On the equivalence of candidate keys with Sperner systems

By J. DEMETROVICS

1. Introduction

The use of the relational data model proposed by E. F. CODD [1—3] is to make many problems mathematically describable. In this model all data are represented by two-dimensional tables with rows representing records, and with coloumns representing attributes. Rows are identified by the values of a subset of attributes, if these are not identical for two different rows. These subsets of attributes are called keys and those keys which contain no further keys as subsets are called candidate keys.

Functional dependencies were introduced in 1970 by Codd, but were investigated mathematically only later [4, 5, 8]. In this paper we prove, that for any Sperner system we can construct a relation the set of candidate keys of which is the same as the Sperner system. It is clear, that apart from trivial cases the set of candidate keys of any relation is a Sperner system. At most $\binom{n}{\left[\frac{n}{2}\right]}$ candidate keys may exist in a relation of n attributes and we prove that this limit can be reached by relations with linear dependencies.

2. Definitions

Definition 1. Given the not necessarily different sets $D_1, D_2, \ldots, D_n$, the relation R of n variables denoted by $R(n)$ is a subset of the Cartesian product $D_1 \times D_2 \times \ldots \times D_n$. We shall call the sets D_i domains.

Definition 2. Indices of the domains of the relation $R(n)$ will be called attributes. Values associated to attributes will be called attribute values.

Remark 1. Though the domains of a relation are not necessarily distinct, their attributes are distinct.

In the present paper all domains are sets of natural numbers and the set of their indices in $R(n)$ are denoted by

$$N \quad (N = \{1, 2, \ldots, n\}).$$

Definition 3. The subset of indices $A \subseteqq N$ will be said to generate the index k, in notation $A \to k$, if in any row of the relation $R(n)$ the values d_j ($j \in A$) determine the value d_k uniquely. If in addition, for all rows in $R(n)$, d_k is a linear combination of d_j's ($j \in A$), than A generates the index k linearly. The subset of indices $B \subseteqq N$ will be said to be generated by A if every index in B is generated by A, denoted by $A \to B$. The link $A \to B$ is called a functional dependency in the relation $R(n)$. If A generates every index in B linearly, we say the functional dependency is linear. The set of all functional dependencies in $R(n)$ is denoted by $\{A_i \to B_i\}$ ($i = 1, 2, \ldots, v$).

Definition 4. Let $A \subseteqq N, A \neq \emptyset$ and $A \to N$. A is called a candidate key in the relation $R(n)$ if $B \to N$ does not hold for any of its nontrivial subsets B.

Definition 5. The functional dependency $A \to B$ is trivial if $B \subseteqq A$. The sets of trivial and nontrivial functional dependencies in the relation $R(n)$ will be denoted by $\mathscr{H}$ and $\mathscr{G}$, respectively.

Remark 2. It is easy to see that in a relation $R(n)$

$$|\mathscr{H}| = 3^n$$

3. The link between candidate keys and Sperner systems

In the present paragraph we shall demonstrate a one-to-one correspondence between the set of candidate keys in a relation $R(n)$ and a Sperner system $\mathscr{S}(n)$ over N.

Definition 6. Let $\mathscr{S} = \{S_1, S_2, \ldots, S_m\} \subseteqq 2^N$. $\mathscr{S}$ will be called a Sperner system if it satisfies the following relations,

$$S_i \subset N \quad \text{for} \quad i = 1, 2, \ldots, m; \tag{1}$$

$$S_i \nsubseteq S_j \quad \text{for} \quad i \neq j, \quad i, j = 1, 2, \ldots, m. \tag{2}$$

Trivially, the set of the candidate keys in every relation is a Sperner system or has only one element N. Conversely consider now the following Sperner system:

$$\mathscr{S} = \begin{pmatrix} S_1 = \{a_{11}, a_{12}, \ldots, a_{1m_1}\} \\ S_2 = \{a_{21}, a_{22}, \ldots, a_{2m_2}\} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ S_m = \{a_{m1}, a_{m2}, \ldots, a_{mm_m}\} \end{pmatrix}$$

with $a_{ij} \in N$ and $\bigcup_{i,j} a_{ij} = N$. This Sperner system is a covering of N.

