MATHEMATICA — REVUE D'ANALYSE NUMÉRIQUE
ET DE THÉORIE DE L'APPROXIMATION
L'ANALYSE NUMÉRIQUE ET LA THÉORIE DE L'APPROXIMATION
Tome 9, N° 2, 1980, pp. 269–281

DATA ORGANIZATION ACCORDING TO THE PROPERTY OF CONSECUTIVE RETRIEVAL

by
LEON ŢÂMBULEA
(Cluj-Napoca)

1. Introduction. Let $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ be a set of descriptors (attributes). Every descriptor A_i may take values from a set $\mathcal{D}_i$. If for any descriptor A_i a certain value from $\mathcal{D}_i$ is fixed, we obtain a recording

$$R = (v_1, v_2, \ldots, v_n),$$

where $v_i \in \mathcal{D}_i$, $i = 1, 2, \ldots, n$. A set of recordings will form a collection of data:

$$\mathcal{C} = \{R_i \mid i = 1, 2, \ldots, m\}.$$

The data collection $\mathcal{C}$ would be stored on a certain medium $\mathcal{S}$ which can be considered as a collection of locations of storage. The medium $\mathcal{S}$ has an access-time function between any two locations, a function which depends on the structure of this medium. The most important use of the collection $\mathcal{C}$ stored on $\mathcal{S}$, is that of retrieval of some recordings when a necessity appears. These necessities are specified by means of some questions, and we denote by $\mathcal{Q}$ their set.

What we are interested in is to find an organization of $\mathcal{C}$ on $\mathcal{S}$, so that for any element $q \in \mathcal{Q}$ one can determine the elements from $\mathcal{C}$ which constitute the answer to the question q (denote by $q(\mathcal{C})$).

2. Kinds of questions. Let $A_i \in \mathcal{A}$ and $v_{ij} \in \mathcal{D}_i$. The pair:

$$K = (A_i, v_{ij})$$

is named key. For a key K we denote by $a(K)$ the descriptor (the left part of the pair), and by $v(K)$ its value (the right part). Therefore:

$$K = (a(K),\ v(K)).$$

If $\alpha = \{A_i,\ i \in I\} \subseteq \mathcal{A}$, we denote by $\Pi_\alpha R$ the projection of recording $R = (v_1, v_2, \ldots, v_n)$ on α, which is defined (by [2]) as:

$$R = (v_i,\ i \in I).$$

The projection of collection $\mathcal{C}$ is defined by:

$$\Pi_\alpha \mathcal{C} = \{\Pi_\alpha R \mid \forall R \in \mathcal{C}\}.$$

Various kinds of questions which may be addressed the collection $\mathcal{C}$ may be found in [1, 3, 6, 12]. All these sorts of questions are integrated in the following:

a). Find the elements of $\mathcal{C}$ which "has s specified keys: $K_1, \ldots, K_s$", i.e. find all recordings of the collection which have a particular value for any of the s specified descriptors. The answer to such a question q is:

$$q(\mathcal{C}) = \{R \in \mathcal{C} \mid \Pi_\beta R = (v(K_1),\ \ldots,\ v(K_s))\},$$

where:

$$\beta = \bigcup_{i=1}^{s} a(K_i).$$

b). Questions described as boolean functions of the keys.

c). Questions with limits for a descriptor $A_i \in \mathcal{A}$, that is to search the elements of $\mathcal{C}$ for which the descriptor A_i takes the values contained between v_1 and v_2, where the values $v_1, v_2 \in \mathcal{D}_i$. The answer is:

$$q(\mathcal{C}) = \{R \in \mathcal{C} \mid v_1 \leq \Pi_{A_i} R \leq v_2\}.$$

In that case one must define a preorder relation (" $\leq$ ") on the set $\mathcal{D}_i$.

d). Questions with the restriction on the distance relative to a descriptor A_i. In this case it is necessary to define a distance d on $\mathcal{D}_i$. The answer can be:

$$q(\mathcal{C}) = \{R \in \mathcal{C} \mid d(\Pi_{A_i} R,\ v_0) \leq \lambda\},$$

where v_0 and λ are the specified values, $v_0 \in \mathcal{D}_i$, $\lambda \in R^+$; or:

$$q(\mathcal{C}) = \}R \in \mathcal{C} \mid d(\Pi_{A_i} R,\ v_0) = \min_{R' \in \mathcal{C}} d(\Pi_{A_i} R',\ v_0)\},$$

where $v_0 \in \mathcal{D}_i$ is a specified value.

