

AN ALGORITHM FOR THE ORGANIZATION OF INFORMATION

G. M. ADEL'SON-VEL'SKIĬ AND E. M. LANDIS

In the present article we discuss the organization of information contained in the cells of an automatic calculating machine. A three-address machine will be used for this study.

Statement of the problem. The information enters a machine in sequence from a certain reserve. The information element is contained in a group of cells which are arranged one after the other. A certain number (the information estimate), which is different for different elements, is contained in the information element. The information must be organized in the memory of the machine in such a way that at any moment a very large number of operations is not required to scan the information with the given evaluation and to record the new information element.

An algorithm is proposed in which both the search and the recording are carried out in $C \lg N$ operations, where N is the number of information elements which have entered at a given moment.

A part of the memory of the machine is set aside to store the incoming information. The information elements are arranged there in their order of entry. Moreover, in another part of the memory a "reference board" [1] is formed, each cell of which corresponds to one of the information elements.

The reference board is a dyadic tree (Figure 1a): each of its cells has no more than one left cell, and no more than one right cell subordinated to it. Direct subordination induces subordination (partial ordering). In addition, for each cell of the tree, all the cells which are subordinate to a left (right) directly subordinate cell, will be arranged further to the left (right) than the given cell. Moreover, we assume that there is a cell (the head) to which all the others are subordinate. By transitivity, the conception "further to the left" and "further to the right" extends to the aggregate of all the cell pairs, and this aggregate becomes ordered. Thus, a given order of cells in a reference board should coincide with the order of arrangement of the estimates of the corresponding information elements (to be specific, we shall consider the estimates as increasing from left to right).

In the first address of each cell of the reference board, a place is indicated where the corresponding information element is located. The addresses of the cells of the reference board, which are directly subordinate on the left and right respectively to the given cell, are located in the second and third addresses. If a cell has no directly subordinate cells on either side, then there is zero in the corresponding address. The head address is stored in a certain fixed cell l .

Let us call the sequence of the cells of the tree a *chain* in which each previous cell is directly

subordinate to the following. For each cell of the tree, we shall designate as the length of the left (right) branch the maximum length of the chain which consists of cells subordinate to the given one and located more to the left (right) than the given cell. Any chain whose length is equal to the length of the branch is called the branch pivot.

The *admissible* tree will be such that for each of its cells, the length of the left branch differs from the length of the right branch by no more than unity (Figure 1b).

Two orders for the information on the branch lengths are distinguished in each cell of the address board. If the left branch is longer than the right, then 1.0 stands in these orders. If the right branch is longer than the left, then 0.1 stands; and if they are equal, then 0.0. Moreover, it is considered that if there are no subordinate cells on either side, then the length of the branch on this side is equal to zero.

