The Size of Projections of Relations Satisfying a Functional Dependency

Erol GELENBE

ISEM Université Paris Sud
Bât 508, 91405 Orsay, France,
and Centre Mondial Informatique

Danièle GARDY

Ecole Polytechnique, Centre
de Mathématiques Appliquées
91000 Palaiseau, France

ABSTRACT

In this paper we consider tabulated data or relations in a data base system which are constrained by functional dependencies. This implies that the data in certain columns of each table is determined by the data contained in some other columns. The problem we address is that of the computation of the size of projections of the data on a subset of the columns. This may be viewed as the projection of data in some k dimensional space into a smaller subspace. We thus extend results we had previously obtained [1] for relations without functional dependencies to the case with functional dependencies.

RESUME

Nous nous intéressons au problème du calcul de la taille des projections de relations dans une base de données relationnelle. Les résultats que nous avons précédemment obtenus [1] pour le cas sans dependences fonctionnelles sont rappelés: une nouvelle formule pour la taille moyenne d'une projection est donnée. Nous obtenons ensuite des formules pour la distribution et la taille moyenne de projections de relations en présence d'une dépendance fonctionnelle.

1. Introduction

Consider τ_k the k-dimentional space of vectors of the form

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

where each t_i takes its value in a finite set D_i. Thus $\tau_k = D_1 \times D_2 \times \ldots \times D_k$.
We shall be interested in subsets of τ_k which may be viewed as *tables* of data points or as *relations* in a rational data base system. Let such a subset be

$$T_{\ell k} \subset \tau_k$$

where $|T_{\ell k}| = \ell$ (i.e. $T_{\ell k}$ contains ℓ vectors of τ_k).

We will examine *projections* of $T_{\ell k}$ into subspaces of τ_k.

The projection $\pi_{j_1 \ldots j_u}$ of a vector $t \in \tau_k$ is the u-dimentional vector

$$\pi_{j_1 \ldots j_u}(t) = (t_{j_1}, \ldots, t_{j_u})$$

where $j_i \in (1, \ldots, k)$. Similarly, we define the projection of the table or relation $T_{\ell k}$ as

$$\pi_{j_1 \ldots j_u}(T_{\ell k}) = \{\pi_{j_1 \ldots j_u}(t) : t \in T_{\ell k}\}$$

As in a previous paper [1], we are interested in the *size of projections.* There are several reasons which motivate this interest. Several operations of interest in data base system often contain the computation of projections as one of their components. The time necessary for the execution of such operations will thus be determined in part by the size of the projections obtained [2,3]. In another application area, related to data analysis (statistical data, physics experiments, etc.) projections of the initial data are obtained in order to examine properties of interest. In other cases, if a graphics output is used, projections of the data into two or three dimensions will often be used in order to obtain a visually meaningful presentation. In all of these cases, the number of data points contained in the projection will have an important influence on the storage necessary or on the run time of the processing algorithms used.

Often the information contained in each of the columns of a given table or relation $T_{\ell k}$ will not be independent. One type of restriction common to data bases are the well known *functional dependencies.*

A *functional dependency* will be denoted by

$$x \rightarrow y \quad \text{or} \quad f(x,y)$$

and expressed by "x implies y" where x and y are subvectors of the vector t:

$$x = (t_{x_1}, \ldots, t_{xX}),\ t_{xi} \in \{t_1, \ldots, t_k\}$$

$$y = (t_{y_1}, \ldots, t_{yY}),\ t_{yj} \in \{t_1, \ldots, t_k\}$$

We say that the $T_{\ell k}$ satisfies $x \rightarrow y$ if and only if for all $t, t' \in T_{\ell k}$,

$$\pi_{x_1, \ldots, xX}(t) = \pi_{x_1, \ldots, xX}(t') \Rightarrow \pi_{y_1, \ldots, yY}(t) = \pi_{y_1, \ldots, yY}(t')$$

A typical exemple would be the case of the data base of a government organisation's employees in which the rank and seniority would determine the salary: (RANK,SENIORITY) $\rightarrow$ (SALARY).

In the general case we can assume that $T_{\ell k}$ satisfies a family F of functional dependencies

$$F = \{f(x,y) : x, y, \text{subvectors of } t\}$$

The problem we shall address in this paper is the following.

Suppose that data tables of the form $T_{\ell k}$ are generated in some "random" manner. Then what is the probability distribution of the size of the projection $\pi_{j_1, \ldots, j_n}(T_{\ell k})$ given that ℓ (the size of the table $T_{\ell k}$) is known ?

