Werk

Label: Article

Jahr: 1972

PURL: https://resolver.sub.uni-goettingen.de/purl?31311157X_0097|log86

Kontakt/Contact

Digizeitschriften e.V.
SUB Göttingen
Platz der Göttinger Sieben 1
37073 Göttingen

info@digizeitschriften.de

Časopis pro pěstování matematiky, roč. 97 (1972), Praha

ON THE NUMBER OF INITIAL SEGMENTS OF A FINITE SET OF SEQUENCES (FINITE LANGUAGE)

KAREL ČULÍK and ANTONÍN VRBA, Praha

(Received December 21, 1970)

Two essentially different methods for the evaluation of the number of (nonempty and mutually different) initial segments of sequences (i.e., words or strings) which belong to a given set of sequences (i.e., to a given finite language) are presented. Several open problems are suggested.

1. INTRODUCTORY MOTIVATION AND BASIC NOTIONS

In the theory of automata one considers *sequential mappings*, i.e. functions whose domain is some subset of the set X^∞ of all words or strings over a (finite) alphabet X, i.e. $X^\infty = \{\xi; \xi = x_1 x_2 \dots x_n \text{ where } x_i \in X \text{ for } i = 1, 2, \dots, n \text{ and } n \geqq 1\}$. The *strings* or *words* are nothing else than finite sequences, their length being called the *length of the string* and denoted by $l(\xi)$ for $\xi \in X^\infty$. Further, we introduce the *empty string* ε characterized on the one hand by $l(\xi) = 0 \Leftrightarrow \xi = \varepsilon$, on the other hand by the fact that it is the unit of the free semigroup over X with respect to the operation of *concatenation* (the concatenation of the string $x_1 x_2 \dots x_n$ with the string $y_1 y_2 \dots y_m$ yields the string $x_1 x_2 \dots x_n y_1 y_2 \dots y_m$), i.e., $\varepsilon\xi = \xi\varepsilon = \xi$ holds for each string $\xi \in X^\infty \cup \{\varepsilon\}$. We shall write $\alpha \prec \beta$ provided the string α is an *initial segment* of the string β, i.e., if there is $\xi \in X^\infty \cup \{\varepsilon\}$ such that $\alpha\xi = \beta$, and $\alpha \overset{\prec}{\neq} \beta$ provided it is a *proper segment*, i.e. $\alpha \neq \beta$. The *maximal common initial segment* of the strings α and β will be denoted by $\alpha \wedge \beta$. Then $\alpha \wedge \beta = \varepsilon \Leftrightarrow \alpha$ and β have different first symbols (from the left).

The following algorithm is used for the synthesis of Mealy's automaton for a finite sequential mapping (see [1] or [2]):

Algorithm. To the given finite sequence of strings $P = (\xi_1, \xi_2, \dots, \xi_m)$ over the alphabet X, where $\xi_i = x_{i1} x_{i2} \dots x_{in_i}$ for $i = 1, 2, \dots, m$, the sequence $Q = (\eta_1, \eta_2, \dots, \eta_m)$ is constructed over the alphabet Y which is the set of all positive integers, i.e. $\eta_i = y_{i1} y_{i2} \dots y_{in_i}$ where $y_{ij} \in Y$, by the following recursive rule:

1) $\eta_1 = 12 \ldots n_1$,

2) if for some $p > 1$ the strings $\eta_1, \eta_2, \ldots, \eta_{p-1}$ have been constructed, then η_p is constructed in the following manner: Among the strings $\xi_1, \xi_2, \ldots, \xi_{p-1}$ we find a string such that (i) it has the common initial segment with ξ_p of the maximal length while (ii) it has the minimal index; if such a (nonempty) string exists, denote its index by $f(p)$ so that $1 \leqq f(p) < p$ and $\xi_{f(p)}$ is the string considered; if there is no string with the required properties, put $f(p) = 0$ and assume (for formal reasons) that $\xi_0 = \varepsilon$; hence the function f is well defined for every $i = 2, 3, \ldots, m$; it determines uniquely the number $d_p = l(\xi_p \wedge \xi_{f(p)})$ for $p = 2, 3, \ldots, m$. Further we put $d_1 = 0$ and $f(1) = 0$ and, finally, if $s \in Y$ is the least positive integer which occurs in no string $\eta_1, \eta_2, \ldots, \eta_{p-1}$, then we put $y_{pi} = y_{f(p)i}$ for $i = 1, 2, \ldots, d_p$ (evidently so far as $d_p > 0$) and $y_{p(d_p+j)} = s + (j - 1)$ for $j = 1, 2, \ldots, n_p - d_p$.

