Inform. Systems Vol. 9, No. 3/4, pp. 231–235, 1984
Printed in the U.S.A.

0306–4379/84 $3.00 + .00
Pergamon Press Ltd.

ON THE SIZES OF PROJECTIONS: A GENERATING FUNCTION APPROACH

DANIÈLE GARDY†
LRI, Université de Paris-Sud, Bât. 490—91405 Irsay Cédex, France

and

CLAUDE PUECH‡
Laboratoire de Recherche en Informatique, E.R.A. 452 du CNRS "Al Khowarizmi", Université de Paris-Sud—Bât. 490—91405 Orsay Cédex, France

(*Received* 11 *October* 1983)

Abstract—We study the distributions of the sizes of projections of tabulated data (relations) on subsets of the columns; this is done, by means of generating functions, under various hypotheses.

1. INTRODUCTION

Given a set of vectors in a k-dimensional space, the study of the sizes of its projections on some of its subspaces is of interest in several contexts:

(i) Several operations in *data base* systems contain the computation of projections as one of their components; the size of the projections obtained has a strong influence on the overall execution time of these operations [2, 7]; some knowledge of the estimated sizes of projections is therefore helpful in order to "optimize" the requests;

(ii) In *data analysis,* projections of the initial data, on well chosen subspaces, are often useful in order to obtain accurate "summaries" of the data;

(iii) Visually meaningful projections into two or three dimensions are also used in *graphic outputs*.

In all of these cases, the number of data points contained in the projection will have an important influence on the storage and on the run time of the processing algorithms.

In [5, 6], Gelenbe and Gardy studied the probability distributions of the sizes of projections of "random" sets of points under various hypotheses; the proofs were of a "counting" nature: the probabilities of given sizes for the projections of a data set were evaluated, using conditional probabilities, and the means of the distributions derived.

We present here a different approach: we give means for deriving the generating functions of the distributions *directly from the definition of* (*and constraints on*) *the set of points* (it is possible to do so both under the hypothesis of uniform distribution of "coordinates" and under an "arbitrary" distribution; we give here the results in the latter case).

This gives alternate proofs for the formulae of [5, 6] giving the probability distributions (in a few cases, it leads to equivalent easier formulae); the moments of the distributions can be computed in a straightforward way from the generating functions; moreover, this approach proved to be fruitful [3, 4] in the study of other elementary operations of relational algebra in data base theory (intersection, equijoin . . .).

We recall the notations of [5, 6]. The k-dimensional space is $\tau_k = D_1 \times D_2 \times \ldots \times D_k$ where each D_i is a finite set (d_i is its size). Subsets of τ_k may be viewed as *tables* of data points or as relations in a relational data base system (D_i is then the domain of the i^{th} attribute). T_{lk} denotes a subset of τ_k of size l, i.e. with l rows, or l records.

We examine *projections* of T_{lk} into subspaces of τ_k. The projection $\prod_{j_1,\ldots,j_u}(t)$ of a vector $t = (t_1, \ldots, t_k)$ is $\prod_{j_1,\ldots,j_u}(t) = (t_{j_1}, \ldots, t_{j_u})$. The projection of a table is the set of projections of its rows. It can be viewed as the table obtained from the original one by deleting some columns and removing, then, duplicated rows. When there is no ambiguity we shorten $\prod_{j_1\ldots j_u}(t)$ in $\prod(t)$ and denote by $\overline{\prod}$ the projection on the product of the domains D_i for $i \notin \{j_1, \ldots, j_u\}$. δ denotes the number of elements of $\prod(\tau_k)$ ($\delta = d_{j_1} \times \ldots \times d_{j_u}$), δ' the number of elements of $\overline{\prod}(\tau_k)$ $\left(\delta' = \dfrac{d_1 \times \ldots \times d_k}{\delta}\right)$.

The probabilistic hypotheses are the following:

(i) distinct components (attributes) of a point (record) are independent;

(ii) the probability of a table (relation) is proportional to the probability of each of its rows (records).

