Автомат Глушкова

Определение 1. *Если регулярное выражение $r \in \mathcal{RE}$ содержит n вхождений букв алфавита Σ, тогда линеаризованное регулярное выражение* $\texttt{Linearize}(r)$ *получается из r приписыванием i-ой по счёту букве, входящей в r, индекса i.*

Пример 1. *Рассмотрим регулярное выражение:*

$$(\texttt{ba} \mid \texttt{b})\texttt{aa}(\texttt{a} \mid \texttt{ab})^*$$

Его линеаризованная версия:

$$(\texttt{b}_1\texttt{a}_2 \mid \texttt{b}_3)\texttt{a}_4\texttt{a}_5(\texttt{a}_6 \mid \texttt{a}_7\texttt{b}_8)^*$$

Определение 2. *Пусть $r \in \mathcal{RE}$, тогда:*

- *множество* First *— это множество букв, с которых может начинаться слово из $\mathscr{L}(r)$ (если $\varepsilon \in \mathscr{L}(r)$, то оно формально добавляется в* First*);*
- *множество* Last *— это множество букв, которыми может заканчиваться слово из $\mathscr{L}(r)$;*
- *множество* Follow(c) *— это множество букв, которым может предшествовать c. Т.е. $\{d \in \Sigma \mid \exists w_1, w_2(w_1 c d w_2 \in \mathscr{L}(r))\}$.*

Пример 2. *Множество* Follow *в теории компиляции обычно определяется иначе — это множество символов, которые могут идти за выводом из определённого нетерминального символа. Два этих определения можно унифицировать, если рассматривать каждую букву в r как «обёрнутую» (в смысле, например, н.ф. Хомского).*

Построим указанные множества для регулярного выражения $r = (\texttt{ba} \mid \texttt{b})\texttt{aa}(\texttt{a} \mid \texttt{ab})^*$. Начнём с исходного регулярного выражения.

Исходное регулярное выражение

- $\text{First}(r) = \{\texttt{b}\}$.
- $\text{Last}(r) = \{\texttt{a}, \texttt{b}\}$.
- $\text{Follow}_r(\texttt{a}) = \{\texttt{a}, \texttt{b}\}$; $\text{Follow}_r(\texttt{b}) = \{\texttt{a}\}$.

Хотя данные множества описывают, как устроены слова из $\mathscr{L}(r)$ локально, однако они не исчерпывают всей информации о языке, поскольку разные вхождения букв в регулярное выражения никак не различаются.

Например, по множествам First и Last можно предположить, что $\texttt{b} \in \mathscr{L}(r)$, хотя это не так.

Вспомним, что $r_{\text{Lin}} = (\texttt{b}_1\texttt{a}_2 \mid \texttt{b}_3)\texttt{a}_4\texttt{a}_5(\texttt{a}_6 \mid \texttt{a}_7\texttt{b}_8)^*$.

Линеаризованное выражение

- $\mathrm{First}(r_{\mathrm{Lin}}) = \{\mathtt{b}_1, \mathtt{b}_3\}$.
- $\mathrm{Last}(r_{\mathrm{Lin}}) = \{\mathtt{a}_5, \mathtt{a}_6, \mathtt{b}_8\}$.
- $\mathrm{Follow}_{r_{\mathrm{Lin}}}(\mathtt{b}_1) = \{\mathtt{a}_2\}$; $\mathrm{Follow}_{r_{\mathrm{Lin}}}(\mathtt{a}_2) = \{\mathtt{a}_4\}$; $\mathrm{Follow}_{r_{\mathrm{Lin}}}(\mathtt{b}_3) = \{\mathtt{a}_4\}$; $\mathrm{Follow}_{r_{\mathrm{Lin}}}(\mathtt{a}_4) = \{\mathtt{a}_5\}$; $\mathrm{Follow}_{r_{\mathrm{Lin}}}(\mathtt{a}_5) = \{\mathtt{a}_6, \mathtt{a}_7\}$; $\mathrm{Follow}_{r_{\mathrm{Lin}}}(\mathtt{a}_6) = \{\mathtt{a}_6, \mathtt{a}_7\}$; $\mathrm{Follow}_{r_{\mathrm{Lin}}}(\mathtt{a}_7) = \{\mathtt{b}_8\}$; $\mathrm{Follow}_{r_{\mathrm{Lin}}}(\mathtt{b}_8) = \{\mathtt{a}_6, \mathtt{a}_7\}$.

