Data Base Parameters: Equijoin and Semijoin

Guy Louchard
Département d'Informatique, Université Libre de Bruxelles

February 7, 1994

[summary by Danièle Gardy]

1. Introduction

This talk is a sequel to the talk *Sizes of relations: a dynamic analysis* by D. Gardy, and we shall make references to its summary [4]. A presentation of the database problem and of the model was given there, along with the dynamic study of the size of a relation obtained by *projection* of an initial relation. This second talk is centred around *joins*, i.e. operations building a "derived" relation from two initial relations. As was the case for projections, the joins can be described by urn models [1]. Although the computations become quite involved, due to the fact that we deal with bi-dimensional processes, the techniques are similar to those used for the projection.

This second talk begins by presenting in detail some points introduced in [4], such as the birth and death processes describing the number of balls in an urn, then turns to join models. The complete demonstrations are given in the full papers [2] (for the projections) and [3] (for the joins).

2. Urn models and processes

2.1. Urn models. We can describe a relation present in the database by an urn model, and its size by a random allocation model counting the number of balls allocated according to a certain scheme (see [4] in these proceedings). Now the equi- or semi-join of two relations R and S can be modelled in a similar way: Let X be the join attribute, and d the number of values X can take; we consider d distinguishable urns labelled by these values.

- We throw r red balls for the relation R, and s blue balls for the relation S, according to the rules reflecting the constraints on these two relations;
- We consider each urn in turn. If an urn has received i red balls and j blue balls, we put ij green balls in the urn for the equijoin, or $iI_{j>0}$ green balls for the semijoin;
- The total number of green balls is now the (equi- or semi-) join size.

Urn models were already encountered in the dynamic analysis of tries (see [5]), but there the capacity of the urn varied in time.

2.2. Indicator functions. We introduce the following definitions (κ is an integer-valued function; in the following it counts the number of balls in an urn):

$$\phi_1(\kappa) = I_{\kappa>0}; \qquad \phi_2(\kappa) = \kappa.$$

We shall use upper indices R and B to make precise the relations, or ball colours, we consider. Thus the functions ϕ_2^R and ϕ_2^B are involved in the equijoin, and the functions ϕ_2^R and ϕ_1^B in the

semijoin. We shall also use

$$E_1^i(\phi) := E[\phi(\kappa_1^i)]; \qquad E_{1,2}^{i,j}(\phi) := E[\phi(\kappa_1^i)\phi(\kappa_2^j)]; \qquad Z_1^i := \Pr[\kappa_1^i = 0].$$

2.3. Bi-dimensional processes. Assume that the stochastic processes describing the initial relations R and S are known. There are three ways to combine them, according to the type of correlation allowed on the relations R and S:

(1) The processes may be correlated, in that, at each step, we either update one (and only one) of the relations, or we make a search (query);
(2) The processes may be independent: at each step and for each relation, we can do an insertion, a deletion or a search. In this model, it is possible to update two relations simultaneously, or to query one relation and to update the other one;
(3) We may extend this second model to allow more general probabilities: we define a probability for query, and eight probabilities for updating one or two relations.

For each type of correlation, we can compute the expectation of the join size and the covariance matrix of the bi-dimensional process *(size of R, size of S)*.

3. Birth and death process

We now return to the processes describing the number of balls in an urn, either finite or infinite.

3.1. Infinite urns. The number of balls in one urn is (asymptotically) given by a birth and death process with birth rate and individual death rate given by[1]

$$\lambda(t) = p_{\mathcal{I}}(t)\frac{n}{d}, \qquad \mu(t) = \frac{p_{\mathcal{D}}(t)}{f_1(t)}.$$

The probability that a ball, present at time t_1, survives at time t_2, is

$$ps_{1,2} = \exp\left[-\int_{t_1}^{t_2} \mu(s)\,ds\right].$$

The total number of balls inserted in one urn between times t_1 and t_2, and not deleted at time t_2, follows a Poisson distribution with parameter

$$\rho_{1,2} = \frac{n}{d}\int_{t_1}^{t_2} p_{\mathcal{I}}(u) ps_{1,2}(u, t_2)\,du.$$

3.2. Bounded urns. Such an urn has δ cells; define $\beta := d\delta/n$. At time t_1, the number of balls in any one urn follows (approximately) a binomial distribution with parameters δ and $f_1(t_1)/\beta$.

The next result is Lemma 3 of [2]:

PROPOSITION 1. *Given that we start with k_1 balls in the urn U_i at time t_1, the number of balls $\kappa(t)$ $(t > t_1)$ in the urn U_i is described asymptotically by a birth and death process starting from k_1, with birth rate $\lambda(t) = [\delta - \kappa(t)]f_4(t)$, where $f_4(t) = p_{\mathcal{I}}(t)/[\beta - f_1(t)]$ is the birth rate in a cell, and individual death rate $f_5(t) = p_{\mathcal{D}}(t)/f_1(t)$.*

We can now analyze the distribution of the number of balls in one urn at time t_2, conditioned by the number of balls in the urn at time t_1 (see [2] for details).

[1] We recall that the expectation of the number of balls is asymptotically equal to $nf_1(t)$, with f_1 varying according to the relation scheme.

