REGULAR EXPRESSIONS FOR INFINITE TREES AND A STANDARD FORM OF AUTOMATA

By A.W.MOSTOWSKI

Math.Institute of the University of Gdańsk , 80-952 Gdańsk , Poland
Wita Stwosza 57

Summary For Rabin pair automata [R1] a standard form is defined /def. 2/ i.e. such that an ordered subset $\{s_1,\dots,s_{2I-1}\}$ of states is distinguished in such a way that a path of a run is accepting /rejecting if for some i even/ odd, $1 \leq i \leq 2I-1$, the s_i appears infinitely often, and all s_j , $j < i$ only finitely many times. The class of standard automata is big enough to represent all f.a. representable sets /th.1/ but has many properties similar to special automata defined in [R1] . A standard regular expression is defined /def. 6/ describing a process of forming of an infinite tree, as well as a process of building of an automaton /analysis and synthesis theorems 3,4/. The standard regular expressions are a generalisation of McNaughtons formula $\cup \alpha \beta^{\omega}$ /cf. [N] /.

1. Notions. We refer the reader to automata on infinite binary trees as defined in [R1] , and to automata on finite trees as given in [TW] . For simplicity reasons the theorems are stated and proved for a binary case, but some minor changes will allow the theorems to be true for mixed trees as given in [M3] .

Only in par. 3 the regular expressions theory is shown explicitely to be working for mixed trees.

An alphabet Σ is fixed. A /valued, infinite, binery/ tree $t = (T,v)$ is a tree $T = \{0,1\}^*$ with a valuation $v : T \longrightarrow \Sigma$. The relation $x \leqslant y$ is an order, where $x \leqslant y$ iff there exists $z \in T$ such that $xz = y$. Maximal chains for $\leqslant$ are called paths of T. For $A,B \in PS(S)$ the relation $A \subset B$ denotes proper inclusion.

A /partial/ table is a triple $\mathcal{T} = \langle S, M, s_{in} \rangle$, where S is a finite set of states $s_{in} \in S$, amd $M : S \times \Sigma \longrightarrow PS(S \times S)$ is a transition function. A run of $\mathcal{T}$ on t is a <u>total</u> function $r : T \longrightarrow S$ such that : $r(\Lambda) = s_{in}$, $\langle r(x0), r(x1) \rangle \in M(r(x), v(x))$ for $x \in T$.

A finite automaton on trees is a pair $\mathcal{A} = \langle \mathcal{T}, W \rangle$ of a table $\mathcal{T}$ and a set $W \leqslant S^\omega$ of accepting paths. A tree t is accepted by the $\mathcal{A}$ iff there exists a run r of $\mathcal{T}$ on t, such that for each path $\pi \leqslant T$, $r/\pi \in W$, i.e. each path of the run is accepting.

Let $\mathcal{L} = \{(U_i, L_i) : 0 \leqslant i < I, U_i, L_i \leqslant S\}$ be a collection of pairs and $\mathcal{F} \leqslant PS(S)$ a collection of subsets. Let $s : \omega \to S$ be a sequence. We shall define a set $[\mathcal{L}]$ of sequences on S as follows: $s \in [\mathcal{L}]$ iff card $(s^{-1}(U_i)) \geqslant \omega$ and card $(s^{-1}(L_i)) < \omega$, for some $i : 0 \leqslant i < I$. We speak of a Rabin automaton or a pair automaton if $W = [\mathcal{L}]$. The $I = I(\mathcal{A})$ is called a Rabin pair index of the automaton. Let us define $s \in [\mathcal{F}]$ iff $\{x : \text{card } (s^{-1}(x)) \geqslant \omega\} \in \mathcal{F}$. For $W = [\mathcal{F}]$ we shall say that $\mathcal{A}$ is a Muller, or set, automaton.

<u>Definition 1.</u> A Rabin automaton is a chain form automaton /c.f.a./ iff $L_0 \subset L_1 \subset \ldots \subset L_{I-1}$ for the $L_0, \ldots, L_{I-1}$ appearing in the $\mathcal{L}$.

