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

A. Károlyi, J. Demeter, I. 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 13 September 1979

The paper gives asymptotic bounds for the maximum number \( N_k \) of non-trivial maximal dependencies in a fixed order. The result shows that these are related to the upper bound of the order.

1. Introduction

Maximal elements are important characteristics of a dependency structure. In a data base relation \( R \), they determine, on average, the number of irreducible functional dependencies in a full family. Moreover, the left-hand sets of attributes in maximal elements of type \( A \rightarrow B \) are keys for the set \( B \). Parallel with the enumeration of the maximal number of keys in a relation of fixed number of attributes \( |A| \), it is natural to consider also the maximal possible number of non-trivial maximal elements, as well. But, while the first problem was easy to answer—the answer is, in fact, explicitly given by Sprague’s theorem [1] and Molodowitch’s refinement [2]—the 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 not results.

2. Definitions

Let \( H = \{ x_1, x_2, \ldots, x_n \} \) be a set of elements ("attributes") and \( 2^H \) be the power set of \( H \). The function \( f : 2^H \rightarrow 2^H \) is called a closure function or closure if for every \( A, B \in 2^H \):

\[ A \subseteq f(A) \]
\[ f(A) = f(f(A)) \]
\[ A \subseteq B \Rightarrow f(A) \subseteq f(B) \]

* The work of the authors was partially supported by the National Scientific Committee under grant number K 717/2045.
If \( R(A) \) we say that \( R \) is functionally dependent on \( A \), this binary relation will be denoted by \( A 
rightarrow R \) and called a functional dependency.

It can be seen that the lattice of functional dependencies defined in this way (with \( R(X) \) as the top element) is not a complete lattice, and conversely that there is a unique closure of \( R(X) \) in the lattice of all closure operators, by Armstrong's Theorem 3, there exists a unique closure of \( R(X) \) which is called the closure of \( R(X) \). This closure operator is denoted by \( R(X) \).

A functional dependency \( A 
rightarrow R \) is called maximal if \( C 
rightarrow R \) and \( C \subseteq A \) implies \( C = A \) for all \( C \). If \( A 
rightarrow R \) is maximal and \( A \nrightarrow R \) then \( A \) is called non-trivial, all other maximal dependencies, i.e. those of the form \( A = A \) are trivial. With other words the non-trivial maximal dependencies are the pairs \( A \nrightarrow R \) where \( A \) satisfies the following two conditions:

(a) There is no \( C \subseteq A \) such that \( C 
rightarrow R \) and \( C \supset A \).

(b) \( A \nrightarrow R \).

Cell base \( A \)'s basis:

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

\[ N = \text{max} \{ N \} \]

where \( N \) is the maximum possible number of non-trivial maximal elements in a relation of \( \text{Card}(R)(Y) \) of fixed order. This number is given by the number of base sets in a closure.

Observe that \( 2^n = N \). Indeed, if \( \emptyset \neq B \) is any fixed attribute and \( \{A\} = \{A\} \) then \( \{A\} \) is a basis and \( \bar{A} = \{A\} \) where \( \{B\} \) is the set of all non-trivial maximal dependencies and \( \bar{A} = \{A\} \). For a big base we thought the extension exist over \( N \). For \( n = 1,2,3 \) and 4, however we shall see, \( N \approx 10 \).

Similarly there is a trivial upper estimation of \( N \) which is \( N_2 = 2^n \).

2. Theorem

**Theorem 1**

\[ \prod_{i=1}^{n} \left( \prod_{j=1}^{m} (a_{ij} \cdot b_{ij} \cdot c_{ij}) \right) = \prod_{i=1}^{n} \left( \prod_{j=1}^{m} (a_{ij} \cdot b_{ij} \cdot c_{ij}) \right) \]

where \( a_{ij}, b_{ij}, c_{ij} \) are positive integers and \( \sum_{i=1}^{n} a_{ij} = n \).

**Proof.** Suppose \( n \). Let us consider a partition of \( B \): \( \{B_1, B_2, \ldots, B_r\} \).
Let $\{A_1, A_2, \ldots, A_n\}$ be a family of sets, where $n = 0(\text{mod} 2)$. Define a chain $f$ as

$$f(A_i) = A_{i+1}$$

where $i$ runs over the integers $1, 2, \ldots, n$ for which $f(A_i) = A_i \rightarrow A_{i+1}$ holds.