R e m a r k. The case d can be extended to the situation of more descriptors $\alpha = \{A_{i_1}, \ldots, A_{i_p}\} \subseteq \mathcal{A}$. For this, the distance d, must be defined on the cartesian product:

$$\mathcal{D}_{i_1} \times \ldots \times \mathcal{D}_{i_p}.$$

e). Questions of the form described above linked together by the logical operators: AND, OR, NOT.

In [7] there are many algorithms which give the answer to the questions of the kind a, for $s = 1$, and if $q(\mathcal{C}) = \Phi$, to insert an new recording in $\mathcal{C}$.

In [6] there is described such a way of the organization of the collection $\mathcal{C}$, that for any question of kind b can be obtained the answer. Besides the data collection $\mathcal{C}$ is considered a new zone, named directory, with the informations of the form:

$$(K_i,\ n_i,\ h_i;\ a_{i1},\ \ldots,\ a_{ih_i}),\ i = 1, 2, \ldots, p,$$

where:

— $K_1, \ldots, K_p$ are the keys which form a question;
— n_i is the number of the elements of $\mathcal{C}$ which contain the key K_i;
— h_i is the number of sublists in which the n_i elements are divided;
— a_{ij} is the address, on the support $\mathcal{S}$, at the beginning of the j-th sublists.

The collection $\mathcal{C}$ with her directory is named a generalized file. For various private values of n_i and h_i we obtain the structure of: inverted file, multilists file, sequential-indexed file.

A certain recording can contain more specified keys, and then, at the inverted file, the address of the recording can appear many times in the directory of the collection. In [12] a "canonical structure", where the address of the recording appears only once, is found.

A way of transforming a search by limits of values (type c) in a search described by a boolean function of keys (type b) is described in [3].

In [1] is given such a way of organizing of the collection $\mathcal{C}$, that one can give the answer to the questions of the type d.

3. An optimal model of organization of data collections

In the section 2 some models of the organization of data collection were indicated. For giving an answer to a question q of $\mathcal{Q}$ a time $t(q)$ is needed. It would be very important to find such as organization for which $t(\bar{q})$ is minimum for any element $q \in \mathcal{Q}$. In the next section we suppose that the medium $\mathcal{S}$ is linear, and the access-time between any storage locations depends on the distance between them.

In general, in various kinds of organizations the needed data for giving the answer to a question q are stored at various addresses of $\mathcal{S}$, and $t(q)$ depends on the distances between the recordings. This time is minimum if all the recordings pertinent to q are stored consecutively on $\mathcal{S}$. If this fact is valid for all the desired recordings of every element from $\mathcal{Q}$, then we will obtain the best organization. If there is such an organization between $\mathcal{C}$ and $\mathcal{Q}$ on $\mathcal{S}$, then we will say that $\mathcal{C}$ has the consecutive retrieval property (CR-property) relative to $\mathcal{Q}$. In [4] are given some conditions which must satisfy $\mathcal{C}$ and $\mathcal{Q}$ for the existance of the CR-property. For other results with regard to this problem one can consult [5, 8, 9, 10].

In section 4 we will give some results about the number of the recordings from a collection $\mathcal{C}$ when CR-property relative to a set $\mathcal{Q}$ of the questions exists. In sections 5 two algorithms are given which determine the order of the recordings from $\mathcal{C}$ on $\mathcal{S}$ when CR-property holds.

4. The number of the recordings from a CR-set

Let $\mathcal{C} = \{R_1, \ldots, R_m\}$ and $\mathcal{Q} = \{q_1, \ldots, q_n\}$. We consider the boolean matrix $B = (B_{ij})$, $i = 1, \ldots, m$; $j = 1, \ldots, n$; where an element B_{ij} is one if the recording R_i is pertinent with the question q_i (that is $R_i \in q_j(\mathcal{C})$), and zero otherwise.

We denote by $u(R_i)$ the number of the values 1 from i-th line of the matrix B and by $M_j = q_j(\mathcal{C})$ the set of recordings pertinent to the question q_j.

