WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Survey on Various Data Mining Based Clustering Techniques

Hemlata Kargaiya*[1], Dr. Harsh Lohiya*[2], Dr. Rajendra Singh Kushwaha*[3]

[*1]Research Scholar, SSSUTMS, Sehore, Madhya Pradesh, India.
[*2] Assistant Professor, SSSUTMS, Sehore, Madhya Pradesh, India.
[*3] Associate Professor, SSSUTMS, Sehore, Madhya Pradesh, India.

ABSTRACT

Clustering means keeping similar objects together. Document clustering is an extension of clustering, which is related to keeping similar text documents together. Clustering plays a vital role in many real world applications. Document clustering is a special application of clustering. Document clustering plays a vital role in development of search engines, where a group of document is required to listed as a result of query in minimum response time. In search engines, document clustering improves search time and search result. This paper elaborates the concept of document cum text clustering. This thesis will provide a survey of recent work done in the field of text clustering. A critical review of modern text clustering techniques will also be provided by this thesis. This thesis presents a methodology for document clustering. This methodology is based on an efficient K means variant. This algorithm makes use of density based connected objects for selecting better clusters. It will result in overall improvement in clustering accuracy.

Keywords- Document Clustering, Search Engine, Data Mining, Forecasting, Clustering Method.

1. Introduction

We provide a comprehensive review of different clustering techniques in data mining. Clustering refers to the division of data into groups of similar objects. Each group, or cluster, consists of objects that are similar to one another and dissimilar to objects in other groups. When representing a quantity of data with a relatively small number of clusters, we achieve some simplification, at the price of some loss of detail (as in lossy data compression, for example). Clustering is a form of data modeling, which puts it in a historical perspective rooted in mathematics and statistics. From a machine learning perspective, clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data concept. Clustering as applied to data mining applications encounters three additional complications: (a) large databases, (b) objects with many attributes, and (c) attributes of different types. These complications tend to impose severe computational requirements that present real challenges to classic clustering algorithms. These challenges led to the emergence of powerful broadly applicable data mining clustering methods developed on the foundation of classic techniques. These clustering methods are the subject of this survey.

1.1 Requirements of Clustering in Data Mining

Here are the typical requirements of clustering in data mining:

- Scalability - We need highly scalable clustering algorithms to deal with large databases.
- Ability to deal with different kind of attributes - Algorithms should be capable to be applied on any kind of data such as interval based (numerical) data, categorical, binary data.
- Discovery of clusters with attribute shape - The clustering algorithm should be capable of detect cluster of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small size.
- High dimensionality - The clustering algorithm should not only be able to handle low- dimensional data but also the high dimensional space.
- Ability to deal with noisy data - Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- Interpretability - The clustering results should be interpretable, comprehensible and usable. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept.

From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others. Presenting data

by fewer clusters necessarily loses certain fine details (loss in data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters.

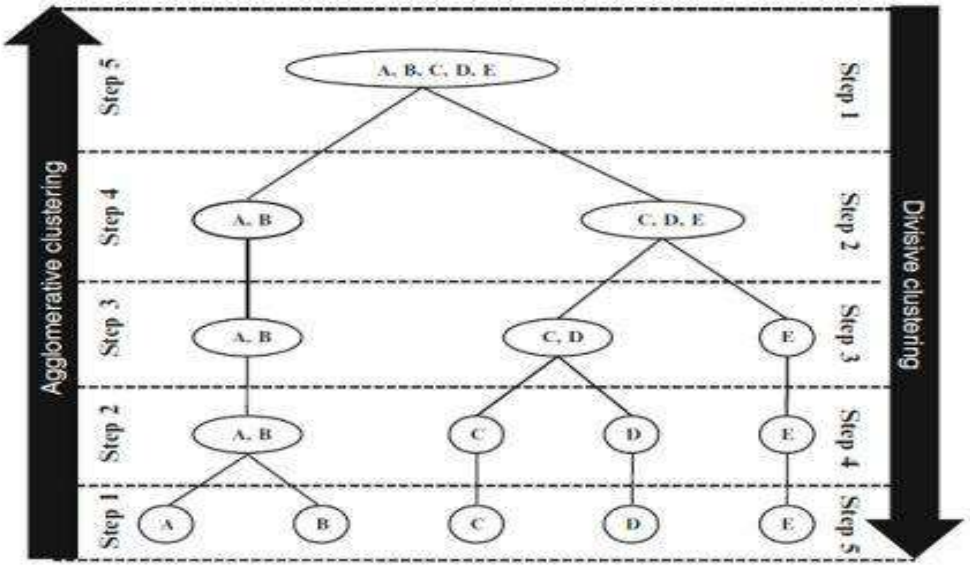
Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Clustering techniques fall into a group of undirected data mining tools. The goal of undirected data mining is to discover structure in the data as a whole. In general, there are two types of attributes associated with input data in clustering algorithms, i.e., numerical attributes, and categorical attributes. Numerical attributes are those with a finite or infinite number of ordered values, such as the height of a person or the x-coordinate of a point on a 2D domain. On the other hand, categorical attributes are those with finite unordered values, such as the occupation or the blood type of a person. Many different clustering techniques have been defined in order to solve the problem from different perspective, i.e. partition based clustering, density based clustering, hierarchical methods and grid-based methods etc.