Let us first show that if $f$ is a chain function, Property (i) holds trivially. Since $A_{i+1} = f(A_i)$ for all indices $i$, (ii) is satisfied, as well. Turning to (iii), it is easy to see that

$$f(B) = \bigcap_{i \in \mathbb{N}} f(A_i)$$

corresponds to $\bigcap_{i \in \mathbb{N}} f(A_i)$.

Now we are going to determine the basic $A$'s i.e. the sets satisfying (iii). Suppose $A_{i+1} = A_i \rightarrow A_{i+2}$

$$f(A_i) = A_i$$

were the left-hand side of (iii) greater than 1, then $f(A_i) = A_i \rightarrow A_i(A_i)$ would be trivially satisfied if $A_i$ were equal to 0. Thus, let the set $\bigcap_{i \in \mathbb{N}} f(A_i)$, where $x \subseteq 0 \in \bigcap_{i \in \mathbb{N}} f(A_i)$. Then $\bigcap_{i \in \mathbb{N}} f(A_i)$ is also trivially satisfied. If, however, (iii) holds then $A$ is obviously finite.

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

$$\#(\bigcap_{i \in \mathbb{N}} f(A_i)) = 1$$

and the number of those satisfying

$$\#(\bigcap_{i \in \mathbb{N}} f(A_i)) = 2$$

the latter number is $\#(\bigcap_{i \in \mathbb{N}} f(A_i))$ while the order $\#(\bigcap_{i \in \mathbb{N}} f(A_i))$. As a special case we may take the case $n = 1$, $0 < x < 1$. Then the lower estimation gives $\bigcap_{i \in \mathbb{N}} f(A_i)$. This is the last example where $x < 1$.

Remark. The idea of this proof was, in fact, a more general consideration. Let $f$ be a chain function on $\mathbb{N}$ given by an arbitrary function $f$, defined over $\mathbb{N}$ (i.e. $f(1) = 1, f(2) = 2$, etc.), then the number of all maximal elements in the set of deppositions $f$ is defined by $M(f)$ and the number of local maximal elements by $L(f)$.

$$M(f) = \bigcup_{i \in \mathbb{N}} f(A_i), \quad L(f) = \bigcup_{i \in \mathbb{N}} f(A_i)$$

is a closure over $\mathbb{N}$, $f$ with the property $M(f) \subseteq M(f)$, $L(f) \subseteq L(f)$.
The non-trivial elements are of cardinality

\[ \mathcal{M} \rightarrow \mathcal{N} = \prod_{\alpha \in \mathcal{M}} R_{\alpha}(\mathcal{N}) \]

In the proof above the choice

\[ \mathcal{D} \supseteq A \]

was made for all \( A \leq B \), \( 1, 2, \ldots, \).

By upper estimation, let \( R \) be the set of all finite sets. If \( A \leq B \), then there is a \( R \) such that \( R \leq A \). As \( A \leq B \), then \( R \leq R \leq A \). Since \( A \leq B \), \( R \) and \( A \) are in \( R \) from the stability condition (2). The set \( R \) is the set of all \( R \) such that \( R \leq A \). A is a different term that only, consequently of limit for \( R \) exists.

For \( n \), implying

\[ |R| = \frac{2^n - 2}{n} \]

equivalent to the desired upper estimation in (3).

The proof of Theorem 1 is completed.

For the estimation \( 2^{n+1} - 2 \) above by considering that the majority among the \( 2^n \) addition of \( R \) and \( A \) is chosen so that one \( R \) can be said about \( (n+1) \) terms only, that we don't go into details of the proof because the gap is comprehendible.