For a c.f.a. the index $I(\mathcal{Q})$ will be called a chain index and denoted by $CI(\mathcal{Q})$.

Definition 2. A Rabin automaton is a standard form automaton /s.f.a./ iff an order on a certain subset F of S is given, say $F = \{s_1, s_2, \ldots, s_{2I-1}\}$, for which $U_i = \{s_{2i+1}\}$, $L_i = \{s_1, \ldots, s_{2i}\}$ for $i = 0,1,\ldots,I(\mathcal{Q}) - 1$. For s.f.a. the index $I(\mathcal{Q})$ will be called a standard form index and denoted by $SFI(\mathcal{Q})$.

Definition 3. For a Muller automaton $\mathcal{Q} = \langle \mathcal{T}, \mathcal{F} \rangle$ let us chose a chain $X_1 < X_2 < \ldots < X_z$ such that $X_{2i-1} \notin \mathcal{F}$, $X_{2i} \in \mathcal{F}$, for $i = 1,2,\ldots,m$ where $z = 2m$, or $z = 2m + 1$, and additionally $X_z \notin \mathcal{F}$ for $z = 2m + 1$. The maximal m for all chains with this property is called a Muller index of the $\mathcal{Q}$ and denoted by $MI(\mathcal{Q})$.

Definition 4. Let the index be one of I, CI, SFI, MI indices for suitable classes of automata, and let V be a set of trees. Index (V) is defined as minimal index $(\mathcal{Q})$ for an automaton from a suitable class representing V, or infinity, when no such automaton in the class exists.

For ω-sequences the MI and I was studied in [W] : for trees the I was studied in [M 1,2].

2. Equivalence of various notions of automata. Muller automata are equivalent to Rabin automata [R1]. We shall prove in the sequel something more stronger, that each automaton is equivalent to a chain automaton and in a consequence to a standard form automaton.

Lemma 1. Each Muller automaton $\mathcal{Q} = \langle \mathcal{T}, \mathcal{F} \rangle$ is equivalent to a chain automaton $\mathcal{Q}' = \langle \mathcal{T}', \mathcal{R} \rangle$ such that $MI(\mathcal{Q}) = CI(\mathcal{Q}')$.

Proof. For $\mathcal{F} = \{\emptyset\}$ the lemma is trivially true. Suppose $\emptyset \in \mathcal{F}$

and define $\mathfrak{X}_0 = \{\emptyset\}$, $\mathcal{F}_0 = \mathcal{F}$, $\mathcal{F}_{i+1} = (PS(S) \setminus \mathcal{F}_i) \cup \mathfrak{X}_i$, $\mathfrak{X}_{i+1} = \{X : PS(X) \subset \mathcal{F}_{i+1}\}$. Then $PS(S) = \mathfrak{X}_z = \mathfrak{X}_{z+1} = \ldots$ where $z = 2m$ or $z = 2m + 1$, where $m = MI(\mathcal{A})$. Moreover $\mathcal{F} = \bigcup_{i=1}^{m} (\mathfrak{X}_{2i} \setminus \mathfrak{X}_{2i-1})$.

A strategy of the proof is as follows. For each $\mathfrak{X}_i$, $i = 1,\ldots,z-1$ we shall construct a suitable deterministic special automaton $\mathcal{C}_i$ testing whether or not a set of the frequent states on the paths /for a run of the table $\mathcal{T}$ / does not belong to $\mathfrak{X}_i$. Then we shall use a product automaton of $\mathcal{T}$ and the automata $\mathcal{C}_i$. Let $Y_1^{(i)},\ldots, Y_{n(i)}^{(i)}$ be a set of maximal respective to $\prec$ elements to the $\mathfrak{X}_i$. A set of states of the $\mathcal{C}_i$ is $\{\overline{0},0,1,\ldots,n(i)-1\}$. For the product $\mathcal{T}' = \mathcal{T} \times \mathcal{C}_1 \times \ldots \times \mathcal{C}_{z-1}$ the initial state is $(s_{in},0,\ldots,0)$. The transition function is :

