Discrete Mathematics 30 (1980) 83–88.

ON THE NUMBER OF MAXIMAL DEPENDENCIES IN A DATA BASE RELATION OF FIXED ORDER

A. BÉKÉSSY, J. DEMETROVICS, L. HANNÁK

Computer and Automation Institute, Hungarian Academy of Sciences, Budapest, Hungary

P. FRANKL, Gy. KATONA*

Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary

Received 21 September 1979

The paper gives asymptotic bounds for the maximum number N_n of non-trivial maximal elements in a data base relation of given order. The result shows that there exist relations which are very rich in maximal elements.

1. Introduction

Maximal elements [2] are important characteristics of the dependency structure in a data base relation [1]. They determine, as shown by Armstrong [2], all functional dependencies in a full family. Moreover, the left-hand sets of attributes in maximal elements of type $A \to B$ are *keys* for the set B. Parallel with the enumeration of the maximal number of keys in a relation of fixed number of attributes [4], it was obvious to inquire after the maximal possible number of non-trivial maximal elements, as well. But, while the first problem was easy to answer—the answer was, in fact, implicitly given by Sperner's theorem [3] and Armstrong's theorem [2]—this second one turned out hard and no exact figure in terms of the order n has yet been found; some asymptotic lower and upper bounds are our results.

2. Definitions

Let $\Omega = \{a_1, a_2, \ldots, a_n\}$ be a set of n elements ("attributes") and 2^Ω its power set. The function $f: 2^\Omega \to 2^\Omega$ is called a *closure function* or *closure* iff for every $A, B \in 2^\Omega$

$$\begin{aligned} &\text{(a)} \quad A \subseteq f(A), \\ &\text{(b)} \quad f(f(A)) = f(A), \\ &\text{(c)} \quad A \subseteq B \Rightarrow f(A) \subseteq f(B). \end{aligned} \tag{1}$$

* The work of this author was partially supported by the National Science Foundation under grant number MCS77-22985.

If $B \subseteq f(A)$ we say that B is *functionally dependent* on A; this binary relation will be denoted by $A \to B$ and called a *functional dependency.*

It can be seen that the lattice of functional dependencies defined in this way satisfy Armstrong's axioms [2], and conversely that there is a unique closure to any lattice satisfying these axioms hence, by Armstrong's theorem 5, there exists a relation of Codd's type to any closure.

A functional dependency $A \to f(A)$ is called *maximal* if $C \to f(A)$ and $C \subseteq A$ implies $C = A$ for all C's. If $A \to f(A)$ is maximal and $A \neq f(A)$ then it is called *non-trivial*, all other maximal dependencies, i.e. those of the form $A \to A$ are *trivial*. With other words the non-trivial maximal dependencies are the pairs $A \to f(A)$ where A satisfies the following two conditions:

$$\begin{aligned} &\text{(a)} \quad \text{There is no } C \subseteq \Omega \text{ such that } C \subset A \text{ and } f(C) = f(A); \\ &\text{(b)} \quad A \neq f(A). \end{aligned} \tag{2}$$

Call these A's *basic.*

Let the number of non-trivial maximal elements in the set of all functional dependencies generated by a closure be denoted by $N(f)$. Now let us consider all closures over Ω and the number

$$N_n = \max_f N(f) \tag{3}$$

i.e. *the maximum possible number of non-trivial maximal elements in a relation of Codd's type of fixed order n.* This number is equal to the number of basic sets in a certain closure.

Observe that $2^{n-1} \leq N_n$. Indeed, if $x \in \Omega$ is any fixed attribute and $f(A) = A \cup \{x\}$ then all pairs $A \to A \cup \{x\}$ where $x \notin A$ form non-trivial maximal dependencies and their number is 2^{n-1}. For a long time we thought this estimation exact since $N_n = 2^{n-1}$ for $n = 1, 2, 3$ and 4. However, as we shall see, $N_5 > 16$.

