

IT-math F2003 : Supplementary Material

Episode 12, April 29, 2003

Automata, Regular Expressions, and Grammars

1. DEFINITION. A (non-deterministic) automaton M is a quintuple of the form (A, S, T, I, R) , where

- A is a finite alphabet;
- S is finite a non-empty set (whose elements are called *states*);
- $T \subseteq S$ (elements of T are called *accepting states*);
- $I \in S$ is a distinguished state, called the *start* or *initial state*;
- $R \subseteq (S \times A) \times S$ is a relation referred to as the *transition relation*.

If the transition relation is known to be a function from $S \times A$ to S then the automaton is called *deterministic*¹ and the transition relation is called *transition function*.

If M is a non-deterministic automaton then one can extend the transition relation R to a relation \tilde{R} between $S \times A^*$ and S in the following inductive manner:

$$(s, \varepsilon) \tilde{R} t \iff s = t, \quad \text{and}$$

$$(s, wa) \tilde{R} t \iff \text{there is } s' \in S \text{ such that } (s, w) \tilde{R} s' \text{ and } (s', a) R t.$$

Equivalently, for a word $w = a_1 \dots a_n \in A^*$ one has $(s, w) \tilde{R} t$ if and only if there is a sequence s_0, s_1, \dots, s_n of states of M such that $s = s_0, t = s_n$, and $(s_{i-1}, a_i) R s_i$ for each $i \in \{1, \dots, n\}$.

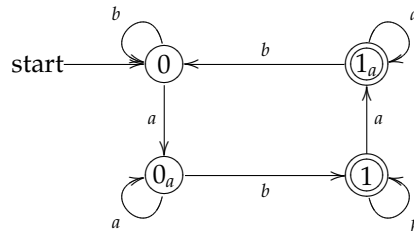
An automaton M *accepts* a word $w \in A^*$ if and only if $(I, w) \tilde{R} t$ for some accepting state $t \in T$. We define the *language of M* as

$$\mathcal{L}(M) = \{w \in A^* \mid M \text{ accepts } w\}.$$

2. EXAMPLES. It is typical to give individual examples of automata by drawing their *transition diagrams*: these are directed graphs (meaning each edge in such a graph has a direction that is indicated by an arrow) where nodes represent states. One draws an arrow labelled by a letter $a \in A$ from the state s to the state t if and only if $(s, a) R t$. The initial state is indicated by having an (unlabelled) arrow (or an arrow marked by the word 'start') pointing to it. The accepting states are indicated by drawing their nodes with a double circle.

Here is a (deterministic) automaton M with

$$\mathcal{L}(M) = \{w \in \{a, b\}^* \mid w \text{ has an odd number of occurrences of the subword } ab\} :$$



3. DEFINITION. Let $L_1, L_2 \subseteq A^*$ be languages over the alphabet A . The *union*, *intersection*, and *complement* (relative to A^*) are merely the set-theoretical operations:

$$L_1 \cup L_2 = \{w \in A^* \mid w \in L_1 \text{ or } w \in L_2\}, \quad L_1 \cap L_2 = \{w \in A^* \mid w \in L_1 \text{ and } w \in L_2\}, \quad \text{and}$$

¹The textbook refers to these as *finite-state machines*.

$$\bar{L} = \{w \in A^* \mid w \notin L\}.$$

Further, we define $L_1 \circ L_2$, the *concatenation* of two languages by

$$L_1 \circ L_2 = \{w_1w_2 \in A^* \mid w_1 \in L_1 \text{ and } w_2 \in L_2\},$$

so that $L_1 \circ L_2$ consists of those words over A that can be split into two parts, the prefix coming from L_1 , and the suffix from L_2 .

L^* , the *Kleene star* of a language L , is defined as

$$L^* = \{w_1w_2 \cdots w_n \in A^* \mid n \in \mathbb{N} \text{ and } w_i \in L \text{ for all } 1 \leq i \leq n\}.$$

Thus L^* arises from L in much the same way as A^* , the set of all words over A , arises from A . In particular, putting $n = 0$ in the above definition, we see that $\varepsilon \in L^*$ no matter what language L is.

Caution: The binary operations \cup and \circ are in most cases meant to be only applied to languages over the same alphabet.

4. DEFINITION. *Regular expressions* (over A) are thought of as a kind of formulas (or strings of symbols) and are defined as follows:

- \emptyset is a regular expression;
- ε is a regular expression;
- a_i is a regular expression for each $a_i \in A$;
- if E_1 and E_2 are regular expressions then $(E_1 + E_2)$, (E_1E_2) , and E_1^* are regular expressions.

The symbols \emptyset and ε above are presumed to lie outside the alphabet A . This is an example of an inductive definition, which means that nothing other than what is mandated by the four clauses above is a regular expression. We shall omit some of the parentheses in individual regular expressions whenever that creates no problems for readability.

