Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
REVIEW ARTICLE
Available Online at www.ijarcs.info

A Survey on STING and CLIQUE Grid Based Clustering Methods

Suman
Department of Computer Science,
Kurukshetra University, Kurukshetra, India

Pinki Rani
Department of Computer Science,
Kurukshetra University, Kurukshetra, India

Abstract: Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity but are very dissimilar to objects in other clusters. Clustering methods can be classified as Partitioning methods, Hierarchical methods, Density-based methods, Grid-based methods and Model-based methods. This paper intends to overview the grid based clustering methods like STING and CLIQUE. The grid based clustering approach uses a multiresolution grid data structure. It quantizes the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. The main advantage of the approach is 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.

Keywords: Clustering, STING, CLIQUE, grid based clustering, hierarchical structure.

I. INTRODUCTION

Grid-based Algorithm define a set of grid-cells, it assign objects to the appropriate grid cell and compute the density of each cell and eliminate cells, whose density is below a defined threshold t. Form clusters from contiguous (adjacent) groups of dense cells (usually minimizing a given objective function). Grid-based algorithm uses multiresolution grid data structure. Clustering complexity depends on the number of populated grid cells and not on the number of objects in the dataset. Some popular Grid Based clustering methods are:

- STING (a STatistical INformation Grid approach)
- CLIQUE (Clustering In QUEst)

II. STING: A STATISTICAL INFORMATION GRID APPROACH

In the STING algorithm, the spatial area is divided into rectangular cells. There are several different levels of such rectangular cells corresponding to different resolution and these cells form a hierarchical structure. Each cell at a high level is partitioned to form a number of cells of the next lower level. Statistical information of each cell is calculated and stored beforehand and is used to answer queries [1].

A. Hierarchical Structure for STING Clustering

The area is divided into rectangular cells and a hierarchical structure is employed. Let the root of the hierarchy be at level 1; its children at level 2, etc. A cell in level *i* corresponds to the union of the areas of its children at level *i* + 1. Each cell (except the leaves) has 4 children and each child corresponds to one quadrant of the parent cell. The root cell at level 1 corresponds to the whole spatial area. The size of the leaf level cells is dependent on the density of objects. In addition, a desirable number of layers can be obtained by changing the number of cells that form a higher level cell. Assuming the space of two dimensions, it is easy to generalize the hierarchy structure to higher dimensional models [2]. In two dimensions, the hierarchical structure is illustrated in Fig. 1.

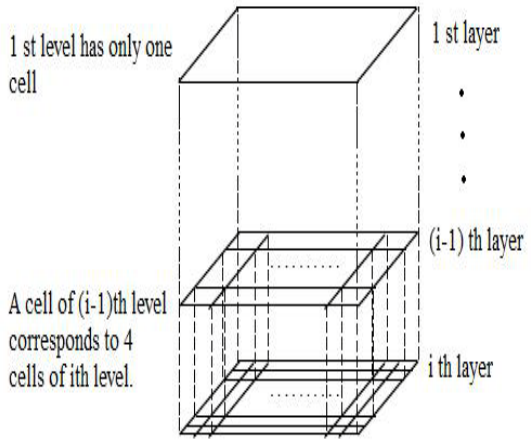

Fig. 1: Hierarchical Structure

If the statistical information stored in the STING hierarchical structure is not sufficient to answer a query, then we have to query the underlying database. However, the statistical information in the STING structure can answer many commonly asked queries very efficiently and we often do not need to access the full database.

The most commonly used query is region query which is to select regions that satisfy certain conditions. Another type of query elects regions and returns some function of the region, e.g., the range of some attributes within the region.

B. STING Algorithm

The algorithm is given below:

1. Determine a layer to begin with.
2. For each cell of this layer, we calculate the confidence interval (or estimated range) of probability that this cell is relevant to the query.
3. From the interval calculated above, we label the cell as *relevant* or *not relevant.*
4. If this layer is the bottom layer, go to Step 6; otherwise, go to Step 5.
5. We go down the hierarchy structure by one level. Go to Step 2 for those cells that form the *relevant* cells of the higher level layer.