$M((s, \ldots, j^{(i)}, \ldots), v) = \langle (s_1, \ldots, j_1^{(i)}, \ldots) , (s_2, \ldots, j_2^{(i)}, \ldots) \rangle$

where $\langle s_1, s_2 \rangle \in M(s,v)$ and for $k = 1,2,$:

For $j^{(i)} \neq \overline{0}$ and $s_k \notin Y^{(i)}_{j^{(i)}+1}$:

$j_k^{(i)} = j^{(i)} + 1$ for $j^{(i)} \neq n(i)-1$, $j_k^{(i)} = \overline{0}$ for $j^{(i)} = n(i)-1$

For $j^{(i)} \neq \overline{0}$ and $s_k \in Y^{(i)}_{j^{(i)}+1}$: $j_k^{(i)} = j^{(i)}$

For $j^{(i)} = \overline{0}$: $j_1^{(i)} = j^{(i)} = 0$.

Now let us observe that for a path of the S-run, the set of frequent states does not belong to the $\mathfrak{X}_i$ means the state $\overline{0}$ appears infinitely often on the i-th coordinate of $\mathcal{C}_1 \times \ldots \times \mathcal{C}_{z-1}$ Moreover if for the i-th coordinate $\overline{0}$ appears only finitely often then for each j : $i < j < z$, $\overline{0}$ appears only finitely often on the j-th coordinate.

Let us define $L_0 = \emptyset$, $L_z =$ the set of all states of $\mathcal{T}'$, and for $i = 1,\dots,z-1$: $L_{z-i} = \{(s,t_1,\dots,t_{z-1}) : t_j = 0$ for some $j : i \leq j < z\}$, Hence $L_0 < L_1 \dots L_{z-1} < L_z$. Now for a path u of the $\mathcal{T}'$ run and its S-projection s , according to that what is said above :

$$s \in [\mathcal{X}_{z-j} \setminus \mathcal{X}_{z-j-1}] \quad \text{iff} \quad u \in [\{(L_{j+1}, L_j)\}] .$$

Hence for $\mathcal{R} = \{(L_1,L_0)(L_3,L_2),\dots,(L_{2m-1},L_{2m-2})\}$ in the case $z = 2m$ and for $\mathcal{R} = \{(L_2,L_1) , (L_4,L_3) , \dots , (L_{2m},L_{2m-1})\}$ in the case $z = 2m + 1$, the $\mathcal{A}' = \langle \mathcal{T}', \mathcal{R} \rangle$ is equivalent to $\mathcal{A}$. It has a chain form, and $CI(\mathcal{A}') = MI(\mathcal{A})$ as required in the lemma.

Note. Let us observe that maximal number of states for the automaton $\mathcal{A}'$ is reached when all sets of even cardinality belongs to the $\mathcal{F}$. Then the number of states is $n \prod_{k=1}^{n+1} \{1 + \binom{n}{k}\}$, where m is the number of states for $\mathcal{T}$. Hence the cardinality of states of $\mathcal{T}'$ can be $\geq \exp(n)$, but surely is $\leq \exp(n^2)$.

Since a c.f.a. can be regarded as a Muller automaton of the same index, we obtain from the lemma an immediate.

Corrolary. For any set V of trees $CI(V) = MI(V)$.

Now we shall prove the crucial for our paper

Theorem 1. Each automaton is equivalent to an automaton in a standard form.

Proof. We shall take an automaton in a chain form, e.g. constructed as in lemma 1. Say $\mathcal{A} = \langle \mathcal{T}, \mathcal{R} \rangle$ where $\mathcal{R} = \{(L_{2i-1},L_{2i-2}) : i = 1,\dots,m\}$ and $L_0 < L_1 < \dots < L_{2m-1}$. We shall order the states of $\mathcal{A}$ eventually adding some new nonaccesible states. Suppose some

states are ordered as $s_1,\ldots,s_{2J-1}$ and

$\bigcup_{k=1}^{J}[(\{s_{2k-1}\},\{s_1,\ldots,s_{2k-2}\})] = \bigcup_{i=1}^{N}[(L_{2i-1}, L_{2i-2})]$

