On the numbers of shortest keys in relational databases on nonuniform domains

O. SELESNJEW, B. THALHEIM

The use of the relational model of data structures by E. F. Codd is a promising mathematical tool for handling data. In this model the user's data are expressed as relations where the rows denote the records and the columns represent domains or attributes. For the handling of relations the identification of sets of domains, called keys, is suggested. The keys uniquely determine the values of the rest of the domains. Delobel and Casey, Fadous and Forsyth, M. Fernandez, C. L. Lucchesi and S. L. Osborn, J. Demetrovics and V. Thi have given different algorithms for finding the set of all minimal keys in a relational database given by a set of functional dependencies on the database. For characterizing the complexity of this algorithms we need some combinatorial bounds.

In this paper we consider the maximal numbers of shortest keys in relational databases on weighted domains and extend the result of J. Demetrovics who solved the problem for relational databases on uniform domains. [1]

1. Basic notions

We recall briefly some definitions of the theory of relational databases. Given sets $D_1, D_2, \ldots, D_n$, called domains, not necessarily distinct, an n-ary relation R defined over $D_1, \ldots, D_n$ is a subset of the cartesian product $D_1 \times D_2 \times D_n$.

An attribute is a name assigned to a domain of a relation. Any value associated with an attribute is called attribute value. The attribute names must be distinct. The symbol U will be used to denote the set of all n attributes of R.

A set of attributes X, $X \subseteq U$, is called a *key* of R if, for every n-tuple of R, the values of the attributes in X uniquely determine the values of the attributes in U.

Now, suppose we are given some weight function (or complexity measure) $g: U \to \mathbf{N}'$ and a system S_R of keys of R. For $X \subseteq U$ let $g(X) = \sum_{A \in X} g(A)$ the complexity of X. An element K of S_R is called *g-shortest* if there does not exist an element K' of S_R with $g(K') < g(K)$. By $S_R(g)$ we denote the set of all g-shortest elements of S_R and by $s_R(g)$ its cardinality. For $g=1$ the set $S_R(g)$ is called the set of all shortest keys or the set of shortest keys in an unweighted database. It is obvious that

any set $S_R(g)$ is a subset of a set of minimal keys [1]. For any set S of minimal keys there exists a subset $S(g)$ of shortest keys. This is well-known for $g=1$.

Theorem 1. [1] The maximal size of a set of shortest keys in a database with n attributes is $\binom{n}{\left[\frac{n}{2}\right]}$. For any k, $1 \leqq k \leqq \binom{n}{\left[\frac{n}{2}\right]}$, there is an n-ary relation R with k shortest keys.

2. Maximal number of shortest keys in nonuniform databases

In practical cases, keys are of different meaning and complexity. Domains for attributes have very different complexity. This is well-known in practice but it is not taken into consideration in the theory of minimal keys. Therefore, shortest keys are introduced.

Lower and upper bounds for $s_R(g)$ are proved in [4]. The most interesting set of functions g is the set G^+ of functions g with $g(A_i) \neq g(A_j)$ for $i \neq j$. The other cases can be splitted in the case $g(A)=1$ for $A \in X$ and in this case for $A \in U \setminus X$. We introduce the following functions:

$s(g) = \max_R s_R(g)$,

$s(G') = \max_{g \in G'} s(g)$ for sets G' of weight functions.

Using the functions g_1, g_2, g_3 with $g_1(A_i)=2^i$, $g_2(A_i)=3^{[i/2]}$, $g_3(A_i)=i$, for i, $1 \leqq i \leqq n$, by the definitions and the recursion formulas for g_3, we get

Corollary 2. 1. For weight functions g it holds $1 \leqq s(g) \leqq \binom{n}{\left[\frac{n}{2}\right]}$ [1].

2. $s(g_1)=1$, $s(g_2)=2^{[n/2]}$, $s(g_3) \geqq \frac{2^n}{n^2}$ [4].

Our next aim is to prove

Theorem 3. $s(G^+) = \frac{2^n}{\sqrt{\frac{\pi}{6} n^{3/2}}} (1-o(1))$.

From number theory [2] we get that functions g with $s(g)=s(G^+)$ must be regulary. W.l.o.g. we consider a subclass G^* of G^+, the class of all equidistant functions g with the property $g(A_i)-g(A_{i-1})=c$ for some c and any i, $2 \leqq i \leqq n$.

Lemma 4. 1. Given two equidistant functions g, g' from G^+. Then $s(g)=s(g')$.

2. Let g be a function from G^+. There exists an equidistant function g' in G^* such that $s(g)=s(g')$.

Proof. 1. Is obvious.

