Acta Sci. Math. (Szeged)
60 (1995), 213–223

Minimal representations of branching dependencies

JÁNOS DEMETROVICS, GYULA O. H. KATONA* and ATTILA SALI°

Dedicated to Professor László Leindler for his 60th and to Professor Károly Tandori for his 70th birthday

Communicated by F. Gécseg

Abstract. A new type of dependencies in a relational database model is investigated. If b is an attribute, A is a set of attributes then it is said that $b(p,q)$-depends on A, in notation $A \xrightarrow{(p,q)} b$, in a database r if there are no $q+1$ relations in r such that they have at most p different values in A, but $q+1$ different values in b. $(1,1)$-dependency is the classical functional dependency. Let $\mathrm{J}(A)$ denote the set $\{b : A \xrightarrow{(p,q)} b\}$. Using some characterization of the set function $\mathrm{J}(A)$ we give estimates for the minimum number of records in a database system that results the set functions $\mathrm{J}(A)$.

1. Introduction

A relational databases system of the scheme $R(A_1, A_2, \ldots, A_n)$ can be considered as a matrix, where the columns correspond to the *attributes* A_i's (for example name, date of birth, place of birth etc.), while the rows are the n-tuples of the relation r. That is, a row contains the data of a given *individual*. Let Ω denote the set of attributes (the set of the columns of the matrix). Let $A \subseteq \Omega$ and $b \in \Omega$.

Received October 19, 1994, and in revised form April 5, 1995.
AMS Subject Classification (1991): 68P15, 68R05, 05D05.
* Supported by the Hungarian NFS, Grant No. 1909 and by the European Communities (Cooperation in Science and Technology with Central and Eastern European Countries), contract No. CIPACT 930113.
° Partially supported by the Hungarian NFS, Grant No. 4267.

We say that b (*functionally*) *depends* on A (see [1,2]) if the data in the columns of A determine the data of b, that is there exist no two rows which agree in A but different in b. We denote this by $A \longrightarrow b$.

Functional dependencies have turned out to be very useful. All existing data base managing systems are based on this concept. Let us consider the following example. Suppose that $\Omega = \{A_1, A_2, A_3, A_4\}$ and $A_1 \longrightarrow A_2$ and $A_3 \longrightarrow A_4$ hold. If we store the whole matrix in the memory of a computer, then it requires $4N_1N_3$ registers in the worst case, where $N_1(N_3)$ denotes the number of possible different values of $A_1(A_3)$. Indeed, A_1 and A_3 can take values independently, but they determine A_2 and A_4, respectively. Thus, the number of different rows is at most N_1N_3. However, using the given functional dependencies, we can save a lot of memory. Indeed, it is enough to store the matrix consisting of the columns A_1 and A_3 ($2N_1N_3$ registers) together with two little matrices each having two columns. One contains values of A_1 and A_2 in the first and second columns, respectively. The first column contains all possible values of A_1, while the second one contains the values determined by the dependency $A_1 \longrightarrow A_2$. The other small matrix is built up from A_3 and A_4 in the same way. The number of stored values is at most $2N_1N_3 + 2(N_2 + N_4)$, which is usually significantly smaller than $4N_1N_3$.

In the present paper we investigate a more general (weaker) dependency, than the functional introduced in [5]. We illustrate it first a very particular case, then we show the usefulness of the concept. Let $A \subseteq \Omega$ and $b \in \Omega$, we say that $b(1,2)$-depends on A if the values in A determine the values in b in a "two-valued" way. That is, there exist no three rows same in A but having three different values in b. We denote it by $A \overset{(1,2)}{\longrightarrow} b$. Similarly, $A \overset{(1,q)}{\longrightarrow} b$ if there exist no $q+1$ rows each having the same values in columns of A, but containing $q+1$ different values in the column b.

