COL 362 & COL 632

Query Optimization

01 Apr 2023

Statistical Information for Cost Estimation

Analyze
Vacuum

- n_r: number of tuples in a relation *r.*
- b_r: number of blocks containing tuples of *r.*
- l_r: size (length) of a tuple of *r.*
- f_r: blocking factor of *r* — i.e., the number of tuples of *r* that fit into one block.
- If tuples of *r* are stored together physically in a file, then:

$$\boldsymbol{b_r} = \left\lceil \frac{\boldsymbol{n_r}}{\boldsymbol{f_r}} \right\rceil$$

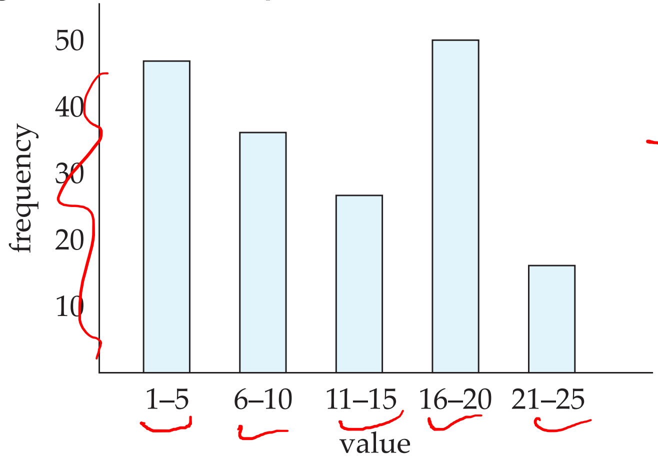
- $\boldsymbol{V(A, r)}$: number of distinct values that appear in *r* for attribute *A;* same as the size of $\boldsymbol{\Pi_A(r)}$

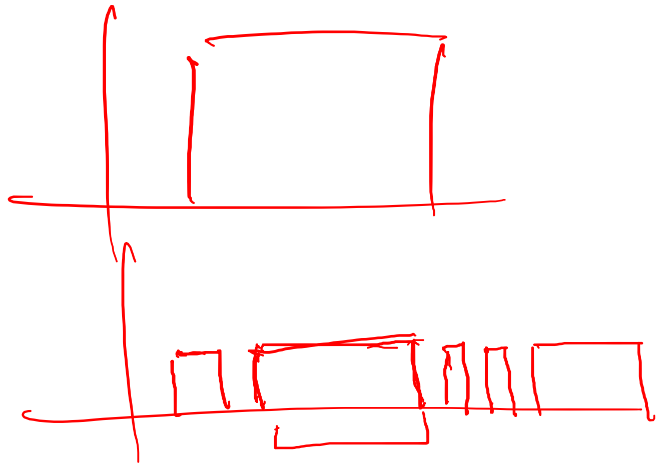
Histograms

frequent values

optimal histogram

- Histogram on attribute *age* of relation *person*

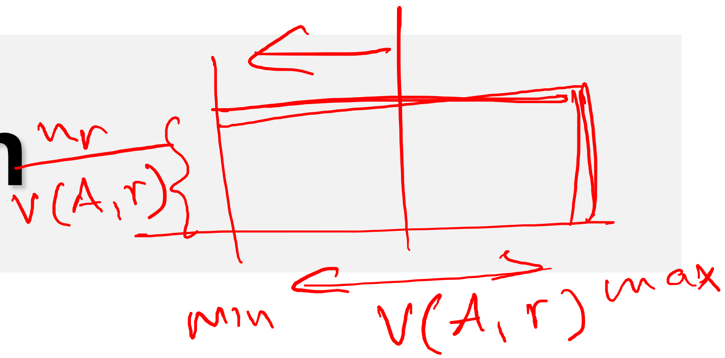




- **Equi-width** histograms
- **Equi-depth** histograms break up range such that each range has (approximately) the same number of tuples
- Many databases also store **n** most-frequent values and their counts
- Histogram is built on remaining values only

piecewise linear, or spline.

Selection Size Estimation



- $\sigma_{A=v}(r)$
  - $\frac{n_r}{V(A,r)}$: number of records that will satisfy the selection
  - If the equality condition on a key attribute: *size estimate* = 1
- $\sigma_{A \le v}(r)$
  - Let c denote the estimated number of tuples satisfying the condition
  - If $\min(A,r)$ and $\max(A,r)$ are available in catalog
  - $c = \begin{cases} 0, & v < \min(A,r) \\ n_r \left(\frac{v - \min(A,r)}{\max(A,r) - \min(A,r)} \right) \end{cases}$
  - **If histograms available**, we can refine above estimate
  - In absence of statistical information *c* is assumed to be $\frac{n_r}{2}$

(case of $\sigma_{A \ge v}(r)$ is symmetric)

