Entropy-based Histograms

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 advantage 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(Nlog(B)), 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.