Let us suppose that the database consists of the rips of an international transport truck, more precisely, the names of the countries the truck enters. For the sake of simplicity, let us suppose that the truck goes through exactly four countries in each trip, (counting the strat and endpoints, too) and does not enter a country twice during one trip. Suppose furthermore, that there are 30 possible countries and one country has at most five neighbours. Let A_1, A_2, A_3, A_4 denote the first, second, third and fourth country as attributes. It is easy to see that

$$A_1 \overset{(1,5)}{\longrightarrow} A_2, \quad \{A_1, A_2\} \overset{(1,4)}{\longrightarrow} A_3 \quad \text{and} \quad \{A_2, A_3\} \overset{(1,4)}{\longrightarrow} A_4.$$

Now, we cannot decrease the size of the stored matrix, as in the case of functional ((1,1)–) dependency, but we can decrease the range of the elements of the matrix. The range of each element of the original matrix consists of 30 values, names of countries or some codes of them (5 bits each, at least). Let us store a little table ($30 \times 5 \times 5 = 750$ bits) that contains a numbering of the neighbours of each country, which assignes to them the numbers 0, 1, 2, 3, 4 in some order. Now we can replace attribute A_2 by these numbers (A_2^*), because the values of A_1 gives the starting country and the value of A_2^* determines the second country with the help of the little table. The same holds for the attribute A_3, but we can decrease the number of possible values even further, if we give a table of numbering the possible third countries for each $A_1 A_2$ pair. In this case, the attribute A_3^* can take only 4 different values. The same holds for A_4, too. That is, while each element of the original matrix could be encoded by 5 bits, now for the cost of two little auxiliary tables we could decrease the lengh of the elements in the second column to 3 bits, and that of the elements in the third and fourth columns to 2 bits.

It is easy to see, that the same idea can be applied in each case when we store the paths of a graph, whose maximal degree is much less than the number of its vertices or when we want to store the sequence of states of a process, where the number of all possible states is much langer, than the number of possible successing states of a state or in any case when there hold many $(1, q)$-dependencies, where q is small.

The general concept we shall study is the (p, q)-dependency ($1 \le p \le q$ integers).

Definition 1.1. Let a relational database system of the scheme $R(A_1, A_2, \ldots, A_n)$ be given. Let $A \subseteq \Omega$ and $b \in \Omega$. We say that b *(p, q)-depends on A* if there are no $q+1$ rows (n-tuples) of r such that they contain at most p different values in each column (attribute) of A, but $q+1$ different values in b.

For a given relation r (or its matrix M) we define a function from the family of subsets of Ω into itself as follows.

Definition 1.2. Let M be the matrix of the given relation r. Let us suppose, that $1 \le p \le q$. Then the mapping $\mathbf{J}_{Mpq}: 2^\Omega \to 2^\Omega$ is defined by

$$\mathbf{J}_{Mpq}(A) = \left\{ b : A \xrightarrow{(p,q)} b \right\}.$$

We collect two important properties of the mapping $\mathbf{J}_{Mpq}$ in the following proposition, see [5].

Proposition 1.3. *Let r, Ω, M, p and q as above. Furthermore, let $A, B \subseteq \Omega$. Then*
(i) $A \subseteq \mathbf{J}_{Mpq}(A)$
(ii) $A \subseteq B \Longrightarrow \mathbf{J}_{Mpq}(A) \subseteq \mathbf{J}_{Mpq}(B)$.

Definition 1.4. Set functions satisfying (i) and (ii) are called *increasing-monotone functions.* We say that such an increasing-monotone function $\mathcal{N}$ is (p,q)-*representable* if there exists a matrix M such that $\mathcal{N} = \mathbf{J}_{mpq}$.

The aim of this paper is to generalize theorems on minimal representations valid for functional dependencies to (p,q)-dependencies. There arise several very interesting combinatorial problems in this context.

2. Minimal representations

In this section we investigate the minimum number of rows of a matrix M that (p,q)-represents a given increasing-monotone set function $\mathcal{N}$, provided such representations exists. We always assume that $p \leq q$.

Definition 2.1. For an increasing-monotone function $\mathcal{N}$ let $s_{pq}(\mathcal{N})$ denote the minimum number of rows of a matrix that (p,q)-represents $\mathcal{N}$. If $\mathcal{N}$ is not (p,q)-representable, then we put $s_{pq}(\mathcal{N}) = \infty$.

Let us note that in the case of $p > 2$ we have examples of the latter equality occuring.