D e f i n i t i o n 1. *$\mathcal{C}$ is a CR-set relative to $\mathcal{Q}$ if the recordngs from $\mathcal{C}$ can be stored on the support $\mathcal{S}$ such that the CR-property holds for every $q_j \in \mathcal{Q}$.*

D e f i n i t i o n 2. *Two questions q_i and q_j are considered distincts if in the matrix B there correspond to them distinct columns.*

D e f i n i t i o n 3. *The characteristic vector of the recording R_i is equal with the i-th line of the matrix B. Two recordings are considered distincts if the characteristic vectors are different.*

We considered that the recordings from the data collections in what follows are different.

THEOREM 1. *If m is the number of recordings from a CR-set relative to a set $\mathcal{Q}$ of n distinct questions, then $m \leq 2n - 1$.*

Proof. The characteristic vector of the recording (different from the first recording) differs from the characteristic vector of the precedent recording by at least one position. There are at least n recordings pertinent for the first time to a question, and at most n recordings for which a question will not be asked. There exists a recording which appears in both subsets.

THEOREM 2. *If $\mathcal{C}$ is a CR-set relative to a set $\mathcal{Q}$ of n distinct questions, then the number of recordings with $u(R)$ even is at most $n - 1$, and the number of recordings with $u(R)$ odd is at most n.*

Proof. We prove the theorem by induction in respect to n. For $n = 2$ we have at most a recording with $u(R)$ even, and at most two recordings with $u(R)$ odd.

We suppose that the theorem is true for the sets $\mathcal{Q}$ with at most $n - 1$ distinct questions. We decompose $\mathcal{C} = \mathcal{C}_p \cup \mathcal{C}_i$, where $\mathcal{C}_p$ is the subset of those recordings with $u(R)$ even, and $\mathcal{C}_i$ with $u(R)$ odd. The two subsets have CR-property relative to a set $\mathcal{Q}$ with n distinct questions.

Let R_1 and R_2 be the first two recordings from $\mathcal{C}_p$ or $\mathcal{C}_i$, R_1 in front of R_2 on $\mathcal{S}$, and $u_1 = u(R_1)$, $u_2 = u(R_2)$.

a). If $u_1 = u_2$, then in the characteristic vector of R_2 appears at least a value 0 in the place of the value 1 from the characteristic vector of the R_1, and then on the respective column does not appear the value 1 at all. Deleting the question corresponding to this column from Q and the recording R_1, the induction hypothesis is true and we obtain the assertion of the theorem.

b). If $u_1 \geq 2 + u_2$, then there are at least two values 0 in the characteristic vector of the R_2 in the place of some values 1 from the characteristic vector of the R_1. Using the reasoning from a, the theorem results.

c). If $u_1 + 2 \leq u_2$, then in the characteristic vector of the R_2 we have $n - u_2$ null elements. The recordings that fallow R_2 on S can be divided in two subsets, which can be nondisjoint: a first subset, in which appear the recordings that have at least a value 1 in the characteristic vector in the place of the value 0 from the characteristic vector of the precedent recording, the number of these recordings being at most $\left[\frac{n - u_2}{2}\right]$ and a second subset, in which appear the recordings that have at least a value 0 in the characteristic vector in the place of the value 1 from the characteristic vector at the precedent recording, with at most $\left[\frac{n-2}{2}\right]$ recordings (we have no recording with the null characteristic vector). Together with R_1 and R_2, in $\mathcal{C}_p$ or $\mathcal{C}_i$, can appear at most:

$$2 + \left[\frac{n - u_2}{2}\right] + \left[\frac{n-2}{2}\right] \leq n - \left[\frac{u_1}{2}\right]$$

recordings. If the recordings are from $\mathcal{C}_i$ then $u_1 \geq 1$, and the maximum number of recordings is n, and if the recordings are from $\mathcal{C}_p$ then $u_1 \geq 2$, and the maximum number of recordings is $n - 1$.

D e f i n i t i o n 4. *A CR-set relative to a set $\mathcal{Q}$ of n distinct question is maximum if it has $2n - 1$ recordings.*

THEOREM 3. *I na maximum CR-set the characteristic vectors of two neighbour recordings differ by a single position.*

Proof. Suppose the opposite: there are two neighbour recordings R and R', for which the characteristic vectors differ by at least two position, which are supposed to be the first two. Let $(e_1, e_2, e_3, \ldots, e_n)$, be the characteristic vector of the R_1, and $(\bar{e}_1, \bar{e}_2, e'_3, \ldots, e'_n)$ of the R_2 where $\bar{e}_i = 1 - e_i$, $i = 1, 2$.