Theorem 1. To every Sperner system $\mathscr{S}$ a relation $R_{\mathscr{S}}(n)$ of n variables can be constructed with the set of the candidate keys equivalent to the Sperner system $\mathscr{S}$.

Proof. First we shall construct the class of sets $\mathscr{M} = \{A_1, A_2, \ldots, A_t\}$. Let A_j belong to $\mathscr{M}$ iff the following conditions hold:

$$A_j \subseteqq N, \quad j = 1, 2, \ldots, t \tag{3}$$

and

$$A_j \cap S_i \neq \emptyset \quad \text{for} \quad i = 1, 2, \ldots, m. \tag{4}$$

We shall choose $\mathscr{F}$ as the set of the elements minimal in $\mathscr{M}$, i.e.

$$A_j \in \mathscr{F} \Leftrightarrow \bar{\exists} A_l \in \mathscr{M}\colon (A_l \subset A_j). \tag{5}$$

From (3), (4) and (5) we have

$$\max\{m_1, m_2, \ldots, m_m\} \leqq |\mathscr{F}| \leqq m_1 \cdot m_2 \cdot \ldots \cdot m_m \tag{6}$$

and

$$A_j \in \mathscr{F} \quad \text{implies} \quad 1 \leqq |A_j| \leqq m. \tag{7}$$

Let us consider the following subsets

$$\mathscr{F}_k (k = 1, 2, \ldots, n) \quad \text{of} \quad \mathscr{F}\colon$$

$$A_j \in \mathscr{F}_k \Leftrightarrow k \in A_j \in \mathscr{F}. \tag{8}$$

We state that if the k'th index of the relation $R_{\mathscr{S}}(n)$ is determined by the function f_k, the latter identical with the class of sets $\mathscr{F}_k$, then the relation obtained satisfies the conditions of Theorem 1, i.e. the class of the candidate keys in $R_{\mathscr{S}}(n)$ is identical with the given system $\mathscr{S}$. Obviously, this last statement is implied by the following three statements:

a) all the sets S_i in the class $\mathscr{S}$ are keys;

b) no proper subset of S_i is key;

c) there is no candidate key beyond $\mathscr{S}$.

To verify these first we consider

a) Each S_i containing a key K_i $(i=1, 2, \ldots, m)$ is a consequence of

$$\bigcup_{k \in S_i} \mathscr{F}_k = \mathscr{F}. \tag{9}$$

This latter is obvious, as every $A_j \in \mathscr{F}$ is constructed so as to contain at least one element of S_i.

Next we show that the key K_i in S_i equals S_i. To do this

$$\forall a (a \in S_i)\colon \bigcup_{k \in \{A_i \setminus \{a\}\}} \mathscr{F}_k \subsetneqq \mathscr{F} \setminus \{A\} \quad \text{with} \quad A \in \mathscr{F} \tag{10}$$

is sufficient.

This follows from the existence of an $A \in \mathscr{F}$ with $A \cap S_i = \{a\}$. Indeed, for $j=1, 2 \ldots m$, every S_j contains either $\{a\}$ or some $\{a'\}$ with $\{a'\} \cap S_i = \emptyset$.

So we have proved that every $S_i \in \mathscr{S}$ is identical with a minimal key in the relation $R_{\mathscr{S}}(n)$. Now all we have left to prove is that $R_{\mathscr{S}}(n)$ has no minimal key K beyond those in $\mathscr{S}$.

For an indirect proof let us suppose the existence of such a minimal key. From Remark 1 we have $S_i \cap \bar{K} \neq \emptyset$ for $i=1, 2, \ldots, m$. Let the set A be determined by the sets c_i so that $A \in \alpha$ and $A \cap c_i \neq \emptyset$. It is easy to see, that at least one such set A exists and it is not contained in any of the columns determined by the candidate key K, i.e.

$$\bigcup_{k \in K} \mathscr{F}_k \subsetneqq \mathscr{F} \setminus \{A\}. \tag{11}$$

This completes the proof of the theorem.