4. Size of a join

4.1. Theorem. Theorem 6.1 of [3] gives a characterisation of the size of a join as a non-Markovian, Gaussian process of known expectation and covariance:

THEOREM 1. *In the join model, the size $S([nt])$ of the join at time nt is asymptotically given by a non-Markovian Gaussian process with*

$$E[S([nt])] \sim nG(t) \qquad \mathrm{Cov}(S_1, S_2) \sim n\Psi_R(t_1, t_2).$$

The relative error in the density is $O(1/\sqrt{n})$.

4.2. Idea of the proof. The principle of the proof is the same as for the projection. A fundamental result is Lemma 1 of [3], which we recall below.

Define $Y_1 = \sum_{i=1}^{d} \phi(\kappa_1^i)\psi(\lambda_1^i)$: the function κ is relative to red balls, i.e., to the relation R, and the function λ is relative to blue balls, i.e., to the relation S. The index i denotes the number of the urn, and the index 1 shows that we consider the situation at time t_1. We define similarly Y_2 for time t_2.

PROPOSITION 2. *The mean of Y_1 and the covariance of Y_1 and Y_2 are given by*

$$\begin{aligned} E(Y_1) &= dE_1^i(\varphi)E_1^i(\psi); \\ \mathrm{Cov}(Y_1, Y_2) &= dE_{1,2}^{i,i}[\varphi]E_{1,2}^{i,i}[\psi] - dE_{1,2}^{i,j}[\varphi]E_{1,2}^{i,j}[\psi] \\ &\quad + d^2\left[E_1^i[\varphi]E_2^j[\varphi]\ C_{1,2}^{i,j}[\psi] + E_1^i[\psi]E_2^j[\psi]\ C_{1,2}^{i,j}[\varphi] + C_{1,2}^{i,j}[\varphi]C_{1,2}^{i,j}[\psi]\right], \end{aligned}$$

with

$$\begin{aligned} C_{1,2}^{i,j}[\varphi] :&= \sum_{k_1} \Pr(\kappa_1^i = k_1)\varphi(k_1) \sum_{k_2} [\Pr(\kappa_2^j = k_2 | \kappa_1^i = k_1) - \Pr(\kappa_2^j = k_2)]\varphi(k_2) \\ &= E_{1,2}^{i,j}[\varphi] - E_1^j[\varphi]E_2^j[\varphi]. \end{aligned}$$

Proposition 2 of [3] gives the covariance for the "static" structure, when the sizes of the initial relations are assumed known ("non-random" case):

$$E(Y_1) \sim n\ F\left(f_1^R(t_1), f_1^B(t_1)\right); \qquad \mathrm{Cov}(Y_1, Y_2) \sim n\ \psi_{NR}(t_1, t_2).$$

As the number of balls in the urns in no longer known with certainty, but is described by a random variable, this introduces a perturbation on the join size, which can be computed. Lengthy computations give the result.

4.3. Example. Let us take an example to illustrate our results: the equijoin when the urns for both relations are unbounded and when, at each step, either we throw a ball, or we delete a ball, or we make a search. Asymptotically, i.e., for $n \to +\infty$, the expectation of the process *size of equijoin* is

$$E[S([nt])] \sim n\frac{\bar{x}_R\bar{x}_B t^2}{\alpha},$$

and its covariance is

$$\mathrm{Cov}(S_1, S_2) \sim n\left[\frac{\bar{x}_R\bar{x}_B t_1^2 ps_{1,2}^R ps_{1,2}^B}{\alpha} + \frac{\bar{x}_B^2 t_1 t_2}{\alpha^2}\sigma_R^2 t_1 - 2\frac{\bar{x}_B\bar{x}_R t_1 t_2}{\alpha^2}\bar{x}_B\bar{x}_R t_1 + \frac{\bar{x}_R^2 t_1 t_2}{\alpha^2}\sigma_B^2 t_1\right].$$

5. Conclusion

A natural direction for further research would be to implement a toolbox in a computer algebra system such as Maple which, given the structure and dependencies of the initial relations, the kind of relational operation, and the update operations on the database, would compute automatically the moments of the process describing the projection or join size.

Bibliography

[1] Gardy (D.). – Join sizes, urn models and normal limiting distributions. *Theoretical Computer Science, Series A*, vol. 131, n° 2, August 1994.

[2] Gardy (D.) and Louchard (G.). – *Dynamic analysis of some relational data base parameters I: projections.* – Technical Report n° 94-6, Laboratoire Prism, University of Versailles, February 1994.

[3] Gardy (D.) and Louchard (G.). – *Dynamic analysis of some relational data base parameters II: equijoins and semijoins.* – Technical Report n° 94-7, Laboratoire Prism, University of Versailles, February 1994.

[4] Gardy (Danièle). – Sizes of relations: a dynamic analysis. In Salvy (B.) (editor), *Algorithms Seminar 1993-1994*. Institut National de Recherche en Informatique et en Automatique, *Research Report.* – 1994. These proceedings.

[5] Louchard (G.). – Trie size in a dynamic list structure. In Gaudel (M.-C.) and Jouannaud (J.-P.) (editors), *TAPSOFT'93. Lecture Notes in Computer Science*, vol. 668, pp. 719–731. – Springer Verlag, 1993.