We will show that in this case between the two recordings one can introduce another one and the CR-proprety remains true, which contradics the fact that the set of recordings is maximum.

a). If $(e_1, e_2) = (0, 1)$ or $(e_1, e_2) = (1, 0)$, we can insert the recording which has the characteristic vector $(1, 1, e_3, \ldots, e_n)$. If this recording would be in the set, the CR-property would not hold.

b). If $(e_1, e_2) = (0, 0)$ or $(e_1, e_2) = (1, 1)$, then:

— if no recording appears with $(e_1, e_2) = (1, 0)$ or $(e_1, e_2) = (0, 1)$, we can insert one with these values;

— if a recording with $(e_1, e_2) = (1, 0)$ (or $(e_1, e_2) = (0, 1)$) appear, we can insert a recording with $(e_1, e_2) = (0, 1)$ (or $(e_1, e_2) = (1, 0)$);

— we can not have two recordings with $(e_1, e_2) = (0, 1)$ and $(e_1, e_2) = (1, 0)$ in this set.

R e m a r k. *The recordings from a maximum* CR-*set verify*:

$$u_1 = 1,\ u_2 = 2,\ u_{2n-2} = 2,\ u_{2n-1} = 1.$$

Proof. If $u_1 \neq 1$ or $u_{2n-1} \neq 1$ then we can add new recordings at the beginning or at the end of the collection $\mathcal{C}$, which contradicts the fact that the set is maximum. From theorem 3 it results that $u_2 = u_{2n-2} = 2$.

THEOREM 4. *The number of maximum* CR-*set relative to a set of n distinct questions is bigger than the number of sequences of lenght* $2n - 1$, *formed with the natural numbers, in which the difference between two consecutive elements of the sequence is* $+1$ *or* -1, *and the conditions from the precedent remark are verified. Any element from the sequence corresponds to the number of elements equal with* 1 *from the characteristic vector of a recording.*

Proof. From the former theorems it results that for a maximum CR-set one can construct a sequence with the given conditions.

Suppose that we have a sequence:

$$S = (s_1, s_2, \ldots, s_{2n-2}, s_{2n-1}),$$

with: $s_1 = s_{2n-1} = 1$, $s_2 = s_{2n-2} = 2$, and $|s_i - s_{i+1}| = 1$, for $i = 1, 2, \ldots, 2n - 2$. Corresponding to this sequence we construct a CR-set thus: R_1 has the characteristic vector $(1, 0, 0, \ldots, 0)$, R_2 with $(1, 1, 0, \ldots, 0)$, again if R_{i-1} has the characteristic vector $(t_1, t_2, \ldots, t_n)$, with $u(R_{i-1}) = s_{i-1}$, then the recording R_i has the characteristic vector $(t_1', t_2', \ldots, t_n')$, where $t_i' = t_i$ for $n - 1$ values of i, and $t_i' = 1 - t_i$ for a value of i: if $s_{i-1} = s_i + 1$, the first value 1 from the characteristic vector of R_{i-1} is changed in 0, and if $s_{i-1} = s_i - 1$, then the first value 0 that follows after a value 1 is changed in 1. For different sequences one obtains, in this way, different sets. But there exist many maximum CR-sets for the same sequence.

THEOREM 5. *The number of maximum* CR-*sets relative to a set of n distinct questions is bigger than*:

$$2 \cdot \frac{2n-3}{n(n-1)} \cdot \binom{2n-4}{n-2}.$$

Proof. We will determine the number of the sequences that appear in the theorem 4. For this, we will construct a labeled binary tree, as follows:

— the root is labeled by 1;

— a node with the label i has as a left subtree a binary tree. This subtree is constructed in the some way, with the root $i - 1$. The right subtree is a subtree with the root $i + 1$;

— the subtree which has a negative number or null for label of the root, is replaced by an empty tree.

The number of the sequences that appear in the theorem 4 is equal wiht the number of paths, in this binary tree, which connects the root wiht a node that has the label 1 on the level $2n - 1$.

If we denote by b_{ij} the number of nodes on the level i having the label j, then:

$$b_{11} = 1,$$

$$b_{ij} = 0 \text{ for } i < j \text{ or } i \cdot j = 0,$$

$$b_{ij} = b_{i-1,j-1} + b_{i-1,j+1} \text{ for } i > 1,\ j \geq 1.$$

One observe that on an even level we have only nodes with even labels, and on an odd level we have only nodes with odd labels.

