Linear Clustering of Objects with Multiple Attributes

H V Jagadish

Abhishek Srivastava
Student ID: 861307778

February 7, 2017
CS 236, Winter 2017

The problem:
The paper notes that multi-dimensional data are often converted to one-dimensional space but the existing methods does not support variety of queries in optimal way and take many disc blocks access to compute the queries.

The contribution:
The author propose "Hilbert Curve" a method of converting multi-dimensional data such as *spatial data* to one-dimensional or linear space which can preserves the "locality" of neighboring points as much as possible, it can be achieved by following snake scan pattern but avoiding as many jump as possible.

The method:
There are multiple ways of converting multi-dimensional data to one-dimension. Author presented few of these methods and has compared them:

- Column Scan : scan single column at a time.
- Column-wise Snake Scan : reverse direction of scan on alternate columns.
- Z-curve : interleaving binary bits on different axises.
- Gray-Code mapping : similar to z-curve but gray code to obtain 1-D coordinates.
- Hilbert Curve : an extension of Peano Curve.

Goodness of different space filling curves are measured by 2 types of queries: **Partial Exact Match Selection** which is selection done using exact values on some but not all attributes and **Range Queries** which is selection on ranges of some may be all attributes. Other criteria is cost measurement which is Number of Disk blocks fetched, Number of non-consecutive disk blocks fetched and Size of linear span for given selection. Author also proposes the extension of Hilbert curve from 2-D to 3-D and so on.

For Patial match queries snake scan, Gray code and Hilbert curve perform best, z-curve performs significantly worse. For Range selection queries Author says that simple column scan can also yield best result along with Hilbert curve, Gray code and z-curve didn't performed good comparatively.

Cost measurement of different mapping algorithm is also performed. Gray and Hilbert performs better in number of blocks fetched as block size is increased but column scan become worse as block size is increased. Gray and Hilbert also performed better in fetching few non-consecutive blocks but all perform same if block size gets very large. For number of hits per block Hilbert performs the best, Gray and z-curve were also respectively good.

Comments:
The paper presents space filling curve to map multi-dimensional data to one dimension while performing better or same in multiple aspects than other existing mapping algorithm. However, it also gives rise to a few problems:

- With increase in the number of dimensions visualization of hilbert curve becomes less intuitive.
- With increase in dimension the space between neighboring points also grows so it becomes harder to maintain locality and loose less information using Hilbert curve i.e curse of dimensionality.