2. Clustering Techniques And Algorithms

Clustering is the main task of Data Mining. And it is done by the number of algorithms. The most commonly used algorithms in Clustering are Hierarchical, Partitioning, Density based, Grid based, Model Based and Constraint based algorithms.

2.1 Hierarchical Algorithms

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. It is the connectivity based clustering algorithms. The hierarchical algorithms build clusters gradually. Hierarchical clustering generally fall into two types: In hierarchical clustering, in single step, the data are not partitioned into a particular cluster. It takes a series of partitions, which may run from a single cluster containing all objects to „n" clusters each containing a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by series of fusions of the „n" objects into groups, and divisive methods, which separate „n" objects successively into finer groupings.



2.2 Partitioning Algorithms

Partitioning algorithms divide data into several subsets. The reason of dividing the data into several subsets is that checking all possible subset systems is computationally not feasible; there are certain greedy heuristics schemes are used in the form of iterative optimization. Specifically, this means different relocation schemes that iteratively reassign points between the k clusters. Relocation algorithms gradually improve clusters.

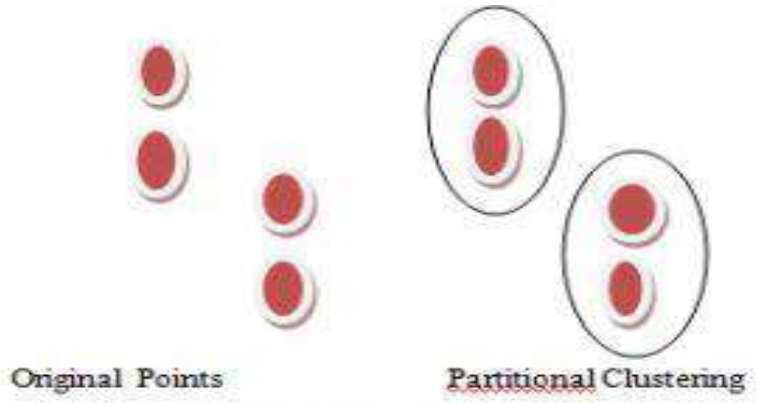


Figure 2: Partitioned Clustering

There are many methods of partitioning clustering; they are k-mean, Bisecting K Means Method, Medoids Method, PAM (Partitioning Around Medoids), CLARA (Clustering LARge Applications) and the Probabilistic Clustering. We are discussing the k-mean algorithm as: In k-means algorithm, a cluster is represented by its centroid, which is a mean (average pt.) of points within a cluster. This works efficiently only with numerical attributes. And it can be negatively affected by a single outlier. The k-means algorithm is the most popular clustering tool that is used in scientific and industrial applications.

2.3 Density-Based Clustering

In density-based clustering, clusters are defined as areas of higher density than the remaining of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points. There are two major approaches for density-based methods. The first approach pins density to a training data point and is reviewed in the sub-section Density-Based Connectivity. In this clustering technique density and connectivity both measured in terms of local distribution of nearest neighbors. So defined density connectivity is a symmetric relation and all the points reachable from core objects can be factorized into maximal connected components serving as clusters.

