EXPECTED SIZE OF THE NATURAL JOIN

By

Don-lin Yang

Computer Science Report #TR85-10

July 5, 1985

Expected Size of the Natural Join

Don-lin Yang

Department of Computer Science
University of Virginia
Charlottesville, VA 22903
U.S.A.

Abstract. The primary cost in processing relational database queries is the cost of joining two or more relations. In order to develop more efficient join algorithms or to optimize query strategies at run time, we must be able to accurately compute the expected size of a join relation. In [6], Rosenthal derives the expected join size formula in terms of the sizes of the join domain and source relations. However, his proof process requires two stringent conditions. First, the distributions of the join attribute values in source relations must be independent and second, at least one of the distributions must be uniform. In this paper, we show that Rosenthal's expression is still valid under much more general conditions through the use of an exact join size formula.

Keywords: Join size, relational database, expectation, statistical parameter

---

This research was supported in part by NSF Grant MCS-83-02654.

1. Introduction

A major cost of processing relational database queries is that of performing the relational join $r \bowtie s$ [6]. Therefore, any procedure that would interactively optimize query processing in a large database system must be able to effectively determine the expected cost of a join. As a simple example, since $\bowtie$ is associative and commutative, there are three distinct sequences of evaluating $q \bowtie r \bowtie s$. The cost of evaluating one sequence may easily differ by an order of magnitude from the cost of a different order.

The cost of performing a join $r \bowtie s$ is, in turn, dependent on several factors [1], such as

a) the actual algorithm used,

b) whether the join attribute(s) are indexed, or

c) the expected number of tuples in s that will join with a single tuple in r.

However, the dominant factor in join cost is always three separate cardinalities, $|r|$, $|s|$, and $|r \bowtie s|$. We assume, as in most databases, that the first two are known. The goal of this paper is to give a reasonable estimate for the latter.

We will first examine and analyze the expected join size formula derived by Rosenthal [6]. Then, we develop the exact size formula of a join and provide a new proof for Rosenthal's expression without using his stringent assumptions.

2. Rosenthal's Expected Join Size Formula

For notational convenience, we assume all joins are over a single attribute A_j which we call the **join attribute** and denote simply by A. Of course, one can join relations over several attributes, but there is no loss of generality in our analysis since they can be regarded as a single attribute mapping into the Cartesian product

of the original attribute domains.

By the size of a relation r, denoted $|r|$, we mean the number of tuples in r, or its cardinality. If we were concerned with storage costs or memory management, we would also have to consider the number of attributes and the nature of their domains.

Rosenthal [6] derives the following expression for the expected size of a join

$$exp(|r \underset{A}{\bowtie} s|) = \frac{|r| \cdot |s|}{|A|} \tag{2.1}$$

where $|r|$, $|s|$, and $|A|$ denote the sizes of relations r, s, and the domain size of attribute A respectively. His proof process requires two conditions to derive this result. First, the distribution of join attribute values in r and s must be independent and second, the distribution of join attribute values must also be fair (uniform) in at least one of the relations.

These are stringent constraints that are seldom realized in practice. Christodoulakis [2,3] shows that these assumptions, which are used in most analytic work, may result in large estimation errors. In fact, they commonly lead to pessimistic estimations of the database cost. Moreover, they are really unneeded. A major result of this paper is to show that expression (2.1) is still valid under much more general conditions. It is accomplished by using the exact join size formula which is developed in the next section.

3. Exact Size of a Join

The results of this section are most easily motivated with a running example, consisting of two small relations r and s with cardinalities $|r| = 6$ and $|s| = 9$. As we know, the size of a resultant join $|r \underset{A}{\bowtie} s|$ is dependent on the size of its initial factors, $|r|$ and $|s|$, together with the distribution of the join attribute A in each. It is the latter which determines "*the expected number of tuples in s that will join with a single tuple in r*". Blasgen and Eswaran [1] used a single factor P, called a **join filter**,

to account for this aspect in their analysis.