We determine $b_{2n-1,1}$, which is the specified value in the text of the theorem. We have: $b_{3,1} = 1$; $b_{5,1} = 2$; $b_{7,1} = 5$; $b_{9,1} = 14$.

From the set of values b_{ij} we construct the set of values d_{ij} as follows:

$$d_{i1} = d_{i2} = 0,\ d_{ij} = 0 \text{ for } j > i,$$

$$d_{i,j+2} = b_{2i,2(i-j+1)} \text{ for } 1 \leq j \leq i,$$

that is we take only the values b_{ij} in which the first index is even, and on such a level we take only the nonull elements (corresponding to the even labels), but in inverse sense.

Using the reccurence relation for b_{ij}, it results:

$$d_{11} = d_{12} = 0,\ d_{13} = 1,\ d_{1j} = 0 \text{ for } j < 3;$$

$$d_{i1} = d_{i2} = 0;$$

$$d_{ij} = d_{i-1,j-2} + 2 \cdot d_{i-1,j-1} + d_{i-1,j} \text{ for } 3 \leq j \leq i + 2;$$

$$d_{ij} = 0 \text{ for } j > i + 2.$$

We must determine $d_{n-1,n+1}$, since: $b_{2n-1,1} = b_{2n-2,2} = d_{n-1,n+1}$.

We consider the system of the numbers c_{ij} defined by:

$c_{11} = c_{12} = 0$, $c_{13} = 1$, $c_{1j} = 0$ for $j > 3$;

$c_{i1} = c_{i2} = 0$;

$c_{ij} = c_{i-1,j-2} + 2 \cdot c_{i-1,j-1} + c_{i-1,j}$ for $j \geq 3$.

We construct the generator function:

$$C(x, y) = \sum_{i,j \geq 1} c_{ij} x^i y^j,$$

which verifies:

$$x \cdot C(x, y) + 2xy \cdot C(x, y) + xy^2 \cdot C(x, y) = C(x, y) - xy^3.$$

We have:

$$C(x, y) = \frac{xy^3}{1 - x(1+y)^2} = \sum_{i,j \geq 0} \binom{2i}{j} \cdot x^{i+1} y^{j+3} = \sum_{i,j \geq 1} \binom{2i-2}{j-3} x^i y^i.$$

Therefore:

$$c_{ij} = \binom{2i-2}{j-3}.$$

On can prove, by induction in respect to i, that for the values: $1 \leq j \leq i + 3$, we have:

$$d_{ij} = c_{ij} - c_{i,j-2},$$

that is the nonull values on the level i. Then:

$$d_{n-1,n+1} = \binom{2n-4}{n-2} - \binom{2n-4}{n-4} = 2 \cdot \frac{2n-3}{n(n-1)} \cdot \binom{2n-4}{n-2}.$$

5. Algorithms. In what follows we will give two algorithms which determine the order in which the sets M_j must be stored on $\mathcal{S}$, when the collection $\mathcal{C}$ is a CR-set relative to a set $\mathcal{Q}$ of questions.

LEMMA 1. *The sets M_i and M_j must be stored consecutively by all means, if: $M_i \cap M_j \neq \Phi$ and $M_i \setminus M_j \neq \Phi$ and $M_i \setminus M_j \neq \Phi$.*

The proof follows immediately.

We construct an undirected graph $G = (Y, U)$ in the following way (by [9]):

$Y = \{M_1, M_2, \ldots, M_n\}$;

$(M_i, M_j) \in U$ if and only if the conditions of lemma 1 are carried out.

We can consider that the graph G is conex (if it has some conex components, we will analyse one by one the conex components).

For determining the first (or the last) set M_j that must be stored on $\mathcal{S}$, we will determine a subset $Y_0 \subseteq Y$ so that:

$$\bigcup_{M \in Y_0} M = \mathcal{C}.$$

LEMMA 2. *If* CR *property exists, then the subgraph G_0 obtained from G eliminating the vertices that do not belong to Y_0, is a path.*

Proof. If G_0 contains a cycle μ or a point M with the degree superior to 2, then one of the subsets of the cycle μ, or one of the adjacent sets with M in G_0, is included in the union of the other sets, which contradicts the mode of construction of Y_0.