2.4 Grid Based Algorithms

Grid-based clustering where the data space is quantized into finite number of cells which form the grid structure and perform clustering on the grids. Grid based clustering maps the infinite number of data records in data streams to finite numbers of grids. Grid based clustering is the fastest processing time that typically depends on the size of the grid instead of the data. The grid based methods use the single uniform grid mesh to partition the entire problem domain into cells and the data objects located within a cell are represented by the cell using a set of statistical attributes from the objects.

2.5 Model-Based Methods

In this method a model is hypothesize for each cluster and find the best fit of data to the given model. This method locates the clusters by clustering the density function. This reflects spatial distribution of the 1111111111111111111111111111data points. This method also serve a way of automatically determining number of clusters based on standard statistics, taking outlier or noise into account. It therefore yield robust clustering methods.

2.6 Constraint-Based Method

In this method the clustering is performed by incorporation of user or application oriented constraints. The constraint refers to the user expectation or the properties of desired clustering results. The constraint gives us the interactive way of communication with the clustering process. The constraint can be specified by the user or the application requirement.

3. Conclusion

This part report clarifies the finish of the directed examination around the content bunching methods. In this circumstance, after examination some significant realities are discovered, that are assists with discovering future expansion of the anticipated work.

In this exploration work, the attention is on Document Clustering which is extremely ongoing innovation, we investigate many existing algorithms and proposed another one. We characterize a methodology that shows original copy bunching strategies for PCs held onto examination. Additionally, we revealed and talked about a digit of down to earth result that can be useful for analysts and specialists of scientific registering. In particular, we saw that grouping calculations in fact will in general prompt bunches framed by either applicable or immaterial reports, consequently adding to improve the master inspector's activity.

Moreover, our assessment of the proposed approach in five certifiable applications show that it can possibly accelerate the PC review process. As bunching assumes an extremely essential job in different applications, numerous explores are as yet being finished. The forthcoming developments are for the most part because of the properties and the attributes of existing techniques. This proposal presents a prologue to the current archive grouping idea alongside the techniques utilized for report bunching. A refreshed grouping method is likewise talked about in detail alongside the model.

We likewise saw that partitioned calculations additionally accomplished top notch result when appropriately instated. Thinking about the methodologies for surmise the quantity of bunches, the relative legitimacy model known as outline has appeared to improved form. In particular, we saw that grouping calculations truly will in general make bunches shaped by either related or irrelevant reports, in this manner adding to improve the master analyst's activity. Besides, our estimation of the anticipated strategy in five genuine applications show that it can possibly increment up the PC review process.

4. Future Scope

The calculations proposed in this theory are simple stage and there are numerous potential upgrades that can be executed, focused on further procedure the utilization of things bunching calculations in similar capacity, an able area for future work possess look at routine methodologies for group marking. The commitment of marks to bunches may allow the expert analyst to arrange the semantic substance of each group all the more rapidly—in the long run even before their substance.

REFERENCES

[1] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, “Mining writeprints from anonymous e-mails for forensic investigation,” Digital Investigation, Elsevier, vol. 7, no. 1– 2, pp. 56–64, 2010.

[2] 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, 2010.

[3] Alam, S., Dobbie, G., Riddle, P. and Naeem, M.A. “Particle Swarm Optimization Based Hierarchical Agglomerative Clustering”, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 2, pp. 64-68, 2010.

[4] 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, 2010.

[5] 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, 2010.

[6] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem, “Towards an integrated e-mail forensic analysis framework,” Digital Investigation, Elsevier, vol. 5, no. 3–4, pp. 124–137, 2009.

[7] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and R. Zunino, “Manuscript clustering for digital forensics analysis,” Computat. Intell. Security Inf. Syst., vol. 63, pp. 29–36, 2009

[8] Pallav Roxy and Durga Toshniwal, “Clustering Unstructured Manuscript Documents Using Fading Function”, International Journal of Information and Mathematical Sciences, Vol. 5, No. 3, pp. 149-156, 2009