NOTE ON THE EXPECTED SIZE OF A JOIN

Arnon S. Rosenthal
Sperry Research Center, Sudbury, MA 01776

Program 134-44-10
Information Science Department
COMPUTER RESEARCH LABORATORY

SRC-RP-80-75
January 1981

This paper is prepared for journal publication. Issued as a Research Paper, it constitutes a preprint and is given a limited distribution to provide early dissemination of current information.

SPERRY RESEARCH CENTER

I. Introduction

The cost of processing queries in relational databases often depends strongly on the size of the join of two relations. The dependence is particularly strong in distributed databases, where the result may need to be shipped to another site for further processing. Because finding the size of a join requires nearly as much work as actually performing the operation, most published work uses a very simple model in calculating join sizes. It is assumed that each value in the joining domain appears an equal or perhaps a "random" number of times in each relation. We will see that the resulting formula remains valid under much weaker assumptions.

II. Definitions

Suppose relations R and R' are to be joined, with attribute A of R joined with attribute A' of R'. It is assumed that A and A' have the same underlying domain D; $|D|$ denotes the cardinality of D, $|R|$ the number of tuples in R. The join (denoted R*R') is $\{(t,t') | t \in R \text{ and } t' \in R' \text{ and } t.A=t'.A'\}$, where t and t' refer to tuples of R and R', and t.A refers to the value of attribute A in tuple t.

Example 1 presents the standard analysis of join size:

Example 1. Assume that every $x \in D$ appears in exactly $|R|/|D|$ tuples of R and $|R'|/|D|$ tuples of R'. Then $|R*R'|=|D|\cdot(|R|/|D|)\cdot(|R'|/|D|) = |R|\cdot|R'|/|D|$.

To provide a framework for weakening the assumption of evenly distributed values, it is necessary to define a probablistic model of relations. Instead of $|R*R'|$, the expectation $E(|R*R'|)$ will be calculated. The utility of expected value analysis is discussed in the last section.

The set of tuples actually present (called the value of R) is considered to be chosen according to some probability distribution on the Cartesian product of the relation's domains. The value of R' is another random variable. For each $x \in D$, the value of R determines the random variable (number of tuples of R such that t.A=x), denoted $n_x(A)$. $n_x(A')$ is defined similarly on R'.

A distribution for R will be called fair (with respect to A) if $E(n_x(A))=|R|/|D|$ for every x. Notice that we are not constraining particular realizations of R, but are only asking that R be fair "in the mean". Observed values of $n_x(A)$ may differ from the mean value and may be dependent. The rigid formula of example 1 was fair; below, we describe more ways that fair relations may be obtained.

Example 2. Each tuple is randomly assigned a value $x \epsilon D$, sampling with replacement. $n_x(A)$ will then be a binomial random variable, with $|R|$ trials and success probability $1/|D|$.

Several papers (e.g. [CHB]) have analyzed join size under the assumption that R and R' obey the randomness conditions of example 2.

Example 3. Tuples are chosen randomly without replacement from the Cartesian product of the domains of R. Every tuple is equiprobable.

Example 4 A single value of x is chosen, with all $x \epsilon D$ equiprobable. This value is then used in every tuple of R.

III. Analysis

"Fairness" is the basis of our generalized analysis:

Theorem: Suppose for all x, $n_x(A)$ and $n_x(A')$ are independent. Suppose also that either R or R' is fair. Then $E(|R*R'|)=|R|\cdot|R'|/|D|$.

Proof: Join commutes, so without loss of generality, suppose R' is fair, i.e. $E(n_x(A')) = |R'|/|D|$, for all $x \epsilon D$. By independence, $E(n_x(A)\cdot n_x(A')) = E(n_x(A))\cdot E(n_x(A'))$. Finally, note that $|R*R'| = \sum [n_x(A)\cdot n_x(A')]$ (All summations in this section sum over $\{ x \epsilon D \}$). The remainder of the proof is arithmetic: $E(|R*R'|)=$
$E\sum[n_x(A)\cdot n_x(A')] = \sum E[n_x(A)\cdot n_x(A')] =$
$\sum [E(n_x(A))\cdot E(n_x(A'))] = \sum [E(n_x(A))\cdot|R'|/|D|] =$
$[\sum E(n_x(A))]\cdot|R'|/|D| = E[\sum n_x(A)]\cdot|R'|/|D| = E(|R|)\cdot|R'|/|D| =$
$|R|\cdot|R'|/|D|$. QED

