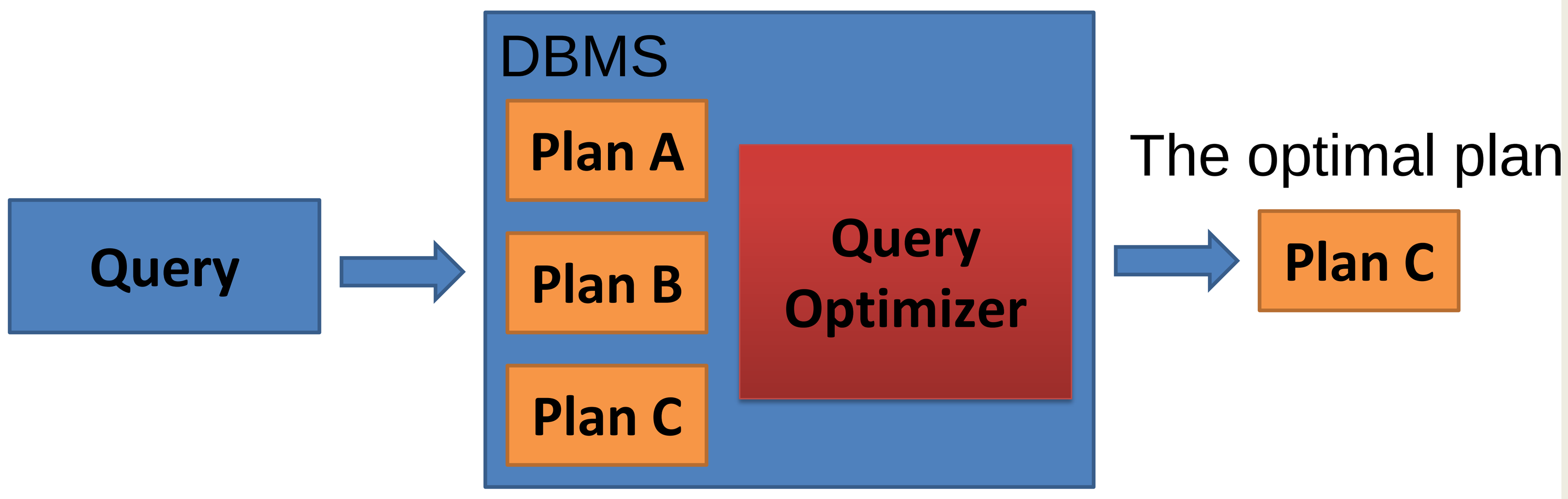
Entropy-based Histograms

Hien To, Kuorong Chiang, Cyrus Shahabi

Motivation

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



Multiple execution plans can be generated for a particular query

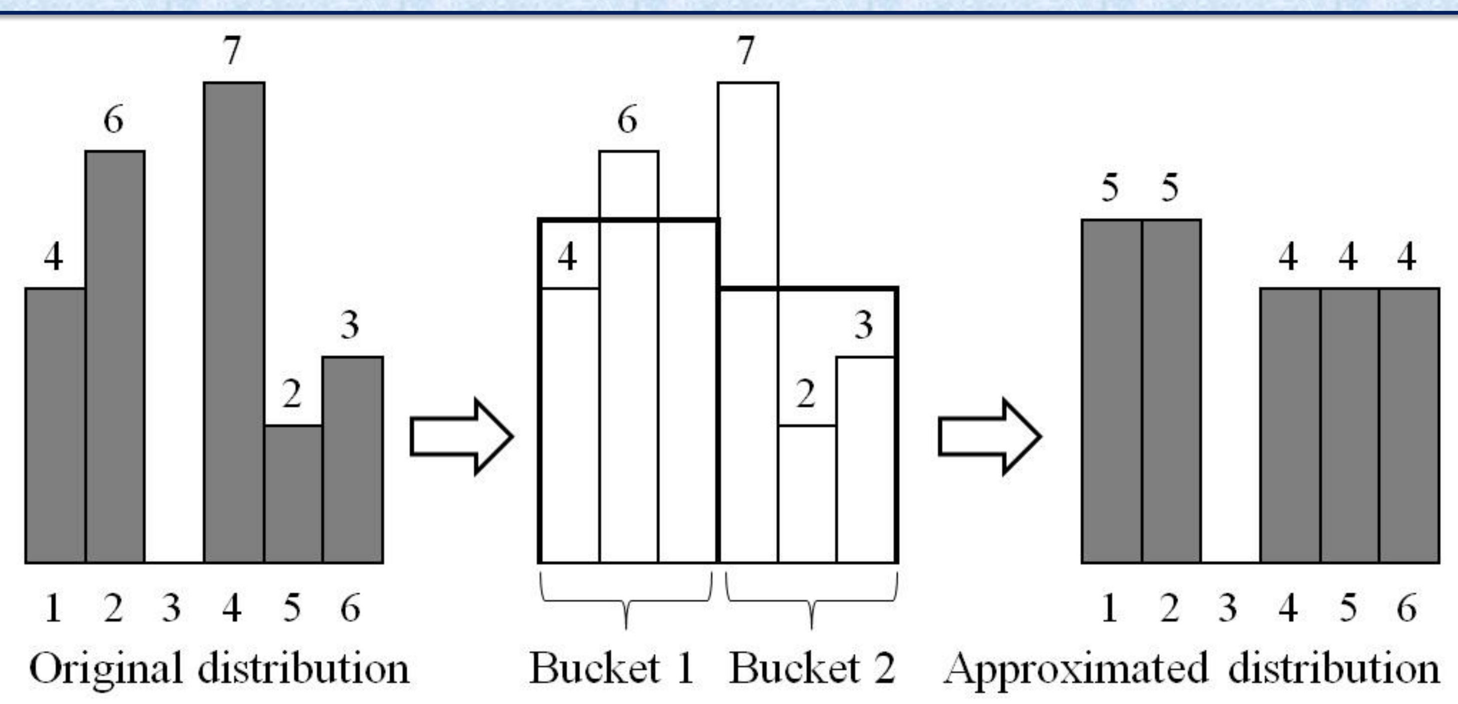
Accurate estimations are crucial to generate optimal execution plans