В описании данных множеств содержится исчерпывающая информация о языке $\mathscr{L}(r_{\mathrm{Lin}})$.

Построение 1.

- *Строим линеаризованную версию r:* $r_{\mathrm{Lin}} = \mathtt{Linearize}(r)$.
- *Находим* $\mathrm{First}(r_{\mathrm{Lin}})$, $\mathrm{Last}(r_{\mathrm{Lin}})$, *а также* $\mathrm{Follow}_{r_{\mathrm{Lin}}}(c)$ *для всех* $c \in \Sigma_{r_{\mathrm{Lin}}}$.
- *Все состояния автомата, кроме начального (назовём его S), соответствуют буквам* $c \in \Sigma_{r_{\mathrm{Lin}}}$.
- *Из начального состояния строим переходы в те состояния, для которых* $c \in \mathrm{First}(r_{\mathrm{Lin}})$. *Переходы имеют вид* $S \xrightarrow{c} c$.
- *Переходы из состояния c соответствуют элементам d множества* $\mathrm{Follow}_{r_{\mathrm{Lin}}}(c)$ *и имеют вид* $c \xrightarrow{d} d$.
- *Конечные состояния — такие, что* $c \in \mathrm{Last}(r_{\mathrm{Lin}})$, *а также S, если* $\varepsilon \in \mathscr{L}(R)$.
- *Теперь стираем разметку, построенную линеаризацией, на переходах автомата. Конструкция завершена.*

Исходное регулярное выражение:

$$(\mathtt{ba} \mid \mathtt{b})\mathtt{aa}(\mathtt{a} \mid \mathtt{ab})^*$$

Линеаризованное регулярное выражение:

$$(\mathtt{b}_1\mathtt{a}_2 \mid \mathtt{b}_3)\mathtt{a}_4\mathtt{a}_5(\mathtt{a}_6 \mid \mathtt{a}_7\mathtt{b}_8)^*$$

Автомат Глушкова:

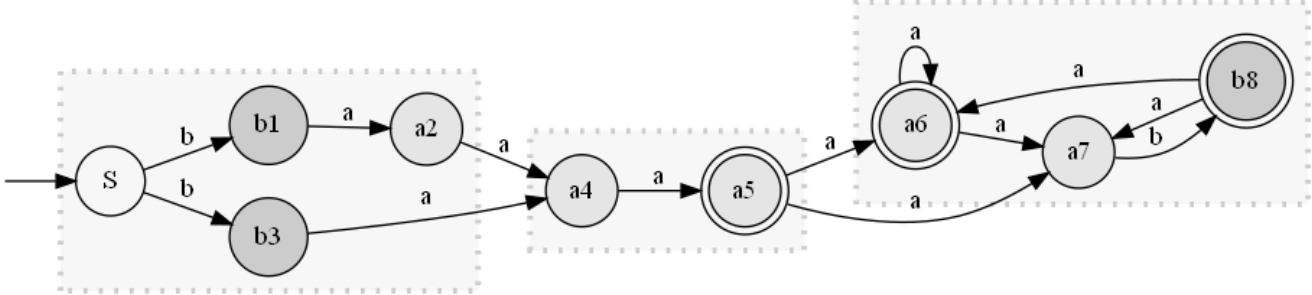

Рисунок 1 — Пример автомата Глушкова

Подграфы, распознающие регулярные выражения, являющиеся подструктурами исходного, не имеют общих вершин. Это свойство автомата Глушкова используется в реализациях `match`-функций некоторых библиотек регулярных выражений.

Свойства автомата Глушкова

- Не содержит ε-переходов.
- Число состояний равно длине регулярного выражения (без учёта регулярных операций), плюс один (стартовое состояние).
- В общем случае недетерминированный.

Пример 3. *Для* 1*-однозначных регулярных выражений* r *автомат* $\texttt{Glushkov}(r)$ *является детерминированным. Эту его особенность активно используют в современных библиотеках регулярных выражений, например, в* RE2*. Выигрыш может получиться колоссальным: например,* $\texttt{Thompson}((a^*)^*)$ *является экспоненциально неоднозначным, а* $\texttt{Glushkov}((a^*)^*)$ *однозначен и детерминирован!*