Similarly there is a trivial upper estimation of N_n, namely $N_n < 2^n$.

3. Theorems

Theorem 1

$$\prod_{i=1}^{k} (2^{q_i} - 1) - \prod_{i=1}^{k} (2^{q_i} - q_i - 1) \leq N_n \leq 2^n \left(1 - \frac{1}{n+1}\right) \tag{4}$$

where $q_1, q_2, \ldots, q_n$ are positive integers and $\sum_{i=1}^{k} q_i = n$.

Proof. (a) *Lower estimation.* Let us consider a partition of Ω: $(\Omega_1, \Omega_2, \ldots, \Omega_k)$,

($\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$) where $|\Omega_i| = q_i$ ($i, j = 1, 2, \ldots, k$). Define a closure f as

$$f(A) = A \cup \bigcup_j \Omega_j \tag{5}$$

where j runs over the indices $1, 2, \ldots, k$ for which $|A \cap \Omega_i| = |\Omega_i| - 1$ holds.

Let us first check if f of (5) is a closure function. Property (1a) holds trivially. Since $|f(A) \cap \Omega_j| \neq |\Omega_j| - 1$ for all indices j, (1b) is satisfied, as well. Turning to (1c), it is easy to see that

$$A \subseteq B \Rightarrow f(A) \cap \Omega_i \subseteq f(B) \cap \Omega_i$$

consequently $f(A) \subseteq f(B)$ holds.

Now we are going to determine the basic A's i.e. the sets satisfying (2a, b). $A \subseteq \Omega$ is basic iff

$$\min_i (|\Omega_i| - |A \cap \Omega_i|) = 1 \tag{6}$$

Were the left hand side of (6) greater than 1, then $f(A) = A$ (i.e. $A \rightarrow f(A)$ would be trivial) while if it were equal to 0 then take the set $A^* = A \backslash x$ where $x \in \Omega_j$, $|\Omega_j| = |A \cap \Omega_j|$. Then $f(A^*) = f(A)$ by (5) contradicting (2a). If, however, (6) holds then A is obviously basic.

So $N(f)$ is equal to the number of A's satisfying (6). But this number can be expressed as a difference between the number of sets satisfying

$$\min_i (|\Omega_i| - |A \cap \Omega_i|) \geqslant 1$$

and the number of those satisfying

$$\min_i (\Omega_i - |A \cap \Omega_i|) \geqslant 2;$$

the first number is $\prod_{i=1}^{k} (2^{q_i} - 1)$ while the other $\prod_{i=1}^{k} (2^{q_i} - q_i - 1)$.

As a special case let us take the case $n = 5$, $q_1 = 3$, $q_2 = 2$. Then the lower estimation gives $17 \leqq N_5$. This is the first example where $2^{n-1} < N_n$.

Remark. The idea of this proof was, in fact, a more general consideration. Let there be disjoint attribute sets $\Omega_{i'}$ given; let an arbitrary closure f_i defined over Ω_i ($i = 1, 2, \ldots, k$); let the number of *all maximal* elements in the set of dependencies $\mathscr{F}_i$ be denoted by $M(f_i)$ and the number of *trivial maximal elements* by $T(f_i)$ ($i = 1, 2, \ldots, k$). Now

$$f(A) = \bigcup_{i=1}^{k} f_i(A \cap \Omega_i), \quad (A \subseteq \Omega)$$

is a closure over $\Omega = \bigcup_i \Omega$ with the property $M(f) = \prod_{i=1}^{k} M(f_i)$, $T(f) = \prod_{i=1}^{k} T(f_i)$

hence the non-trivial elements are of cardinality

$$M(f)-T(f)=\prod_{i=1}^{k} M(f_i)-\prod_{i=1}^{k} T(f_i).$$

In the proof above the choice

$$f_i(A_i)=\begin{cases}\Omega_i & \text{if } |A_i|=|\Omega_i|-1,\\ A_i & \text{otherwise}\end{cases}$$

was made for all $A_i\subseteq\Omega_i$ $(i=1,2,\ldots,k)$.