In [1] we solved this problem in the absence of functional dependencies, and we provided an efficient computational algorithm to obtain this probability distribution. The assumption made was that the tables $T_{\ell k}$ are generated at "random" with a uniform distribution.

Here we shall make a probabilistic assumption but consider the case where functional dependencies hold.

In section 2 we shall recall the main result obtained in [1]; we shall also give a new result providing a closed form expression for the *average size* of a projection in the absence of functional dependencies.

In section 3 we shall consider the simplest case of a single functional dependency. It will be analysed both for uniform and non-uniform distributions of attribute values on the domains. Again, formulae for the average size of the projection will be given together with the probability distribution.

2. Results obtained for a system without functional dependencies

In a previous paper [1] we had derived the probability distribution of the size of

$$\Pi_{j_1,\ldots,j_u}(T_{\ell k})$$

which is the projection of the relation $T_{\ell k}$ on coordinates $(j_1,\ldots,j_u)$.
We had also provided an efficient computational algorithm allowing us to compute any particular value of this distribution in time ℓ^3.

The basic assuption concerning this "probabilistic" analysis was that any $T_{\ell k}$ in τ_k is generated at random by choosing any ℓ distinct vectors $(t_1,\ldots,t_k)$ among the $d = d_1 \ldots d_k$ possibilities[1)] with equal probability. Furthermore it was assumed that any one of the coordinates t_i is uniformly distributed over D_i, and that the coordinates are independent.

In this section we shall conserve the same assuptions. We first recall the main result in [1], and then provide a new formula for the average size of projections.

Throughout this section we assume that all the elements of any given domain D_i are equally likely to occur in tuple t of a relation (uniform distribution assumption). We also assume that no functional dependency constrains the relations.

Let the probability and average value be denoted by:

$$P_{\ell,k}^{j_1,\ldots,j_u}(r) = P[\text{size}(\Pi_{j_1\ldots j_u}(T_{\ell k}))=r]$$

$$E_{\ell,k}^{j_1,\ldots j_u} = E[\text{size}\ (\Pi_{j_1\ldots j_u}(T_{\ell k}))]$$

Then we have the following result proved in [1]:

RESULT 1.

$$P_{\ell,k}^{j_1\ldots j_u}(r) = \frac{\binom{d_{j_1}\ldots d_{j_u}}{r}}{\binom{d}{\ell}} X_{r,\ell-r}(d/d_{j_1}\ldots d_{j_u})$$

where we define

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

The following formula is new. It provides an efficient tool for computing the average size of a projection.

RESULT 2.

Let $\delta = d_{j_1} \ldots d_{j_u}$, $\delta' = d/\delta$. Then

$$E^{j_1 \ldots j_u}_{\ell,k} = \delta \left[1 - \frac{\binom{d-\delta'}{\ell}}{\binom{d}{\ell}}\right] ,$$

$$\cong \ell\left[1 - \frac{1}{2\delta} (\ell - 1)\right]$$

for $\ell \ll \delta \ll d$ (i.e. for "small" relations).

Proof

$$E^{j_1 \ldots j_u}_{\ell,k} = \sum_{r=1}^{\ell} r \, . \, p^{j_1 \ldots j_u}_{\ell,k}(r)$$

$$= \sum_{r=1}^{\ell} r \frac{\binom{\delta}{r}}{\binom{d}{\ell}} X_{r,\ell-r}(\delta')$$

$$= \frac{\delta}{\binom{d}{\ell}} \sum_{r=1}^{\ell} \binom{\delta-1}{r-1} X_{r,\ell-r}(\delta')$$

$$X_{1,\ell-1}(\delta') = \binom{\delta'}{\ell}$$

for $r > 1$: $X_{r,\ell-r}(\delta')$

$$= \sum_{z=0}^{\ell-r} \binom{\delta'}{z+1} X_{r-1,\ell-r-z}(\delta')$$

Hence:

$$E^{j_1 \ldots j_u}_{\ell,k} = \frac{\delta}{\binom{d}{\ell}} \left[\binom{\delta'}{\ell} + \sum_{\substack{2 \le r \le \ell \\ 0 \le z \le \ell - r}} \binom{\delta-1}{r-1} \binom{\delta'}{z+1} X_{r-1,\ell-r-z}(\delta')\right] \Rightarrow$$

$$= \frac{\delta}{\binom{d}{\ell}} \left[\binom{\delta'}{\ell} + \sum_{z=0}^{\ell-2} \binom{\delta'}{z+1} \sum_{s=1}^{\ell-z-1} \binom{\delta-1}{s} X_{s,(\ell-z-1)-s}(\delta')\right]$$