Here $[\mathfrak{R}]$ denotes the set of paths satisfying $\mathfrak{R}$. Let $L_{2N+1} \setminus L_{2N} = \{z_1,\ldots,z_s\}$ and $L_{2N} \setminus L_{2N-1} = \{w_1,\ldots,w_t\}$. We shall define $s_{2J-1+2j} = w_j$, $j = 1,\ldots,t$, $s_{2J-1+2t+2n} = z_n$, $n = 1,\ldots,s$. The new added nonaccessible states are $s_{2J-1+2j-1}$, $j = 1,\ldots,t$ and $s_{2J-1+2t+2n-1}$, $n = 1,\ldots,s$, and the transition function for them $M = \emptyset$. It is easy to observe that $[(L_{2N+1}, L_{2N})] = \bigcup_{k=J+1}^{k=J+2n-2s}[(\{s_{2k-1}\},\{s_1,\ldots,s_{2k-2}\})]$ which gives an inductive proof of the theorem.

Remark to theorem 1. The automaton $\langle \mathcal{T}, \mathfrak{R} \rangle$ where $\mathfrak{R} = \{(F,\emptyset)\}$, $F \subseteq S$ is called special automaton /s.a./ /cf. [R1]/. In [R2] it is proved that not every f.a. is equivalent to a s.a. As theorem 1 shows the situation for s.f.a. is better because they can represent all f.a. reorentable sets. But s.f.a. have much in common with s.a. For both an accepting run starts from s_{in} to an appearance of a state from F for each path. /For s.f.a. the F is the F from definition 2/. Next an accepting run is built from finite trees with runs from some state of F to states of F. For s.a. it is a necessary and sufficient condition for a run to be accepting.

For s.f.a. the additional standard form conditions must be made to the mode of the building of an accepting run.

Theorem 2. For any $N = 0,1,\ldots$ there exists a set of infinite binary trees /respective ω-sequences/ of standard form index N.

Proof. Let us choose an N letter alphabet $\Sigma = \{\sigma_1,\ldots,\sigma_N\}$ and define : a tree $t \in W$ iff each path of t is of the form

$\sigma_1^*,\ldots,\sigma_{i-1}^*\sigma_i^\omega$ for some $i = 1,\ldots,N$. Let the states of an automaton be $s_1,s_2,\ldots,s_{2N+1} = s_{in}$, where $s_{2i-1} = \sigma_i$, $i = 1,\ldots,N$ and the partial transition function $M(s_{in},\sigma_i) = \langle\sigma_i,\sigma_i\rangle$, $M(\sigma_i,\sigma_j) = \langle\sigma_j,\sigma_j\rangle$ for $i \leq j$, and $U_i = \{s_{2i+1}\}$, $L_i = \{s_1,\ldots,s_{2i}\}$, $i = 0,\ldots,N-1$. The automaton is a deterministic s.f.a. representing W, of index N. Moreover the whole Situation can be coded for trees over a fixed, say two letter alphabet, not depending on N.

Now suppose some nonneccesarilly deterministic s.f.a. of index $< N$ represents W. Hence two paths $\sigma_1^*,\ldots,\sigma_{i-1}^*\sigma_i^\omega$ and $\sigma_1^*,\ldots,\sigma_{j-1}^*\sigma_j^\omega$, $i<j$ from two accepted trees must belong to one $[(\{s_{2k-1}\},\{s_1,\ldots,s_{2k-2}\})]$. Then by grafting on s_{2k-1} muttually suitable parts of accepting runs of the trees, we obtain an accepting run of a tree containing a path $\sigma_1^*,\ldots,\sigma_{j-1}^*(\sigma_j^+\sigma_i^+)^\omega$, which is impossible.

According to the remark to theorem 1, let us observe that using a method similar to that of [M3] and [TW] concerning regular expressions theory one can prove that a computable function $f(N,M,n)$ exists such that if there exists s.f.a. of index M, equivalent to a s.f.a. of index N having n states, then an other equivalent s.f.a exists of index M, having at most $f(N,M,n)$ states. The proof is easy but tedious. For s.a. a similar result is given in [K].

From the decidability of the equivalence problem we obtain Theprem 3. A standerd form index of a set of trees represented by an automaton is a computable function of the automaton.