ALGORITHM A. We will give an order of storing of the sets M_i, $i = 1, \ldots, n$, on the medium S, such that the recordings from a certain set M_i be stored consecutively (by [11]).

Step 1. We determine the graph G, starting from the sets $M_1, M_2, \ldots, M_n$;

Step 2. $v := 1$;
For $i = 1, 2, \ldots, m$ do: $p_i := 0$;

Step 3. For any conex component of the graph G do the steps 4 — 10, and after that proceed to the step 11;

Step 4. For a specified conex component, we determine the subgraph G_0 defined in the lemma 2;

Step 5. We find an extremity M_0 of the G_0;

Step 6. For all recordings $R_k \in M_0$ do: $p_k := v$;

Step 7. We take a certain set M from this conex component, that has not been taken into consideration until now. If all the sets M have been taken into account, go to the step 4 with other conex component;

Step 8. For M fixed, we determine $X = \{p_k | R_k \in M, p_k \neq 0\}$. If X is empty, go to the step 7, and the set M will be considered later. We determine $s = \min X$ and $t = \max X$;

Step 9. For every R_j from the conex component do:
if $R_j \in M$ and: a) $p_j = 0$ then $p_j := v$;
b) $p_j \neq 0$ then $p_j := p_j + 1$;
if $R_j \notin M$ and $p_j \geq t$ and $s < t$ then $p_j := p_j + 2$;

Step 10. $v := v + 2$. Go to the step 7;

Step 11. The recordings R_k are stored in such a mode that the associated values p_k are in an increasing order.

LEMMA 3. *$\mathcal{C} = M_1 \cup M_2$ is a* CR*-set realtive to $\mathcal{Q} = \{q_1, q_2\}$, with $M_i = q_i(\mathcal{C})$, $i = 1, 2$.*

Proof. The recordings from $\mathcal{C}$ can be stored in the order: $M_1 \setminus M_2$; $M_1 \cap M_2$; $M_2 \setminus M_1$.

LEMMA 4. *$\mathcal{C} = M_1 \cup M_2 \cup M_3$ is a* CR*-set relative at $\mathcal{Q} = \{q_1, q_2, q_3\}$, with $M_i = q_i(\mathcal{C})$, $i \in I = \{1, 2, 3\}$, if and only if one of the next three assertions is true:*

$P_1(i, j) : (\exists i, j \in I, i \neq j : M_i \subseteq M_j)$;
$P_2(i, j) : (\exists i, j \in I, i \neq j : M_i \cap M_j = \Phi)$;
$P_3(i, j, k) : (i, j, k \in I, \ i \neq j, j \neq k, i \neq k : M_i \cap M_j \subseteq M_k \subseteq M_i \cup M_j)$.

Proof. The necessity. Let $C = M_1 \cup M_2 \cup M_3 = \{R_k, k = 1, \ldots, p\}$ be a CR-set relative at $\mathcal{Q} = \{q_1, q_2, q_3\}$. We suppose that M_i is delimited by the recordings with indexes L_1^i and L_2^i, $1 \leq L_1^i \leq L_2^i \leq p$, $i \in I$. Without losing from the generality, we can suppose: $1 \leq L_1^1 \leq L_2^1 \leq L_3^1 \leq p$.

a). If $L_2^1 < L_1^2$ it results $P_2(1, 2)$;
b). If $L_2^1 \geq L_1^2$ then:
b1). If $L_2^1 < L_1^3$ it results $P_2(1, 3)$;
b2). If $L_2^1 \geq L_1^3$ then:
b21). If $L_2^2 \leq L_2^1$ it results $P_1(2, 1)$;
b22). If $L_2^2 > L_2^1$ then:
b221). If $L_2^3 \leq L_2^2$ it results $P_1(2, 3)$;
b222). If $L_2^3 > L_2^2$ it results $P_3(1, 3, 2)$.

