Query Optimization: Exercise

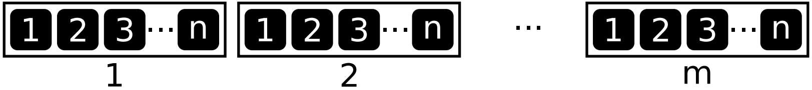
Session 13

Bernhard Radke

January 28, 2019

Direct, Uniform, Distinct: Yao

Given m pages with n tuples on each page, e.g. a total of $N = m \cdot n$ tuples:



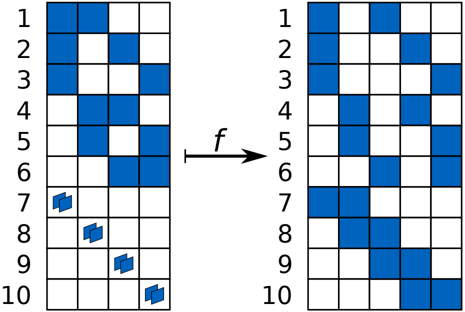
- How many distinct subsets of size k exist? $\binom{N}{k}$
- How many distinct subsets of size k exist, where a page does not contain any of the chosen tuples? Choose k from all but one page, i.e. from $N - n$ tuples: $\binom{N-n}{k}$
  So the probability that a page contains none of the k tuples is

$$p := \frac{\binom{N-n}{k}}{\binom{N}{k}}$$

- What is the probability that a certains page contains at least one tuple? $1 - p \ldots$ unless all pages have to be involved $(k > N - n)$.
- Multiplied by the number of pages, we get the number of qualifying pages, denoted $\overline{\mathcal{Y}}_n^{N,m}(k)$.

Direct, Uniform, Non-Distinct: Cheung

- Now *with replacement*: How many distinct *multisets* exist chosing k from n?
  As many as there are distinct sets chosing k from $n + k - 1$!
- Bijection between multisets and sets. From multiset to set:
  $f : (x_1, x_2, \ldots, x_k) \mapsto (x_1 + 0, x_2 + 1, \ldots, x_k + (k - 1))$
- Example: Choose 2 from 4
  - \# sets: $\binom{4}{2}$
  - \# multisets: $\binom{4+2-1}{2}$



- Like Yao, but not necessarily distinct
- Same formula as Yao, but:
  - No special case for $k > N - n$
  - We substitue N by $N + k - 1$ to compute $\tilde{p}$

Sequential, Uniform, Distinct

- Estimate the distribution of distance between two qualifying tuples
- Bitvector B, b bits are set to 1
- First, the distribution of the number of j zeros
  - before first 1
  - between two consecutive 1s
  - after last 1
- Bitvectors having a 1 at position i followed by j zeros: $\binom{B-j-2}{b-2}$
- $B-j-1$ positions for i
- every bitvector has $b-1$ sequences of a form $10\ldots01$
- $\mathcal{B}_b^B(j) = \frac{(B-j-1)\binom{B-j-2}{b-2}}{(b-1)\binom{B}{b}} = \frac{\binom{B-j-1}{b-1}}{\binom{B}{b}}$
- now, the expected number of 0s: $\frac{B-b}{b+1}$
- then, the expected total number of bits between first bit and the last 1: $B - \frac{B-b}{b+1} = \frac{Bb+b}{b+1}$

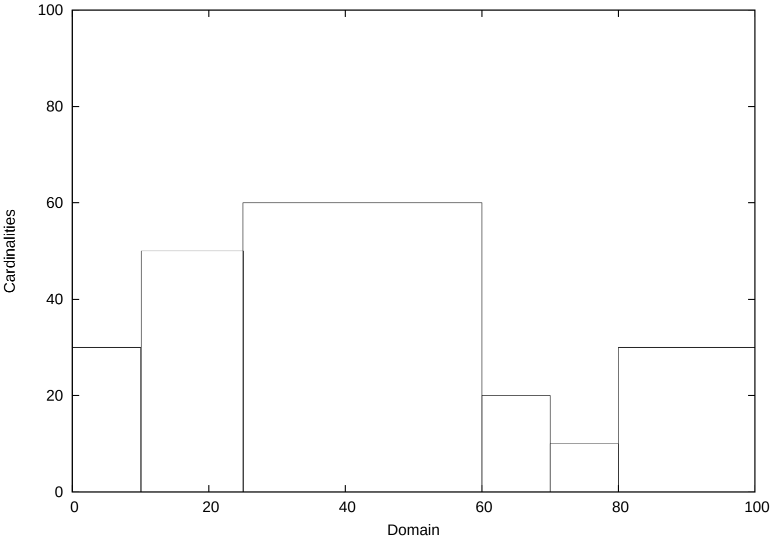
Histograms

A histogram $H_A : B \to \mathbb{N}$ over a relation R partitions the domain of the aggregated attribute A into disjoint buckets B, such that

$$H_A(b) = |\{r | r \in R \wedge R.A \in b\}|$$

and thus $\sum_{b \in B} H_A(b) = |R|$.

A rough histogram might look like this:



Given a histogram, we can approximate selectivities as follows:

$$A = c \qquad \frac{\sum_{b \in B : c \in b} H_A(b)}{\sum_{b \in B} H_A(b)}$$

$$A > c \qquad \frac{\sum_{b \in B : c \in b} \frac{\max(b) - c}{\max(b) - \min(b)} H_A(b) + \sum_{b \in B : \min(b) > c} H_A(b)}{\sum_{b \in B} H_A(b)}$$

$$A_1 = A_2 \qquad \frac{\sum_{b_1 \in B_1, b_2 \in B_2, b' = b_1 \cap b_2 : b' \neq \emptyset} \frac{\max(b') - \min(b')}{\max(b_1) - \min(b_1)} H_{A_1}(b_1) \frac{\max(b') - \min(b')}{\max(b_2) - \min(b_2)} H_{A_2}(b_2)}{\sum_{b_1 \in B_1} H_{A_1}(b_1) \sum_{b_2 \in B_2} H_{A_2}(b_2)}$$

Given the following histogram of an integer attribute $R.a$:

bucket	$[0, 20)$	$[20, 40)$	$[40, 60)$	$[60, 80)$	$[80, 100)$
count	1	3	4	2	0

Estimate the number of elements for which $R.a >= 55$ holds true.

- Slides: db.in.tum.de/teaching/ws1819/queryopt
- Exercise task: gitlab
- Questions, Comments, etc:
  mattermost @ mattermost.db.in.tum.de/qo18
- Bonus sheet due: 9 AM next monday