3. Regular expressions. McNaughton introduced in [N] regular expressions $\bigcup \alpha \beta^{\omega}$ where α and β are classical regular expressions over sequences and the sum is finite. The $\alpha \beta^{\omega}$ describes a language $L(\alpha \beta^{\omega}) = \{ x = x_0 x_1 x_2 \dots , x_0 \in L(\alpha) , x_i \in L(\beta) , x_i \neq \Lambda , i > 0 \}$, as well as a nondeterministic automaton on sequences. Thatcher Wright [TW] generalised classical expressions from sequences to finite trees labelled by some states on the frontier. Instead of . and * there are new operations : . s /s-concatenation/ and ^{s} /s-closure/ . The regular expressions for infinite trees, with two kinds of iterations : star iteration /finite/ and omega iteration /infinite/ are investigated in [M3] .

Now we shall introduce an infinite operation which must be only once used in the expression, similarly as the infinite operation $^{\omega}$ is once used in McNaughton's expression $\alpha \beta^{\omega}$.

Definition 5. Let $B_0, B_1, \dots, B_{2I-1}$ be a collection of sets of finite trees indexed in the frontier by some finite set S containing an ordered subset $F = \{ s_1, \dots, s_{2I-1} \}$. The result of a standard omega iteration on $B_0, B_1, \dots, B_{2I-1}$, is any tree $t = \lim_{n \to \infty} t_n$, where the sequence t_n of trees satisfies the following :

1° $t_0 \in B_0$, t_{n+1} is obtained from t_n by grafting on each frontier node of t_n , labelled by some $s_i \in F$, some, nonsingle node, tree belonging to B_i . Different trees can be grafted on different nodes with the same label. For nodes labelled by a $s \notin F$, nothing required by grafted.

2° For each infinite path of the tree t , there exists an i : $0 \leq i < I$ such that the label s_{2i+1} appears infinitely many

times and the labels $s_1,\ldots,s_{2i}$ appear only finitely many times.

The set of all trees obtained in a such way will be denoted by $L(B_o,B_1/s_1,\ldots,B_{2I-1}/s_{2I-1})$ or by $L(B_o,\ldots,B_{2I-1})$ when no confusion can arise.

Let us observe that for $F \neq S$ some mixed or finite trees can arise, but for $F = S$ the result of the standard omega iteration is either the empty set of trees, or contains infinite trees only.

The operation is a natural generalisation of McNaughton's operation $\bigcup AB^{\omega}$. Especially for sets $B_o = A_1 \cup \ldots \cup A_I$, $B_1,\ldots,$ B_{2I-1}, of finite sequences, where the A_i and B_{2i-1} are labelled by s_i for $i = 1,\ldots,I$, the result of standard omega iteration $L(B_o,B_1,\ldots,B_{2I-1}) = A_1B_1^{\omega} \cup A_2B_3^{\omega} \cup \ldots \cup A_IB_{2I-1}^{\omega}$.

For a set B_o of finite trees, $L(B_o) = B_o$, sice the set F is empty. The following definition will now be justified

<u>Definition 6.</u> Let $\varphi_o,\varphi_1,\ldots,\varphi_{2I-1}$ be regular expressions over a species $\Sigma = \Sigma_o \cup \Sigma_1 \cup \ldots \cup \Sigma_n$, cf. [TW]. The Σ_o is a set of 0-argument functions, i.e. the labels of frontier nodes. We can always suppose that Σ_o contains an ordered subset $F = \{s_1,\ldots,s_{2I-1}\}$. <u>The standard regular expression</u> $SRE(\varphi_o, \varphi_1/s_1,\ldots, \varphi_{2I-1}/s_{2I-1})$ on /infinite/ trees is : φ_o and a mapping of the subset $F = \{s_1,\ldots,s_{2I-1}\} \leq \Sigma_o : s_1 \longrightarrow \varphi_1$, $s_2 \longrightarrow \varphi_2$, ..., $s_{2I-1} \longrightarrow \varphi_{2I-1}$.