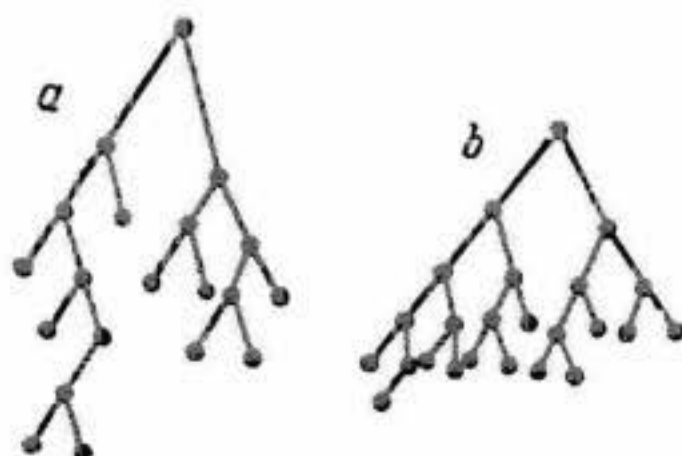


Figure 1

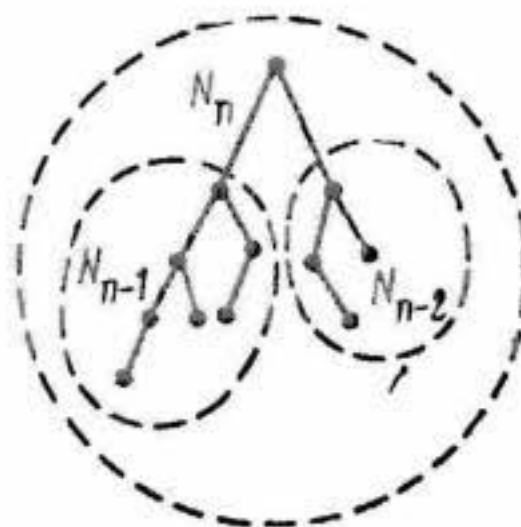


Figure 2

The recording algorithm is such that at each moment, the reference board is an admissible tree.

Lemma 1. *Let the number of cells of the admissible tree be equal to N . Then the maximum length of the branch is not greater than $(3/2) \log_2 (N + 1)$.*

Proof. Let us denote by N_n the minimum number of cells in the admissible tree when the given maximum length of the branch is n . Then it can be easily proven (see Figure 2) that $N_n = N_{n-1} + N_{n-2} + 1$.

When we solve this equation in finite remainders, we get

$$N_n = \left(1 + \frac{2}{\sqrt{5}}\right) \left(\frac{1 + \sqrt{5}}{2}\right)^n + \left(1 - \frac{2}{\sqrt{5}}\right) \left(\frac{1 - \sqrt{5}}{2}\right)^n - 1.$$

Whence

$$n < \log_{\frac{1 + \sqrt{5}}{2}} (N + 1) < \frac{3}{2} \log_2 (N + 1),$$

q.e.d.

The search algorithm of the information element with the given estimate consists in the following. We compare the given estimate with the estimate of the information element which corresponds to the head. Depending on the result of the comparison, we now compare the given estimate with the estimate of the information element which corresponds to the left or right directly subordinate cell head. Let k comparison steps be made of the given estimate m with the estimate m_u of the information element which corresponds to a certain cell u of the reference board. If $m < m_u$ ($m > m_u$), then at step number $(k + 1)$, a comparison is made of the estimate m with the estimate of the information element corresponding to the cell directly subordinate to u on the left (right). If $m = m_u$, the search is complete.

From the uniqueness of the head, it follows that if the information element is among the information gathered, the estimate of which element is equal to the given estimate, then it can be found at any step. In addition, the number of comparisons will be equal to the number of cells of a certain chain of the tree (considering that the comparison gives three answers). Since the number of operations is proportional to the number of such comparisons, it follows from Lemma 1 that the number of the operations which are necessary in this algorithm for the search, is not greater than $C \log_2(N + 1)$.

Let us now describe the algorithm of the structure of the reference board in the form of an admissible tree.

The tree is constructed as the information enters in the following fashion. When the first information element enters a certain cell in the first address, the location of this element is noted in the memory. Zeros are recorded in the remaining addresses and in the separate code orders. This cell is called the head. Accordingly, its address is recorded in a cell which has been isolated for this purpose. Let an admissible tree be constructed for N information elements and let the $(N + 1)$ th information element enter. We apply the search algorithm to the estimate m of this element, remembering the cell addresses of the chain along which we proceed with this algorithm. Henceforth, we shall refer to this chain as *recorded*. If it turns out that the new element is already contained in the former information, the tree does not change. Otherwise, we will come to the cell u which has the following characteristics. If $m < m_u$, where m_u is the estimate of the information element corresponding to u , cell u has no direct subordinate on the left (right). The new cell v is then added to the reference board which is directly subordinate on the left (right) to cell u . A place is indicated in the first address of cell v where the new information element is located; there are zeros in the remaining addresses and in the separate orders. The address of cell v is located in the corresponding address of cell u .

Thus the reference board, supplemented with cell v , is still a tree, but perhaps inadmissible. Moreover, the separate orders in the cells of the recorded chain must be corrected.