In Table 3.1, we display only the join attribute values for the six tuples of r and nine tuples of s. The remaining attributes of each tuple have been suppressed. For simplicity, we limit the domain of the join attribute A to the three integers, {1, 2, 3}. Finally, we provide five different, and arbitrary, instances of each of these two relations. In each pair the distribution of attribute values will determine the size of the join.

The frequency distributions of attribute values, μ_r and μ_s, for each pair r, s in sets (1) - (5) are shown in Table 3.2. For example, in set (1) all six elements (tuples) of r have value 3 for the join attribute, and similarly for all nine elements of s. The resulting join for case (1) must be the Cartesian product $r \times s$ of the two relations with size $|r| \cdot |s| = 6 \cdot 9 = 54$; as it is. In case (5) the two relations have no common join attribute values, so readily $|r \bowtie s| = 0$. Cases (2) through (4) simply illustrate other possible distributions between these extremes.

Distributions of join attribute, A, values
for relations r and s

(1)		(2)		(3)		(4)		(5)	
$r(A)$	$s(A)$	$r(A)$	$s(A)$	$r(A)$	$s(A)$	$r(A)$	$s(A)$	$r(A)$	$s(A)$
3	3	1	1	1	1	1	1	3	1
3	3	2	2	1	1	2	1	3	1
3	3	3	2	2	1	3	1	3	1
3	3	3	3	2	2	3	1	3	1
3	3	3	3	3	2	3	1	3	1
3	3	3	3	3	2	3	1	3	1
	3		3		3		2		1
	3		3		3		2		1
	3		3		3		3		1

Table 3.1 Attribute value distributions

Now, the question is how to account for the variability of distributions in the join size formula. Our solution is to include a correction term based on the correlation coefficient c and standard deviations δ_r, δ_s as shown in the following theorem.

Theorem 3.1 Let r and s be two relations with a common attribute A. Let c be the correlation coefficient between the distributions of join attribute values in r and s, and δ_r, δ_s be their corresponding standard deviations. Also let $|A|$ denote the number of attribute values in A. Then

$$|r \underset{A}{\bowtie} s| = \frac{|r| \cdot |s|}{|A|} + |A| c \delta_r \delta_s$$

Proof :

Let $\{a_1, a_2, ..., a_k, ..., a_{|A|}\}$ be the join attribute values in A. Let $\mu_r(a_k)$ and $\mu_s(a_k)$ be the number of tuples having join attribute value a_k in r and s, respectively. We know that

Distributions of attribute value frequencies

Join Domain Values	(1)		(2)		(3)		(4)		(5)	
	μ_r	μ_s	μ_r	μ_s	μ_r	μ_s	μ_r	μ_s	μ_r	μ_s
1	0	0	1	1	2	3	1	6	0	9
2	0	0	1	2	2	3	1	2	0	0
3	6	9	4	6	2	3	4	1	6	0
Join Size	54		27		18		12		0	

Table 3.2 Results of the natural-join

$$|r \underset{A}{\bowtie} s| = \sum_{a_k \in A} (\mu_r(a_k) \cdot \mu_s(a_k)) = |A| \cdot \frac{\sum_{a_k \in A} (\mu_r(a_k) \cdot \mu_s(a_k))}{|A|}$$

$$= |A| \cdot exp(\mu_r \cdot \mu_s) \tag{3.1}$$

Let X = μ_r and Y = μ_s. The standard definition of the correlation coefficient [7] is

$c = \frac{exp((X - exp(X)) \cdot (Y - exp(Y)))}{\delta_x \delta_y}$. Then, we have the following expression for the covariance

$$\begin{aligned} c\,\delta_x \delta_y &= exp((X - exp(X)) \cdot (Y - exp(Y))) \\ &= exp(XY - X\,exp(Y) - exp(X)Y + exp(X)exp(Y)) \\ &= exp(XY) - exp(X\,exp(Y)) - exp(exp(X)Y) + exp(exp(X)exp(Y)) \\ &= exp(XY) - exp(X)exp(Y) - exp(X)exp(Y) + exp(X)exp(Y) \\ &= exp(XY) - exp(X)exp(Y) \end{aligned}$$

That is, $exp(XY) = exp(X)exp(Y) + c\,\delta_r \delta_s$.