The numbers d_i, $i = 1, 2, \ldots, m$ found during the algorithm determine the number

$$a(P) = \sum_{i=1}^{m} (n_i - d_i) \tag{1}$$

which obviously shows the number of positive integers which occur in the strings $\eta_1, \eta_2, \ldots, \eta_m$.

Example 1. For $X = \{0, 1\}$ and $P = (\xi_1, \xi_2, \xi_3, \xi_4)$ the algorithm yields successively the values of the function f, the numbers d_p and, finally, $a(P)$ in the following way:

$$\begin{aligned}
&\xi_1 = 10011\,, \quad n_1 = 5\,; \quad f(1) = 0\,, \quad d_1 = l(\xi_1 \wedge \xi_0) \quad = l(\varepsilon) \quad = 0\,;\\
&\xi_2 = .1010\,, \quad n_2 = 4\,; \quad f(2) = 1\,, \quad d_2 = l(\xi_2 \wedge \xi_{f(2)}) = l(10) \quad = 2\,;\\
&\xi_3 = 01101\,, \quad n_3 = 5\,; \quad f(3) = 0\,, \quad d_3 = l(\xi_3 \wedge \xi_{f(3)}) = l(\varepsilon) \quad = 0\,;\\
&\xi_4 = 01110\,, \quad n_4 = 5\,; \quad f(4) = 3\,, \quad d_4 = l(\xi_4 \wedge \xi_{f(4)}) = l(011) = 3\,,
\end{aligned}$$

which implies $a(P) = (5 - 0) + (4 - 2) + (5 - 0) + (5 - 3) = 14$.

Evidently $\eta_1 = 12345$, $\eta_2 = 1267$, $\eta_3 = 89\ 10\ 11\ 12$, $\eta_4 = 89\ 10\ 13\ 14$.

The number $a(P)$ is of essential importance in the theory of automata, namely, it gives the maximal number of the inner states of Mealy's automaton which realizes the considered sequential mapping. Hence we may expect that $a(P)$ does not depend on the order of terms of the sequence P. This conjecture is supported also by the following assertion.

Theorem 1. *The number $a(P)$ gives the number of all (nonempty and mutually different) initial segments of the strings which occur in the finite sequence P of strings.*

Proof. Let $P = (\xi_1, \xi_2, \ldots, \xi_m)$ where $\xi_i = x_{i1}x_{i2} \ldots x_{in_i}$ and $x_{ij} \in X$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n_i$. We prove the theorem by mathematical induction

with respect to $k = \sum_{i=1}^{m} n_i$. For $k = 1$ there must be precisely one $n_i = 1$ while the other strings are empty, hence $d_i = 0$ and consequently, $a(P) = 1$. In this case the number of initial segments is indeed equal to one. Suppose the theorem is true for $k - 1$ and prove it for $k > 1$. Construct P^* from P in the following way: In the last nonempty string ξ_h delete the last symbol x_{hn_h}, i.e., denoting $P^* = (\zeta_1, \zeta_2, \ldots, \zeta_m)$, it is $\zeta_i = \xi_i$ for $i \neq h$, $i = 1, 2, \ldots, m$ while $\zeta_h x_{hn_h} = \xi_h$. Let f^* and d_i^* be evaluated by applying the algorithm to P^* and put $n_i^* = l(\zeta_i)$. Then obviously $n_i = n_i^*$ for $i \neq h$, $i = 1, 2, \ldots, m$ while $n_h = n_h^* + 1$ and, similarly, $f(i) = f^*(i)$ for $i \neq h$, $i = 1, 2, \ldots, m$, which implies $d_i^* = d_i$ for $i \neq h$, $i = 1, 2, \ldots, m$. Hence only numbers d_h and d_h^* are to be examined.

