Entropy-based Histograms

Hien To, Kuorong Chiang, Cyrus Shahabi

Motivation

Query optimization plays an important role in a database management system (DBMS).

The optimal plan

Multiple execution plans can be generated for a particular query. Accurate estimations are crucial to generate optimal execution plans.

Abstract

Histograms have been extensively used for selectivity estimation by academics and have successfully been adopted by database industry. However, the estimation error is usually large for skewed distributions and biased attributes, which is typical with real-world datasets. In this paper, we therefore propose effective models to measure bias and selectivity based on information entropy. These models together with the principles of maximum entropy are then used to develop a class of entropy-based histograms. In addition, taking ad-vantage of the fact that entropy can be computed incrementally, we present incremental variations of our algorithms that reduce the complexities of entropy-based histograms from O(N^2) to O(NlogB), where N is the number of distinct values and B is the number of histogram buckets. We conducted numerous experiments with both synthetic and real-world datasets to compare the accuracy and efficiency of our proposed techniques with many other histogram-based techniques, showing the best overall performance of our entropy-based approaches for both equality and range queries.

Background

Histogram construction

Entropy of a bucket

Algorithm 1: Maximum Entropy (ME)

Algorithm 2: Minimum Selectivity Error (MSE)

Algorithm 3: Maximum Reduction in Bias (MRB)

Performance Evaluation

Construction cost

Table 2: Construction complexity

Table 3: Construction times in seconds (uniform.zip dataset)

Estimation error

Figure 11: Average estimation errors of equality query over all synthetic datasets

Figure 12: Average estimation errors of range query over all synthetic datasets

In sum, a good histogram for equality queries does not necessarily performs well for range queries and vice versa. However; our entropy-based histograms IMSE and IME are generally the best histograms for equality and range queries, respectively.