Estimating Join Selectivities using Bandwidth-Optimized Kernel Density Model

M. Kiefer, M. Heimel, S. Breß, V. Markl

Problem & Motivation

- Task: Estimating selectivities for queries of the form $|\sigma_{c_1}(R_1) \bowtie R_1.A = R_2.A \sigma_{c_2}(R_2)|$
- 1D statistics + Independence Assumption (IA) are commonly used in practice
  - But IA is often violated in real-world data
  - Potentially results in suboptimal query plans
- Lifting IA requires multidimensional statistics that are
  - accurate
  - efficiently computable
  - easy to maintain and construct
- Prior approaches do not provide all these characteristics
- Generalization to n-way joins and selections on join predicate are covered in the paper

Method 1: Join Model

- KDE model constructed from the join result
  - Sampling from the join result
  - Requires exact or a good estimate of the join size
- Joins: Handled implicitly by sampling
- Selections: Handled explicitly during estimation

Method 2: Table Model

- Estimates are computed by combining base table KDE models
- Joining the estimated distributions
- Joins: Handled explicitly during estimation
- Selections: Handled explicitly during estimation

Table Model: Pruning Techniques

- Sample Pruning
  - Computes invariant contributions for every sample point
  - Removes sample points with negligible contributions
  - Reduces the number of input tuples to the cross product
- Cross Pruning
  - Computes cross contribution only for sufficiently close points
  - For each sample point in Sample 1
    - a binary search locates sufficiently close tuples in Sample 2
  - Join with range predicate instead of cross product

Evaluation

- Baselines
  - PostgreSQL: 1D Statistics + Independence Assumption
  - AGMS: Sketch-based approach
  - Table/Join Sample: Uniform sample evaluation
  - Correlated Sample: Nonuniform sample evaluation
- KDE-based estimators trained on 100 workload queries
- KDE-based estimators
  - ... outperform plain samples for smaller models
  - ... converge with plain samples for larger models
  - ... outperform other baselines in most cases

KDE for Selectivity Estimation

- Kernel Density Estimation (KDE)
  - Multivariate probability density estimation
  - Based on a uniform data sample
  - Smoothing by centering kernel functions on sample points
  - Smoothing controlled by bandwidth parameter
- KDE has been applied to range filters over base tables
  - Good accuracy
  - Hybrid between sampling and histograms
  - Bandwidth selection based on query feedback
  - Learning estimator
  - Efficiently trainable and evaluable
  - Suitable for GPU acceleration
- In this publication: Extension to joins that are subject to selections

Sample & Data on GitHub: martinkiefer/join-kde

Contact: martin.kiefer@tu-berlin.de

Funding

The work received funding through the EU Horizon2020 project SAGE (671500) and from the German Ministry for Education and Research as Berlin Big Data Center BBDC (G1814013A) and Software Campus (01IS17052).