The only two cases which may occur are the following ones:

either ξ_h is an initial segment of a string ξ_i for $1 \leqq i < h$ so that, on the one hand, P^* and P have obviously the same number of initial segments, while, on the other hand, $d_h^* = d_h - 1$ holds so that $a(P^*) = a(P)$. The assumption of induction that $a(P^*)$ gives the number of initial segments of the strings in P^* implies that $a(P)$ has the same meaning for P;

or ξ_h is not an initial segment of the strings ξ_i, $1 \leqq i < h$ so that, on the one hand, P has one initial segment more than P^* and, on the other hand, it is evident that $f(h) = f^*(h)$ as well as $d_h = d_h^*$ which means that $a(P^*) + 1 = a(P)$. Again it follows from the assumption of induction that $a(P)$ gives the number of initial segments from P. The proof is complete.

Another proof of Theorem 1. It is immediately seen from the algorithm that two initial segments of strings from P are different if and only if the corresponding initial segments of strings from Q are different. Let us order all initial segments of strings from Q into a sequence $\{y_{11}, y_{11}y_{12}, \ldots, y_{11}y_{12} \ldots y_{1n_1}, y_{21}, y_{21}y_{22}, \ldots y_{m1}, y_{m1}y_{m2}, \ldots, y_{m1}y_{m2} \ldots y_{mn_m}\}$. If some member appears more than once, let it stay only at its first occurrence and delete all its repeatings. Hence we obtain a sequence of all mutually different initial segments of strings from Q. Evidently the algorithm is constructed so that the number y_{ij} gives the position of the initial segment $y_{i1}y_{i2} \ldots \ldots y_{ij}$ in the sequence. However, we know that the greatest one of numbers y_{ij} is equal to $a(P)$.

2. FOREST OF SEQUENCE OF STRINGS

If with each string $\eta_i = y_{i1}y_{i2} \ldots y_{in_i}$ of the resulting sequence $Q = (\eta_1, \eta_2, \ldots, \eta_m)$ formed by applying the algorithm of Sec. 1 to the given sequence of strings $P = (\xi_1, \xi_2, \ldots, \xi_m)$ where $\xi_i = x_{i1}x_{i2} \ldots x_{in_i}$, an auxiliary oriented graph $G_i = \langle V_i, \varrho_i \rangle$ is associated where $V_i = \{y_{i1}, y_{i2}, \ldots, y_{in_i}\}$ and $\varrho_i = \{(y_{i1}, y_{i2}), (y_{i2}, y_{i3}), \ldots, (y_{in_i-1}, y_{in_i})\}$ for $i = 1, 2, \ldots, m$, then an oriented graph $G = \langle V, \varrho \rangle$ where $V = \bigcup_{i=1}^{m} V_i$ and $\varrho = \bigcup_{i=1}^{m} \varrho_i$ may be determined.

Moreover, the described algorithm guarantees that if $y_{ij} = y_{hk}$, then $j = k$ and that $y_{ip} = y_{hp}$ as well as $x_{ip} = x_{hp}$ for $p = 1, 2, \ldots, j$, hence the binary relation $g = \{(y_{ij}, x_{ij});\ 1 \leqq i \leqq m \text{ and } 1 \leqq j \leqq n_i\}$ being a function. The domain of the function g is obviously the set of all vertices of the graph G and the range is a subset of the alphabet X over which all strings from P are formed.

The description of the algorithm implies immediately that the *oriented graph with labeled vertices* $\langle V, \varrho, X, g\rangle$, g being the labeling of vertices and X the set of values of the vertices, fulfils the following conditions:

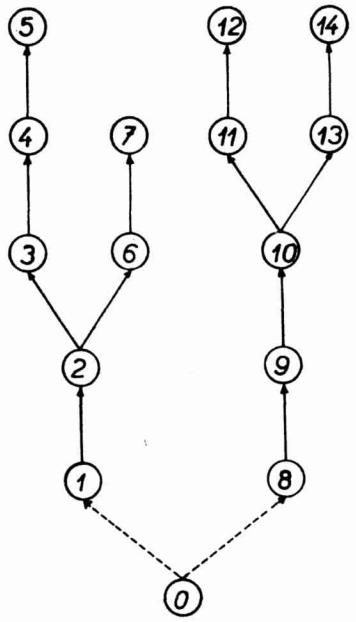

Fig. 1.