It was proved in [5] that an increasing-monotone function $\mathcal{N}$ with $\mathcal{N}(\emptyset) = \emptyset$ is (p,q)-representable if (i) $1 = p < q$, (ii) $p = 2$ and $3 < q$ or (iii) $2 < p$ and $p^2 - p - 1 < q$. From the proof one can easily deduct the following general upper bound.

Theorem 2.2. *Let $\mathcal{N}$ be an increasing-monotone function with $\mathcal{N}(\emptyset) = \emptyset$ and let (p,q) satisfy one of* (i) - (iii) *above. Then*

$$s_{pq}(\mathcal{N}) \leq q(n+1)2^n.$$

The above bound is quite coarse, which is caused by the lack of knowledge about the structure of an increasing-monotone function. However, there is a nice structure theory of closures (that are special increasing-monotone functions), furthermore, it is shown in [4] that for $p = q$ the resulting set function is indeed a closure.

Definition 2.3. An increasing-monotone function $\mathcal{N}$ satisfying

$$\mathcal{N}(\mathcal{N}(A)) = \mathcal{N}(A) \qquad \forall A \subseteq \Omega$$

is called a *closure*.

In the rest of this section we will consider closures only. First, we prove a direct product theorem analogous to the $(1, 1)$ case considered in [3].

Definition 2.4. Let $\mathcal{L}$ and $\mathcal{N}$ be closures on ground sets U and V, respectively, with $U \cap V = \emptyset$. *The direct product of* $\mathcal{L}$ *and* $\mathcal{N}$ is the closure on ground set $U \cup V$ defined by

$$\text{For } A \subseteq U \cup V \qquad (\mathcal{L} \times \mathcal{N})(A) = \mathcal{L}(A \cap U) \cup \mathcal{N}(A \cap V).$$

This direct product plays an important role in the theory and practice of relational database systems.

Theorem 2.5. *Let* $\mathcal{L}$ *and* $\mathcal{N}$ *be closures on ground sets* U *and* V, *respectively. Then*

$$s_{pq}(\mathcal{L} \times \mathcal{N}) \leq s_{pq}(\mathcal{L}) + s_{pq}(\mathcal{N}) - p.$$

Proof. The statement is trivial if some of $\mathcal{L}$ and $\mathcal{N}$ is not (p, q)-representable. Thus, we may assume that both $s_{pq}(\mathcal{L})$ and $s_{pq}(\mathcal{N})$ are finite. Let M_1 be a minimal representation matrix for $\mathcal{L}$ and let M_2 be that for $\mathcal{N}$. We form the following matrix:

$$\mathbf{M} = \begin{pmatrix} Q & W \\ R & T \\ Y & P \end{pmatrix},$$

where Q is obtained from M_1 by dropping the last p rows. W consists of the first row of M_2 taken as many times as the number of rows of Q, R consists of the last p rows of M_1, while T consists of the first p rows of M_2. P is obtained from M_2 by dropping its first p rows, finally Y contains the last row of M_1 in as many copies as the number of rows of P. We claim that $\mathbf{J}_{\mathbf{M}_{pq}} = \mathcal{L} \times \mathcal{N}$.

Let us suppose first that $y \notin (\mathcal{L} \times \mathcal{N})(A)$ for some $A \subset U \cup V$. We may assume without loss of generality that $y \in U$. This implies that M_1 contains $q + 1$ rows that have at most p different values in columns of A, but are all distinct in y. These $q + 1$ rows occur in matrix $\mathbf{M}$, as well, namely in the part $\begin{pmatrix} Q & W \\ R & T \end{pmatrix}$.

Thus, they have at most p different values in their extensions to V, which implies $y \notin \mathbf{J}_{\mathbf{M}_{pq}}(A)$.