(b) *Upper estimation.* Let $\mathcal{H}$ be the set of all basic sets. Iff $A\in\mathcal{H}$ then there is a B such that $|B|=|A|+1$, $A\subseteq B\subseteq f(A)$. Then $f(A)\subset f(B)\subseteq f(f(A))=f(A)$. Since $f(A)=f(B)$ and $A\subset B$ the set B does not satisfy condition (2a). The set B can be obtained from at most n different sets A only, consequently at least for $|\mathcal{H}|/n$ sets $B\notin\mathcal{H}$, implying

$$|\mathcal{H}|+\frac{|\mathcal{H}|}{n}\leqslant 2^n$$

equivalent to the desired upper estimation in (4).

The proof of Theorem 1 is completed.

Note that the upper bound could be improved to $2^n(1-2/n)$ about by considering that the majority among the 2^n subsets of Ω has about $\frac{1}{2}n$ elements so that one B can be used about $\frac{1}{2}n$ times only. But we don't go into details of the proof because the gain is inconsiderable.

Next we want to have a lower bound of N_n in terms of n. For this purpose the numbers q_i on the left side of (4) will be chosen in a special way. The result can be written as

Theorem 2

$$\left(1-\frac{4}{\log_2 \mathrm{e}}\frac{\log\log_2 n}{\log_2 n}(1+\mathrm{o}(1))\right)\leqslant\frac{N_n}{2^n}\leqslant\left(1-\frac{1}{n+1}\right). \tag{7}$$

Proof. Define the integer number q as

$$q=q(n)=(\log n-\log\omega(n)) \tag{8}$$

where

$$\omega(n)=\frac{1}{\log \mathrm{e}}(\log\log n-\log\log\log n-\log\log e-1) \tag{9}$$

and log means the logarithm of base 2, $[x]$ is the integral part of x. Divide n by q, so let $k(n)$, $r(n)$ be defined by

$$n=qk(n)+r(n) \tag{10}$$

where $k(n)$ is a non-negative integer and $0\leqslant r(n)<q$.

Let the q_i's be chosen in the following way:

$$\begin{aligned} q_1 = q_2 = \cdots = q_r = q+1, \\ q_{r+1} = q_{r+2} = \cdots = q_k = q. \end{aligned} \tag{11}$$

Hence, making use of the inequalities of the elementary calculus

$$(1-x)^y \geqslant 1-xy \quad (0 \leqslant |x| \leqslant 1,\ y=0 \quad \text{or} \quad y \geqslant 1)$$

and

$$(1-x)^y \leqslant e^{-xy} \quad (0 \leqslant |x| \leqslant 1,\ y \geqslant 0)$$

we have

$$\begin{aligned} \frac{1}{2^n} \prod_{i=1}^{k} (2^{q_i}-1) &= \left(1-\frac{1}{2^{q+1}}\right)^{r(n)} \left(1-\frac{1}{2^q}\right)^{k(n)-r(n)} \\ &\geqslant \left(1-\frac{r(n)}{2^{q+1}}\right)\left(1-\frac{k(n)-r(n)}{2^q}\right) \\ &\geqslant 1-\frac{k(n)-r(n)}{2^q}(1-o(1)) \end{aligned}$$

by taking account of

$$\begin{aligned} \frac{r(n)}{k(n)-r(n)} &\leqslant \frac{r(n)q}{n-r(n)-r(n)q} \\ &\leqslant \frac{r(n)q}{n-q-q^2} = o(1) \quad (\text{for } n \to \infty). \end{aligned}$$

