Theoretical Computer Science 78 (1991) 377-381
North-Holland

Note

On the number of databases and closure operations

G. Burosch

Wilhelm Pieck Universität, 2500 Rostock, Universitätsplatz 1, FRG

J. Demetrovics*

Computer and Automation Institute, Hungarian Academy of Sciences, Budapest, Kende u. 13-17., 1111 Hungary

G.O.H. Katona*

Mathematical Institute, Hungarian Academy of Sciences, Budapest, P.f. 127, 1364 Hungary

D.J. Kleitman

Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

A.A. Sapozhenko

Department of Computer Science, Moscow State University, Moscow, U.S.S.R.

Communicated by G. Ausiello
Received February 1988
Revised July 1989

Abstract

Burosch, G., J. Demetrovics, G.O.H. Katona, D.J. Kleitman and A.A. Sapozhenko, On the number of databases and closure operations (Note), Theoretical Computer Science 78 (1991) 377-381.

Closure operations are considered as models of databases. Estimates on the number of closure operations on n elements (or equivalently, on the number of databases with n attributes) are given.

* The work of these authors is supported by the Hungarian National Foundation for Scientific Research, grant number 1066.

Codd [3] and Armstrong [1] introduced the system of functional dependencies as a model of a database. We, however, prefer another equivalent variant, the closure operation (see. e.g. [2]). Let X be the (finite) set of *attributes*, that is, the set of types of data. The elements of X are words like "name", "date of birth", "age", etc. Some of the data determine some other data uniquely. For instance, the date of birth determines the age. Let $A \subseteq X$, $a \in X$. We say that A *determines* a and write $A \to a$ iff the set of data in A determines the data in a, more precisely, there are no two individuals having the same data in A and different in a. The function $\mathcal{L}: 2^X \to 2^X$ is defined by

$$\mathcal{L}(A) = \{a : A \to a\}.$$

This function obviously possesses the following properties:

$$A \subseteq \mathcal{L}(A),$$

$$A \subseteq B \text{ implies } \mathcal{L}(A) \subseteq \mathcal{L}(B),$$

$$\mathcal{L}(\mathcal{L}(A)) = \mathcal{L}(A).$$

Such a function is known as a *closure operation* or briefly a *closure*. Therefore a closure is a possible model of a database.

$$\mathcal{L}(\emptyset) = \emptyset$$

is a rather natural assumption for closures formed from databases. In the present paper we will use the name closure for the functions satisfying this additional condition.

Let A, $B \subseteq X$. Following [1], we say that A *determines* B iff the set of data in A determines the data in B uniquely, more precisely, iff there are no two individuals having the same data in A but different in B. We write $A \to B$ in this case and $A \to B$ is called a *functional dependency*. The functional dependencies satisfy four natural conditions, the so called *Armstrong axioms*. In the present paper we add one more axiom:

$$\emptyset \to B \text{ implies } B = \emptyset.$$

A set of pairs $A \to B$ satisfying these five axioms is called a *system of functional dependencies*.

It is easy to see that

$$A \to B \text{ iff } B \subseteq \mathcal{L}(A)$$

holds for the system of functional dependencies and the closure, respectively, defined by a given database. It is easy to see [4] that this is a bijection between the set of closures and the set of all systems of functional dependencies defined on the same groundset. That is, we have the right to consider the closures only, instead of the systems of functional dependencies.

In the present note we investigate a very natural question: what is the number of closures on an n-element set?

Let $\mathcal{L}$ be a closure. The *closed sets* C are defined by $\mathcal{L}(C)=C$. It is known that the family $Z=Z(\mathcal{L})$ of closed sets possesses the following properties:

$$\emptyset, X \in Z, \tag{1}$$

$$A, B \in Z \text{ implies } A \cap B \in Z. \tag{2}$$

The families satisfying (1) and (2) are called *intersection semi-lattices*. It is shown in [4] that the function $f: \mathcal{L} \to Z(\mathcal{L})$ is a bijection between the set of closures and the set of intersection semi-lattices. Thus the number of closures is equal to the number of intersection semi-lattices.

If Z is an intersection semi-lattice, let $\mathcal{M}(Z)$ denote the family of those members $C \in Z$ which are not intersections of two other members of Z, that is, $A \neq C \neq B$, $A, B \in Z$ imply $A \cap B \neq C$. It is obvious that $\mathcal{M}=\mathcal{M}(Z)$ satisfies the following properties:

$$X \in \mathcal{M}, \tag{3}$$

$$\bigcap_{A \in \mathcal{M}} A = \emptyset, \tag{4}$$

$$A = \bigcap_{i=1}^{r} A_i, \quad A, A_1, \ldots, A_r \in \mathcal{M}, \ (r \geq 1)$$

$$\text{imply } A = A_i \quad \text{for some } i \ (1 \leq i \leq r). \tag{5}$$

The families satisfying (3), (4) and (5) are called *intersection-free families*. It is also proved in [4] that the function $f: Z \to \mathcal{M}(Z)$ is a bijection between the set of intersection semi-lattices and intersection-free families. Thus the number of closures is equal to the number of intersection-free families.

Kleitman [6] proved that if the family $\mathcal{A} \subset 2^X$ ($|X|=n$) contains no three distinct members A, B, C satisfying $A \cap B = C$ then

$$|\mathcal{A}| \leq \binom{n}{n/2}(1+o(1)). \tag{6}$$

This inequality already implies that the number of intersection-free families cannot be too large. The following proposition should be used, only, which can be proved by a straightforward but tedious calculation.