On the other hand, let $y \in (\mathcal{L} \times \mathcal{N})(A)$ and $r_0, r_1, \ldots, r_q$ be such rows that they contain at most p different values in columns of A. Assume again that $y \in U$. If there are two among $r_0, r_1, \ldots, r_q$ such that both of them are in the part (YP), or one of them is in (YP) and the other one is the last row of (RT), then these two rows agree on y, so they cannot contain $q+1$ distinct values in y. However, if at most one of the r_i's from the "lower" part of $\mathbf{M}$, then there exist $q+1$ distinct rows of M_1 that have same values in A and y as r_i's, namely the restriction of $r_0, r_1, \ldots, r_q$ to U. Thus, $y \in \mathcal{L}(A \cap U)$ implies that r_i's have at most q different values in y, i.e., $y \in \mathbf{J}_{\mathbf{M}_{pq}}(A)$. ∎

Next we will calculate certain $s_{pq}(\mathcal{L})$ values for the following well-studied closures.

Definition 2.6. Let $\mathcal{L}_n^k$ denote the following closure on Ω:

$$\mathcal{L}_n^k(X) = \begin{cases} X & \text{if } |X| < k \\ \Omega & \text{otherwise.} \end{cases}$$

First we need a general lemma.

Lemma 2.7. *Let us assume that $\mathcal{L}_n^k$ is (p,q)-representable. Then*

$$\binom{s_{pq}(\mathcal{L}_n^k)}{q+1} \geq \binom{n}{k-1}.$$

Proof. Let M be a matrix of $s_{pq}(\mathcal{L}_n^k)$ rows (p,q)-representing $\mathcal{L}_n^k$. For any $k-1$ element subset B of Ω there exist $q+1$ rows of M that contain at most p values in B but $q+1$ different values in some other column. If rows $\{r_0, r_1, \ldots, r_q\}$ belong to $B \subset \Omega$ and rows $\{t_0, t_1, \ldots, t_q\}$ belong to $C \subset \Omega$, then

$$\{r_0, r_1, \ldots, r_q\} \notin \{t_0, t_1, \ldots, t_q\}.$$

Indeed, if the same $q+1$-tuple of rows belonged to B and C, then

$$\mathbf{J}_{Mpq}(B \cup C) \notin \Omega$$

would hold, a contradiction. Thus, the number of possible different $q+1$-tuples of rows is at least as large as the number of $k-1$-subsets of Ω. ∎

Proposition 2.8.

$$s_{pq}(\mathcal{L}_n^1) = q + 1.$$

Proof. The inequality $s_{pq}(\mathcal{L}_n^1) \geq q + 1$ follows from Lemma 2.7. The inequality in the other direction and the representability is proved by the following construction:

$$\begin{matrix} 0 & 0 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 1 \\ 2 & 2 & 2 & \dots & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q & q & q & \dots & q \end{matrix}$$

■

Theorem 2.9. *If $n > 3$, then*

$$s_{22}(\mathcal{L}_n^2) = 2n.$$

Proof. We construct a matrix M of $2n$-rows $(2,2)$-representing $\mathcal{L}_n^2$ as follows. Rows $2i-1$ and $2i$ will contain 0 in column i and $2i-1$ and $2i$ in other columns, respectively. If $A \subset \Omega$ has more than one element, then there exist no three rows of M containing at most two different values in columns of A implying $\mathbf{J}_{M22}(A) = \Omega$. On the other hand, for any pair of one element subsets $\{i\}$ and $\{j\}$ of Ω rows $2i-1, 2i$ and $2j$ shows that $\{i\} \xrightarrow{(2,2)} j$ in M.

In order to prove that we need at least $2n$ rows to $(2,2)$-represent $\mathcal{L}_n^2$, let us assume that M is a representing matrix of minimum number of rows. As in the proof of Lemma 2.7, for every column there exist three rows that contain at most two different values in that column. That is, for every column, there is a pair of rows that agree on that column. We claim that these pairs are disjoint for different columns.

Case (i) There exist columns i and j such that the same pair of rows are chosen above, say r and s. Then $\{i,j\} \xrightarrow{(2,2)} \Omega$ implies that the two rows r and s must contain identical entries, which contradicts the minimality of M. Indeed, if r and s contained different elements in a column k we could choose a third row t that contains a third different value in column k (by $\mathcal{L}_n^2(\emptyset) = \emptyset$ any column contains at least three different entries), the triplet (r,s,t) would show $\{i,j\} \xrightarrow{(2,2)} \{k\}$, a contradiction.