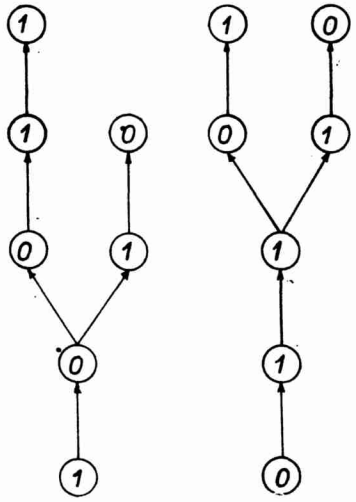

Fig. 2.

Theorem 2. *Every (connected) component of the graph* $\langle V, \varrho, X, g\rangle$ *is a rooted oriented tree, i.e., an oriented tree with exactly one vertex — the so called root — in which no edge ends. If* $v \in V$, $v^* \in V$ *are vertices such that* $v \neq v^*$ *and either both are roots of components or there is a vertex from which edges start to both* v *and* v^*, *then* $g(v) \neq g(v^*)$.

However, a graph whose all connected components are trees is called a *forest.* Therefore any graph $\langle V', \varrho', X', g'\rangle$ which is label isomorphic with $\langle V, \varrho, X, g\rangle$ will be called an *oriented forest of the sequence of strings P.* Here an isomorphism is said to be a *label isomorphism* if it preserves the labeling, i.e. if it is a one-to-one map ι of the set V' onto V such that $(v, v^*) \in \varrho' \Leftrightarrow (\iota v, \iota v^*) \in \varrho$ and at the same time $g'(v) = g(\iota v)$ for all $v, v^* \in V'$.

Fig. 1 shows the usual (planary) representation of the graph $G = \langle V, \varrho \rangle$ corresponding to the resulting sequence of strings Q from Example 1, which evidently has two components. The vertex 0 marked by dotted lines which is the only root of the graph extended in this way, could correspond to associating number zero with the auxiliary string $\xi_0 = \varepsilon$ from the algorithm. The advantage of such a formal extension is that the graph obtained would be immediately a tree and not generally a forest.

Fig. 2 shows the representation of the forest of the sequence P from Example 1. In this representation no proper names of vertices are introduced as usual since the vertices are distinguished by different positions of the corresponding circles.

3. INVARIANCE OF THE NUMBER $a(P)$

Theorem 1 implies that the number $a(P)$ given by formula (1) on the basis of the algorithm is independent of the order of strings in the sequence P. The following lemma provides a proof of the invariance of $a(P)$ on the order of strings which is independent of the meaning of $a(P)$ (i.e., not refering to Theorem 1).

Lemma 1. *If P^* is an arbitrary ordering of all members of the sequence of strings P, then $a(P^*) = a(P)$.*

Proof. It is well known that we can pass from P^* to P by means of a finite number of exchanges of two adjacent members of the sequence. Hence it is sufficient to prove Lemma 1 for the particular case when P^* differs from P just by an exchange of two adjacent members, i.e., if $P = (\xi_1, \xi_2, \ldots, \xi_m)$ then $P^* = (\zeta_1, \zeta_2, \ldots, \zeta_m)$ where $m \geqq 2$ and there exists a positive integer p, $1 \leqq p < m$ such that $\xi_i = \zeta_i$ for $i = 1, 2, \ldots, p-1, p+2, p+3, \ldots, m$ while $\xi_p = \zeta_{p+1}$ and $\xi_{p+1} = \zeta_p$.

Further, let us assume that the algorithm was applied also to the sequence P^* and that the symbols f^*, d_i^* and n_i^* have the analogous meaning for P^* as f, d_i and n_i have for P. Then it follows immediately from the above assumptions that $f(i) = f^*(i)$ for $i = 1, 2, \ldots, p-1$; further, $n_i^* = n_i$ for $i = 1, 2, \ldots, p-1, p+2, p+3, \ldots, m$ while $n_p^* = n_{p+1}$ and $n_{p+1}^* = n_p$ and, finally, $d_i = d_i^*$ for $i = 1, 2, \ldots, p-1, p+2, p+3, \ldots, m$. However, this means according to (1) that to verify the equality $a(P^*) = a(P)$ it is sufficient to show e.g. that $d_p^* + d_{p+1}^* = d_p + d_{p+1}$. Indeed, we shall succeed in proving this identity in all cases.