Also we have

$$\begin{aligned} \frac{1}{2^n} \prod_{i=1}^{k} (2^{q_i}-q_i-1) &= \left(1-\frac{q+2}{2^{q+1}}\right)^{r(n)} \left(1-\frac{q+1}{2^q}\right)^{k(n)-r(n)} \\ &\leqslant \exp\left\{-\frac{n+k(n)-r(n)-\frac{1}{2}qr(n)}{2^q}\right\}. \end{aligned}$$

The expression $k(n)-r(n)-\frac{1}{2}qr(n)$ is, by (10), certainly positive for sufficiently large n thus

$$\frac{1}{2^n} \prod_{i=1}^{k} (2^{q_i}-q_i-1) \leqslant \exp\left\{-\frac{n}{2^q}\right\}.$$

Therefore our intermediate result is, by Theorem 1,

$$1-\frac{k(n)-r(n)}{2^q}(1+o(1)) - \exp\left\{-\frac{n}{2^q}\right\} \leqslant N_n \leqslant 1-\frac{1}{n+1}. \tag{12}$$

Further on,

$$\begin{aligned} \exp\left\{-\frac{n}{2^q}\right\} &\geqslant \exp\left\{-\frac{n}{2^{\log n - \log \omega(n)}}\right\} \\ &= \exp\{-\omega(n)\} = \frac{2}{\log e} \frac{\log \log n}{\log n} \end{aligned}$$

and

$$\frac{k(n)-r(n)}{2^q} \leqslant \frac{n}{q2^q} \leqslant \frac{2\omega(n)}{q}$$
$$= \frac{2}{\log e} \frac{\log \log n}{\log n}(1+o(1)).$$

These inequalities together with (12) imply Theorem 2.

We guess that neither estimation in Theorem 2 is exact, the true value of N_n lies somewhere between. The trivial corollary of Theorem 2 is $N_n/2^n \to 1$, $(n \to \infty)$ which was our first result.

In the construction of the lower bound for N_n the cardinalities of the basic sets tend to infinity with n as can be seen from the proof of Theorem 2. In fact, it is necessary, otherwise $N(f)$ cannot grow as large as found:

Theorem 3. *If $f^*(n)$ $(n = n_0,\ n_0+1, \ldots;\ n_0 = const. > 0)$ is a sequence of closures over $\Omega_m^* = \{a_1, a_2, \ldots, a_n\}$ having a fixed maximal dependency $A \to B$ $(|A| = s, |B| = t > s)$ common for all n, then*

$$N(f_n^*)/2^n < c < 1$$

where c does not depend on n.

Proof. Without loss of generality we can assume that $A = \{a_1, a_2, \ldots, a_s\}$, $B = A \cup \{a_{s+1}, \ldots, a_t\}$. The possible maximal dependencies in $f^*(n)$ are all of the form

(a) $A \cup X \to B \cup X'$, or

(b) $A' \cup Y \to Y'$

where (a): $a_i \notin X$ for $i \leqslant t$ and (b): $A' \subset A$, $a_i \notin Y$ for $i \leqslant s$. Since $|X| \leqslant n-t$ there are 2^{n-t} maximal elements of type (a) at most, and similarly, by $|A'| \leqslant s-1$ and $|Y| \leqslant n-s$, the highest number of maximal elements of type (b) is $(2^s-1)2^{n-s}$. Hence

$$N(f_n^*) \leqslant 2^{n-t} + 2^n - 2^{n-s} \leqslant c2^n$$

as stated.

References

[1] E.F. Codd, A relational model of data for large shared data banks, Comm. Assoc. Comput. Mach. 13 (6) (June 1970) 377–387.

[2] W.W. Armstrong, Dependency structures of data base relationships, Information Processing 74 (North-Holland, Amsterdam, 1974) 580–583.

[3] E. Sperner, Ein Satz über Untermengen einer endlichen Menge, Mathematische Zeitschrift 27 (1928) 544–548.

[4] A. Békéssy and J. Demetrovics, Contribution to the theory of data base relations, Discrete Mathematics 27 (1979) 1–10.