Лекция 9

Теория формальных языков (I)

(Конспект: А. Ю. Волков)

Эта лекция почти целиком состоит из определений.

9.1 Формальные языки

Будем называть *алфавитом* произвольное конечное множество (например, $\{0, 1\}$ — алфавит). Строкой в алфавите Σ будет называться конечная последовательность его элементов (пустой строкой, обозначаемой ϵ, будет называться последовательность из нуля элементов).

Языком в алфавите Σ называется множество некоторых строк в алфавите Σ. Например: $\{\epsilon, 1, 00, 01\}$. Или: $\{\underbrace{0\ldots0}_{n}\underbrace{1\ldots1}_{n} | n \in \mathbb{N}\}$.

Далее определим *регулярные выражения* в алфавите Σ. Они будут определяться индуктивно:

- $\emptyset$ — регулярное выражение;
- $\{\epsilon\}$ — регулярное выражение;
- $\{a\}$ — регулярное выражение (для каждого $a \in \Sigma$);
- Если A, B — регулярные выражения, то $A \cup B$ — тоже регулярное выражение;
- Если A, B — регулярные выражения, то $A \cdot B$ — тоже регулярное выражение;
- Если A — регулярное выражение, то A^* — тоже регулярное выражение.

Это определение исчерпывает все возможные регулярные выражения.

Каждое регулярное выражение определяет некоторый язык. Для большинства пунктов определения очевидно, какой язык имеется в виду; оставшиеся пункты:

- для данных языков A и B язык $A \cdot B$ состоит из строк вида ab, где $a \in A$, $b \in B$ (значок операции "точка" — конкатенации строк — часто опускается);
- $A^* = \{\epsilon\} \cup A \cup A \cdot A \cup A \cdot A \cdot A \ldots$ (все конечные $A \cdot \ldots \cdot A$).

Например, $(\{0\} \cup \{11\})^* \cdot \{000\}$ обозначает множество всех последовательностей нулей и пар единиц, заканчивающихся на три нуля.

9.2 Формальные грамматики

Формальной *грамматикой* называется упорядоченная четверка (N, Σ, S, R), где

- N — множество *нетерминальных* символов (будут обозначаться заглавными латинскими буквами),
- Σ — множество *терминальных* символов (будут обозначаться маленькими латинскими буквами и цифрами),
- S — стартовый символ ($S \in N$),
- $R \subseteq (N \cup \Sigma)^* N (N \cup \Sigma)^* \times (N \cup \Sigma)^*$ — множество правил (вида $\alpha \to \beta$, где α, β — произвольные строки терминалов и нетерминалов, но α содержит хотя бы один нетерминал).

Пример 9.1. $N = \{S, A\}$, $\Sigma = \{0, 1\}$. Правила: $S \to 0S1$, $0S \to 1A$, $A \to \epsilon$. □

Для данной грамматики введем отношение *выводимости*:

- $\gamma \Rightarrow \delta$, если $\gamma = \gamma_1 \alpha \gamma_2$, $\delta = \gamma_1 \beta \gamma_2$ и существует правило $\alpha \to \beta$ из R;
- $\Rightarrow^*$ — транзитивно-рефлексивное замыкание $\Rightarrow$.

Пример 9.2. Продолжая пример 9.1: $S \Rightarrow 0S1 \Rightarrow 00S11 \Rightarrow 01A11 \Rightarrow 0111$; $S \Rightarrow^* 01A11$, $S \Rightarrow^* 0111$, $S \Rightarrow^* S$. □

Строка α, для которой $S \Rightarrow^* \alpha$, называется *выводимой* в данной грамматике. *Язык $L(G)$, порождаемый грамматикой G*, состоит из всех строк терминалов, выводимых в ней, т.е. $L(G) = \{\alpha \in \Sigma^* | S \Rightarrow^* \alpha\}$.

Пример 9.3. В примере 9.2 указано три выводимых строки, но лишь одна из них (0111) попадает в язык, порождаемый рассматриваемой грамматикой (туда еще много строк попадает: он бесконечный). □

Контекстно-зависимой (*неукорачивающей*) грамматикой называется грамматика, для каждого правила которой правая часть — не короче левой.

Грамматика называется *бесконтекстной* (*контекстно-свободной*), если все ее правила имеют вид $A \to \alpha$ $(A \in N)$ (т.е. у каждого правила слева — ровно один символ).

Грамматика называется *праволинейной*, если каждое ее правило имеет вид $A \to \alpha$, $A \to \alpha B$ (где $\alpha \in \Sigma^*$; $A, B \in N$). (Замечание: не забудем, что $\epsilon \in \Sigma^*$.)

Две грамматики называются *эквивалентными*, если они порождают одинаковые языки.

Лемма 9.1. *Для любой праволинейной грамматики G_1, такой что $\epsilon \notin L(G_1)$, существует грамматика G_2, эквивалентная G_1, в которой имеются только правила вида $A \to a$ и $A \to aB$ (где $a \in \Sigma$; $A, B \in N$).*

Доказательство. Введем обозначение $\to^*$ для транзитивно-рефлексивного замыкания отношения $\to$. Пусть $S(A) = \{B \neq A \mid B \to^* A\}$; $S(\epsilon) = \{A \mid A \to^* \epsilon\}$.

Будем заменять правила G_1:

- для каждого правила $A \to \gamma$, где $\gamma \notin N$, и для каждого $B \in S(A)$, добавим правило $B \to \gamma$;
- для каждого правила $A \to \gamma B$, где $\gamma \neq \epsilon$, и для каждого $B \in S(\epsilon)$, добавим правило $A \to \gamma$;
- выкинем все правила вида $A \to B$, $A \to \epsilon$ (где $A, B \in N$).

Далее, если осталось правило $A \to ab\gamma$ (где $a, b \in \Sigma$), заменим это правило на два правила: $A \to aZ$, $Z \to b\gamma$, введя новый нетерминал Z. Так, постепенно укорачивая правила, мы заменим все правила на правила вида, требуемого в условии леммы.

Легко убедиться в том, что полученная грамматика эквивалентна исходной. □

9.3 Конечные автоматы

Недетерминированный конечный автомат — упорядоченная пятерка $(Q, \Sigma, q_S, F, \delta)$, где

- Q — конечное множество состояний,
- Σ — алфавит,
- $q_S \in Q$ — стартовое состояние,
- $F \subseteq Q$ — множество конечных состояний,
- $\delta : Q \times (\Sigma \cup \{\epsilon\}) \to 2^Q$ — функция перехода.

Обозначение: будем писать $q_1 \xrightarrow{a} q_2$, если $q_2 \in \delta(q_1, a)$ (здесь $a \in \Sigma \cup \{\epsilon\}$).

Недетерминированному конечному автомату дают строку в алфавите Σ; он, по очереди считывая ее символы, переходит из своего текущего состояния q (начинает он с $q = q_S$) в одно из состояний из множества $\delta(q, a)$ (считав символ $a \in \Sigma$) или одно из состояний из множества $\delta(q, \epsilon)$ (не считав ничего — такой переход называется *ϵ-переходом*).

Если таким образом он <u>может</u>[1] попасть в одно из состояний из F (полностью считав данную ему строку и, быть может, сделав несколько ϵ-переходов), то говорят, что он *принимает* данную строку. Множество всех строк в алфавите Σ, принимаемых данным автоматом $\mathcal{A}$, называется *языком, принимаемым этим автоматом*, и обозначается $L(\mathcal{A})$.

(Продолжение секции 9.3 следует...)

[1]На каждом шаге у него может быть несколько возможностей; нас интересует "наилучший вариант".