Let us distinguish the following cases. First of all, denote $\alpha = \xi_p \wedge \xi_{p+1}$ and consider the possibility $\alpha = \varepsilon$ (i.e., if ξ_p and ξ_{p+1} are nonempty strings then they have not the same first symbol). Then obviously $f(p) = f^*(p+1)$ and $f(p+1) = f^*(p)$ so that $\xi_{f(p)} \wedge \xi_p = \zeta_{f^*(p+1)} \wedge \zeta_{p+1}$ and $\xi_{f(p+1)} \wedge \xi_{p+1} = \zeta_{f^*(p)} \wedge \zeta_p$ which implies obviously $d_p = d_{p+1}^*$ and $d_{p+1} = d_p^*$ and hence also $d_p^* + d_{p+1}^* = d_p + d_{p+1}$.

Therefore, let $\alpha \neq \varepsilon$ in the sequel. If $\alpha \wedge \xi_{f(p)} = \varepsilon$ then also $\xi_{f(p)} \wedge \xi_p = \varepsilon$, i.e. $f(p) = 0$. However, this means that $\alpha \wedge \xi_i = \varepsilon$ for $i = 1, 2, \ldots, p-1$ so that

the only possibility is $f(p+1) = p$. Then it is seen immediately that also $f^*(p) = 0$ and $f^*(p+1) = p$ so that $\xi_{f(p)} \wedge \xi_p = \varepsilon = \zeta_{f^*(p)} \wedge \zeta_p$ and $\xi_{f(p+1)} \wedge \xi_{p+1} = \alpha = \zeta_{p+1} \wedge \zeta_p = \zeta_{p+1} \wedge \zeta_{f^*(p+1)}$ which again implies $d_p = d_p^*$ and $d_{p+1} = d_{p+1}^*$.

Thus, let $\alpha \wedge \xi_{f(p)} \neq \varepsilon$ hold in the sequel. Then obviously also $\alpha \wedge \xi_{f(p+1)} \neq \varepsilon$ and even — according to the definition of the function f — it holds $\alpha \prec \xi_{f(p+1)}$. Denoting $\beta = \xi_{f(p+1)} \wedge \xi_{p+1}$ we have obviously $\alpha \prec \beta$ and it remains to distinguish two cases.

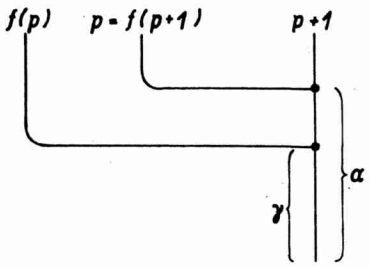

Fig. 3.

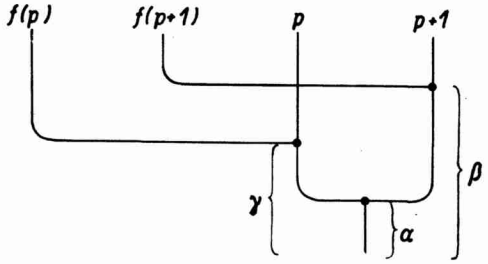

Fig. 4.

If $f(p+1) = p$ then evidently $\alpha = \beta$ and if we denote $\gamma = \xi_p \wedge \xi_{f(p)}$, it must be $\gamma \prec \alpha$ and $\gamma \neq \alpha$ which can be shown in the following way: It holds $\alpha \prec \xi_p$ as well as $\gamma \prec \xi_p$ and hence one of the possibilities $\alpha \preceq \gamma$ $\alpha = \gamma$, $\gamma \preceq \alpha$ must occur. However, if it were $\alpha \prec \gamma$ then $\xi_{p+1} \wedge \xi_{f(p)} = \alpha = \xi_{f(p+1)} \wedge \xi_{p+1}$ which is a contradiction with the definition of the function f (namely, with the requirement (ii)), since $f(p) < p = f(p+1)$. The case $\gamma \prec \alpha$ and $\gamma \neq \alpha$ is illustrated by Fig. 3 (which is a partial and sketchy representation of the graph of the sequence of strings) which makes it easy to see that $f^*(p) = f(p)$ and $f^*(p+1) = p = f(p+1)$. Hence, similarly as above, $d_p^* = d_p$ as well as $d_{p+1}^* = d_{p+1}$.

