ON THE SIZE OF PROJECTIONS: I

Erol GELENBE
LRI Université de Paris-Sud, 9495 Orsay, France

Danièle GARDY
Ecole Polytechnique, Centre de Mathematiques Appliqués, 91128 Palaiseau, France

Received 9 October 1981; revised version received 26 November 1981

Relational model, projections, size, data base operations

1. Introduction

Consider the following problem which arises in various areas of application (physics experiments, census data, etc.) where the collection of a large number of data is involved.

The result of the data collection processes is a set

$$T_{\ell k} = \{(t_{ii}, ..., t_{ik}), ..., (t_{\ell 1}, ..., t_{\ell k})\}$$

of vectors where each t_{ij}, $1 \leqslant i \leqslant \ell$, is an element of the set D_j, $1 \leqslant j \leqslant k$. Thus we may consider that T is an ℓ-row and k-column matrix; the elements of the j-th column all being elements of some set D_j.

For instance, as a result of a physics experiment the mass, position, and velocity of one or more particles may be measured. We would then have a table $T_{\ell 3}$ where ℓ is the number of distinct (mass, position, velocity) vectors encountered, the i-th row of $T_{\ell 3}$ being

(mass, position, velocity).

Once such a table has been obtained from the physics experiment it is of interest to compute projections of this table along certain of its columns or coordinates. For instance, one may be interested in obtaining the set of all distinct values of the pair (mass, position). Thus, from $T_{\ell 3}$ we would obtain a new table

$$\pi_3(T_{\ell 3})$$

which would contain ℓ' rows ($\ell' \leqslant \ell$) and 2 columns, the last (velocity) column being removed.

Just as for $T_{\ell 3}$, all of the rows of $\pi_3(T_{\ell 3})$ would be distinct. Thus, for instance, if the measurements yield

$$T_{33} = \begin{bmatrix} 1, & 0, & 0 \\ 1, & 0, & 1 \\ 1, & 1, & 5 \end{bmatrix},$$

we would have

$$\pi_3(T_{33}) = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

The following problem is of interest.

Problem 1. For a given $T_{\ell k}$ and a given projection π_x along the columns $(1, ..., x-1, x+1, ..., k)$ estimate the (size or) member of distinct rows of $\pi_x(T_{\ell k})$.

Let us now consider a seemingly more general problem closely related to this one. The result of the experiment could be a table $T_{\ell 4}$, where each row would be

(number, mass, position, velocity)

giving the number of particles counted which have the same (mass, position, velocity) characteristic. We might then be interested in obtaining the total number of particles having the same (mass, position). Thus with the example given above, our initial data might have been

$$T_{34} = \begin{bmatrix} 100, & 1, & 0, & 0 \\ 25, & 1, & 0, & 1 \\ 45, & 1, & 1, & 5 \end{bmatrix},$$

while the result of interest would be

$$T' = \begin{bmatrix} 125, & 1, & 6 \\ 45, & 1, & 1 \end{bmatrix}.$$

We see in this case that the number of lines of T′ is the same as that of $\pi_3(T_{33})$. In fact the problem of estimating the size of T′ is simply an instance of Problem 1, and we see that the estimation of the size of projections is, in fact, quite general in various problems of data handling.

It is of course also an important issue in data base theory (see for instance [1–4]). We shall address and solve it in a specific simplified mathematical context.

2. The formal problem and its solution

Let D_j, $1 \leqslant j \leqslant k$, the set of values which may be taken by an element of $T_{\ell k}$, be a finite set. We take $\ell \leqslant |D_i|$ for all $1 \leqslant i \leqslant k$.

A table $T_{\ell k}$ is a set of ℓ distinct elements of $D_1 \times D_2 \times \cdots \times D_k$. We shall assume that an element $t \in D_1 \times \cdots \times D_k$ is generated in the following manner:

$$t = (t_1, \ldots, t_k),$$

where t_j is equally likely to be any of the elements of D_j. That is, we treat D_j as a sample space to which we associate a uniform distribution. Furthermore, we assume that t_j is independent of t_m if $j \neq m$.

Remark 1. Let $d_i = |D_i|$, and let $h_{\ell k}$ denote the number of distinct tables $T_{\ell k}$. Then [1]

$$h_{\ell k} = \binom{d}{\ell},$$

where $d = d_1, d_2 \cdots d_k$. This is simply the number of distinct ℓ-row tables.