So that, $exp(\mu_r \cdot \mu_s) = exp(\mu_r) \cdot exp(\mu_s) + c\,\delta_r \delta_s$.

Then, substituting into (3.1) we have

$$|r \underset{A}{\bowtie} s| = |A| \cdot (exp(\mu_r) \cdot exp(\mu_s) + c\,\delta_r \delta_s)$$

$$= |A| \cdot \left(\frac{\sum_{a_k \in A} \mu_r(a_k)}{|A|} \cdot \frac{\sum_{a_k \in A} \mu_s(a_k)}{|A|} + c\,\delta_r \delta_s \right)$$

$$= |A| \cdot \left(\frac{|r|}{|A|} \cdot \frac{|s|}{|A|} + c\,\delta_r \delta_s \right)$$

$$= |A| \cdot \frac{|r|}{|A|} \cdot \frac{|s|}{|A|} + |A|\,c\,\delta_r \delta_s$$

$$= \frac{|r| \cdot |s|}{|A|} + |A| c \delta_r \delta_s \quad \square$$

To illustrate the formula of Theorem 3.1, consider the same five attribute distributions that were shown in Table 3.2. In Table 3.3 (a) row, we display the rounded value of correlation coefficient, c, for each distribution set of join attribute values between r and s. In row (b), $|A| c \delta_r \delta_s$ represents the correction factor of the join size. The result of $|r| \cdot |s| / |A|$ is shown in row (c) and is equal to the average join size over the entire possible experiment space (1540 distribution sets). The sum of (b) and (c) shown in (d) matches the exact join size as we expected.

Notice that the correction term $|A| c \delta_r \delta_s$ vanishes when the correlation coefficient or one of the standard deviations is equal to zero. The implications of these observations are emphasized by the following two corollaries. First, if we know the frequencies of join attribute values in the source relations are independent variables, then Rosenthal's expected join size expression is the same as the exact join size expression because the correlation coefficient is zero [5]. Note that we may have $c = 0$, and yet join attribute distributions need not be independent.

	Distribution sets				
	(1)	(2)	(3)	(4)	(5)
(a) Correlation: c	1.0	.98	0	-.65	-.50
(b) Correction: $\|A\| c \delta_r \delta_s$	36	9	0	-6	-18
(c) Expected size: $\|r\| \cdot \|s\| / \|A\|$	18	18	18	18	18
(d) Actual size: (b) + (c)	54	27	18	12	0

Table 3.3 Results of join size parameters

Corollary 3.1 **If the frequency distributions of the join attribute values in relations r and s are independent, then Rosenthal's expected join size expression is the same as the exact join size expression. That is,**

$$|r \underset{A}{\bowtie} s| = \frac{|r| \cdot |s|}{|A|}$$

Second, if it is known that there is a perfectly uniform distribution of join attribute values in either r or s, then the following corollary shows that the same result holds regardless of the distribution of join attribute values in the other relation!

Corollary 3.2 **If the join attribute values of either relation have a perfectly uniform distribution, that is, if for all values a_k in join attribute A, either $\mu_r(a_k) = \frac{|r|}{|A|}$ or $\mu_s(a_k) = \frac{|s|}{|A|}$, then we have the exact equality:**

$$|r \underset{A}{\bowtie} s| = \frac{|r| \cdot |s|}{|A|}$$

Proof :