Case (ii) There exist columns i and j such that the pairs chosen for them have one common row, i.e., these two columns contain at most two different values in certain three rows. Then again $\{i,j\} \xrightarrow{(2,2)} \Omega$ implies that all other columns contain at most two different entries in these columns, so for $n > 3$ we get back to Case (i). Thus, we have proved that the pairs of rows belonging to the columns must be pairwise disjoint if $n > 3$. Let us note that $s_{22}(\mathcal{L}_n^2) = 4$ and $s_{22}(\mathcal{L}_3^2) = 5$. The lower bounds follow from Lemma 2.7 either directly or by easy argument. The upper bounds are given by the following constructions.

$$s_{22}(\mathcal{L}_2^2) = 4 : \qquad \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 2 \\ 2 & 2 \end{pmatrix}$$

$$s_{22}(\mathcal{L}_3^2) = 5 : \qquad \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 2 & 2 \\ 2 & 2 & 3 \\ 3 & 1 & 0 \end{pmatrix}$$

■

The next theorem is an interesting application of a theorem of Lovász.

Theorem 2.10.

$$s_{pp}(\mathcal{L}_n^n) = \min\{v : \binom{v-1}{p} \geq n\}.$$

Proof. Let us first prove the upper bound by a construction. Assume that $\binom{v-1}{p} \geq n$. Construct a matrix M of v rows and n columns as follows. The first row consists of all 0's. Then assign a distinct p-element subset of the remaining $v-1$ rows to every column, and put the numbers $1, 2, \ldots, p$ in them, respectively. The remaining entries are 0s. We show the case $p = 2$, $n = 6$ and $v = 5$.

$$\begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 1 & 0 \\ 0 & 2 & 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 0 & 2 & 2 \end{matrix}$$

Let us now assume that $b \notin A \subset \Omega$. Then there are $p+1$ distinct entries in column b in row 0 and the p rows assigned to b, while 0 occurs at least twice in these rows in columns of A. This means that $b \notin \mathbf{J}_{Mpp}(A)$, i.e. every subset A of Ω is closed under $\mathbf{J}_{Mpp}$, so $\mathbf{J}_{Mpp} = \mathcal{L}_n^n$.

On the other hand, let M (p,p)-represent $\mathcal{L}_n^n$ and let V be its set of rows. Every $n-1$-element set is closed in $\mathcal{L}_n^n$, thus there exist $p+1$ rows for any column $b \in \Omega$ such that they contain $p+1$ different entries in b but at most p distinct ones in each of the remaining columns. Thus these $p+1$-element row sets are all different, let S_b denote the one belonging to column b. We may assume without loss of generality that for every b the numbers $0,1,2,\ldots,p$ are standing in b and in the rows of S_b. Now let us change all entries of M which is not between 0 and p (inclusive) to 0. It is easy to see that the obtained matrix still (p,p)-represents $\mathcal{L}_n^n$, but now exactly $p+1$ different entries occur in each column. Let us consider the hypergraph $\mathcal{V} = (V, \{S_b : b \in \Omega\})$. $\mathcal{V}$ is $p+1$-uniform and there exists a partition of the vertex set V into $p+1$ classes for every edge S_b that completely cuts S_b but does not cut completely any other edge. This latter partition can be constructed according to the numbers occuring in column b. Such a hypergraph is called $p+1$-*forest.* Lovász [6] proved that the maximum number of edges of a k-forest on m vertices is $\binom{m-1}{k-1}$. Now $\mathcal{V}$ is a $p+1$-forest on v points with n edges, so Lovász's result gives

$$n \leq \binom{v-1}{p}.$$

■

Let us note that Theorem 2.10 shows that equality does not necessarily occur in Theorem 2.5. Indeed, one has only to observe that $\mathcal{L}_n^n = \mathcal{L}_{n-1}^{n-1} \times \mathcal{L}_1^1$.

The following inequality is another example of the "strange" behaviour of $s_{pp}(\mathcal{L}_n^k)$. It shows that the function is not monotonic in k.

