Bifocal Sampling for Skew-Resistant Join Size Estimation

Sumit Ganguly, Phillip B. Gibbons, Yossi Matias, Avi Silberschatz

Outline

- Problem Definition
- Algorithm Overview
- Sparse-Any Estimation Procedure
- Proof of Sparse-Any Estimator
- Probabilistic Elimination
- Cost analysis

Problem Definition

Estimate the size of an equi-join between two relations.
Useful for:

- Compare costs of alternate join plans
- Distribute load in a distributed environment
- Sometimes the count is interesting: Financial audits and statistical studies

Algorithm Bifocal Sampling

Algorithm Bifocal-Sampling(n, m_1, m_2, δ)

$//n$ is the number of tuples in each relation.
$//m_1$ is the sample size for the 1st procedure.
$//m_2$ is the sample size for the 2nd procedure.
$//\delta$ is a threshold used in the 1st procedure.
$\bigcup := NONE;$
$\hat{A}_d := Dense\_dense\_Estimation(n, m_1, m_2, \delta);$
$\hat{S}_{s_1} := Sparse\_Any\_Estimation(R, S, n, m_2);$
$\hat{S}_{s_1} := Sparse\_Any\_Estimation(S, R, n, m_2);$
$\hat{A} := \hat{A}_d + \hat{A}_{s_1} + \hat{A}_{s_2};$
if $\hat{A} < nlgn$ then $\bigcup := nlgn;$
return $(\hat{A}, \bigcup);$

For trials used: $m_1 = (\sqrt{n} + lg\ n) * lg\ n$, $m_2 = \sqrt{n} + lg\ n$, $\delta = lg\ n$

Sparse-Any Estimation Procedure

Sparse_Any_Estimation(R, S, n, m_2)

// n is the number of tuples in each relation.

// m_2 is the sample size of each relation.

$A^* := 0$;

Select m_2 random tuples from S, and let S^* be the resulting random sample;

//If determining mult$_R(v)$ is slow then use probabilistic elimination to suppress dense v;

if avoid_dense_mult then begin

Select m_2 random tuples from R, and let R^* be the resulting random sample;

For each join attribute value v appearing in R^* do

Remove from S^* all tuples with value v;

end;

//Compute the estimate on the (remaining) values that are sparse in R:

For each tuple τ in S^* do begin

Determine the number, x, of tuples in R that join with τ;

if $x < n/m_2$ then A^* := $A^* + x$;

end;

$\hat{A}_s := nA * /m_2;$

return $\hat{A}_s$;

Sparse-Any Estimation Lemma

- Analyzing for $m_2 = \sqrt{n}$
- $\hat{S}^*$ is the set of the remaining values in S^* after (optional) probabilistic elimination.
- Let A_S be the number of pairs in either sparse-sparse or sparse-dense SubJoins.
- Compute $\hat{A}_S$ as an estimate for A_S.

Lemma $E(\hat{A}_S) = A_S$, and with high probability $\hat{A}_S = \theta(A_S) + O(n\ lg\ n)$. Also, The probe cost is $O(\sqrt{n})$.

Sparse-Any Estimation Proof (sketch)

Let $\varepsilon(\tau)$ be the contribution of tuple τ in S; i.e., $\hat{A}_S = \Sigma_{\tau \in S} \varepsilon(\tau)$:

$$\varepsilon(\tau) = \begin{cases} \frac{n}{m_2} * \text{mult}_R(v) & \text{if } \tau \in \hat{S}^* \wedge \text{mult}_R(v) < \frac{n}{m_2} \\ 0 & \text{otherwise} \end{cases}$$

If the value of a tuple $\tau \in S$ is dense in R then $\varepsilon(\tau) = 0$.

Consider tuple $\tau \in S$, $\tau \in$ SubJoin(v), such that $\text{mult}_R(v) < n/m_2$. Since $Pr(\tau \in \hat{S}^*) = m_2/n$, we have $E(\varepsilon(\tau)) = \frac{m_2}{n} * \frac{n}{m_2} * \text{mult}_R(v) = \text{mult}_R(v)$

Probe Cost $= m_2 = \sqrt{n}$

Lemma $E(\hat{A}_S) = A_S$, and with high probability $\hat{A}_S = \theta(A_S) + O(n\ lg\ n)$. Also, The probe cost is $O(\sqrt{n})$.

Probabilistic Elimination

Definition Let m be a positive integer. A join attribute value v is defined to be m-dense in R if mult$_R \geq n/m$, and otherwise defined to be m-sparse in R. An (m-dense) join attribute value v is defined to be very m-dense in R if mult$_R(v) \geq 2n\ lg\ n/m$.

Show sample size $m_2 = \sqrt{n}$ satisfies the following requirements:

(i) With high probability, all tuples whose value is very dense in R are eliminated.

- Enables us to assume that the maximum contribution of a single tuple is bounded by $2n\ lg\ n/m$. This is useful for bounding the tuple cost.

(ii) Let τ be a tuple in S with join attribute value v such that SubJoin(v) is of type sparse-sparse or sparse-dense, and not eliminated by R^*. The expected contribution of each such τ to the total estimate is within a small error of mult$_R$(v).