If finally $f(p+1) < p$ then the same argument as above leads to $\alpha \prec \gamma$ or $\gamma \prec \alpha$. However, if it were $\gamma \prec \alpha$ and $\gamma \neq \alpha$ then it would hold $\xi_p \wedge \xi_{f(p+1)} = \alpha$ which is again a contradiction with the definition of the function f. The last case $\alpha \prec \gamma$ is illustrated by Fig. 4 which shows easily that again $f^*(p) = f(p+1)$ and $f^*(p+1) = f(p)$. Hence $\gamma = \xi_{f(p)} \wedge \xi_p = \zeta_{f^*(p+1)} \wedge \zeta_{p+1}$ as well as $\beta = \xi_{f(p+1)} \wedge \xi_{p+1} = \zeta_{f^*(p)} \wedge \zeta_p$ and, consequently, $d_p = d_p^*$ and $d_{p+1} = d_{p+1}^*$ which completes the proof.

Problem 1. The nature of the above proof shows that the assertion of Lemma 1 is apparently a simple consequence of some identity in the sequential algebra, which is an algebraic structure including a certain semilattice as well as a free semigroup satisfying at least the following axioms:

1. $\prec$ is a partial ordering, i.e., it is a reflexive, antisymmetric and transitive binary relation with the least element ε;

2. l is a function associating with every string ξ a non-negative number $l(\xi)$ so that $\xi \prec \eta$ implies $l(\xi) \leqq l(\eta)$ and $l(\xi) = 0 \Leftrightarrow \xi = \varepsilon$;

3. $\wedge$ is a binary operation defined everywhere which is idempotent (i.e., $\xi \wedge \xi = \xi$), commutative, associative and whose neutral element is ε, i.e., $\varepsilon \wedge \xi = \xi \wedge \varepsilon = \varepsilon$;

4. the binary operation of concatenation is associative and ε is its unit;

5. for any three strings ξ, ζ and η it holds either $\xi \wedge \zeta \preccurlyeq \xi \wedge \eta$ or $\xi \wedge \zeta = \xi \wedge \eta$ or $\xi \wedge \eta \preccurlyeq \xi \wedge \zeta$;

6. for any three strings ξ, ζ and η, $\xi \wedge \zeta \succ \xi \wedge \eta$ implies $\zeta \wedge \eta = \xi \wedge \eta$; and

7. for any three strings ξ, ζ and η it holds $\xi(\zeta \wedge \eta) = \xi\zeta \wedge \xi\eta$ but generally does not hold $\xi \wedge \zeta\eta = (\xi \wedge \zeta)(\xi \wedge \eta)$.

Lemma 2. *If P^* is a sequence of strings obtained from the sequence of strings P by inserting a string, which is an initial segment of some string in P or is empty, between two adjacent members or in front of the first or behind the last member of the sequence P, then $a(P^*) = a(P)$.*

Proof. According to Lemma 1 we may assume that P and P^* are ordered in the following way: The string whose initial segment was put into P to form P^* is in the first place both P and P^* (or, if there are more such strings, any one of them); the inserted string is in the second place in P^* and, finally, the i-th string in P is the $(i+1)$-st string in P^* for $i = 2, 3, \ldots, m$. Hence it holds $n_1 = n_1^*$ and $n_i = n_{i+1}^*$ for $i = 2, 3, \ldots, m$ so that the length of the sequence P is m while that of P^* is $m + 1$. In accordance with the above notation we may write $d_1 = d_1^* = 0$, $d_2^* = n_2^*$ and $d_i = d_{i+1}^*$ for $i = 2, 3, \ldots, m$ which evidently implies $a(P) = (n_1 - d_1) + \sum_{i=2}^{m}(n_i - d_i) = (n_1^* - d_1^*) + (n_2^* - d_2^*) + \sum_{i=3}^{m+1}(n_i^* - d_i^*) = a(P^*)$.

Lemmas 1 and 2 together imply

Theorem 3. *If P^* is a sequence of strings such that each its member is an initial segment of a string of a finite sequence P, then $a(P^*) \leqq a(P)$.*

Moreover, Lemma 1 implies that the above algorithm associates with every finite set of strings M over an alphabet X a number $a(M)$ which is equal to the number $a(P)$ for an arbitrary ordering P of the set M.