Proposition. *The number of families $\mathcal{A} \subset 2^X$ of subsets of an n-element set X, satisfying*

$$|\mathcal{A}| \leq c\binom{n}{n/2}$$

is at most

$$2^{\frac{1}{2}c\binom{n}{n/2}\log n(1+o(1))}. \tag{7}$$

Equations (6) and (7) prove that the number of intersection-free families is at most

$$2^{\binom{n}{n/2}\log n(1+o(1))}.$$

We can, however, improve this upper estimate.

Theorem. *The number* $\alpha(n)$ *of the closures (databases, intersection semi-lattices, intersection-free families) satisfies the following inequalities*:

$$2^{\binom{n}{\lfloor n/2\rfloor}} \leqslant \alpha(n) \leqslant 2^{2\sqrt{2}\binom{n}{\lfloor n/2\rfloor}(1+o(1))}. \tag{8}$$

Proof. Any family consisting of $\lfloor n/2\rfloor$-element members is intersection-free, therefore the number of intersection-free families exceeds the left hand side estimate of the theorem.

To prove the right hand side, partition the ground-set into two subsets X_1 and X_2 of the same cardinality (suppose that n is even). Let $\mathcal{F}$ be an intersection-free family. Define

$$\mathcal{F}_1 = \{F\colon F \in \mathcal{F},\ \nexists G \in \mathcal{F} \text{ s.t. } F \cap X_1 = G \cap X_1 \text{ and } F \cap X_2 \subset G \cap X_2\} \tag{9}$$

and

$$\mathcal{F}_2 = \{F\colon F \in \mathcal{F},\ \nexists G \in \mathcal{F} \text{ s.t. } F \cap X_2 = G \cap X_2 \text{ and } F \cap X_1 \subset G \cap X_1\}. \tag{10}$$

Suppose that $F \notin \mathcal{F}_1$ and $F \notin \mathcal{F}_2$ but $F \in \mathcal{F}$. By (9) and (10) there exist two subsets G_1 and G_2 in $\mathcal{F}$ satisfying $F \cap X_1 = G_1 \cap X_1$, $F \cap X_2 \subset G_1 \cap X_2$, $F \cap X_2 = G_2 \cap X_2$ and $F \cap X_1 \subset G_2 \cap X_1$. $F = G_1 \cap G_2$ is obvious and contradicts the assumption that $\mathcal{F}$ is intersection-free. This proves

$$\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2. \tag{11}$$

Let F and G be members of $\mathcal{F}_1$ such that $F \cap X_1 = G \cap X_1 = A$. By (9) we have $F \cap X_2 \not\subset G \cap X_2$ and $F \cap X_2 \not\supset G \cap X_2$ that is,

$$\mathcal{F}_1(A) = \{B\colon B \subseteq X_2,\ A \cup B \in \mathcal{F}_1\}$$

is inclusion-free (no member contains another member as a proper subset) for any $A \subseteq X_1$. It is known ([5], see also the sharper result [7]) that the number of inclusion-free families on an n-element set is at most

$$2^{\binom{n}{\lfloor n/2\rfloor}(1+o(1))}.$$

This implies that the number of possible families $\mathcal{F}_1(A)$ (for a fixed A) is at most

$$2^{\binom{n/2}{\lfloor n/4\rfloor}(1+o(1))}.$$

$\mathcal{F}_1$ is determined by the families $\mathcal{F}_1(A)$. In the worst case they can be chosen independently, so the number of possible families $\mathcal{F}_1$ is at most

$$2^{\binom{n/2}{\lfloor n/4\rfloor}2^{n/2}(1+o(1))} = 2^{\sqrt{2}\binom{n}{n/2}(1+o(1))}. \tag{12}$$

The same is true for the number of choices of $\mathcal{F}_2$. By (11) $\mathcal{F}$ is determined by $\mathcal{F}_1$ and $\mathcal{F}_2$. Therefore the number of possible intersection-free families $\mathcal{F}$ is upper-bounded by the square of (12). The statement is proved for even n. The case of odd n is analogous. □

Conjecture. *The constant* $2\sqrt{2}$ *can be omitted in the exponent of the right hand side of* (8).

Remark. If the condition $\mathscr{L}(\emptyset)=\emptyset$ is omitted from the definition of the closure then the total (real) number of closures can be expressed as

$$\sum_{i=0}^{n}\binom{n}{n-i}\alpha(i).$$

This expression satisfies the estimates of the Theorem, again.

References

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

[2] G. Burosch, J. Demetrovics and G.O.H. Katona, The poset of closures as a model of changing databases, *Order* **4** (1987) 127–142.

[3] E.F. Codd, A relational model of data for large shared data banks, *Comm. ACM* **13** (1970) 377–387.

[4] J. Demetrovics and G.O.H. Katona, Combinatorial problems of database models, *Coll. Math. Soc. J. Bolyai*, **42**. *Algebra, Combinatorics and Logic in Computer Science* (Győr, Hungary, 1983) 331–353.

[5] D.J. Kleitman, On Dedekind's problem: the number of monotone Boolean functions, *Proc. Amer. Math. Soc.* **21** (1969) 677–682.

[6] D.J. Kleitman, Extremal properties of collections of subsets containing no two sets and their union, *J. Combinatorial Theory* (*A*) **20** (1976) 390–392.

[7] A.D. Korshunov, On the number of monotone Boolean functions, *Problemy Kibernet.* **38** (1981) 5–108 (in Russian).