontology based k-means algorithm [7] presented how ontological domains are used in clustering documents. In [8] datasets are tasted for analyzing efficiency of K-means algorithm when crime documents are given as an input. K-means clustering algorithm is considered as the best algorithm for document categorization. The paper analysed[9] the performance of three major clustering algorithms k means, hierarchical clustering and density based clustering algorithms and compared the performance of these three major clustering algorithms on aspects of correctly class wise cluster building ability of algorithm. Performance of these three techniques was compared using a clustering weka tool. [10]compared the three basic algorithms f or model-based clustering on high-dimensional discrete variable data sets and based on the log-marginal-likelihood criterion EM algorithm performs best of all other algorithm. The paper [16] discusses various algorithm to overcome the short comings of k-means clustering. N-cut method is a divisive hierarchical clustering and the output results in more than two clusters. This algorithm is widely used where multiple clusters are required. SLINK and CLINK are examples of hierarchical clustering. Various pros and cons of above discussed algorithm is shown in Table1.

[4] discussed density based clustering algorithm for large datasets and discussed DBSCAN algorithm which is used to find clusters of arbitrary shapes and sizes yet may have trouble with clusters of varying density. The density- and grid-based clustering technique CLIQUE [3] has been proposed for data mining in high-dimensional data space. Input parameters are the size of the grid and a global density threshold for clusters. The major difference between this algorithm and all other clustering approaches is that this method also detects subspaces of the highest dimensionality as high-density clusters exist in those subspace. The paper [13] compared two algorithm CLARANS nad DBSCAN in terms of effectivity by visual insepection as both the algorithms are of different type,and have no common quantitative measure of the classification accuracy. The experimental results show that run time of DBSCAN is slightly higher than linear in number of points and the run time of CLARANS is close to quadratic in number of points. DBSCAN outperforms CLARANS by a factor of 100. The paper [12] presents a technique TDCT(triangle-density based clustering technique) for efficient clustering of spatial data. The polygon approach was accessed to execute the clustering where the number of points inside of a polygon is calculated using barycentre formula. Due to smaller space dimension, it is efficient to partition the datasets in triangular shape compared to any other polygonal shape. The paper[14] presented an algorithm OPTICS-ordering points to identify the clustering structure which does not produce clustering of datasets explicitly but instead create an augmented ordering of database representing its density-based clustering structure. It automatically extract not only traditional clustering info but also the intrinsic clustering structure. The paper [15] presents a review of various density based clustering algorithm such as DBSCAN, OPTICS, VDBSCAN, GMDBSCAN with there various pros and cons. VDBSCAN works on the major drawback of DBSCAN, it can work well with clusters of varying density. Table 1 shows various pros and cons of above discussed algorithms.

Algorithm	Complexity	Pros	Cons
k-means	Time complexity of algorithm is O(nkl). Space complexity is O(k+n).	Produce tighter clusters.	Difficult to know about number of clusters in advance.
Spectral Clustering Algorithm	Computational complexity is $O(n^3)$ where n is number of data points.	Local optima is not considered, work faster.	High computational complexity
CLARANS	Quadratics in total performance	Can handle points objects as well as polygon objects efficiently.	Doesn't work well with high dimensional data.
Agglomerative Hierarchical	Time complexity is $O(n^3)$ which makes them too slow for large datasets.	Generate small clusters, produce ordering of objects.	Different distance metrics generate

All rights reserved by www.ijsrd.com

Algorithm			different result.
Divisive Hierarchical Algorithm	Time complexity is O(2n) which is even worse.	Ease of handling any forms of similiarity or distance.	Vagueness in termination criteria.
SLINK	Time complexity is $O(n^2)$	Can handle non global shape.	Small clusters have small distance but elements at opposite ends much farther from each other.
CLINK	Time comlexity is $O(n^2)$.	Have compact clusters with small diameter.	Pays too much attention to outliers
DBSCAN	Time complexity of algorithm is O(n)	Resistant to nosie, can handle clusters of various size and shapes.	doesn't work well for environments with different densities.
OPTICS	Time complexity is $O (n)^2$.	Work well with clusters of varying density.	It has the problem of overlapped clusters.
VDBSCAN	Time complexity is O (time complexity of DBSCAN*i), where i is the number of iterations.	Work well with clusters of varying density.	Requires k as an input parameter from user deteriorates the accuracy of algorithm.
Grid Based Algorithm	Time complexity is O (number of populated grid cells and have limited shape to union of grid cells.)	Fastest processing time.	Low accuracy.
STING	Computational complexity is O(k) where k is number of grid cells.	Query independent approach	Efficient for low dimension
EM Algorithm	Computional complexity is O(n*k).	Produce good clusters with huge data sets and sensitive to noise.	Less accuracy and is highly complex in nature.