6. If the specification of the query is met, go to Step 8; otherwise, go to Step 7.
7. Retrieve those data fall into the *relevant* cells and do further processing. Return the result that meet the requirement of the query. Go to Step 9.
8. Find the regions of *relevant* cells. Return those regions that meet the requirement of the query. Go to Step 9.
9. Stop [3].

C. Advantages

i) It is a query-independent approach since the statistical information exists independently of queries.
ii) The computational complexity is $O(K)$, where K is the number of grid cells at the lowest level. Usually, $K << N$, where N is the number of objects.
iii) Query processing algorithms using this structure are trivial to parallelize.
iv) When data is updated, we need not recompute all information in the cell hierarchy. Instead, we can do an incremental update [4].

III. CLIQUE: A DIMENSION GROWTH SUBSPACE CLUSTERING METHOD

CLIQUE (Clustering in QUEst) is a bottom-up subspace clustering algorithm that constructs static grids. It uses apriori approach to reduce the search space. CLIQUE is a density and grid based i.e. subspace clustering algorithm and find out the clusters by taking density threshold and number of grids as input parameters. CLIQUE operates on multidimensional data by not operating all the dimensions at once but by processing a single dimension at first step and then grows upward to the higher one [5].

A. CLIQUE Working

The clustering process in CLIQUE involves:

1. CLIQUE partitions the d- dimensional data space into non-overlapping rectangular units called grids according to the given grid size and then find out the dense region according to a given threshold value. A unit is dense if the data points in this are exceeding the threshold value.
2. Clusters are generated from the all dense subspaces by using the apriori approach [6]. CLIQUE algorithm generates minimal description for the clusters obtained by first determining the maximal dense regions in the subspaces and then minimal cover for each cluster from that maximal region. It repeats the same procedure until covered all the dimensions.

A k-dimensional cell c (k>1) can have at least l points only if every (k-1)-dimensional projection of c, which is a cell in a (k-1)-dimensional subspace, has at least l points, where l is the density threshold. Consider the fig. where the embedding data space contains three dimensions: age, salary, and vacation. A 2-D cell, say in the subspace formed by age and salary, contains l points only if the projection of this cell in every dimension, that is age and salary, respectively, contains at least l points [7].

The following figures show that dense units found with respect to age for the dimensions salary and vacation are intersected to provide a candidate search space for dense units of higher dimensionality [8].

