Query Optimization: Exercise

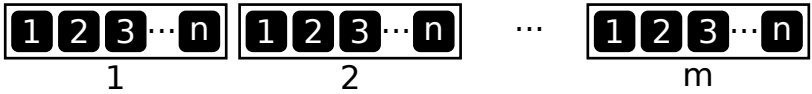
Session 12

Bernhard Radke

January 21, 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 certain page contains at least one tuple? $1 - p$... unless all pages have to be involved ($k > N - n$).
- Multiplied by the number of pages, we get the number of qualifying pages, denoted $\overline{\mathcal{Y}}_n^{N,m}(k)$.

Let $m = 50,\ n = 1000 \Rightarrow N = 50k,\ k = 100$

Yao (exact) : $p = \frac{\binom{N-n}{k}}{\binom{N}{k}} = \prod_{i=0}^{k-1} \frac{N-n-i}{N-i} = \prod_{i=0}^{99} \frac{49k-i}{50k-i} = 13.2\%$

Waters : $p \approx (1 - \frac{k}{N})^n \approx 13.5\%$

- Given a relation with 3 pages and two tuples per page, compute the average number of accessed pages when reading 2 tuples.

$$\overline{\mathcal{Y}}_n^{N,m}(k) = \overline{\mathcal{Y}}_2^{6,3}(2) = 3 \cdot (1 - \frac{\binom{6-2}{2}}{\binom{6}{2}}) = 1.8$$

- Given a relation with 100 pages with 10 tuples each, plot the expected number of accessed pages when reading 1, 2, ..., 1000 tuples using Yao's formula. Also plot the approximations of Bernstein et al. and Waters.

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