TECHNICAL NOTE

A Note on Estimating the Cardinality of the Projection of a Database Relation

RAVI MUKKAMALA
Old Dominion University
and
SUSHIL JAJODIA
George Mason University

The paper by Ahad et al. [1] derives an analytical expression to estimate the cardinality of the projection of a database relation In this note, we propose to show that this expression is in error even when all the parameters are assumed to be constant. We derive the correct formula for this expression.

Categories and Subject Descriptors: H.2.1 [**Database Management**]: Logical Design; H.2.4 [**Database Management**]: Systems—*query processing*

General Terms: Design, Performance

Additional Key Words and Phrases: Block access estimation, query cost-estimation, relational databases

The paper by Ahad et al. [1] derives an analytical expression to estimate the cardinality of the projection of a database relation. This result on estimation of cardinalities derived in this paper may be described as [1]:

> Let $R'(A, B)$ be a relation with attributes A and B, where $|Dom(A)| = m$, and $|Dom(B)| = n$. Further, assume that for any instance of R', the number of distinct A values that occur with a given B value is p, and the number of distinct B values that occur with a given A value is q. Also, let Q be a unary relation of k distinct A values, and let R be a natural join of Q and R'. Now, the expected number of distinct values of B in R is expressed as

$$E_s(|R[B]|) = n\left(1 - \prod_{i=1}^{kq} 1 - \frac{p}{np - (i-1)}\right). \quad (1)$$

In this note, we show that Eq. (1) is not correct for the *expected number* of distinct values of B in R *even when* p *and* q *are constants*. In addition, we derive the correct formula for $E_s(|R[B]|)$.

First, we show a counterexample to Eq. (1) when p and q are constants (as opposed to random variates with a mean of p and q, respectively). Let us consider a case where $|Dom(A)| = m = 3$ and $|Dom(B)| = n = 3$. Also, let $p = q = 2$. Thus, the relation R' has $np = mq = 6$ tuples. Let $k = 2$, so that the relation Q has two out of three of the A-values.

With up to nine permutations of the values of A and B, there is only one relation R' which meets these conditions. There are three possible relations Q. In all cases, R will have three distinct values for B, so

$$E_s(R[B]) = 3. \tag{2}$$

But Eq. (1) gives

$$E_s(R[B]) = 2.8. \tag{3}$$

From Eqs. (2) and (3), it is clear that Eq. (1) does not give the correct expected value even when p and q are assumed to be constants (as opposed to random variates). Equation (1) is also incorrect in the special cases $q = 0$ and $q = n$.

We now derive the correct formula for $E_s(R[B])$ when p and q are constants.

PROPOSITION 1. *Given the exact values for* m, n, p, k, *with* $np = mq$, $E_s(R[B])$ *is given by*

$$E_s(R[B]) = n\left(1 - \frac{\binom{m-p}{k}}{\binom{m}{k}}\right). \tag{4}$$

PROOF. For any value b of B, the probability that the set P of p values of A and the set K of k values of A have an empty intersection in a universe of m values of A is given by

$$P_0 = \frac{\binom{m-p}{k}}{\binom{m}{k}}. \tag{5}$$

Hence, the probability that one particular value b, of B will appear in R is $1 - P_0$, and, by symmetry, the expected value is given by $n(1 - P_0)$. □

Note. The notation of Eq. (4) takes into account the case when $m - p + 1 \leq k \leq m$: $E_s(R[B]) = n$.

The discrepancy we correct in this note was also pointed out by T. H. Merrett in his review of [1]. Luc Devroye derived Eq. (4) in the form given here.

REFERENCES

1. AHAD, R., BAPA RAO, K. V., AND MCLEOD, D. On estimating the cardinality of a database relation *ACM Trans. Database Syst. 14*, 1 (March 1989), 28-40.

Received June 1989; revised February 1990; accepted May 1990