The first few lines of the proof established the fact that even if fairness does not hold, expected join size depends only on the means, not on how each $n_x(A)$ varies about its mean. More precisely:

Corollary: $E(|R*R'|)$ depends only on the expected values $\{ E(n_x(A)), E(n_x(A')) \}$ for all $x \epsilon D$, not on the entire distributions.

IV. Discussion

The theorem states that it is not necessary to have perfect symmetry in values of x, but only to have independence and for one relation to be evenly distributed "on the average". We now discuss the validity of our assumptions and the utility of expected value analysis.

Fairness of R is difficult to verify -- in general one would need a large collection of independent values for R. Example 4 is particularly difficult to show fair, since each sample has only one value for the attribute. In contrast, if one hypothesizes that Example 2 describes the creation of R, then each tuple gives an independent sample of t.A, and one can test whether the observed value of R is consistent with the hypothesis that the attribute values are properly distributed (i.e. according to a multinomial distribution with $|R|$ trials, and probability $1/|D|$).

The in ependence assumption is often quite reasonable. For example, if R is ENROLLMENT(STUDENT, COURSE) and R' is READING_LIST(COURSE, TEXTBOOK) then one would not expect the number of textbooks used in a course used to have a great effect on enrollment (assuming workload was reflected in the credit assigned).

On the other hand, suppose R IS EMPLOYEE(EMP_NAME, DEPT) and R' is PROJECT(DEPT, PROJ_NAME). Departments which have only a small number of employees might tend to have relatively few projects. Hence $n_x(A)$ and $n_x(A')$ are positively correlated, so our estimated join size would be too small.

Expected value analysis is appropriate if "real world" costs depend approximately linearly on query processing time. The time to process a query involving R*R' will generally be a function of $|R*R'|$. This function may grow faster than linearly (e.g. for sorting), but generally this effect will not be too great. A potentially catastrophic nonlinearity can occur in applications where the real world cost is a grossly nonlinear function of query processing time, e.g. in process control, where very long processing times may be intolerable, even if rare. The use of $E(|R*R'|)$ can lead to a low estimate of expected real world cost. Of course, the deterministic model of Example 1 shares the failing, since it produces the same result. To really estimate query costs in these situations, one needs more information about the variance and distribution of $|R*R'|$. [MO] provides formulas for variance in a more complicated model, but was forced to make much stronger assumptions.

Proper definition of D (and hence $|D|$) is crucial to the accuracy of the estimator. The probabilistic model permits D to include values which do not actually occur among the tuples of the relation, but only actual candidates should be included, e.g. social security numbers of employees in one company, rather than all social security numbers.

Recently, Richard [Ri] described a parameter system for calculating the expected size of arbitrary relational algebra expressions. His results are more sweeping than ours, but also much more complicated. He assumes all relations are fair, and that "compatible" tuples of the same relation are independent. The assumptions about interrelational independence and the size estimates also differ from ours.

V. Summary

We showed that the restrictive assumptions underlying the usual analysis of join size are somewhat less serious than they seem at first glance, because the join size calculated from a deterministic model actually may be used for a much wider class of situations. The assumptions were weakened as much as seemed possible without losing analytic tractability. The limitations of the analysis were then discussed.

References

[CHB] D.M. Chiu, Y.C. Ho, and P.A. Bernstein, "On optimization of query processing strategies," Technical Report 672, Harvard University Division of Applied Sciences, Cambridge Massachusetts, 1979.

[MO] T. Merrit and E. Otoo, "Distribution models of relations", Proceedings of tne conference on Very Large Data Bases, IEEE 1979. pp. 418-425.

[Ri] P. Richard, "Evaluation of the size of a query expressed in relational algebra," Proceedings of SIGMOD Conference, ACM 1981.

Acknowledgement

Murray Eaelberg and David Reiner contributed substantially to the rigor and clarity of this note.