2. W.l.o.g. we consider only functions g from G^+ with $g(A_i)<g(A_{i+1})$ for i, $1 \leqq i < n$. We prove the assertion by induction. For $n=2$ the assertion is obvious. Let n be a fixed number. Now we assume that for a fixed function g there is no equi-

distant function $g' \in G^*$ such that $s(g) \leqq s(g')$. Let be S_R a key system with $s(g) = s_R(g)$. Now we define $S_1 = \{K \in S_R / A_n \notin K\}$, $S_2 = \{K \setminus \{A_n\} / K \in S_R, A_n \in K\}$. By precondition of induction, we get for $g' = g|_{U \setminus \{A_n\}}$ an equidistant function g'' such that $s(g') \leqq s(g'')$. It follows that there is an equidistant function g^+ in G^* such that $g^+|_{U \setminus \{A_n\}} = g''$ and $s(g) \leqq s(g^+)$. That is a contradiction.

W.l.o.g. we can consider for $s(G^+)$ the function g_3 of Corollary 2. Now we define independent random variables ξ_k with two-point distribution for $k = 1, 2, \ldots, n$:

$$\xi_k = \begin{cases} k \\ 0 \end{cases}$$

and consider the distribution of $S_n = \sum_{i=1}^{n} \xi_i$.

Corollary 5. $P\left(S_n = \left[\frac{n(n+1)}{4}\right]\right) \cong \frac{1}{2^n} s(g_3)$ for probability $P(S_n = m)$.

For the expectation ES_n and the variance DS_n of S_n we get

$$M_n = ES_n = \sum_{k=1}^{n} E\xi_k = \sum_{k=1}^{n} \frac{k}{2} = \frac{n(n+1)}{4}$$

$$B_n^2 = DS_n = \sum_{k=1}^{n} D\xi_k = \frac{n(n+1)(2n+1)}{24} \sim \frac{n^3}{12} \quad (n \to \infty).$$

We shall say that the sequence $\{S_n\}$ satisfies a local limit theorem iff

$$\sup_m |B_n P(S_n = m) - \varphi(x_{nm})| \to 0 \quad (n \to \infty)$$

where $x_{nm} = \frac{m - M_n}{B_n}$, $z_n = \frac{S_n - M_n}{B_n}$, φ is the standard normal distribution density. We denote

$$\alpha(a, q, N) = \frac{1}{q^2} \sum_{-q/2 < r \leqq q/2} r^2 P(a\tilde{\xi}_k = r \pmod q, \quad |\xi_k| \leqq N) \qquad (+)$$

for $\tilde{\xi}_k = \xi_k - \xi_k'$ symmetrized random variable, where ξ_k' is a random variable independent of ξ_k and having the same distribution as ξ_k, relatively prime integers a, q with $a \leqq \frac{q}{2}$ and $1 < q \leqq 2N$.

In [3] is proved the following: If the distribution function of the sum of unboundedly increasing number of random variables converges to the standard normal distribution function,

$$\text{i.e.} \quad z_n \xrightarrow{D} N(0, 1), \qquad (1)$$

$$N_n \exp\left\{-\frac{1}{2} \min_{a,q} \sum_{k=1}^{n} \alpha_k(a, q, N_k)\right\} \to 0 \quad (n \to \infty), \qquad (2)$$

$$N_n \text{ is selected such that } \lim_{n \to \infty} \frac{1}{B_n^2} \sum_{k=1}^{n} \int_{|x| \leqq N_n} x^2 \, dF_{\tilde{\xi}_k}(x) = l > 0, \qquad (3)$$

then the sum satisfies a local limit theorem.

Let $N_n=n$. Then we get

$$l = \lim_{n\to\infty} \frac{1}{B_n^2} \sum_{k=1}^{n} D\xi_k^2 = 1 > 0,$$

$$P(\tilde{\xi}_k = k) = P(\tilde{\xi}_k = -k) = \frac{1}{4}, \quad P(\tilde{\xi}_k = 0) = \frac{1}{2}.$$

By summation of (+) over representatives of q we get $|\tilde{\xi}_k| \leqq n$ for $k \in \{1, \ldots, n\}$. Observe that if $\tilde{\xi}_k = 0$ then $r=0$ and this summand can be eliminated and that if $\tilde{\xi}_k = k$ then $ak = r_k + ql_k$ for the unique representative of q. We get

$$\alpha_k(a, q, n) = \frac{1}{q^2} \sum_{-q/2 < r \leqq q/2} r^2 P(a\tilde{\xi}_k = r \pmod q) =$$

