


Cluster Analysis

- Grid Based Clustering
 - STING
 - CLIQUE

1



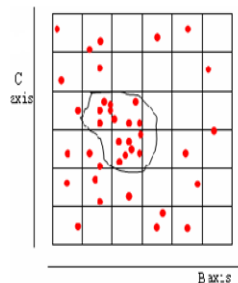
GRID BASED CLUSTERING

- Uses a grid data structure
- Quantizes space into a finite number of cells that form a grid structure
- Several interesting methods
 - **STING** (a S**T**atistical **I**Nformation **G**rid approach) by Wang, Yang and Muntz (1997)
 - **CLIQUE** (C**L**ustering In **Q**U**E**st): Agrawal, et al.

2

GRID BASED CLUSTERING

- Example: We have a set of records and we want to cluster with respect to two attributes, then we divide the related space into a grid structure.



3

GRID BASED CLUSTERING

- Fast processing time:
 - No distance computations
 - Clustering is performed on summaries and not individual objects; complexity is usually $O(\# \text{-populated-grid-cells})$ and not $O(\# \text{objects})$
 - Easy to determine which clusters are neighboring

4

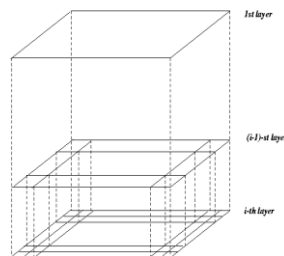
WHAT IS SPATIAL DATA ?

- Spatial data may be thought of as features located on or referenced to the Earth's surface.
- Anything that can be mapped.
- The area that encompasses the locations of all the spatial data is called **spatial area**.

5

STING

- Wang, Yang and Muntz
- It is used for performing clustering on spatial data.
- The spatial area is divided into rectangular cells
- There are several levels of cells corresponding to different levels of resolution



6



STING

- Each cell at a high level is partitioned into a number of smaller cells in the next lower level
- Parameters of higher level cells are computed from those at lower levels.
- Statistical attributes are stored in cell.
 - Count, Mean, Maximum, Minimum

7



STING

Used a top-down approach to answer spatial data queries:

- Start from a pre-selected layer—typically with a small number of cells
- For each cell in the current level compute the confidence interval indicating a cell's relevance to a given query.
- The confidence interval is calculated by using the statistical parameters of each cells.

8

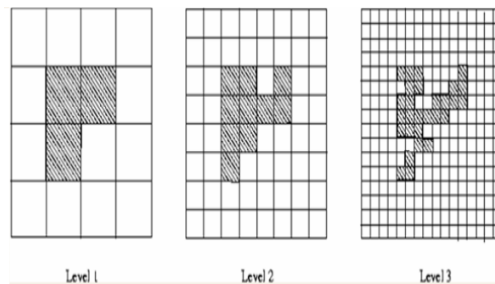
STING

- Remove irrelevant cells from further considerations.
- When finished the current layer, proceed to next lower level.
- Processing of the next lower level examines only the remaining relevant cells.
- Repeat this process until the bottom layer is reached.
- Return the regions of relevant cells that satisfy the query.

9

STING

- Different grid levels during query processing



10



STING

- Advantages:
 - Very efficient.
 - $O(K)$, where K is the number of grid cells at the lowest level
- Disadvantages:
 - All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected

11



SAMPLE QUERY EXAMPLES

- Select the maximal regions that have at least 100 houses per unit area and at least 70% of the house prices are above \$400K.
- Select the range of age of house houses in those maximal regions where there are at least 100 houses per unit area and at least 70% of the houses have price between \$150K and \$300K with area at least 100 units in California.

12



CLIQUE

- Agrawal, Gehrke, Gunopulos, Raghavan
- **Grid-based:** It discretizes the data space through a grid and estimates the density by counting the number of points in a grid cell.
- **Density-based:** A cluster is a maximal set of connected dense units in a subspace.
- CLIQUE identifies the dense units in the subspaces of high dimensional data space and uses these subspaces to provide more efficient clustering.

13



CLIQUE

CLIQUE can be considered as both density-based and grid-based.

- It partitions each dimension into the same number of equal length interval.
- It partitions an m-dimensional data space into non-overlapping rectangular units.
- Identify the subspaces that contain clusters using the Apriori principle.

14

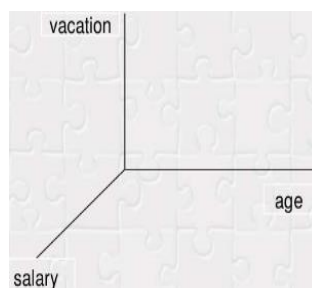
CLIQUE

- Identify clusters:
 - Determine dense units in all subspaces of interests
 - Determine connected dense units in all subspaces of interests.
- Generate minimal description for the clusters
 - Determine maximal regions that cover a cluster of connected dense units for each cluster
 - Determination of minimal cover for each cluster

15

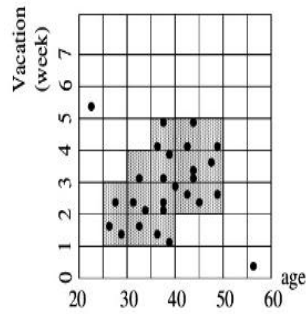
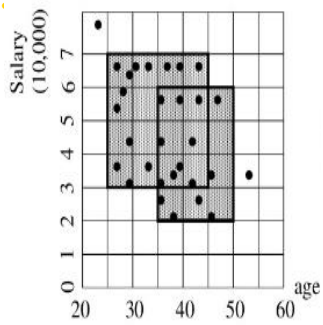
CLIQUE

- Records that have three attributes: salary, vacation and age.
- The data space for the this data would be 3-dimensional.



16

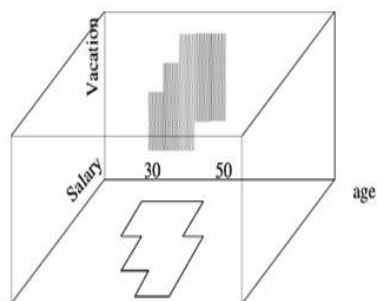
CLIQUE



17

CLIQUE

- Try to visualize the dense units of the two planes on the following 3-d figure:



18



CLIQUE

- Strength
 - Automatically finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces
 - Insensitive to the order of records in input and does not presume some canonical data distribution
 - Scales linearly with the size of input and has good scalability as the number of dimensions in the data increases
- Weakness
 - The accuracy of the clustering result may be degraded at the expense of simplicity of the method