From Theorem 3.1 we have

$$|r \underset{A}{\bowtie} s| = \sum_{a_k \in A} (\mu_r(a_k) \cdot \mu_s(a_k))$$

Assume the distribution of join attribute values in r is perfectly uniform, then we have $\mu_r(a_k) = \frac{|r|}{|A|}$. So that

$$|r \underset{A}{\bowtie} s| = \sum_{a_k \in A} (\frac{|r|}{|A|} \cdot \mu_s(a_k))$$

$$= \frac{|r|}{|A|} \cdot \sum_{a_k \in A} \mu_s(a_k)$$

$$= \frac{|r| \cdot |s|}{|A|} \quad \square$$

4. Expected size of a join

We know that the join size depends primarily on the sizes of two joining relations and their join domain size. This expression (2.1) was asserted in [6] as an expectation. We have refined this earlier work by introducing a correction term reflecting the actual distribution of the join attribute values. This allows us to show that under the conditions Roselthal proved were sufficient for (2.1) to be an expectation, are independently sufficient for the expression to denote the *actual* size of the resultant join.

In real applications it is often not practical to derive the exact join size by computing the correlation coefficient and standard deviations. An expectation is usually sufficient. Corollary 4.1 below gives sufficient conditions for expression (2.1) to be an *expectation*.

Corollary 4.1 Let r and s be two randomly generated relations with a common attribute A. Then

$$exp(|r \underset{A}{\bowtie} s|) = \frac{|r| \cdot |s|}{|A|}$$

Proof :

From Theorem 3.1 we know

$$|r \underset{A}{\bowtie} s| = \frac{|r| \cdot |s|}{|A|} + |A|\, c\delta_r \delta_s \quad \text{and}$$

$$c\delta_r \delta_s = exp[(\mu_r - exp(\mu_r)) \cdot (\mu_s - exp(\mu_s))]$$

$$= \frac{1}{|A|} \sum_{k=1}^{|A|} [(\mu_r(a_k) - exp(\mu_r(a_k))) \cdot (\mu_s(a_k) - exp(\mu_s(a_k)))]$$

So that, $|r \underset{A}{\bowtie} s| = \frac{|r| \cdot |s|}{|A|} + \sum_{k=1}^{|A|} [(\mu_r(a_k) - exp(\mu_r(a_k))) \cdot (\mu_s(a_k) - exp(\mu_s(a_k)))]$

To find the expected join size, we consider computing the average join size from all possible combinations of join attribute, A, value distributions for any two relations of sizes $|r|$ and $|s|$ respectively. Let m denote the number of all possible A-attribute distributions in the relation r. Since we assume $|r|$ and $|A|$ are finite, m is finite; and $m = C(|A|+|r|-1, |r|)$ [4]. Similarly, let n denote the number of all possible A-attribute distributions in the relation s and $n = C(|A|+|s|-1, |s|)$. Let r_i and s_j denote the ith and jth join attribute value distributions of the respective relations, where $i = 1, 2, \ldots m$ and $j = 1, 2, \ldots n$. Then,

$$exp(|r \underset{A}{\bowtie} s|) = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} (|r_i \underset{A}{\bowtie} s_j|)$$

$$= \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} [\frac{|r| \cdot |s|}{|A|} + \sum_{k=1}^{|A|} [(\mu_{r_i}(a_k) - exp(\mu_{r_i}(a_k))) \cdot (\mu_{s_j}(a_k) - exp(\mu_{s_j}(a_k)))]]$$

$$= \frac{|r| \cdot |s|}{|A|} + \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{|A|} [(\mu_{r_i}(a_k) - exp(\mu_{r_i}(a_k))) \cdot (\mu_{s_j}(a_k) - exp(\mu_{s_j}(a_k)))]]$$

Now, we will show that for any j

$$\sum_{i=1}^{m} \sum_{k=1}^{|A|} [(\mu_{r_i}(a_k) - exp(\mu_{r_i}(a_k))) \cdot (\mu_{s_j}(a_k) - exp(\mu_{s_j}(a_k)))]] = 0.$$