If $t = (t_1, t_2, \ldots, t_k)$, the probability $p(t)$ of t is equal to the product of the probabilities of the t_i. We denote by $p_1, p_2, \ldots, p_\delta$ the probabilities of the δ distinct points of $\prod(\tau_k)$, whereas $\bar{p}_1, \bar{p}_2, \ldots, \bar{p}'_\delta$ denote the probabilities of the points of $\overline{\prod}(\tau_k)$.

† Also at Centre Mondial Informatique et Ressources Humaines, 22 Avenue Matignon—75008 Paris, France.

‡ Also at Ecole Normale Supérieure, 1 rue Maurice Arnoux—92120 Montrouge, France.

In Section 2, we examine the case when there is no functional dependency, and, in Section 3, the case when there is a single functional dependency $x \to y$ (the values of a record on some subset x of the attributes determine in an unique way the values on the "complementary" subset of attributes y).

2. RELATIONS WITHOUT ANY FUNCTIONAL DEPENDENCY

The formal polynomial:

$$P = K \prod_{t \in \tau_k} (1 + p(t)x_t),$$

where

$$K = \frac{1}{\prod_{i=1}^{d} (1 + p_i)}$$

"describes" all the possible tables on D with no functional dependency, in the following way: the monomial $x_{t_1}x_{t_2} \ldots x_{t_l}$ "represents" the table of size 1 whose rows are the tuples $t_1, t_2, \ldots, t_l$; its coefficient in P, $Kp(t_1)p(t_2) \ldots p(t_l)$ is, as mentioned in the introduction, the probability of that table.

We can derive easily from P the value of the probability for a "random" table to be of size l: if we substitute x for each x_t in P, every monomial representing a table of size l is transformed in x^l, so that the coefficient of x^l in the image polynomial is the sum, over all possible tables of sizes l, of the probabilities of these tables. In other words:

Property 1

The probability for a table to be of size l is:

$$\frac{[x^l] \prod_{i=1}^{d} (1 + p_i x)}{\prod_{i=1}^{d} (1 + p_i)}$$

(we denote by $[x^l]f$ the coefficient of x^l in f).

This is a trivial result. But the same method of proof will enable us to prove many other results.

To obtain the generating function for the sizes of projections according to the sizes of the initial relations, we first substitute $x_{\Pi_{j_1 \ldots j_u}(t)}$ for x_t in P: the image of $x_{t_1}x_{t_2} \ldots x_{t_l}$ is $x_{\Pi_1}^{i_1}x_{\Pi_2}^{i_2} \ldots x_{\Pi_r}^{i_r}$, where r is the size of $\prod_{j_1 \ldots j_u}(t)$, $x_{\Pi_1}, \ldots, x_{\Pi_r}$ are the distinct values of $x_{\Pi(t_1)}, \ldots, x_{\Pi(t_l)}$, and $i_1 + i_2 + \cdots + i_r = 1$. Then, we substitute αx^i for $x_{\Pi(t)}^i$ (for all $i \geq 1$) in the resulting polynomial: the image of $x_{\Pi_1}^{i_1}, x_{\Pi_2}^{i_2}, \ldots, x_{\Pi_r}^{i_r}$ is $x^l\alpha^r$, so that the coefficient of $x^l\alpha^r$ in the image polynomial is the probability for a table to be of size l and to have a projection $\prod_{j_1 \ldots j_u}(T_k)$ of size r.

The first substitution $(x_t \to x_{\Pi_{j_1} \ldots \Pi_{j_u}(t)})$ in P gives:

$$K \prod_{\Pi \in D_j \times \ldots \times D_{j_u}} \prod_{\substack{t \in \tau_k \\ \Pi(t) = \Pi}} (1 + p(t)x_{\Pi_i}).$$

