

The Igrid Index: Reversing the Dimensionality Curse for Similarity Indexing in High Dimensional Space

Authors:
C. Aggarwal, IBM
Phillip Yu, IBM

Overview

- Introduction the Problem
- Key Contribution
- Relevant Prior Work
- Methodology
- Results
- Our opinion of the paper

Problem

- Similarity Searching in high dimensions (lots of features) is usually not so easy.
- Two real problems
 - Takes a lot of time in high dimensional space
 - Typical similarity measures are not so great, one large difference in one feature can skew result

Key Contributions

- New Similarity Measure
- Definition of a “new” access structure for computing similarity in high dimensions.

Related Work

- Spatial based access structures
 - R-Trees
 - KDB-Trees
- Problem is that you often have to visit EVERY node in these trees to do a similarity calculation.

New Similarity Measure

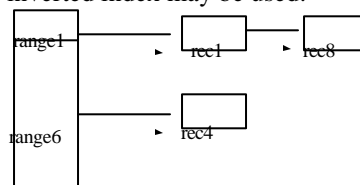
- $IDIST(X, Y) =$
- $SUM(1 - (|x_i - y_i| / (u_i - l_i))^p)^{1/p}$
- Typical measure: Minkowski (Eq. 8.7)
- $SUM(|x_i - y_i|^p)^{1/p}$
- New measure is skewed for u(upper) and l (lower) bounds.

Use Equidepth Ranges

- Now partition each attribute into equidepth ranges. Each range has approximately same number of values in it. Ranges are not same size.
- Now compute similarity measure as sum across all ranges for all dimensions. With lower and upper bounds on each range.

Index structure

- Many ranges will have no values in the range. To efficiently represent this an inverted index may be used.



To compute similarity

- For a given record, compute the range for each dimension.
- Now use the inverted index to access all records that have a value in that range.

Results

- MUCH faster than other approaches. See graph.