4. ANOTHER METHOD OF EVALUATING NUMBER $a(M)$

If M is the set of strings over an alphabet X, denote by $b(M)$ the number of all (mutually different) elements from X which appear in the first places of strings from M. Evidently

$$0 \leqq b(M) \leqq |X| \quad \text{and} \quad b(M) = 0 \Leftrightarrow M = \emptyset \quad \text{or} \quad M = \{\varepsilon\} \tag{2}$$

holds.

It is seen immediately that the number $b(M)$ gives the number of (connected) components of the forest of the set of strings M, i.e., of the forest of the language M.

Let again $M = \{\xi_1, \xi_2, \ldots, \xi_m\}$ where $\xi_i \in X^\infty$, $l(\xi_i) = n_i$ for $i = 1, 2, \ldots, m$ and put $k = \max_{1 \leqq i \leqq m} (n_i - 1)$. Further, define

(3) $$M_\alpha = \{\xi;\ \xi \in X^\infty \text{ and } \alpha\xi \in M\} \quad \text{for all} \quad \alpha \in X^\infty$$

and

(4) let $\tilde{M}$ denote the set of all proper (and nonempty) initial segments of strings from M;

(5) $$c(M) = b(M) + \sum_{x \in X} c(M_x) \quad \text{and} \quad c(\emptyset) = 0\,;$$

(6) $$d(M) = b(M) + \sum_{\alpha \in \tilde{M}} b(M_\alpha)\,.$$

It follows immediately from (5) that

(7) $$c(M) = b(M) + \sum_{x \in X} b(M_x) + \sum_{xy \in X^2} b(M_{xy}) + \ldots + \sum_{x_1 x_2 \ldots x_k \in X^k} b(M_{x_1 x_2 \ldots x_k})$$

holds and since $\{M_\alpha;\ \alpha \in \tilde{M}\} \subset \{M_x;\ x \in X\} \cup \{M_{xy};\ xy \in X^2\} \cup \ldots \cup \{M_{x_1 x_2 \ldots x_k};\ x_1 x_2 \ldots x_k \in X^k\}$, it is evident that $d(M) \leqq c(M)$.

If, to the contrary, $\alpha \in X^h$ where $1 \leqq h \leqq k$ but $\alpha \notin \tilde{M}$, then $M_\alpha = \emptyset$ and hence $b(M_\alpha) = 0$ according to (2). This implies

Lemma 3. *$d(M) = c(M)$ for every set of strings M.*

Theorem 4. *$d(M) = a(M)$ for every set of strings M.*

Proof. Let $M = \{\xi_1, \xi_2, \ldots, \xi_m\}$ be an arbitrary set of strings, $n_i = l(\xi_i)$ for $i = 1, 2, \ldots, m$ and let us use mathematical induction with respect to $n = \sum_{i=1}^{m} n_i$ to prove the theorem. For $n = 1$ it is obviously $d(M) = a(M)$. Accepting the assumption of induction for $n - 1$, we prove the same identity for $n > 1$. To this purpose, form the set M^* from M by omitting in the string $\xi_m = \xi_m^* x$ its last symbol x so that, denoting by asterisk the quantities concerning the set M^*, it obviously holds $n_i = n_i^*$ and $d_i = d_i^*$ for $i = 1, 2, \ldots, m - 1$ and $n_m = n_m^* + 1$, assuming without further notice that the algorithm was applied to both sets M and M^*, in the ordering mentioned above. Let us distinguish several cases.

a) If $n_m > 1$, i.e. $\xi_m^* \neq \varepsilon$, then $\xi_m^* \in M^*$, $m^* = m$ and even $b(M) = b(M^*)$ since the set of the first symbols in strings from M does not change when passing to M^*. The following two cases may occur:

a1) $d_m = d_m^*$; in this case we obtain immediately $a(M) = a(M^*) + 1$ according to (1) (since $n_m = n_m^* + 1$) and, on the other hand, $d_m = d_m^*$ implies that ξ_m is not an