$$= \frac{1}{4q^2}(r_k^2 + r_{-k}^2) \geqq \frac{1}{4q^2} r_k^2.$$

From number theory it is known that if $\{x\}$ form a full system of representatives of q then $\{ax\}$ form a full system of representatives. Now $\lambda_n \geqq \min\limits_{a,q} \sum\limits_{k=1}^{n} \alpha_k(a, q, n) \geqq$ $\geqq \min\limits_{q} \frac{1}{4q^2} \sum\limits_{k=1}^{n} r_k^2$. Assume that $q=2m$. (For odd q proof is analogous.) Let $0 < \alpha < \frac{1}{2}$. If $\alpha_n \leqq m \leqq n$ then

$$\sum_{k=1}^{n} r_k^2 \geqq \sum_{k=1}^{m} k^2 \geqq cm^3 \geqq c\alpha^3 n^3$$

for the full system of representatives $r_k = -(m-1), \ldots, 0, 1, \ldots, m$ and therefore

$$\lambda_n \geqq \min_q \frac{1}{4q^2} c\alpha^3 n^3 \geqq \frac{c\alpha^3}{4\alpha^2 4} \frac{n^3}{n^2} = \beta n, \quad \beta > 0.$$

If $1 < m < \alpha n$ then the full system of representatives $\{r\}_{-(m-1)}^{m}$ is contained in $\{1, \ldots, n\}$ at least $\left[\frac{n}{q}\right]$ times. Consequently we get

$$\lambda_n \geqq \min_{a,q} \frac{1}{4q^2} \sum_{k=1}^{n} r_k^2 \geqq \min_{a,q} \frac{\left[\frac{n}{q}\right] \sum\limits_{k=1}^{m} k^2}{4q^2} \geqq$$

$$\geqq \min_{a,q} \left(\frac{n}{q} - 1\right) \frac{\sum\limits_{k=1}^{m} k^2}{4q^2} \geqq \min_{q} (n-q) \frac{\sum\limits_{k=1}^{m} k^2}{4q^3} \geqq \min_{q} (n - 2\alpha n)c =$$

$$= c(1-2\alpha)n = \beta n > 0, \quad \beta > 0.$$

We get that (2) holds because $\Delta_n = n \exp\left\{-\frac{1}{2}\lambda_n\right\} \leqq n \exp\left\{-\frac{\beta}{2}n\right\} \to 0 (n \to \infty)$. Summarizing corollary 5, lemma 4 and the properties of S_n we get

$$s(g_3) = 2^n P\left(S_n = \left[\frac{n(n+1)}{4}\right]\right) = \frac{2^n}{B_n} \varphi\left(x_{n\left[\frac{n(n+1)}{4}\right]}\right) \underset{n\to\infty}{\sim}$$

$$\underset{n\to\infty}{\sim} \frac{2^n}{\sqrt{2\pi \frac{n(n+1)(2n+1)}{24}}} = \frac{2^n}{\sqrt{\frac{\pi}{6}} n^{3/2} \left(1 + \frac{3}{2n} + \frac{1}{2n^2}\right)^{1/2}} \sim \frac{2^n}{\sqrt{\frac{\pi}{6}} n^3}.$$

The proof of theorem 3 is complete.

It is of interest to compare this result with $s(g_4) \sim \frac{2^n}{\sqrt{\frac{\pi}{2}} \sqrt{n}}$ for $g_4(A) = 1$ for $A \in U$.

Using a central limit theorem we get further

$$\left| s(G^+) - \frac{2^n}{\sqrt{\frac{\pi}{6} n^3 \left(1 + \frac{3}{2n} + \frac{1}{2n^2}\right)}} \right| \leqq \frac{c}{\sqrt{n}}$$

for some constant c.

O. SELESNJEW
MOSCOW STATE UNIVERSITY
DEPT. OF MATHEMATICS AND
MECHANICS
117 234 MOSCOW
USSR

B. THALHEIM
DRESDEN UNIVERSITY OF TECHNOLOGY
DEPT. OF MATHEMATICS
COMPUTER SCIENCE DIVISION
8027 DRESDEN
GDR

References

[1] J. DEMETROVICS, On the equivalence of candidate keys with Sperner systems. Acta Cybernetica 4 (1979), 247—252.
[2] K. KNOPP, I. SCHUR, Elementarer Beweis einiger asymptotischer Formeln der additiven Zahlentheorie. Mathematische Zeitschrift 24, 1925, 559—574.
[3] А. А. Миталаускас, В. Л. Статулявичус, Локальные предельные теоремы и асимптотические разломение для сумм независимых решетчатых случайных величин. Литовский математический сборник 1966, Т. 6, N. 4, 569—583.
[4] B. THALHEIM, Abhängigkeiten in Relationen. Dissertation B, Technische Universität Dresden, Dresden 1985.

(Received March 4, 1987)