Abstract

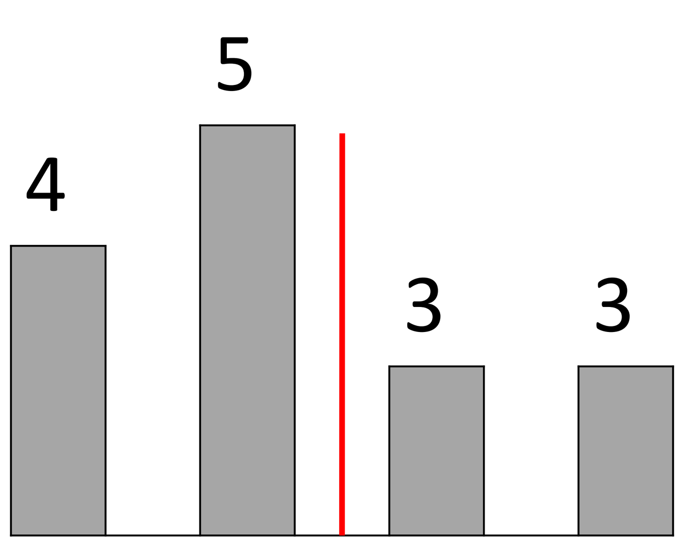
Histograms have been extensively used for selectivity estimation by academics and have successfully been adopted by database industry. However, the estimation error is usually large for skewed distributions and biased attributes, which is typical with real-world datasets. In this paper, we therefore 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



$$H(X) = -\sum_{i=1}^{N} p_i(x) log(p_i(x))$$

1. Group similar frequencies in the same bucket
2. Maximize information content

Entropy-based Histograms

Algorithm 1: Maximum Entropy (ME)

- Maximize total entropy
- Approximated algorithms
  - Maximum entropy (ME)
  - Incremental maximum entropy (IME)

$$\sum_{i=1}^{B} W(b_i)H(b_i)$$

Algorithm 2: Minimum Selectivity Error (MSE)

Selectivity of an equality query: $s_i = \dfrac{1}{dv_i}$ vs $s = 2^{-H}$

- Minimizes the total mean squared error
- Approximated algorithm
  - Minimum selectivity error (MSE)
  - Incremental minimum selectivity error (IMSE)

$$\sum_{i=1}^{B} W(b_i)E(b_i)$$

Algorithm 3: Maximum Reduction in Bias (MRB)

- Minimizes the total weighted bias
- Approximated algorithm
  - Maximum reduction in bias (MRB)
  - Incremental maximum reduction in bias(IMRB)

$$\sum_{i=1}^{B} W(b_i)BF(b_i)$$

Performance Evaluation

Construction cost

Method	**Time**	**Ref**
VODP	$O(N^2B)$	[14]
ME, **MSE**, **MRB**	$O(N^2)$	this
MHIST	$O(B(N+logB))$	[14]
VOII	$O(NB)$	[21]
IME, **IMSE**, **IMRB**	$O(NB)$	this
MD	$O(NlogB)$	[21]
EH	$O(N)$	[20]
EW	$O(B)$	[20]