The sufficient. If $P_1(i, j)$ is true, the recordings can be stored in the order:

$$M_j \backslash (M_k \cup M_i);\ M_i \backslash M_k;\ M_i \cap M_k;\ M_j \cap (M_k \backslash M_i);\ M_k \backslash M_j.$$

If $P_2(i, j)$ is true, the recordings can be stored in the order:

$$M_i \backslash M_k;\ M_i \cap M_k;\ M_k \backslash (M_k \cap (M_i \cup M_j));\ M_k \cap M_j;\ M_j \backslash M_k.$$

If $P_3(i, j, k)$ is true, the order of the recordings may be:

$$M_i \backslash M_k;\ M_k \backslash M_j;\ M_i \cap M_j;\ M_k \backslash M_i;\ M_j \backslash M_k.$$

From the lemmas 3 and 4 it results that if $\mathcal{C}$ is a CR-set relative to a set of questions $\mathcal{Q} = \{q_1, \ldots, q_n\}$, then it can be decomposed in disjoint sets: $T_1, T_2, \ldots, T_p$, whose union gives $\mathcal{C}$, and if $\mathcal{C}$ as stored on $\mathcal{S}$ in the order $T_1, \ldots, T_p$, in T_i the order of the recordings being arbitrary, then CR-property is true for the set $\mathcal{Q}$ of questions.

For the collection considered in the lemma 3 it results that: $p=3$ and $T_1 = M_1 \backslash M_2$, $T_2 = M_1 \cap M_2$, $T_3 = M_2 \backslash M_1$, some of the subsets T_i may be empty.

Suppose that $\mathcal{C} = M_1 \cup M_2 \cup \ldots \cup M_n$ is a CR-set relative at $\mathcal{Q} = \{q_i, i = 1, \ldots, n\}$, with $M_i = q_i(\mathcal{C})$, $i = 1, \ldots, n$, and the independent subsets described above are: $T_1, \ldots, T_p$. One may put the problem if by adding of a new question q, whose answer is the set M of recordings, the collection $\mathcal{C} \cup M$ has the CR-property relative to $\mathcal{Q} \cup \{q\}$. The solution of this problem is given by the next theorem.

THEOREM 6. *If $\mathcal{C}$ has the* CR*-property relative to $\mathcal{Q}$, and T_i, $i \in I = \{1, 2, \ldots, p\}$ are subsets previously defined, then* CR*-property holds for the collection $\mathcal{C} \cup M$ relative to $\mathcal{Q} \cup \{q\}$, where $q(\mathcal{C} \cup M) = M$, if and only if one of the following conditions is true:*

C1. $M \cap T = \Phi$, $i \in I$;

C2. $\exists i, j \in I : M \subseteq \bigcup_{k=1}^{j} T_k$,

$$\textit{and if } i - j \geq 2, \textit{ then } \bigcup_{k=i+1}^{j-1} T_k \subseteq M;$$

C3. $\exists i \in I : \bigcup_{k=1}^{i-1} T_k \subseteq M$, $T_i \nsubseteq M$, $M \cap T_j = \Phi$ *for* $j > i$;

$$\left(\textit{if } i = 1, \text{ then} \sum_{k=1}^{i-1} T_k = \Phi\right);$$

C4. $\exists j \in I : \bigcup_{k=j+1}^{p} T_k \subseteq M$, $T_j \nsubseteq M$, $M \cap T_i = \Phi$, for $i < j$,

$$\left(\text{if } j = p, \text{ then} \bigcup_{k=j+1}^{p} T_k = \Phi\right);$$

C5. $\bigcup_{k \in I} T_k \subseteq M$.

Proof. We will show that if one of the five conditions is true, then CR-property holds for $\mathcal{C} \cup M$ relative to $\mathcal{Q} \cup \{q\}$ (we will construct the new disjoint subsets $T_1', \ldots, T_r'$.

If C1 holds, then: $r = p + 1$, $T_i' = T_i$, $i \in I$; $T_r' = M$.

If C2 is true, then subsets $T_1', \ldots, T_r'$ will be the noempty subsets:

$$T_1, \ldots, T_{i-1}, T_i \setminus M, T_i \cap M, T_{i+1}, \ldots, T_{j-1}, T_j \cap M, T_j \setminus M, T_{j+1}, \ldots, T_p.$$

In the case when C3 is true, the subsets will be:

$$M \setminus \mathcal{C},\ T_1, \ldots, T_{i-1}, M \cap T_i, T_i \setminus M, T_{i+1}, \ldots, T_p.$$

When C4 is true, the subsets will be:

$$T_1, \ldots, T_{j-1}, T_j \setminus M, T_j \cap M, T_{j+1}, \ldots, T_p, M \setminus \mathcal{C}.$$

and for C5 true they will be:

$$T_1, \ldots, T_p, M \setminus \mathcal{C}.$$

We still must show that if $\mathcal{C} \cup M$ has the CR-property relative to $\mathcal{Q} \cup \{q\}$, then it results that one of the specified conditions is true:

a). If $M \cap \mathcal{C} = \Phi$ it results C1;

b). If $M \cap \mathcal{C} \neq \Phi$, then:

b1). If $\mathcal{C} \subseteq M$ it results C5;

b2). If $M \subseteq \mathcal{C}$ it results C2;

b3). If $M \setminus \mathcal{C} \neq \Phi$ and $\mathcal{C} \setminus M \neq \Phi$ it results C3 or C4, because the recordings from $\mathcal{C} \cup M$ are stored consecutively on $\mathcal{S}$.