By noting that:

$$\sum_{s=1}^{\ell} P^{j_1 \ldots j_u}_{\ell,k}(s)$$

We obtain:

$$\sum_{s=1}^{\ell-z-1} \binom{\delta-1}{s} X_{s,\ell-z-1-s}(\delta') = \binom{\delta'(\delta-1)}{\ell-z-1}$$

Then :

$$E^{j_1 \ldots j_u}_{\ell,k} = \frac{\delta}{\binom{d}{\ell}} \left[\binom{\delta'}{\ell} + \sum_{z=0}^{\ell-2} \binom{\delta'}{z+1} \binom{d \;\; \delta'}{\ell-z-1}\right]$$

$$= \frac{\delta}{\binom{d}{\ell}} \cdot \sum_{z=1}^{\ell} \binom{\delta'}{z} \binom{d-\delta'}{\ell-z}$$

Now, by applying the binominal formula to $(1+X)^d$ written as $(1+X)^{\delta'} . (1+X)^{d-\delta'}$, we obtain for $0 \le \ell \le d$:

$$\sum_{\substack{0 \le s \le \delta' \\ 0 \le r \le d-\delta' \\ s + r = \ell}} \binom{\delta'}{s} \binom{d-\delta'}{r} = \binom{\delta}{\ell}$$

For $\delta \ge \ell$, we can write it as:

$$\binom{d}{\ell} = \sum_{s=0}^{\ell} \binom{\delta'}{s} \binom{d-\delta'}{\ell-s}$$

Hence the result:

$$E_{\ell,k}^{j_1 \cdots j_u} = \frac{\delta}{\binom{d}{\ell}} \left[\binom{d}{\ell} - \binom{d-\delta'}{\ell}\right]$$

The proof of the approximate formula is then obtained as follows.
Clearly

$$\binom{d-\delta'}{\ell} \Big/ \binom{d}{\ell} = \frac{(d-\delta'-\ell+1)\ldots(d-\delta')}{(d-\ell+1)\ldots d}$$

$$= (1-\frac{\delta'}{d})^{\ell} \frac{(1-\frac{1}{d-\delta'})\ldots(1-\frac{\ell-1}{d-\delta'})}{(1-\frac{1}{d})\ldots(1-\frac{\ell-1}{d})}$$

$$\cong (1-\frac{\ell}{\delta} + \frac{\ell(\ell-1)}{2\delta^2})(1-\frac{\ell(\ell-1)}{2(d-\delta')})(1+\frac{\ell(\ell-1)}{2d})$$

$$\cong (1-\frac{\ell}{\delta} + \frac{\ell(\ell-1)}{2\delta^2})(1-\frac{\ell(\ell-1)}{2d}(1+\frac{1}{\delta})(1+\frac{\ell(\ell-1)}{2d})$$

$$\cong (1-\frac{\ell}{\delta} + \frac{\ell(\ell-1)}{2\delta^2})(1-\frac{\ell(\ell-1)}{2d_\delta})$$

where we have used the assumption $\ell << \delta << d$. The approximate formula then follows directly.

3. The case of a single functional dependency

In this section we shall consider a relation $T_{\ell k}$ satisfying a single functional dependency $x \to y$. Without loss of generality we assume that x and y are disjoint. We shall examine both the case of uniform and non-uniform distributions of the values of the attributes on the domains.

The problem of computing the size of the projection $\Pi_{xy}(T_{\ell k})$ on the set of the columns (x,y) is identical to the case without functional dependencies; this is in fact the case for any projection of the form $\Pi_{xyz}(T_{\ell k})$.

Thus in this section we shall concentrate on the size of $\Pi_y(\Pi_{xy}(T_{\ell k}))$. Using the formula for conditional probabilities we have

$$P[|\Pi_y(\Pi_{xy}(T_{\ell k}))| = r]$$

$$= P[|\Pi_y(\Pi_{xy}(T_{\ell k})) = r / |\Pi_{xy}(T_{\ell k})| = j]$$

$$\cdot P[|\Pi_{xy}(T_{\ell k})| = j]$$

where the second term on the right hand side is available from RESULT 1. Thus it suffices to compute the conditional probability. The formulae derived in this section provide this conditional probability in the case of uniform and non-uniform distributions of attribute values over the domains.