$\Rightarrow$ for each value $V(A,r)$: $f(V(A,r))$

$\Rightarrow$ 10%-le chunks

$\Rightarrow$ mean, mode, median

$\underline{A} = v_1$ AND $\underline{B} = v_2$

Size Estimation of Complex Selections

- The **selectivity of a condition** θ_i is the probability that a tuple in the relation *r* satisfies θ_i .
  If s_i is the number of satisfying tuples in *r*, the selectivity of θ_i is given by s_i / n_r.
- Conjunction: $\sigma_{\theta_1 \wedge \theta_2 \wedge \cdots \wedge \theta_n}(r)$ *Assuming independence*, estimate of tuples in the result is:

$$\frac{n_r (s_1 \times s_2 \times \cdots \times s_n)}{n_r^n}$$

$$n_r \left(\frac{s_1}{n_r} \cdot \frac{s_2}{n_r} \cdots \right)$$

- Disjunction: $\sigma_{\theta_1 \vee \theta_2 \vee \cdots \vee \theta_n}(r)$ Estimated number of tuples (again assuming independence):

$$n_r \left(1 - \left(\left(1 - \frac{s_1}{n_r}\right) \times \left(1 - \frac{s_2}{n_r}\right) \times \cdots \times \left(1 - \frac{s_n}{n_r}\right) \right) \right)$$

- Negation: $\sigma_{\neg\theta}(r)$

Estimation of the Size of Joins

r (R)
s (S)

- The Cartesian product $r \times s$ contains $n_r \cdot n_s$ tuples; each tuple occupies $s_r + ss$ bytes.
- If $R \cap S = \varnothing$, then $r \bowtie s$ is the same as $r \times s$
- If $R \cap S$ is a key for R, then a tuple of s will join with at most one tuple from r n_s
  - the number of tuples in $r \bowtie s$ is no greater than the number of tuples in s
- If $R \cap S$ *in* S is a foreign key in S referencing R, then the number of tuples in $r \bowtie s$ is the same as the number of tuples in s.
  - The case for $R \cap S$ being a foreign key referencing S is symmetric.

Estimation of the Size of Joins

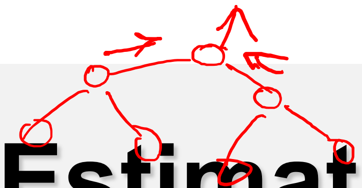
two relations
r, s.
$V(A, s)$
distinct A in s

- If $R \cap S = \{A\}$ is not a key for R or S

  If we assume that every tuple *t* in *r* produces tuples in $r \bowtie s$, the number of tuples in $r \bowtie s$ is estimated to be: $\frac{n_r \cdot n_s}{V(A,s)}$

  If the reverse is true, the estimate obtained will be: $\frac{n_r \cdot n_s}{V(A,r)}$

  The lower of these two estimates is probably the more accurate one.
- Can improve on above if histograms are available
  - Use formula similar to above, for each cell of histograms on the two relations

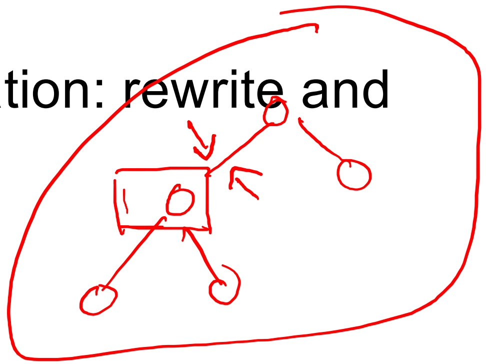


Size Estimation for Other Operations

- Projection: estimated size of $\prod_A(r) = V(A,r)$
- Aggregation : estimated size of ${}_G\gamma_A(r) = V(A,r)$

Accurate values

- Set operations
  - For unions/intersections of selections on the same relation: rewrite and use size estimate for selections
    - E.g. $\sigma_{\theta 1}(r) \cup \sigma_{\theta 2}(r)$ can be rewritten as $\sigma_{\theta 1 \vee \theta 2}(r)$
  - For operations on different relations:
    - estimated size of $r \cup s$ = size of r + size of s.
    - estimated size of $r \cap s$ = minimum size of r and size of s.
    - estimated size of $r - s$ = r.
    - All the three estimates may be quite inaccurate, but provide upper bounds on the sizes.



Optimization of Updates

- **Halloween problem** (https://en.wikipedia.org/wiki/Halloween_Problem)

```
update R set A = 5 * A
where A > 10
```

- If index on A is used to find tuples satisfying A > 10, and tuples updated immediately, same tuple may be found (and updated) multiple times
- Solution 1: *Always defer updates*
  - collect the updates (old and new values of tuples) and update relation and indices in second pass
  - Drawback: extra overhead even if e.g. update is only on R.B, not on attributes in selection condition
- Solution 2: *Defer only if required*
  - Perform immediate update if update does not affect attributes in where clause, and deferred updates otherwise.