Due to theorem 6 we can construct the following algorithm:

ALGORITHM B. As before, this algorithm, will give the order of storing the sets $M_1, \ldots, M_n$ on the medium $\mathcal{S}$, such that the recordings of on arbitrary set M_i are stored consecutively.

In this algorithm we use the variables:

— y_i, $i = 1, 2, \ldots, n$ where $y_i = 1$ if the set M_i has been stored on the medium $\mathcal{S}$, and 0 otherwise;
— X — the set stored recordings;
— $T_1, \ldots, T_p$ — sets mentioned in the theorem 6;
— INC is a variable which shows a position in $T_1, \ldots, T_p$, defined by: for $i <$ INC and $j \geq$ INC the sets T_i and T_j are disjoint.

Step 1. For $i = 1, 2, \ldots, n$ do: $y_i := 0$;
Step 2. $p := 1$, $T_1 := M_1$; $X := T_1$; INC $:= 1$;
Step 3. $i := 2$;
Step 4. If $y_i = 0$ and $M \cap X \neq \Phi$ go to the step 7, else $i := i + 1$. If $i \leq n$ go to the step 4, else $i := 2$;
Step 5. If $y_i = 0$ go to the step 6, else $i := i + 1$. If $i \leq n$ go to the step 5, else stop ($T_1, \ldots, T_p$ gives the order of storing the recordings on the medium).
Step 6. $p := p + 1$; INC $:= p$; $T_p := M_i$; $y_i := 1$; $X := X \cup M_i$; go to the step 3;
Step 7. This step verifies which one of the conditions of the theorem 6 are verified relative to the subsets $T_{\text{INC}}, \ldots, T_p$. Starting with these subsets and according to the verified condition, we construct the subsets $T_{\text{INC}}, \ldots, T_r$ as the proof of the theorem 6 indicates;
Step 8. $p := r$; $X := X \cup M_i$; $y_i := 1$; go to the step 3.

REFERENCES

[1] Burkhard, W. A., Keker, R. M., *Some Approaches to Best-Match File Searching.* Comm. ACM, **16**, 4, 230–236 (1973).
[2] Delobel, C., *Aspects théoretiques sur la structure de l'information dans une base de données.* R.I.R.O., **5**, B–3, 37–64 (1971).
[3] Diaconescu, G., *The SA-V algortihm.* Economic Computation and Economic Cybernetics Studies and Research. **1**, 1975 (Bucharest).
[4] Ghosh, S. P., *File Organization: The Consecutive Retrieval Property.* Comm. ACM, **15**, 9, 802–808 (1972).
[5] Gorski, J., *On the optimal arrangement for consecutive storage of objects.* ICS PAS REPORTS, 279, 1977, Warszawa.

[6] Hsiado, D., Harary, F., *A Formal System fo r Information Retrieval from Files* Comm. ACM, **13**, 2, 67–73 (1970).
[7] Knuth, D. E., *The art of Computer Programming* III: *Sorting and Searching.* Addison-Wesley, Reading, MA, 1973.
[8] Kou, L. T., *Polynomial complete consecutive information retrieval problems.* SIAM J. COMPUT., **16**, 1, 67–75 (1977).
[9] Lipski, W., Marek, W., *On information storage and retrieval systems.* CC PAS REPORTS, 200, 1975, Warszawa.
[10] Traczyk, T., *On consecutive retrieval from drum-type storage.* CC PAS REPORTS, 285, 1977, Warszawa.
[11] Tâmbulea, A., *A model of organization of data collections.* Proc. of the 5-th Symposium „Informatics and management", Cluj-Napoca 1979.
[12] Wong, E., Chiang, T.C., *Cannonical Structure Attribute Based File Organization.* Comm. ACM, **14**, 9, 593–597 (1971).

Received 4. XII. 1979