initial segment of a string ξ_i for $1 \leqq i < m$ (if it were, it would be $d_m = n_m > n_m^*$) which means $\tilde{M} = \tilde{M}^* \cup \{\xi_m^*\}$ while evidently $b(M_{\xi_m^*}) = 1$ since ξ_m^* is a proper segment of only one string in M, namely, the string ξ_m. Then we may write $d(M) = b(M) + \sum_{\alpha \in \tilde{M}} b(M_\alpha) = b(M^*) + \sum_{\alpha \in \tilde{M}^*} b(M_\alpha^*) + b(M_{\xi_m^*}) = d(M^*) + 1$. However, according to the assumption of induction it is apparently $\sum_{i=1}^{m^*} n_i^* = n - 1$ and hence $d(M^*) = a(M^*)$ and the preceding two identities imply $d(M) = a(M)$;

a2) $d_m = d_m^* + 1$; in this case we obtain immediately $a(M) = a(M^*)$ according to (1). On the other hand, it means that ξ_m is an initial segment of a string ξ_i with $1 \leqq i < m$ so that $\xi_m^* \in \tilde{M}^*$ and hence $\tilde{M} = \tilde{M}^*$ which again implies according to (6) that $d(M) = d(M^*)$. Making use of the assumption of induction $d(M^*) = a(M^*)$ (as in the previous case) we obtain $d(M) = a(M)$;

b) If $n_m = 1$, i.e. $\xi_m^* = \varepsilon$, $m^* + 1 = m$, $n_m^* = 0$ and $M^* \cup \{\xi_m\} = M$, then we shall again distinguish two cases:

b1) $d_m = 0$; this means that ξ_m is not an initial segment of a string ξ_i, $1 \leqq i < m$; therefore $b(M) = b(M^*) + 1$ while $\tilde{M} = \tilde{M}^*$ so that $d(M) = d(M^*) + 1$ according to (6). On the other hand, (1) yields immediately that $a(M) = \sum_{i=1}^{m-1} (n_i - d_i) + \sum_{i=1}^{m^*} (n_i^* - d_i^*) + 1 = a(M^*) + 1$. The assumption of induction $d(M^*) = a(M^*)$ implies now $d(M) = a(M)$;

b2) $d_m = 1$; this means that ξ_m is an initial segment of a string ξ_i with $1 \leqq i < m$; consequently, $b(M) = b(M^*)$ as well as $\tilde{M} = \tilde{M}^*$. According to (6) we obtain $d(M) = d(M^*)$; on the other hand, (1) implies $a(M) = a(M^*)$ which together with the assumption of induction gives again $d(M) = a(M)$ which completes the proof.

Another proof of Theorem 4. Denote by R^q the set of all mutually different initial segments of the length q of strings from M and $R = R^1 \cup R^2 \cup \ldots \cup R^k$ (number k was introduced in the introductory part of Chap. 4). Obviously $|R^1| = b(M)$ and $|R^{w+1}| = \sum_{\alpha \in R^w} b(M_\alpha)$ for $w = 1, 2, \ldots, k$. Hence $a(M) = \sum_{j=1}^{k+1} |R^j| = b(M) + \sum_{j=1}^{k} \sum_{\alpha \in R^j} b(M_\alpha) = b(M) + \sum_{\alpha \in R} b(M_\alpha)$. If $\alpha \in R - \tilde{M}$, then $b(M_\alpha) = 0$ and hence

$$a(M) = b(M) + \sum_{\alpha \in \tilde{M}} b(M_\alpha) = d(M) .$$

Example 1 (continued). Determine $d(P)$ for the sequence of strings from Example 1 according to the rule (6). Evidently $b(P) = 2$ and $\tilde{P} = \{1, 0, 10, 01, 100, 101, 011, 1001, 0110, 0111\}$ so that we find successively $P_1 = \{0011, 010\}$, $P_0 = \{1101, 1110\}$, $P_{10} = \{011, 10\}$, $P_{01} = \{101, 110\}$, $P_{100} = \{11\}$, $P_{101} = \{0\}$, $P_{011} = \{01, 10\}$, $P_{1001} = \{1\}$, $P_{0110} = \{1\}$ and finally $P_{0111} = \{0\}$. According to (6), $d(P) = b(P) + b(P_1) + b(P_0) + b(P_{10}) + b(P_{01}) + b(P_{100}) + b(P_{101}) + b(P_{011}) + b(P_{1001}) + b(P_{0110}) + b(P_{0111}) = 2 + 1 + 1 + 2 + 1 + 1 + 1 + 2 + 1 + 1 + 1 = 14$.