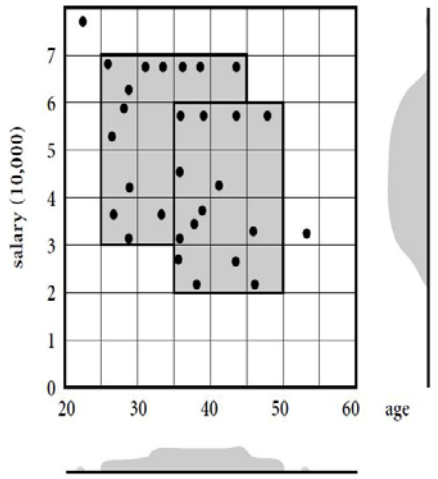

Fig. 2(a): Identification of clusters along (age, salary) plane

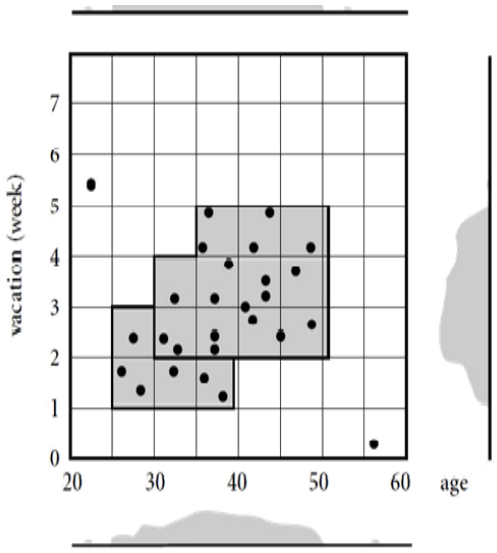

Fig. 2(b): Identification of clusters along (age, vacation) plane

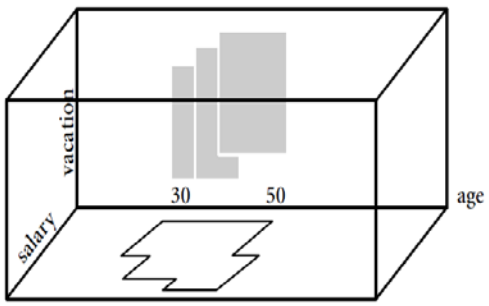

Fig. 2(c): Final clusters in three dimensional (age, salary, vacation) space

The subspaces representing these dense units are to form a candidate search space in which dense units of higher dimensionality may exist. In the second step, CLIQUE generates a minimal description for each cluster as follows. For each cluster, it determines the maximal region that covers the cluster of connected dense units. It then determines a minimal cover (logic description) for each cluster. CLIQUE automatically finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces. It scales linearly with the size of input

and has good scalability as the number of dimensions in the data is increased. However, obtaining meaningful clustering results is dependent on proper tuning of the grid size and the density threshold. This is particularly difficult because the grid size and density threshold are used across all combinations of dimensions in the data set. Thus, the accuracy of the clustering results may be degraded at the expense of the simplicity of the method. Moreover, for a given dense region, all projections of the region onto lower dimensionality subspaces will also be dense. This can result in a large overlap among the reported dense regions. Furthermore, it is difficult to find clusters of rather different density within different dimensional subspaces.

B. Characteristics of CLIQUE

- CLIQUE allows finding clusters of arbitrary shapes.
- CLIQUE is also able to find any number of clusters in any number of dimensions and the number is not predetermined by a parameter.
- Clusters may be found in any subspace means in a single or overlapped subspace.
- The clusters may also overlap each other meaning that instances can belong to more than one cluster [9].

IV. CONCLUSION

Grid based clustering method develops hierarchical Structure out of given data and answer various queries efficiently. STING goes through the database once to compute the statistical parameters of the cells, and hence the time complexity of generating clusters is O(n), where n is the total number of objects. After generating the hierarchical structure, the query processing time is O(g), where g is the total number of grid cells at the lowest level, which is usually much smaller than n. CLIQUE is insensitive to the order of input objects and does not presume any canonical data distribution. It scales linearly with the size of the input and has good scalability as the number of dimensions in the data is increased [10].

REFERENCES

[1] W.,Yang J., Muntz R. STING: A statistical information grid approach to spatial data mining. Proc. 23rd Int. conf. on very large data bases. Morgan Kaufmann, 1997, pp.186-195.

[2] Dr.R.Sabitha, Ms.T.Mythili, Ms.R.D. Priyanka **,** **"**Statistical Information Grid Approach to Segment Large Dataset", International Journal for Research & Development in Technology, Volume-5, Issue- 4 (Apr-16) **.**

[3] Wei Wang, Jiong Yang, and Richard Muntz," STING : A Statistical Information Grid Approach to Spatial Data Mining.

[4] Jaiwei Han and Micheline Kamber, "Datamining: Concepts and Techniques", Morg Kaufman Publishers, 2001.

[5] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications", ACM SIGMOD inter- national conference on Management of data, vol.27, no.2, pp.94-105, June 1998.

[6] Hans-Peter Kriegel and Arthur Zimek, "Subspace clustering, Ensemble clustering, Alternative clustering, Multiview clustering: What can we learn from each other", In Proc. 1st Int'l workshop on discovering, summarizing and using multiple clusterings, 2010.

[7] R.Agrawal, J.Gehrke, D.Gunopulos, P. Rag- havan, "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications", Proceedings of 1998 ACM-SIGMOD, pp. 94-105, 1998.

[8] E.Schikuta,"Grid Clustering: An Efficient Hierarchical Clustering Method for Very Large Data Sets", Proceedings of the 13thInternational Conference on Pattern Recognition,Vol. 2, pp. 101-105, 1996.

[9] Jyoti Yadav, Dharmender Kumar, " Sub space Clustering using CLIQUE: An Exploratory Study", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3, Issue 2, February 2014.

[10] Anne Patrikainen and Marina Meila, "Comparing Subspace Clusterings", IEEE Transactions on knowledge and data engineering, vol.18, no.7, pp.902-916, July 2006.