Remark 3. Let us observe that the proof can be carried out the same way if such a class $\mathcal{L}$ of subsets in $\mathcal{M}$ is taken that $\mathcal{M} \supset \mathcal{L} \supset \mathcal{F}$ is fulfilled instead of $\mathcal{F}$. Out of theset he one of minimal cardinality was taken for our proof. If another have been taken, the set of functional dependencies of a form different from $x \to N$ would be changed and the set of candidate keys $\mathcal{S}$ would be unchanged.

The preceding statements can be interpreted as follows: let different prime numbers correspond to each set in the class $\mathcal{F}$, i.e. let $\mathcal{F}=\{p_1, p_2, \ldots, p_h\}$ be in ascending order for simplicity. So the sets in the classes $\mathcal{F}_k$ have their correspondants as well. Let then the function f_k of $|\mathcal{F}_k|$ variables equal the product of the corresponding primes to the sets in $\mathcal{F}_k$.

For example, let $n=5$ and

$$\mathcal{S}=\{\{1,2,3\}=s_1, \{3,4,5\}=s_2, \{1,3,4\}=s_3\}.$$

Then $\mathcal{F}=\{\{3\}=p_1, \{1,4\}=p_2, \{1,5\}=p_3, \{2,4\}=p_4\}$ and $f_1=p_2 \cdot p_3$, $f_2=p_4$, $f_3=p_1$, $f_4=p_2 \cdot p_4$, $f_5=p_3$. Some rows of the relation $R_{\mathcal{S}}(n)$ corresponding to $\mathcal{S}$ are represented in Fig. 1 for

$$\mathcal{F}^1=\{2,3,5,7\}$$
$$\mathcal{F}^2=\{2,5,7,11\}$$
$$\mathcal{F}^3=\{3,5,7,11\}$$
$$\mathcal{F}^4=\{2,5,7,13\}$$

1	2	3	4	5
15	7	2	21	5
35	11	2	55	7
35	11	3	55	7
35	13	2	65	7

Fig. 1

i.e. $R_{\mathcal{S}}(5) \in \{(15,7,2,21,5), (35,11,2,55,7), (35,11,3,55,7), (35,13,2,65,7)\}$.

4. On the maximal number of candidate keys and on linear relations

Definition 7. We shall call the relation $R(n)$ linear provided all the functional dependencies in it are linear.

First we recall here Lemmas 1 and 2 and a Theorem from [8] in stronger forms. Namely, the result of the construction in the proof of Lemma 2 is a linear relation, therefore we can formulate both of them and the Theorem (as a consequence of the two Lemmas) as follows.

Lemma 1. A relation $R(n)$ may have at most $\binom{n}{\left[\frac{n}{2}\right]}$ candidate keys.

Lemma 2. There exists a linear relation $R(n)$ with $\binom{n}{\left[\frac{n}{2}\right]}$ candidate keys.

Theorem 2. There are linear relations $R(n)$ with as many candidate keys as $\binom{n}{\left[\frac{n}{2}\right]}$ and there is no relation $R(n)$ with more candidate keys.

Lemma 3. In a linear relation $R(n)$ all candidate keys have the same length.

Proof. Let A_k be a candidate key. As a consequence of the fact, that the functional dependency $A \to N$ is linear, we have a linear equation system

$$\sum_{j \in A_k} a_{ij}^k x_j = x_i \quad (i = 1, 2, \ldots, n),$$

which is satisfied by every row in $R(n)$. This is true for every candidate key A_k $(k=1, 2, \ldots, m)$, so we have the system

$$\sum_{j \in A_k} a_{ij}^k x_j = x_i \quad (i = 1, 2, \ldots, n;\ k = 1, 2, \ldots, m)$$

with the solution $R(n)$ in the preceding sense. Obviously, the set of indices of an independent set of variables $x_{i_1}, x_{i_2}, \ldots, x_{i_t}$ in this system composes a candidate key in $R(n)$ and conversely. Moreover, independent sets of variables have the same cardinality t, which completes the proof of Lemma 3.