Since $exp(\mu_{r_i}(a_k)) = \frac{|r|}{|A|}$ and $exp(\mu_{s_j}(a_k)) = \frac{|s|}{|A|}$, we use $\overline{\mu}_r$ and $\overline{\mu}_s$ to denote $exp(\mu_{r_i}(a_k))$ and $exp(\mu_{s_j}(a_k))$ respectively. Then,

$$\sum_{i=1}^{m} \sum_{k=1}^{|A|} [(\mu_{r_i}(a_k) - \overline{\mu}_r) \cdot (\mu_{s_j}(a_k) - \overline{\mu}_s)] = \sum_{k=1}^{|A|} [(\mu_{s_j}(a_k) - \overline{\mu}_s) \cdot \sum_{i=1}^{m} (\mu_{r_i}(a_k) - \overline{\mu}_r)]$$

$$= \sum_{k=1}^{|A|} [(\mu_{s_j}(a_k) - \overline{\mu}_s) \cdot (\sum_{i=1}^{m} \mu_{r_i}(a_k) - m\overline{\mu}_r)]$$

Because we consider all possible combinations of join attribute value distributions, then the occurrences of various frequencies for the relations $r_1, r_2, \ldots, r_m$ must be the same for each attribute value $a_k \in A$. That is,

$$\sum_{i=1}^{m} \mu_{r_i}(a_f) = \sum_{i=1}^{m} \mu_{r_i}(a_g) = d \quad \text{where} \quad f, g = 1, 2, \ldots, |A|$$

Therefore, $(\sum_{i=1}^{m} \mu_{r_i}(a_k) - m\overline{\mu}_r) = d - m\overline{\mu}_r$. Let e denote this constant $(d - m\overline{\mu}_r)$.

Then

$$\sum_{i=1}^{m}\sum_{k=1}^{|A|}[(\mu_{r_i}(a_k)-\overline{\mu}_r)\cdot(\mu_{s_j}(a_k)-\overline{\mu}_s)] = \sum_{k=1}^{|A|}[(\mu_{s_j}(a_k)-\overline{\mu}_s)\cdot e]$$

$$= e\cdot\sum_{k=1}^{|A|}(\mu_{s_j}(a_k)-\overline{\mu}_s)$$

$$= e\cdot(\sum_{k=1}^{|A|}\mu_{s_j}(a_k)-|A|\cdot\overline{\mu}_s)$$

$$= e\cdot(|s|-|A|\cdot\frac{|s|}{|A|})$$

$$= 0$$

Hence $exp(|r \underset{A}{\bowtie} s|) = exp(\frac{|r|\cdot|s|}{|A|} + \frac{1}{m\cdot n}\cdot 0) = \frac{|r|\cdot|s|}{|A|}$ □

Note that we only require that the source relations are randomly generated. Neither the independent distribution of join attribute values, nor the uniform distribution is assumed. The result of this corollary is important because it shows that Rosenthal's expression holds for any kind of join attribute value distributions. That is to say, one need not to know the frequency distributions of join attribute values in order to use the expected join size expression.

A detailed discussion of some other expected join size formulae, in which the source relations have unequal number of unique join attribute values or selection operations are performed before the join, can be found in [8].

References

[1] Blasgen, M.W. and Eswaran, K.P. "Storage and access in relational data bases", IBM Systems Journal, v.16.4 1977 (pp. 363-377)

[2] Christodoulakis, S. "Estimating selectivities in databases" Ph.D. dissertation, Rep. CSRG-136, Computer Science Dept, U. of Toronto, 1981

[3] Christodoulakis, S. "Implications of certain assumptions in database performance evaluation", ACM TODS, v.9.2 June 1984 (pp. 163-186)

[4] Liu, C.L. "Introduction to combinatorial mathematics", McGraw-Hill Book Company, 1968 (p. 13)

[5] Meyer, P.L. "Introductory probability and statistical applications", 2nd edition, Addison-Wesley Publishing Co., Inc., 1965 (p. 144)

[6] Rosenthal, A.S. "Note on the expected size of a join", ACM SIGMOD Record, v.11.4 July 1981

[7] Walker, H.M. and Lev, J. "Statistical Inference", Holt, Rinehart and Winston Inc., 1953 (p. 248)

[8] Yang, Don-lin "Expectations associated with compound selection and join operations", Ph.D. dissertation, University of Virginia, May 1985