Result 2. The number of tables of the form $T_{\ell k}$ whose projection $\pi_j(T_{\ell k})$ along the j-th column contains (exactly) $(\ell - y)$ rows, $0 \leqslant y \leqslant \ell - 1$, is

$$Q_{\ell k}^{j,y} = \binom{d/dj}{\ell - y} \sum_{\substack{n_1,\ldots,n_{\ell-y} \geqslant 0 \\ \Sigma_1^{\ell-y} n_m = y}} \prod_{m=1}^{\ell-y} \binom{d_j}{n_m + 1} \quad 1 \leqslant j \leqslant k. \tag{1}$$

Proof. Consider the table $T^j_{\ell-y,k-1}$ each of whose $(\ell - y)$ rows have the form

$$(t_1, \ldots, t_{j-1}, t_{j+1}, \ldots, t_k).$$

There are $h^j_{\ell-y,k-1}$ such tables, where

$$h^j_{\ell-y,k-1} = \binom{d(j)}{\ell - y} \qquad d(j) = d/d_j.$$

To any such table add n_1 replicates of its first row, n_2 of its second row, ..., $n_{\ell-y}$ replicates of its last row, where $n_m \geqslant 0$ and

$$\sum_{m=1}^{\ell-y} n_m = y$$

to obtain an 'intermediate table'. Then reconstruct a table of the form $T_{\ell k}$ by introducing the j-th column. For the first $(\ell - y)$ positions of the new j-th column any element of D_j may be used. There are thus $(d_j)^{\ell-y}$ possibilities. Thus for the n_m replicates of the m-th row of the 'intermediate table', distinct choices will have to be made: $(d_j - 1)$ for the first replicate, $(d_j - 2)$ for the second and so on, $d_j - n_m$ for the last replicate.

There are thus

$$\prod_{m=1}^{\ell-y} \binom{d_j}{n_m + 1}$$

ways of reconstructing a table of the form $T_{\ell k}$ from a given intermediate table. The total number of tables having k columns and ℓ distinct rows and which yield the same $T^j_{\ell-y,k-1}$ table is therefore

$$\sum_{\substack{0 \leqslant n_1,\ldots,n_{\ell-y} \\ \Sigma_{m=i}^{\ell-y} n_m = y}} \prod_{m=1}^{\ell-y} \binom{d_j}{n_m + 1},$$

hence the result.

[1] $\binom{u}{v}$, $u \geqslant v$, $u!/v!(u - v)!$.

Consequence 3. The probability that the projection $\pi_j(T_{\ell k})$ will be of dimension (no. of lines) x is

$$p^j_{\ell k}(x) = \frac{Q^{j,\ell-x}_{\ell k}}{h_{\ell k}}\ .$$

Proof. It is simply the proportion of tables of the type $T_{\ell k}$ whose projection along the j-th column is of type $T^j_{x,k-1}$.

We would now like to evaluate the size of the projection of a table $T_{\ell,k}$ into a subspace composed of the following columns

$$(1, ..., j_1 - 1, j_1 + 1, ..., j_{s-1}, j_{s+1}, ..., k)$$

obtained by removing columns $(j_1, ..., j_s)$ from $T_{\ell k}$. We shall call this projection

$$\pi_{j_1,...,j_s}(T_{\ell k}).$$

Result 4. The number of tables of the form $T_{\ell k}$ whose projection $\pi_{j_1,...,j_s}(T_{\ell k})$ contains exactly $(\ell - y)$ rows, $0 \leqslant y \leqslant \ell - 1$, is

$$R^{(j_1,...,j_s),y}_{\ell k} = \binom{d/(d_{j_1}, ..., d_{j_s})}{\ell - y} \times \sum_{\substack{n_1,...,n_{\ell-y} \geqslant 0 \\ \Sigma_1^{\ell-y} n_m = y}} \prod_{m=1}^{\ell-y} \binom{d_{j_1}, ..., d_{j_s}}{n_m + 1}. \quad (2)$$

Proof. This is merely a consequence of Result 2. It suffices to notice that the set of tables of the form $T_{\ell k}$ is isomorphic to the set of tables

$$T^{(j_1,...,j_s)}_{\ell,k-s+1}$$

obtained by replacing the columns $j_1, ..., j_s$ of $T_{\ell k}$ by a single column whose elements are chosen from the set $D_{j_1} \times \cdots \times D_{j_s}$. The computation of $\pi_{j_1,...,j_s}(T_{\ell k})$ is then identical to the projection of

$$T^{(j_1,...,j_s)}_{\ell,k-s+1}$$

by the removal of this particular column.

Result 5. The probability that the projection $\pi_{j_1,...,j_s}(T_{\ell k})$ is of dimension (no. of lines) x is

$$p^{j_1,...,j_s}_{\ell k}|(x) = \frac{R^{(j_1,...,j_s),\ell-x}_{\ell,k}}{h_{\ell,k}}\ .$$

3. Computational algorithms

Formulae (1) and (2) (the latter having essentially the same form as (1)) are not computationally very efficient. Let us define

$$X_{a,b}(v) \equiv \sum_{\substack{n_1,...,n_a \geqslant 0 \\ \Sigma_1^a n_m = b}} \prod_{m=1}^{a} \binom{v}{n_m + 1} \quad (3)$$

so that from (1) we have

$$Q^{j,y}_{\ell,k} = \binom{d/d_j}{\ell - y} X_{\ell-y,y}(d_j) \quad (4)$$

and from (2)

$$R^{(j_1,...,j_x),y}_{\ell,k} = \binom{d/d_{j_1}, ..., d_{j_x}}{\ell - y} X_{\ell-y,y}(d_{j_1} d_{j_2}, ..., d_{j_x}). \quad (5)$$