Each regular expression E (over A) is assigned a language $\mathcal{L}(E) \subseteq A^*$ as follows:

- $\mathcal{L}(\emptyset) = \emptyset$, the empty language which contains no words at all;
- $\mathcal{L}(\varepsilon) = \{\varepsilon\}$, the language which only contains the empty word and no other words;
- $\mathcal{L}(a_i) = \{a_i\}$ the language which only contains the one-letter word $\{a_i\}$ and nothing else;
- $\mathcal{L}(E_1 + E_2) = \mathcal{L}(E_1) \cup \mathcal{L}(E_2)$;
- $\mathcal{L}(E_1E_2) = \mathcal{L}(E_1) \circ \mathcal{L}(E_2)$;
- $\mathcal{L}(E_1^*) = \mathcal{L}(E_1)^*$.

Since the operations of union and concatenation are clearly associative, this allows us to further cut down on the number of parentheses in regular expressions.

5. EXAMPLES.

$$\mathcal{L}(a^*b^*) = \{a^n b^m \in \{a, b\}^* \mid n, m \in \mathbb{N}\}$$

$$\mathcal{L}(abc) = \{abc\}, \quad \mathcal{L}(a + b + c) = \{a, b, c\}$$

$$\mathcal{L}(a(\varepsilon + b)a) = \{aa, aba\}, \quad \mathcal{L}(\emptyset a^*) = \emptyset$$

$$\mathcal{L}((b^*ab^*ab^*)^*) = \{w \in \{a, b\}^* \mid w \text{ has an even number of occurrences of the symbol } a\}$$

6. DEFINITION. A (*context-free*) *grammar* G is a quintuple of the form (V, A, S, P) , where

- V is a finite alphabet (of *non-terminal* symbols or *variables*. We typically use upper case letters to denote these);
- A is a finite alphabet (of *terminal* symbols which we tend to write in lower case) such that $V \cap A = \emptyset$;
- $S \in V$ is a distinguished non-terminal symbol called the *start* symbol;
- $P \subseteq V \times (V \cup A)^*$ is a finite set of *productions*². We write productions in the form $X \rightarrow w$ ($X \in V, w \in (V \cup A)^*$).

²The textbook allows a more liberal form of productions than we do.

A grammar is *regular*³ if $P \subseteq \{\varepsilon\} \cup A \circ V$, that is only the empty word and the two-symbol words consisting of a terminal letter followed by a non-terminal one are allowed into the right hand side of productions.

Assume a word $w \in (V \cup A)^*$ has the form vXu where $X \in V, v, u \in (V \cup A)^*$, and that $X \rightarrow t$ is a production of G . Then the word vtu can be obtained from vXu by an application of the production $X \rightarrow t$. One writes $w \xrightarrow{G} vtu$. A sequence of words w_0, w_1, \dots, w_n is a *derivation* in G if $w_0 = S$ (the word consisting of the single start symbol), and for each $i \in \{0, \dots, n-1\}$ the word w_{i+1} can be obtained from w_i by one of the productions in P . In this case we say that G *derives* w_n . We define the *language of* G as

$$\mathcal{L}(G) = \{w \in A^* \mid G \text{ derives } w\}.$$

7. EXAMPLES. Consider the context-free grammar G with $V = \{S\}, A = \{a, b\}, P = \{S \rightarrow aSb, S \rightarrow \varepsilon\}$, and the starting symbol S . Then the following sequence is a derivation in G of the word $aabb$:

$$S, \quad aSb, \quad aaSbb, \quad aabb.$$

In the first two steps one uses the first production, and the second production in the last step. We have $\mathcal{L}(G) = \{a^n b^n \mid n \in \mathbb{N}\}$.

Consider the regular grammar G with $V = \{S, T\}, A = \{a, b\}$,

$$P = \{S \rightarrow aS, S \rightarrow bT, S \rightarrow \varepsilon, T \rightarrow aT, T \rightarrow bS\},$$

and the starting symbol S . Then the following sequence is a derivation in G of the word $aabab$:

$$S, \quad aS, \quad aaS, \quad aabT, \quad aabaT, \quad aababS, \quad aabab.$$

We have $\mathcal{L}(G) = \{w \in \{a, b\}^* \mid w \text{ has an even number of occurrences of } a\}$.

8. THEOREM. Let $L \subseteq A^*$ be a language over a finite alphabet A . The following are then equivalent:

- (i) There exists a deterministic automaton M such that $\mathcal{L}(M) = L$;
- (ii) There exists a regular expression E such that $\mathcal{L}(E) = L$;
- (iii) There exists a regular grammar G such that $\mathcal{L}(G) = L$;
- (iv) There exists a non-deterministic automaton N such that $\mathcal{L}(N) = L$. ■

9. DEFINITION. A language $L \subseteq A^*$ is *regular* if it satisfies any of the conditions (i)–(iv) of Theorem 8.

10. PROPOSITION. If $L_1, L_2 \subseteq A^*$ are regular languages then so are $L_1 \cup L_2, L_1 \cap L_2, \bar{L}_1, L_1 \circ L_2$, and L_1^* . ■

³Called *right-regular* in the textbook.