- Guarantees that the expected total contribution of all sparse-sparse and sparse-dense SubJoins is within a small error of their actual total sum of sizes.

(iii) Some dense tuples in R might not be eliminated by R^*. Nevertheless, each dense tuple in R has constant expected contribution to the total tuple cost.

- Guarantees that tuples that are not sparse in R (but are not necessarily very dense) would not significantly affect the expected total tuple cost.

Probabilistic Elimination Lemma

Let C be the total tuple cost. $C(\tau)$ be the tuple cost resulting from tuple τ.

Let n_d be the number of tuples in S whose value is dense in R.

Lemma We have $A_S/e \leq E(\hat{A}_S) \leq A_S$, and with high probability, $\hat{A}_S = \Theta(A_S) + O(n\ lg\ n)$. Also, the expected tuple cost, $E(C)$, is at most $n_d/e + A_S/\sqrt{n}$.

Prob. Elim. Lemma Proof (sketch)

Consider tuple $\tau \in S^*$ with join attribute value v.

$$
\begin{aligned}
Pr(& \tau \text{ is not eliminated from } S^*) \\
& = (1 - \text{mult}_R(v)/n)^m \approx e^{-(m/n)\text{mult}_R(v)},
\end{aligned}
$$

assuming (without loss of generality) that mult$_R(v) \leq n/10$.
Thus, since $Pr(\tau \in S^*) = m/n$,

$$
\begin{aligned}
Pr(\tau \in \hat{S}^*) & = Pr(\tau \in \hat{S}^* | \tau \in S^*) * Pr(\tau \in S^*) \\
& \approx \frac{m}{n} * e^{-(m/n)\text{mult}_R(v)}
\end{aligned}
$$

Therefore, if mult$_R(v) < n/m$ then

$$
\begin{aligned}
E(\varepsilon(\tau)) & = \text{mult}_R(v) * (n/m) * Pr(\tau \in \hat{S}^*) \\
& \approx \text{mult}_R(v) * e^{-(m/n)\text{mult}_R(v)}
\end{aligned}
$$

Prob. Elim. Lemma Proof (sketch)

$E(\varepsilon(\tau)) \approx \text{mult}_R(v) * e^{-(m/n)\text{mult}_R(v)}$

Since v is sparse, $1/e < e^{-(m/n)\text{mult}_R(v)} \leq 1$
Therefore: $\text{mult}_R(v)/e < E(\varepsilon(\tau)) \leq \text{mult}_R(v)$.

We've met requirement (ii):

(ii) Let τ be a tuple in S with join attribute value v such that SubJoin(v) is of type sparse-sparse or sparse-dense, and not eliminated by R^*. The expected contribution of each such τ to the total estimate is within a small error of mult$_R$(v).

Also since A_S is the sum of $\text{mult}_R(v)$ for all values v that are sparse in R, the Lemma is not far of: $A_S/e \leq E(\hat{A}_S) \leq A_S$

Weak Cost Model Analysis: Very-Dense

The probe cost is $O(m)$. The tuple cost is $\Sigma_{v \in \hat{S}^*}$ $\text{mult}_R(v)$

Let $C(\tau)$ be the tuple cost resulting from tuple τ. If $\tau \in \hat{S}^*$ then $C(\tau) =$ $\text{mult}_R(v)$.

Therefore

$$E(C(\tau)) \approx \text{ mult}_R(v) * \frac{m}{n} * e^{-(m/n)\text{mult}_R(v)}$$

If v is very-dense in R ($\text{mult}_R(v) \geq 2n\ lg\ n/m$) then

$$e^{-(m/n)\text{mult}_R(v)} \leq e^{-(m/n)(2nlg\ n/m)} = n^{-2lge}$$

therefore

$$E(C(\tau)) \leq m * n^{-2lge} = n^{.5-2lg\ e}$$

Weak Cost Model Analysis: Very-Dense

$E(C(\tau)) \leq n^{.5-2lg\ e}$

Markov Inequality: $Pr(X \geq a) \leq \frac{E(X)}{a}$

$$Pr(C(\tau) > 0) \leq n^{.5-2lg\ e}$$

The probability that a tuple that is very-dense in R contributes to the tuple cost is very small: requirement (i).

(i) With high probability, all tuples whose value is very dense in R are eliminated.

Weak Cost Model Analysis: Dense

If v is dense but not very-dense then $1/n \leq e^{-(m/n)\text{mult}_R(v)} \leq 1/e$ and therefore $E(C(\tau)) \leq 1/e$. Which was requirement (iii).

(iii) Some dense tuples in R might not be eliminated by R^*. Nevertheless, each dense tuple in R has constant expected contribution to the total tuple cost.

Thus the total expected contribution to the tuple cost of dense tuples in R is at most n_d/e.

Weak Cost Model Analysis: Sparse

For sparse v, $e^{-(m/n)\text{mult}_{R(v)}} \leq 1$ Therefore $E(C(\tau)) \leq \text{mult}_R(v) * (m/n)$ so the total expected contribution to the tuple cost of sparse tuples in R is

$$(m/n) * A_S = A_S/\sqrt{n}$$

From the Lemma: The expected tuple cost, $E(C)$, is at most $n_d/e + A_S/\sqrt{n}$. $\omega\theta\theta\tau$