As a consequence of this lemma, for linear relations Theorem 1 does not hold. Neither exist linear relations to every Sperner system $\mathcal{S}$ with the set of their candidate keys equivalent to it, as e.g. for $n=4$ and the Sperner system

$$\mathcal{S} = \{\{1, 2\}, \{1, 3\}, \{3, 4\}\}.$$

Considering a linear equation system as in the proof of Lemma 3 which has all subsets of the variables with the cardinality t independent, we have proved:

Theorem 3. There exists a linear relation $R(n)$ with $\binom{n}{t}$ candidate keys with t being their length.

In [5] it was proved that provided the number of dependencies $k \leqq \sqrt{n}$, a relation $R(n)$ exists with as many candidate keys as $\sqrt{n}!$.

S. Osborne and F. Tompa have recently proved (draft paper) that at most $k!$ candidate keys can be deduced from k dependencies and for each k a relation R_k exists with exactly $k!$ candidate keys.

Each of the papers uses a system of derivation axioms which were introduced in [7] and [4], respectively. The first of them consists of 7 and the second of 4 axioms. Next we shall give a system of 3 axioms which is equivalent to the ones mentioned above.

Definition 8. The functional dependency $A \to B$ is deductible from the set of lunctional dependencies $\mathcal{F} = \{A_i \to B_i,\ i=1, 2, \ldots, k\}$ if it can be obtained from the latter using the derivation rules a; b; and c; a finite number of times.

a; $A \to A'$ with $A \supseteqq A'$ is deductible from all $\mathcal{F}$,

b; $(A \to B) \in \mathcal{F}$ and $(B \to C) \in \mathcal{F}$ imply $(A \to C) \in \mathcal{F}$,

c; $(A \to B) \in \mathcal{F}$ and $(A \to C) \in \mathcal{F}$ imply $(A \to (B \cup C)) \in \mathcal{F}$.

By Theorem 4 an example is recalled from [8] in which the number of the undeductible functional dependencies is relatively high and this does not essentially diminish the number of candidate keys.

Theorem 4. Let $k=\binom{n}{\left[\frac{n}{2}\right]}$. A relation T of $n+1$ attributes exists with k undeductible functional dependencies and with the same number of candidate keys.

COMPUTER AND AUTOMATION INSTITUTE
HUNGARIAN ACADEMY OF SCIENCES
KENDE U. 13—17.
BUDAPEST, HUNGARY
H—1502

References

[1] CODD, E. F., A relational model of data for large shared data banks, *Comm. ACM*, v. 13, 1970, pp. 377—387.

[2] CODD, E. F., Normalized data base structure: A brief tutorial, *Proc.* ACM-SIGFIDET Workshop on Data Description, Access and Control 1971.

[3] CODD, E. F., Further normalization of the data base relational model, Courant Computer Science Symposia 6 *Data Base System,* Prentice Hall, Englewood Cliffs, N. J., 1971, pp. 33—64.

[4] ARMSTRONG, W. W., Dependency structures of data base relationships, *Information Processing 74,* North-Holland Publ. Co., 1974, pp. 580—583.

[5] YU, C. T., D. T. JOHNSON, On the complexity of finding the set of candidate keys for a given set of functional dependencies, *Inform. Process. Lett.*, v. 5, 1976, No. 4, pp. 100—101.

[6] SPERNER, E., Eine Satz über Untermengen einer endlichen Menge, *Math. Z.* v. 27, 1928, pp. 544—548.

[7] DELOBEL, C., R. C. CASEY, Decomposition of a data base and the theory of boolean switching functions, *IBM J. Res. Develop.*, v. 17, 1973, pp. 374—386.

[8] DEMETROVICS, J., On the number of candidate keys, *Inform. Process. Lett.*, v. 7, 1978, No. 6. pp. 266—269.

(Received April 2, 1979)