Indeed, we notice that if $T_{\ell k}$ satisfies $x \to y$, then the size of $\Pi_{xy}(T_{\ell k})$ is the same as that of $\Pi_x(T_{\ell k})$. Therefore it suffices to replace $|\Pi_{xy}(T_{\ell k})| = j$ by $|\Pi_x(T_{\ell k})| = j$ in the above formula.
The probability

$$P[|\Pi_x(T_{\ell K})| = j]$$

is then simply computed by setting $x = (t_{j_1}, \ldots t_{j_u})$, r=j, in RESULT 1.

3.1. Uniform distributions

In this section we assume that all attribute values are equally likely (uniform distributions).

RESULT 3.

The number of distinct tables of size m on columns (t_i, t_j) satisfying $t_i \to t_j$, whose projection on the j-th is of size n is

$$\alpha^{ij}_{mn} = \binom{d_j}{n}\binom{d_i}{n} \sum_{\substack{m_1,\ldots,m_n \geq 1 \\ \Sigma_1^n m_i = n}} \frac{m}{m_1!\ldots m_n!}$$

Proof : There are $\binom{d_j}{n}$ possible choices of the column on (t_j).
Once this is done we can choose any m distinct elements among the d_i: the number of distinct choices is $\binom{d_i}{m}$.
We will then have to associate $m_1 \geq 1$ of these to the first element of the (t_j) column,...,$m_n \geq 1$ to the n-th element of the (t_j) column. Clearly we must have $m_1+\ldots+m_n = m$ and the number of distinct possibilities is simply for a fixed choice of $m_1,\ldots,m_n$:

$$\binom{m}{m_1}\binom{m-m_1}{m_2}\cdots\binom{m-\ldots-m_{n-1}}{m_n} = \frac{m!}{m_1!\ldots m_n!}$$

hence the result.

COROLLARY 4.

Let x,y be subvectors of t such that $x \cup y = t$, $x \cap y = \emptyset$.
Let $T_{\ell k}$ be a relation (on t) satisfying $x \to y$. Assuming that, for a given ℓ, all the $T_{\ell k}$ are equally likely to occur, the probability that $\Pi_y(T_{\ell k})$ is of size r is:

$$P^y_{\ell k}(r) = \frac{\binom{d_y}{r}}{(d_y)^\ell} \sum_{\substack{m_1,\ldots m_r \geq 1 \\ \Sigma_1^r m_i = \ell}} \frac{\ell!}{m_1!\ldots m_r!}$$

(where $d_y = \prod_{t_i \in y} d_i$, $d_x = \prod_{t_i \in x} d_i$)

Proof: This is in fact the consequence of RESULT 3 since there are

$$\binom{d_x}{\ell} \quad (d_y)^\ell$$

distinct such tables $T_{\ell k}$. Therefore

$$P^y_{\ell k}(r) = \frac{\alpha^{xy}_r}{\binom{d_x}{\ell}(d_y)^\ell}$$

RESULT 5 $E^y_{\ell k}$ the average size of $\Pi_y(T_{\ell k})$ if $x \cup y = t$, $x \cap y = \emptyset$, and the $T_{\ell k}$ is a relation on t, is

$$E^y_{\ell k} = d_y \left[1 - \left(\frac{d_y - 1}{d_y}\right)^\ell\right]$$

$$\cong \ell - \frac{\ell^2}{2}\frac{1}{d_y} \quad \text{for } \ell \ll d_y$$

so that the relative reduction in size is, on the average

$$\frac{1}{\ell}(\ell - E^y_{\ell k}) \approx \frac{\ell}{2}\cdot\frac{1}{d_y} \quad \text{for } \ell \ll d_y$$

Proof: This formula can be derived somewhat laboriously directly from COROLLARY 4: in fact this is exactly how we have initially discovered it. We shall give a simple indirect proof, however. Let D_y denote the domain of y, and let e_y be any one of its elements. Clearly we may write

$$E^{y}_{\ell,k} = \sum_{e_y \in D_y} E(1(e_y \in \Pi_y(T_{\ell k})))$$

where E(.) denotes the expectation operator, and 1(.) is the characteristic function taking the value 1 if its argument is true and 0 otherwise.

If $T_{\ell k}$ satisfies $x \to y$ we know that all of the elements of its x-column must be distinct: otherwise if any two elements were the same, the corresponding y-column elements would have to be the same and $T_{\ell k}$ would contain two identical rows which is impossible. On the other hand there may be an arbitrary number of repetitions in the y-column.

Thus the y-column of $T_{\ell k}$ is obtained simply by drawing ℓ elements e_y from D_y with repetitions allowed.