Table 2: Construction complexity

Algorithm	**Input size (N)**		
	1000	**10000**	**20000**
VODP	235	2325	5423
MSE	1.034	132	543
MRB	0.845	114	435
ME	0.755	75	303
MHIST	0.547	49	211
VOII	0.047	2.376	8.885
IMSE	0.078	0.453	0.855
IMRB	0.062	0.28	0.553
IME	0.047	0.265	0.395
MD	0.015	0.015	0.025
EH	0	0.001	0.001
EW	0	0.001	0.001

Table 3: Construction times in seconds (uniform_zipf dataset)

Estimation error

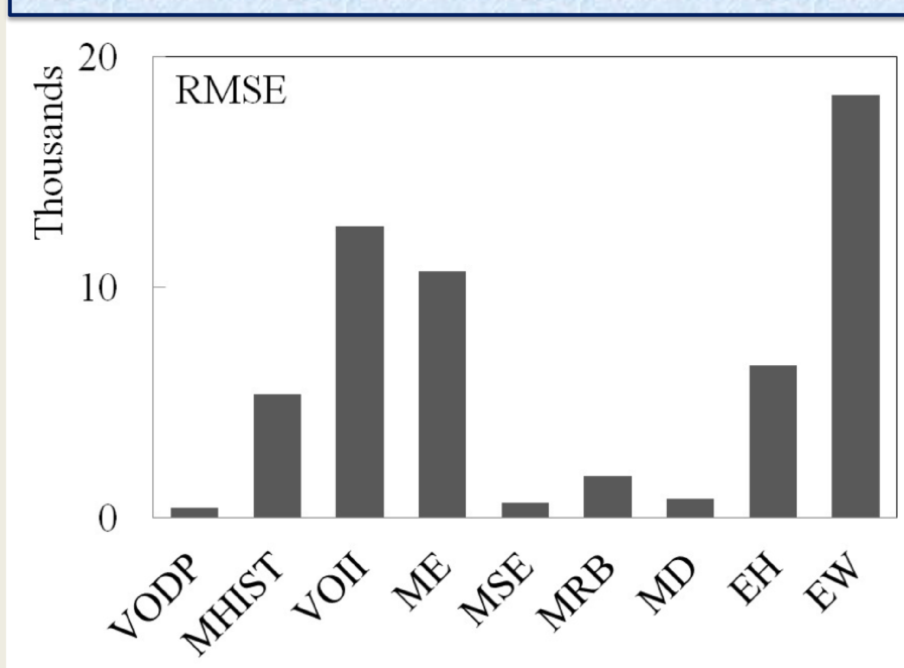


Figure 11: Average estimation errors of equality query over all synthetic datasets

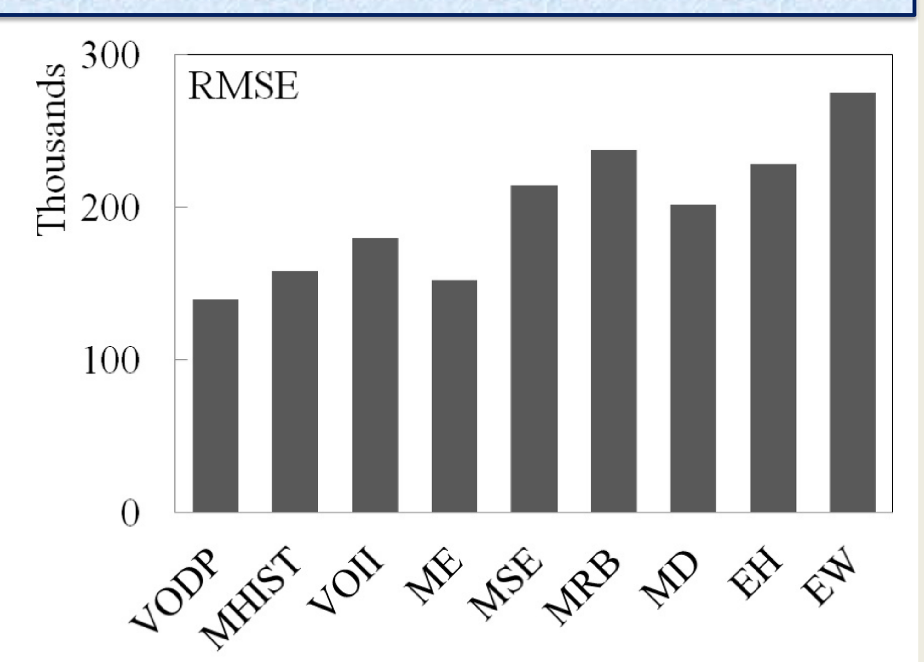


Figure 12: Average estimation errors of range query over all synthetic datasets

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.