The language $L(SRE(\varphi_o, \varphi_1/s_1,\ldots, \varphi_{2I-1}/s_{2I-1}))$ is defined as $L(L(\varphi_o), L(\varphi_1)/s_1, \ldots, L(\varphi_{2I-1})/s_{2I-1})$.

Exampel. For $\Sigma = \Sigma_0 \cup \Sigma_1, \Sigma_0 = \{\lambda\}, \Sigma_1 = \{6_1,\ldots,6_n$ we are dealing with finite sequences. Let Thatcher-Wright regular expressions α and β be built over Σ. Let $s_0, s_1 \in \Sigma$. In [TW] formalism $\alpha\beta^\omega$ will be written as follows $\alpha\beta^\omega = \mathrm{SRE}(\alpha \cdot \lambda\ s_1, \beta \cdot \lambda\ s_1/s_1)$.

In a general case, when forming a tree, we start from φ_0 and moving along a path we move trough the $\varphi_1,\ldots,\varphi_{2I-1}$ in some order. The next φ depends on a moment in which we decide to leave the φ and by a label with which we leave it. The condition 2° from definition 5 gives a strategy which must be fulfilled simultaneously for all paths. According to the condition 2° some runs must be rejected. Hence the standard regular expression describes the way of forming a run and an automaton as well.

One must only construct the tables $\mathcal{T}_i$ representing φ_i, with initial states $s_1,\ldots,s_{2I-1}$ and accepting states s_{in}, $s_1,\ldots,s_{2I-1}$. /th. 2, lemmas 1o-12 [TW] p.62,69-71. The reader please note that in Thatcher-Wright formalism initial means final in our terminology, and accepting means initial in our terminology/.

During the construction we must take care that the inner states of tables are mutually disjoint. Then the sum of tables $\mathcal{T}_i$ $i = 0,1,\ldots,2I-1$ with s_{in} as an initial state, /here in our terminology/, is a table of an automaton representing $L(\mathrm{SRE}(\varphi_0,\varphi_1/s_1,\ldots,\varphi_{2I-1}/s_{2I-1}))$. The conditions are $\{(L_{2i+1},L_{2i}) : 0 \ i \ 1\}$ where $L_j = \{s_1,\ldots,s_j\}$.

This gives the following

Theorem 3. /Synthesis theorem/ For each standard regular expression one can effectively construct an automaton, in a standard form,

representing the set of trees described by the expression.

Theorem 1 and the remark to theorem 1 together with analysis theorem for regular expressions over finite trees / [TW] th.9 p.67/ gives immediately the following

Theorem 4. /Analysis theorem/ For each automaton one can effectively construct the standard regular expression in the sense of definition 6, describing the set represented by the automaton.

Note that both theorems are valid for infinite trees as well as for mixed trees.

4. Bibiography.

[K] Karpiński M., Decidability of the weak definiability in the s2s theories. 3-d symposium of MFCS, Zaborów, January 21-26 1980, ICS PAS Reports 411 p.47-48.

[N] McNaughton R., Testing and generating infinite sequences by a finite trees and inequalities between various Rabin pair indices, Information proccessing letters 15 1983 , pl.59-163

[M2] Mostowski A.W., Classes of automata of a given Rabin pair index. Proccedings of workshop on algorithms and computing theory, September 7-1o,1981 ed. by M.Karpiński and Z.Habasiński, Poznań 1981.

[M3] Mostowski A.W., On differences on automata on infinite trees and those on sequences, Report on the 1-st GTI Workshop, Lutz Prietze ed. Universitat Padderborn, Reiche Teoretische Informatik Marz 1982.

[R1] Rabin M.O., Decidability and definiability in Second-order Theories, Actes Congress Intern. Math., 197o Tome 1 p.239-244.

[R2] Rabin M.O., Weakly definiable relations and special automata. Math.Logic Foundations Set Theory, North Holland 197o p.1-23.

[TW] Thatcher J.W. Wright J.B., Generalised Finite Automata Theory with an Application to Second-order Logic. Math. Systems Theory, vol.2, p.57-81.

[W] Wagner K., On ω-regular Sets. Information and Control 43 1979 , p.123-177.