We know that

$$e_y \notin \Pi_y(T_{\ell k}) \iff e_y \notin [\text{y- column of } T_{\ell k}]$$

so that the probability of these two events is the same. Hence

$$P[e_y \notin \Pi_y(T_{\ell k})] = (1-\frac{1}{d_y})^{\ell}$$

which is the probability that e_y will not be drawn in the ℓ trials, since $1/d_y$ is the probability of drawing e_y. But we then have

$$E(1(e_y \in \Pi_y(T_{\ell k}))) = P[e_y \in \Pi_y(T_{\ell k})]$$

$$= 1 - P[e_y \notin \Pi_y(T_{\ell k})] = 1 - (1-\frac{1}{d_y})^{\ell}$$

Hence

$$E^{y}_{\ell,k} = \sum_{e_y \in D_y} [1-(1-\frac{1}{d_y})^{\ell}]$$

$$= d_y[1-(1-\frac{1}{d_y})^{\ell}]$$

since $|D_y| = d_y$. The approximate formula for $\ell << d_y$ follows from a second order expansion.

3.2. Non-uniform distributions

In many cases of interest uniform distributions over all the tuples are not justified. Take for instance the case of $T_{\ell k}$ with $x \to y$, $x \cup y = t$, $x \cap y = \emptyset$. We can think of x as being a key or numbering, while y can represent a content. In this case a uniform distribution on D_y is difficult to justify.

Here we shall generalize the results of Section 3.1 to the case where we are given an arbitrary distribution on the elements of D_y:

$$p(e_y), \quad e_y \in D_y$$

We have an immediate generalization of COROLLARY 4; the proof is very similar.

RESULT 7

$$E^{y}_{\ell,k} = \sum_{e_y \in D_y} [1-(1-p(e_y))^{\ell}]$$

The probability distribution of the size of $\Pi_y(T_{\ell k})$ can also be obtained:

RESULT 8

$$P^{y}_{\ell,k}(r)$$

$$= \sum_{\substack{(e^1,\dots e^r) \\ \in (D_y)^r}} \sum_{\substack{n_i \geq 1 \\ \Sigma_1^r n_i = \ell}} \frac{\ell!}{n_1!\dots n_r!} \prod_{i=1}^{r} (p(e^i))^{n_i}$$

where $(D_y)^r = D_y \times\dots\times D_y$ r times and $(e^1,\dots,e^r)$ is any vector of r distinct elements of D_y. Notice that this reduces to RESULT 3 when $p(e^i) = 1/d_y$.

Proof: $P^{y}_{\ell,k}(r)$ is the probability that $\Pi_y(T_{\ell k})$ contains exactly any r elements of D_y where $r \in \ell$. The probability that it contains n_i replicates of a given $e^i \in D_y$, $1 \leq i \leq r$, is

$$\binom{\ell}{n_1}\left(p(e^1)\right)^{n_1}\binom{\ell-n_1}{n_2}\left(p(e^2)\right)^{n_2}\dots\binom{\ell-n_1-\dots-n_{r-1}}{n_r}$$

$$\cdot\left(p(e^r)\right)^{n_r}$$

We must have $n_i \geq 1$, $\Sigma_1^r n_i = \ell$. Hence the result.

4. Conclusions

Further results on the size of projections are necessary in the case of more complex systems of functional dependencies.

We think that such results can be obtained. However the price to be paid will reside in some further assumptions concerning the manner in which information is presented in the relations. The recent work of N. SPYRATOS [4] toward the formal representation of data base views provides a promising approach which should be explored.

Another problem which we shall examine in subsequent work is the computation of projections from a dynamic representation of the relation's evolution under the effect of updates.

Acknowledgements

The authors thank Ph. RICHARD, M. SPYRATOS for their comments. The case examined in Section 3.2 was suggested by G. JOMIER.

References

[1] E. GELENBE, D. GARDY, "On the size of projections I" to appear in Information Processing Letters.

[2] Ph. RICHARD, "Evaluation of the size of a query expressed in relational algebra", Proceedings ACM-SIGMOD International Conference on Manegement of Data,April 1981, pp. 1555-163.

[3] R. DEMOLOMBE, "Estimation of the number of tuples satisfying a query expressed in predicate calculus language", VLDB Conference Proceedings (Montreal), October 1980.

[4] N. SPYRATOS, "An operational approach to data bases", ACM SIGACT-SIGMOD Conference on Principles of Data Base Systems, Los Angeles, March 1982.