First, the separate orders in cell u are corrected. There are two possibilities: 1) 1 0 (0 1) stood in the separate orders of cell u , and 2) 0 0 stood in these orders. In the first case, a new branch with a free side is added to cell u , whereupon the lengths of both branches become identical (equal to 1), and the length of any other branch in the tree does not change. By placing 0 0 in the separate orders of cell u , we complete the recording, since the tree obtained is admissible. In the second case, we change the separate orders of u by 1 0, if $m < m_u$; and by 0 1, if $m > m_u$.

Lemma 2. *Chain \mathcal{E} of the admissible tree, if 0 0 stands in the separate orders in every cell of this chain and the last cell has no subordinates (zeros stand in the second and third addresses), is the branch pivot for the cell which directly precedes the first cell w of the chain \mathcal{E} .*

Proof. Let us assume that this lemma is incorrect. From the aggregate of all the chains which begin with w and have a length greater than the length of \mathcal{E} , let us select chain \mathcal{D} which with \mathcal{E} has a maximum number of common cells. Let t be the last common cell of the chains \mathcal{E} and \mathcal{D} . Then, its branches have a different length; therefore, 0 0 cannot stand in the separate orders of cell t . Thus, Lemma 2 is proved.

Let us now examine the maximum connected part of \mathcal{E} of the recorded chain, which goes upward, beginning with cell u , and which consists of cells in the separate orders of which 0 0 stands. The correction of the separate orders in the cells of \mathcal{E} consists in the fact that 0 0 changes to 1 0 or 0 1 depending on whether the following cell of chain \mathcal{E} is directly subordinate on the left or right.

It follows from Lemma 2 that: 1) cells which are subordinate to any cell from \mathcal{C} , form an admissible tree; 2) the correction of the separate orders has been carried out correctly for the cells from \mathcal{C} and 3) the length of any branch which contains cells from the recorded chain increased by 1.

Let us note, moreover, that the cells subordinate to any cell which does not belong to the recorded chain, form an admissible tree which did not change by the addition of cell v .

Three cases are possible:

1) \mathcal{C} is the entire recorded chain. Then we already have an admissible tree.

2) \mathcal{C} lies on the short branch of cell s_0 which directly precedes it. Then, after the addition of cell v , an admissible tree is obtained; and it is only necessary to place 0 0 in the separate orders of cell s_0 .

3) \mathcal{C} lies on the long branch of cell s_0 . After the addition of cell v , this branch of cell s_0 became twice as long as the short branch. Let us examine three possible arrangements (to within symmetry) of the cells subordinate to s_0 , as shown in Figure 3. A, B, C, D, E, F are the branches which are not shown in the figure. The numbers in parentheses designate the length of the corresponding branch.