Table 1: Review of Various Clustering Algorithm

IV. Issues and Challenges

The major issues and challenges associated with document categorization based on clustering are:-

1) Identification of distance/similarity measures: distance/ similarity measures reflect the closeness or separation of target object. Identification of measures for numerical or categorical data type is difficult.
2) Number of clusters: identification of the number of clusters in a data set is a common problem in data clustering.
3) Structure of database: data in real life may not contain identifiable clusters. The orders in which tuples are arranged also affect the result when executing the algorithm.
4) Types of attributes in a database: A database contains both numerical as well as categorical attributes. All types of categorical type so that simple calculations can be made.
5) Selection of algorithm: The main challenge is the selection of algorithm for clustering process based on type of dataset, time requirement, efficiency needed, accuracy required, error tolerance etc
6) Choosing the initial clusters: Partition based algorithms found it difficult to discover the number of clusters in advance.

V. Conclusion

This paper deals with study of clustering process and clustering algorithms. It first defines the clustering which is process of grouping similar objects under a cluster whose members contain some kind of resemblance. After that this paper discusses the various details of clustering algorithms. This paper highlights the issues and challenges concerned with clustering which may be helpful for the upcoming researchers to carry on their work.

References

[1] Osama Abu Abbas, "Comparison of data clustering algorithm" The International Arab Journal of Information Technology, volume 5,no. 3, pp: 320-325,july 2008
[2] Manpreet kaur, "Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection", International Journal of Advanced Research in Computer Science and Software Engineering, (ijaarcsse), volume 3, pp:1454-1459,july 2013.
[3] K.Kameshwaran, K.Malarvizhi,"Survey on Clustering Techniques in Data Mining", International Journal of Computer Science and Information Technologies (IJCSIT), Volume 5 (2) ,pp: 2272-2276 ,2014.
[4] Lovely Sharma, "A Review on Density Based Clustering Algorithms for Very Large Datasets", International journal of emerging technology and advanced engineering, volume 3,PP: 398-403,dec. 2013.
[5] Zonghu Wang, Zhijing Liu, Donghui Chen, Kai Tang,"A New Partitioning Based Algorithm For Document Clustering",Eighth International Conference on Fuzzy Systems and Knowledge Discovery,pp: 1741 - 1745 IEEE,20 I I.

All rights reserved by *www.ijsrd.com*

[6] S.C. Punitha, R. Jayasree anddr. M. Punithavalli, "Partition Document Clustering using Ontology Approach", Multimedia and Expo, 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 04 06,pp: 1-5, 2013.
[7] Lus Filipe da Cruz Nassif and Eduardo Raul Hruschka, Document Clustering for Forensic Analysis: "An Approach for Improving Computer Inspection," IEEE transactions on information forensics and security, Volume 8 ,pp: 46 - 54 Jan 2013.
[8] Tapas Kanungo,David M. Mount,Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" IEEE transactions on pattern analysis and machine intelligence, volume 24, NO. 7,pp :881- 892 JULY 2002.
[9] Bharat Chaudhari and Manan Parikh "A Comparitive Study of Clustering Algorithms Using Weka Tool" International Journal of Applications or Innovation in Engineering and Management ISSN: 2319-4817, Volume 1, Issue 2, October 2012.
[10] Marina Maile, David Herkerman,:"An Experimental Comparsion of Model Based clustering Methods",pp: 9-29, 2001.
[11] Pramod Bide," Improved Document Clustering using K-means Algorithm",IEEE,2015.
[12] Hrishav Bakul Barna, Dhiraj Kumar Das,Sauravjyoti Sarmah," a density based clustering technique for large spatial data using polygon approach,IOSR journal of computer engineering(IOSRJCE), volume 3,pp: 1-9,2012.
[13] Mertin Ester, Hans-Peter Kriegel,Jorg Sander,Xiaowei Xu," A Density Based Algorithm For Discovering Clusters In Large Spatial Databases With Nosie", pp:226-231,1996
[14] Mihael A nkerst,Markns M. Breuning, Hanspeter Kriegel,Jorg Sander,"OPTICS:ordering points to identify the clustering structure", Int. Conference on management of data, 1999.
[15] Noticewala Maitry,Dinesh Vaghela,"survey of different density based algorithms on spatial dataset", International Journal of Advanced Research in Computer Science and Management Studies, volume 2, issue 2, pp:362-366 feb. 2014.
[16] Ashwini Gulhane ,Prashant L.Paikrao, D.S.Chaudhari, " a review of image data clustering techniques", volume 2, issue 1,ISSN 1 pp:212-215,march 2012.
[17] Suman, Pooja Mittal, "Comparsion And Analysis Of Various Clustering Methods In Data Mining On Education Datasets Using Weka Tool",International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 2, ISSN 2278-6856,pp: 240-244, April 2014

All rights reserved by *www.ijsrd.com*