Clearly

$$\binom{y + \ell - y - 1}{\ell - y - 1} = \binom{\ell - 1}{\ell - y - 1}$$

terms have to be computed and added in order to obtain $X_{\ell,y}(\infty, v)$ using (3), and this can be extremely large even for moderate values of ℓ. For instance, for $\ell = 50$ and $y = 20$ we obtain approximately 2.83×10^{13} which is properly astronomical! It is therefore useful and even essential to seek a more efficient computational procedure for $X_{\ell,y}(v)$.

Notice that

$$X_{a,b}(v) = \sum_{n_a=0}^{b} \binom{v}{n_a + 1} \sum_{\substack{n_1,...,n_{a-1} \\ \Sigma_1^{a-1} n_m = b - n_a}} \prod_{m=1}^{a-1} \binom{v}{n_m + 1}$$

$$= \sum_{z=0}^{b} \binom{v}{z + 1} X_{a-1,b-z}(v). \quad (6)$$

3.1. Computation of $\{p^j_{\ell,k}(x)\}_{1 \leqslant x \leqslant \ell}$

The computation of the probability distribution $\{p^j_{\ell,k}(x)\}_{1 \leqslant x \leqslant \ell}$ giving the probability that $\pi_j(T_{\ell k})$ contains x rows requires (see Consequence 3) the computation of $Q^{j,\ell-x}_{\ell,k}$ and hence that of

$$X_{x,\ell-x}(d_j) \qquad 1 \leqslant x \leqslant \ell\ . \quad (7)$$

The proposed algorithm is as follows. After setting all $X(d_j) \leftarrow 0$:

```
$X_{1,0}(d_j) \leftarrow \binom{d_j}{1}$;        $X_{1,1}(d_j) \leftarrow \binom{d_j}{2}$;

for a = 2 to ℓ do $X_{a,0}(d_j) \leftarrow \binom{d_j}{1} * X_{a-1,0}(d_j)$

for a = 2 to ℓ do for b = 1 to ℓ − a do
  for z = 0 to b do

$X_{a,b}(d_j) \leftarrow X_{a,b}(d_j) + \binom{d_j}{z+1} * X_{a-1,b-z}(d_j)$; end all.
```

This algorithm will obtain all $X_{a,b}(d_j)$ with $b \leqslant \ell - a$, and $1 \leqslant a \leqslant \ell$.

There are, of course, $\ell^2/2$ values of the pair (x, y), $1 \leqslant x \leqslant \ell$, $0 \leqslant y \leqslant \ell - x$, for which (7) has to be evaluated in order to compute the probability distribution, $\{p^j_{\ell,k}(x)\}_{1 \leqslant x \leqslant \ell}$. This is obviously because by (1) and (4)

$$p^j_{\ell,k}(x) = \binom{d/d_j}{x} X_{x,\ell-x}(d_j)/h_{\ell,k}. \tag{8}$$

However, for any value of (a, b), $X_{a,b}(d_j)$ is obtained from the values $X_{a-1,b-z}(d_j)$, $0 \leqslant z < b$, in b computational steps. We therefore have a total number of computational steps proportional to

$$\sum_{a=1}^{\ell} \sum_{b=0}^{\ell-a} b = \sum_{a=1}^{\ell} \frac{(\ell - a)(\ell - a + 1)}{2}$$
$$= \tfrac{1}{6}\ell(\ell - 1)(\ell + 1).$$

4. Conclusion

In this note we have examined a problem of interest, and probably of importance in data handling systems or in data base systems: the size of projections of a set of data from a k-dimensional space in which it is given into a smaller subspace. An enumerative approach provides results in the case of uniformly distributed independent samples on a finite dimensional set of possible data values. Certain extensions, in particular to dynamically varying data sets and to certain cases of dependence (e.g. 'functional dependence' in data bases), will be treated in subsequent papers.

Acknowledgements

This work was motivated by discussions with W.A. Armstrong and with Ph. Richard.

References

[1] Ph. Richard, Thèse de Doctorat de 3ème cycle, Orsay (1980).

[2] Ph. Richard, Evaluation of the size of a query expressed in relational algebra, Proc. ACM–SIGMOD International Conf. on Management of Data (1981) pp. 155–163.

[3] R. Demolombe, Estimation of the number of tuples satisfying a query expressed in predicate calculus language, Proc. 6th VLDB Conf. (IEEE Press, 1980) pp. 55–63.

[4] T.H. Merrett and E. Otoo, Distribution models of relations, Proc. 5th VLBD Conf. (IEEE Press, 1979) pp. 418–425.