The second one $(x_{\Pi_i}^i \to \alpha x^i)$ leads to the polynomial, in the variables α and x:

$$Q(\alpha, x) = \Theta_{\alpha;x_{\Pi_1},x}\Theta_{\alpha;x_{\Pi_2},x} \ldots \Theta_{\alpha;x_{\Pi_\delta},x} \times \{K \prod_{\Pi \in D_{j_1} \times \ldots \times D_{j_u}} \prod_{\substack{t \in \tau_k \\ \Pi(t) = \Pi_i}} (1 + p(t)x_{\Pi_i})\}$$

where $\Theta_{\alpha;x_j,x}(R)(\alpha, x)$ is the polynomial obtained from polynomial $R(x_j)$ by substituting αx^i for x_j^i (for all $i \geq 1$), i.e.

$$\Theta_{\alpha;x_j,x}(R)(\alpha, x) = R(0) + \alpha[R(x) - R(0)]$$

(we abbreviate $\Theta_{\alpha;x_j,x}$ in Θ_{α,x_j} in the sequel of Section 2).

Let $P_{l,r}$ denote the probability for a table of size l to have a projection of size r. As a conditional probability $P_{l,r}$ can be written as the quotient of the probability for a table to be of size l *and* to have a projection of size r (which we proved to be equal to $[x^l\alpha^r]Q(\alpha, x)$) by the probability for a table to be of size l, which can be viewed, as is usual, as the coefficient of x^l in $Q(l, x)$. Thus, by an easy rewriting (as $p(t) = p(\prod(t))\bar{p}(\prod(t))$):

Property 2

The probability, for a table of size l, to have a projection of size r is:

$$P_{l,r} = \frac{[x^l\alpha^r]Q(\alpha, x)}{[x^l]Q(1, x)} = \frac{[x^l\alpha^r]\Theta_{\alpha,x_1} \ldots \Theta_{\alpha,x_\delta}\left[\prod_{i=1}^{\delta}\prod_{j=1}^{\delta'}(1 + p_ip_jx_i)\right]}{[x^l]\prod_{i=1}^{\delta}\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j)}.$$

The numerator can be evaluated in two different ways.

First, we can remark that if $R(x_1, \ldots, x_n)$ can be written as a product of polynomials in one single variable:

$$R(x_1, x_2, \ldots, x_n) = R_1(x_1)R_2(x_2) \ldots R_n(x_n)$$

then:

$$\Theta_{\alpha,x_1}\Theta_{\alpha,x_2} \ldots \Theta_{\alpha,x_n}(R) = \Theta_{\alpha,x_1}(R_1)\Theta_{\alpha,x_2}(R_2) \ldots \Theta_{\alpha,x_n}(R_n)$$

As a consequence:

$$\Theta_{\alpha,x_1} \ldots \Theta_{\alpha,x_\delta}\left[\prod_{i=1}^{\delta}\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x_i)\right] = \prod_{i=1}^{\delta}\left[1 - \alpha + \alpha\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x)\right]$$

$(\Theta_{\alpha,x}[\prod(1 + p(t)x)] = 1 + \alpha[\prod(1 + p(t)x) - 1])$.

This proves the following expression for $P_{l,r}$:

Property 3

The probability for a table of size l to have a projection of size r is:

$$P_{l,r} = \frac{[x^l\alpha^r]\prod_{i=1}^{\delta}\left[1 - \alpha + \alpha\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x)\right]}{[x^l]\prod_{i=1}^{\delta}\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x)}.$$

Another way of evaluating

$$\Theta_{\alpha,x_1} \ldots \Theta_{\alpha,x_\delta}\left[\prod_{i=1}^{\delta}\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x_i)\right]$$

is by using the following lemma:

Lemma 1

If $R(x_1, \ldots, x_n)$ is a polynomial in the variables $x_1, \ldots, x_n$:

$$\Theta_{\alpha;x_1,x} \ldots \Theta_{\alpha;x_n,x}(R) = \sum_{k=0}^{n} R_k(x)\alpha^k(1 - \alpha)^{n-k}$$

where $R_k(x)$ is the sum, over all possible choices of $i_1 < i_2 < \cdots < i_k$, of the polynomials obtained from R by substituting x for x_i if $i \in \{i_1, \ldots, i_k\}$, 0 for x_i otherwise.

The lemma can be proved easily from the definition of Θ by induction on the number of variables in R.