Lemma 2.11. *For any positive integer r*

$$s_{22}(\mathcal{L}_n^3) \leq \frac{r}{2r-1}n + 2\frac{r}{2r-1} + 2r - 1.$$

Lemma 2.7 gives a lower bound of the order of magnitude $n^{2/3}$, thus we do not know yet the right order for $s_{22}(\mathcal{L}_n^3)$.

Proof of Lemma 2.11. We give a construction that proves the upper bound. Let M be a matrix of $m+2r-1$ $(2r-1 \leq m)$ rows defined as follows. For each column we assign r pairs of rows that will contain the numbers $1,2,\ldots,r$, respectively, while the other entries of the columns are all distinct and different from $1,2,\ldots,r$. Let us index the rows in two parts, $1,2,\ldots,m$ for the first m rows and $1,2,\ldots 2r-1$ for the last $2r-1$ rows. (Thus we have row "first 1" and row "last 1" and so on...) The pairs we assign will consist of a "first" row and a "last" row. First, we order all possible "first-last" pairs in the following way: $(1,1),(2,2),\ldots,(2r-1,2r-1)$, $(2,1),(3,2),\ldots,(2r-1,2r-2),(1,2r-1)$, $(3,1),(4,2),\ldots$. Then the first r pairs are assigned to the first column, the second r pairs to the second column, etc. It is easy to see that the pairs assigned to one column are pairwise disjoint. Now, any 1-element subset b of Ω is closed, because for any other column c contains distinct elements in any pair assigned to b, and we can add a third row which contains a third different element in c. If $A=\{a,b\}$ is a 2-element subset of Ω, then the pairs assigned to a and b have at most $2r-1$ different second coordinate, but there are $2r$ pairs altogether, so two of them have to have the same second coordinate. These determine three rows, which contain at most 2 different values in columns of A. If there were three columns that contain at most two different values in certain three rows, then one of the rows is from the "first" part of the matrix and two of them from the "last" part, or vice versa, but in both case we would obtain that a pair is assigned to two different columns, a contradiction. Thus, every 2-element subset of Ω is closed, but any 3- or more element subset $(2,2)$-implies the whole Ω. M has $m+2r-1$ rows and $n=\lfloor \frac{2r-1}{r}m\rfloor$ columns. The trivial inequality

$$m+2r-1 \leq \frac{r}{2r-1}\lfloor \frac{2r-1}{r}m\rfloor + \frac{r}{2r-1} + 2r-1$$

completes the proof for n of the form $\lfloor \frac{2r-1}{r}m\rfloor$. If n is not of this form then $n+1$ is, one of the columns is deleted. This gives an additional term $\frac{r}{2r-1}$. ■

We pose the following problem:

Open problem. Determine the value $s_{22}(\mathcal{L}_n^3)$!

References

[1] W. W. ARMSTRONG, Dependency Structures of database Relationships, *Information Processing* **74**, North Holland, Amsterdam, 1974, 580–583.

[2] E. F. CODD, A Relational Model of Data for Large Shared Data Banks, *Comm. ACM*, **13** (1970), 377-387.

[3] J. DEMETROVICS, Z. FÜREDI and G. O. H. KATONA, Minimum matrix representation of closure operations, *Discrete Appl. Math.*, **11** (1985), 115–128.

[4] J. DEMETROVICS and G. O. H. KATONA, Extremal Combinatorial Problems of Database Models, *Proc. MFDBS* **87**, Dresden, Springer-Verlag, *LNCS*, **305** (1987), 99–127.

[5] J. DEMETROVICS, G. O. H. KATONA and A. SALI, The characterization of branching dependencies, *Discrete Appl. Math.*, **40** (1992), 139–153.

[6] L. Lovász, Topological and algebraic methods in graphs theory, *Graph Theory and Related Topics,* Proc. Conf. Univ. Waterloo, Ontario 1977, Academic Press, NY, 1979, 1–14.

J. Demetrovics, Computer and Automatization Institute, HAS, Budapest, Kende u. 13-17, H-1111 Hungary
Gy. O. H. Katona, Mathematical Institute of HAS, Budapest P.o.b. 127, H-1364 Hungary
A. Sali, Mathematical Institute of HAS, Budapest P.o.b. 127, H-1364 Hungary