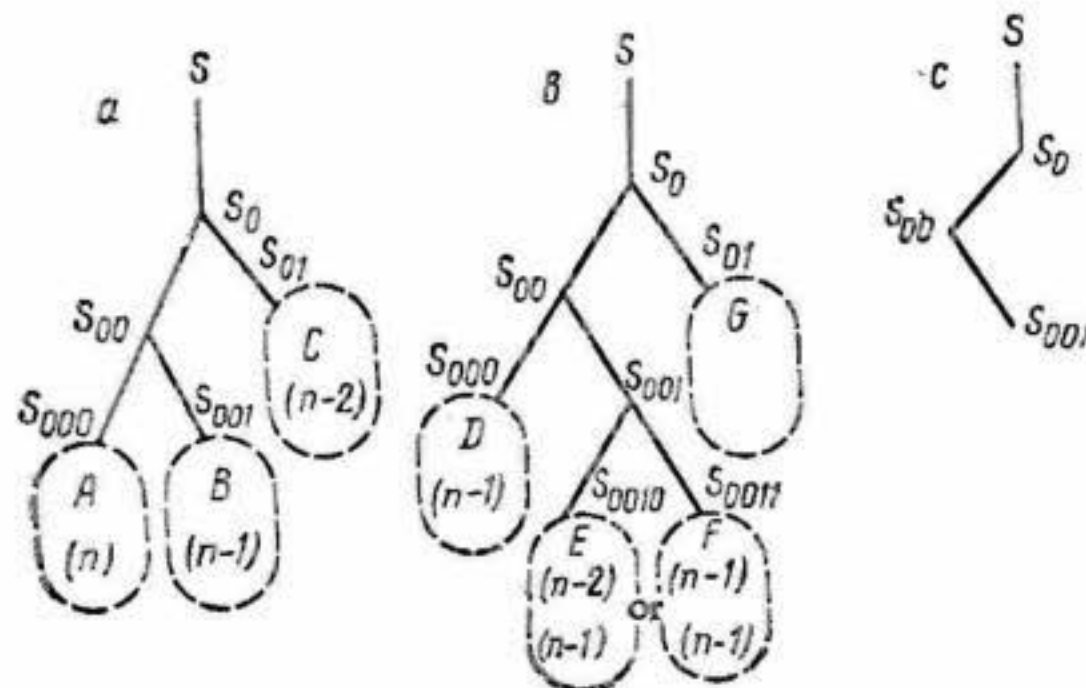


Figure 3

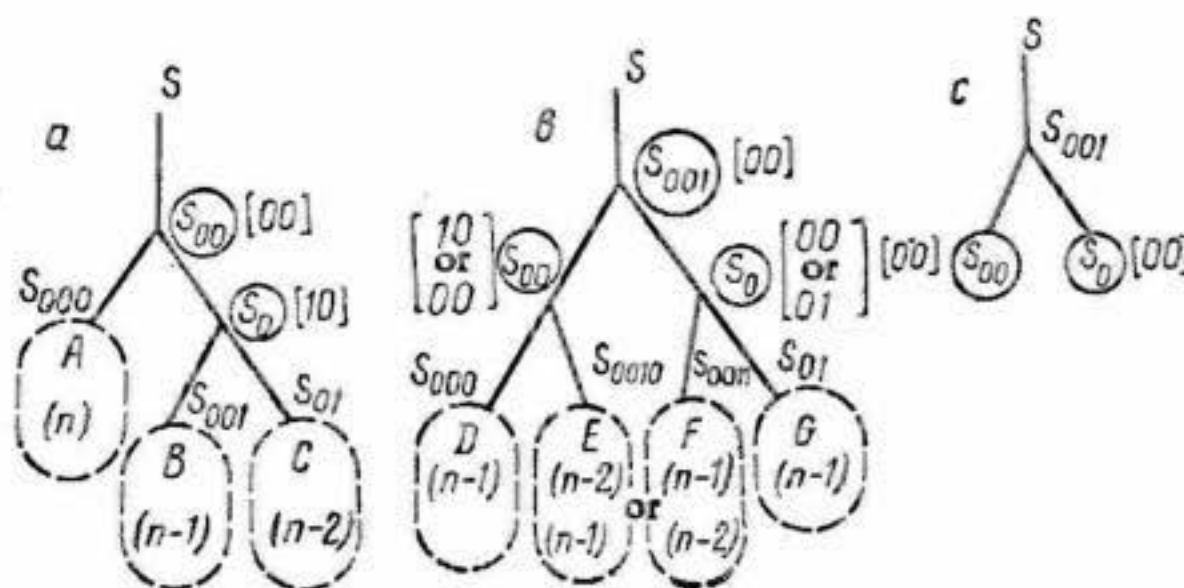


Figure 4

In this case, the tree is reorganized, as shown in Figure 4. s is the cell which directly precedes s_0 . The encircled cells are those whose designations of the directly subordinate cells do not change. We indicated in brackets the values of the separate orders in the corresponding cells, if they must be changed. After this reorganization, as the figure shows, the tree becomes admissible. If s_0 is the head (whereupon, there is no cell s), it is then necessary to change again the fixed cell in which the head of the address is indicated.

Since the search requires $C \log_2(N + 1)$ operations, the movement up along the recorded chain to cell s_0 requires no more than $C_1 \log_2(N + 1)$ operations and, finally, the reorganization of the tree requires a constant number of operations. In all, $C \log_2(N + 1)$ operations are required.

Received 13/APR/62

BIBLIOGRAPHY

- [1] P. F. Windley, Comput. J. 3 (1960/61), 84.

Translated by:
Myron J. Ricci