It leads to:

$$Q(\alpha, x) = K\sum_{k=0}^{\delta} Q_k(x)\alpha^k(1 - \alpha)^{\delta-k}$$

with:

$$Q_k(x) = \sum_{1\le i_1<i_2<\cdots<i_k\le\delta}\ \prod_{\substack{1\le i\le\delta\\ i\in\{i_1,i_2,\ldots,i_k\}}}\ \prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x).$$

This implies a new expression for $P_{l,r}$:

Property 4

The probability for a table of size l to have a projection of size r is:

$$P_{l,r} = \sum_{k=0}^{r}(-1)^{r-k}\binom{\delta - k}{r - k} \times \frac{[x^l]\left\{\sum_{1\le i_1<\cdots<i_k\le\delta}\prod_{m=1}^{k}\prod_{j=1}^{\delta'}(1 + p_{i_m}\bar{p}_j x)\right\}}{[x^l]\prod_{i=1}^{\delta}\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x)}.$$

The generating function approach gives very easily by successive differentiations the moments of the distributions. As particular cases, the mean M_l and variance V_l of the distribution of the sizes of projections for tables of size l are given by:

$$M_l = \frac{[x^l]\frac{\partial}{\partial\alpha}Q(\alpha, x)\,|_{\alpha=1}}{[x^l]Q(1, x)}$$

$$V_l = \frac{[x^l]\frac{\partial^2}{\partial\alpha^2}Q(\alpha, x)\,|_{\alpha=1}}{[x^l]Q(1, x)} + \frac{[x^l]\frac{\partial}{\partial\alpha}Q(\alpha, x)\,|_{\alpha=1}}{[x^l]Q(1, x)} - \left[\frac{[x^l]\frac{\partial}{\partial\alpha}Q(\alpha, x)\,|_{\alpha=1}}{[x^l]Q(1, x)}\right]^2.$$

As:

$$Q(\alpha, x) = K\sum_{k=0}^{\delta} Q_k(x)\alpha^k(1 - \alpha)^{\delta-k}$$

the values of Q and its derivatives at $\alpha = 1$ take simple forms:

$$Q(1, x) = KQ_\delta(x)$$

$$\frac{\partial}{\partial\alpha}Q(\alpha, x)\,|_{\alpha=1} = K[\delta Q_\delta(x) - Q_{\delta-1}(x)]$$

$$\frac{\partial^2}{\partial\alpha^2}Q(\alpha, x)\,|_{\alpha=1} = K[\delta(\delta - 1)Q_\delta(x) - 2(\delta - 1)Q_{\delta-1}(x) + 2Q_{\delta-2}(x)].$$

By substituting these expressions in M_l and V_l, we obtain:

Property 5

The distribution of the sizes of projections of tables of size l has mean:

$$M_l = \delta - \frac{[x^l]\left[\sum_{m=1}^{\delta}\prod_{i\neq m}\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x)\right]}{[x^l]\prod_{i=1}^{\delta}\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x)}$$

and variance:

$$V_l = \frac{[x^l]\sum_{m=1}^{\delta}\prod_{i\neq m}\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x)}{[x^l]\prod_{i=1}^{\delta}\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x)} - \left\{\frac{[x^l]\sum_{m=1}^{\delta}\prod_{i\neq m}\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x)}{[x^l]\prod_{i=1}^{\delta}\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x)}\right\}^2 + 2.\frac{[x^l]\sum_{1\leq m_1<m_2\leq\delta}\prod_{i\neq m_1 i\neq m_2}\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x)}{[x^l]\prod_{i=1}^{\delta}\prod_{j=1}^{\delta'}(1 + p_i\bar{p}_j x)}$$

As particular case, when the values of the attributes are uniformly distributed over their domains (i.e. $p_i = 1/\delta$ for all i, $\bar{p}_j = 1/\delta'$ for all j) we obtain the following properties:

Property 6

Under the uniform hypothesis, the probability for a table of size l to have a projection of size r is

$$P_{l,r} = \frac{\binom{\delta}{r}}{\binom{d}{l}}\sum_{k=0}^{r}(-1)^{r-ka}\binom{r}{k}\binom{k\delta'}{l}.$$