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

**Theorem 2**

\[ \left( \frac{1}{1 + \log n} \right) \leq q \leq \left( \frac{1}{1 + \log (n+1)} \right) \]

**Proof:** Define the integer number \( q \) as

\[ q = \left( \log n + \log (n+1) \right) \]

where

\[ a \leq 1 \]

and \( b \) means the logarithm of base \( 2 \), \( (2) \) is the integral part of \( a \). Divide \( n \) by \( q \), so that, may be defined by

\[ a = q \cdot (\log n + \log (n+1)) \]

where \( q \) is a non-negative integer and \( 0 < \log (n+1) \).
Let the $q^*$ be chosen in the following way:

$$q^* = q_n, \ldots, q_3, q_2, q_1.$$  

(11)

Shades, noting use of the inequality of the elementary calculus

$$-x < x^2 - 1 < x$$ \quad (x > 0, \text{ or } \quad y > 1)$$

and

$$-x < x^2 - 1 < x$$ \quad (x > 0, \text{ or } \quad y > 1)$$

we have

$$\int_0^1 (2^q - 1) \left(1 - \frac{x}{2^q - 1}\right)^{\frac{1}{2^q - 1}} \, dx \leq \frac{1}{2^q - 1} \left(1 - \frac{1}{2^q - 1}\right)^{\frac{1}{2^q - 1}}$$

by taking partial of

$$\frac{1}{2^q - 1} \left(1 - \frac{1}{2^q - 1}\right)^{\frac{1}{2^q - 1}}$$

Also we have

$$\int_0^1 (2^q - 1) \left(1 - \frac{x}{2^q - 1}\right)^{\frac{1}{2^q - 1}} \, dx \geq \frac{1}{2^q - 1} \left(1 - \frac{1}{2^q - 1}\right)^{\frac{1}{2^q - 1}}$$

The expression $\int_0^1 (x^q - 1) \, dx$ is, by (10), certainly positive for sufficiently large $x$:

$$\int_0^1 (2^q - 1) \left(1 - \frac{x}{2^q - 1}\right)^{\frac{1}{2^q - 1}} \, dx \geq \frac{1}{2^q - 1} \left(1 - \frac{1}{2^q - 1}\right)^{\frac{1}{2^q - 1}}.$$

Therefore our intermediate result is, by Theorem 1,

$$\int_0^1 (2^q - 1) \left(1 - \frac{x}{2^q - 1}\right)^{\frac{1}{2^q - 1}} \, dx = \frac{1}{2^q - 1} \left(1 - \frac{1}{2^q - 1}\right)^{\frac{1}{2^q - 1}}.$$

Further on

$$\int_0^1 \frac{x^q}{2^q - 1} \, dx = \frac{2^q - 1}{2^q - 1} - \frac{1}{2^q - 1} \int_0^1 \log x \, dx = \log 2.$$
These inequalities together with (12) imply Theorem 2.

We now prove one corollary to Theorem 1, that the true value of $N_k$ for any $k$ lies somewhere between the trivial lower bound of $N_k \geq 1$ and the upper bound of $N_k = n-1$ which was our first result.

In the construction of the lower bound for $N_k$, the construction of the basis set used to simplify the game is most important. We now prove one corollary to Theorem 1. In fact, it is not at all clear that the lower bound is the best possible.

**Theorem 3.** If $P(n_1, n_2, \ldots, n_k)$ is a sequence of elements of $A^k$ and $A - B$ is the maximal dependence of $A(n_1, n_2, \ldots, n_k)$ common for all $k$, then

$$N_k(n_1, n_2, \ldots, n_k) \leq 1$$

where $n_k$ does not depend on $n$.

**Proof.** Without loss of generality we can assume that $A(n_1, n_2, \ldots, n_k, B) = A(n_1, n_2, \ldots, n_k)$. The possible maximal dependencies of $P(n)$ are all of the form

$$A(n_1, n_2, \ldots, n_k) = B(n_1, n_2, \ldots, n_k)$$

where $A, B$ are both $2$-confluent. Since $A(n_1, n_2, \ldots, n_k)$ and $B(n_1, n_2, \ldots, n_k)$ are types $2^i$-elemental elements of type $2$, and similarly, the $W(n_1, n_2, \ldots, n_k)$ are the highest number of maximal elements of type $2$ in $2^i \leq 2^i$.

Hence

$$N_k(n_1, n_2, \ldots, n_k) \leq 1$$

as stated.

**References**


Netherlands 7 (1978) 1-10.