Property 7

Under the uniform hypothesis, the distribution of the sizes of projections of tables of size 1 has mean:

$$M_1 = \delta\left[1 - \frac{\binom{d-\delta'}{l}}{\binom{d}{l}}\right]$$

and variance:

$$V_1 = \delta^2\left[\frac{\binom{d-2\delta'}{l}}{\binom{d}{l}} - \frac{\binom{d-\delta'}{l}^2}{\binom{d}{l}^2}\right] + \delta\left[\frac{\binom{d-\delta'}{l}}{\binom{d}{l}} - \frac{\binom{d-2\delta'}{l}}{\binom{d}{l}}\right]$$

3. THE CASE OF A SINGLE FUNCTIONAL DEPENDENCY.

As in [6], we consider relations T_{lk} satisfying a single functional dependency $x \rightarrow y$ (where x and y are disjoint sets of attributes which partition the set of all the attributes, in short "$x \cap y = \phi$, $x \cup y = t$") and we study the distribution of the sizes of projections on the y attributes (because of the functional dependency $\prod_x(T_{lk})$ has the same size than T_{lk}).

The formal polynomial

$$P_1 = K_1\prod_{i=1}^{\delta}(1 + p_i x_i(p_1 y_1 + \bar{p}_2 y_2 + \cdots + \bar{p}'_{\delta} y'_{\delta}))$$

where

$$K_1 = \prod_{i=1}^{\delta}\frac{1}{(1 + p_i)}$$

describes all the tables on D satisfying the functional dependency: a monomial $x_{i_1}y_{j_{i_1}}x_{i_2}y_{j_{i_2}}\ldots x_{i_l}y_{j_{i_l}}$ represents the table of size l whose rows are the tuples associated to (i_1, j_{i_1}), (i_2, j_{i_2}), $\ldots$, (i_l, j_{i_l}); every x is associated to one and only one y, which expresses the functional dependency; the coefficient of the monomial associated to a table is equal to the probability of the table.

To obtain the generating function for the sizes of projections on y, we first substitute x for every x_i in P_1 (x "marks" the size of the tables), then substitute α for every y_i^j (α marks the size of the projections).

This leads to consider the polynomial:

$$Q_1(\alpha, x) = \Theta_{\alpha;y_1,1}\,\Theta_{\alpha;y_2,1}\ldots\Theta_{\alpha;y_{\delta'},1}\left[\prod_{i=1}^{\delta}\left(1 + p_i x\sum_{j=1}^{\delta'}\bar{p}_j y_j\right)\right].$$

The probability, for a table of size l, to have a projection on y of size r is:

$$P_{1,r}^l = \frac{[x^l\alpha^r]Q_1(\alpha, x)}{[x^l]\prod_{i=1}^{\delta}(1 + p_i x)}.$$

By linearity of Θ:

$$[x^l\alpha^r]Q_1(\alpha, x) = (\sum p_{i_1}, \ldots, p_{i_l})[\alpha^r]\,\Theta_{\alpha;y_1,1}\ldots\Theta_{\alpha;y_{\delta'},1}\,(p_1y_1 + \cdots + \bar{p}_{\delta'}y_{\delta'})^l$$

so that the expression of $P_{l,r}^1$ reduces to the simpler form given in:

Property 8

The probability, for a table of size l, with functional dependency $x \rightarrow y (x \cap y = \phi, x \cup y = t)$, to have a

projection on y of size r is:

$$P^1_{l,r} = [\alpha^r]Q_{\alpha;y_1,1}, \ldots, \Theta_{\alpha;y_1,1}$$
$$[(\bar{p}_1 y_1 + \cdots + \bar{p}'_{\delta} y_{\delta'})^l]$$

By lemma 1, this can be expressed as:

Property 9

The probability, for a table of size l, with functional dependency $x \to y (x \cap y = \phi, x \cup y = t)$, to have a projection on y of size r is:

$$P^1_{l,r} = \sum_{k=1}^{r} (-1)^{r-k} \binom{\delta' - k}{r - k} \sum_{1 \le i_1 < \cdots < i_k \le \delta'} (\bar{p}_{i_1} + \cdots + \bar{p}_{i_k})^l$$

The mean and variance of the distribution can be calculated as in Section 1. We obtain:

Property 10

The distribution of the sizes of projections on y, for tables of size l (with a single functional dependency $x \to y; x \cap y = \phi; x \cup y = t$) has mean:

$$M_l = \delta' - \sum_{j=1}^{\delta'} (1 - p_j^-)^l$$

and variance:

$$V_l = \sum_{j=1}^{\delta'} (1 - p_j^-)^l - \left[\sum_{j=1}^{\delta'} (1 - \bar{p}_j)^l\right]^2 + 2 \sum_{1 \le i < j \le \delta'} (1 - \bar{p}_i - \bar{p}_j)^l.$$

When the values of the attributes are uniformly distributed over their domains, $\bar{p}_j = 1/\delta'$ and the above properties "reduce" to:

Property 11

Under the uniform hypothesis, the probability, for a table of size l, with functional dependency $x \to y (x \cap y = \phi, x \cup y = t)$, to have a projection on y of size r is:

$$P^1_{l,r} = \frac{\binom{\delta'}{r}}{\delta'^l} \cdot \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} k^l$$
$$= \frac{\delta'(\delta' - 1) \ldots (\delta' - 2 + l)}{\delta'^l} \cdot S_{l,r}$$

where $S_{l,r}$ is a Stirling number of second kind ([1], p. 204). (This result can be given a combinatorial proof in view of the fact that $r!S(l, r)$ enumerates the surjective maps from $\{1, \ldots, l\}$ into $\{1, \ldots, r\}$).

Property 12

Under the uniform hypothesis, the distribution of the sizes of projections on y, for tables of size l (with a single functional dependency $x \to y; x \cap y = \phi; x \cup y = t$) has mean:

$$M_l = \delta' \left[1 - \left(1 - \frac{1}{\delta'}\right)^l\right]$$

and variance:

$$V_1 = \delta'^2 \left[\left(1 - \frac{2}{\delta'}\right)^l - \left(1 - \frac{1}{\delta'}\right)^{2l}\right] + \delta' \left[\left(1 - \frac{1}{\delta'}\right)^l - \left(1 - \frac{2}{\delta'}\right)^l\right].$$

4. CONCLUSION.

The generating function approach proved very useful because it allows to construct in a *systematic* way the generating functions involved.

Projection is only one among the basic operations of relational data bases. Other ones (intersection, join . . .) can also be studied using the same approach [3, 4].

Another problem of interest is the study of the sizes of relations obtained from an initial one by a sequence of operations. Generating functions seem to be useful in this context too.

Lastly let us mention that some changes in the probabilistic model could be reflected into the generating functions.

Acknowledgements—The authors thank Erol Gelenbe and Philippe Flajolet for helpful discussions.

REFERENCES

[1] L. Comtet: *Advanced Combinatorics*. Reidel (1974).
[2] R. Demolombe: Estimation of the number of tuples satisfying a query expressed in predicate calculus language. *VLDB* 80, Montréal (1980).
[3] D. Gardy: Evaluation de résultats d'opérations de l'algèbre relationnelle. Thèse de 3ème cycle, Université de Paris-Sud (Février 1983).
[4] D. Gardy and C. Puech: Operations of relational algebra and sizes of relations. Proc. 11th ICALP, Antwerp, Belgium, pp. 174–186 (July 1984).
[5] E. Gelenbe and D. Gardy: On the sizes of projections I. *IPL* **14**(1), pp. 18–21 (1982).
[6] E. Gelenbe and D. Gardy: On the sizes of projections II. *VLDB* 82, Mexico (Sept. 82).
[7] Ph. Richard: Evaluation of the size of a query expressed in relational algebra. *ACM-SIGMOD Int. Conf. on Management of Data*, pp. 155–163 (April 1981).