Логи

Лучшая команда разработчиков по ТФЯ

2022 г.

Линеаризация

Определение

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

Пример

Рассмотрим регулярное выражение:

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

Его линеаризованная версия:

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

Множества First, Last, Follow

Определение

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 $\mathsf{Follow}(c)$ — это множество букв, которым может предшествовать c. Т.е. $\{d \in \Sigma \mid \exists w_1, w_2 (w_1 c d w_2 \in \mathscr{L}(r))\}$.

First, Last, Follow — пример

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

Исходное регулярное выражение

- $\mathsf{First}(\mathtt{r}) = \{\mathtt{b}\}$.
- $\mathsf{Last}(\mathtt{r}) = \{\mathtt{a}, \mathtt{b}\}$.
- $\mathsf{Follow}_{\mathtt{r}}(\mathtt{a}) = \{\mathtt{a}, \mathtt{b}\}$; $\mathsf{Follow}_{\mathtt{r}}(\mathtt{b}) = \{\mathtt{a}\}$.

Хотя данные множества описывают, как устроены слова из $\mathscr{L}(\mathtt{r})$ 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 $\mathtt{b} \in \mathscr{L}(\mathtt{r})$, хотя это не так.

First, Last, Follow — пример

Построим указанные множества для регулярного выражения
$\mathtt{r} = (\mathtt{ba} \mid \mathtt{b})\mathtt{aa}(\mathtt{a} \mid \mathtt{ab})^*$.
Вспомним, что $\mathtt{r}_{\mathsf{Lin}} = (\mathtt{b}_1\mathtt{a}_2 \mid \mathtt{b}_3)\mathtt{a}_4\mathtt{a}_5(\mathtt{a}_6 \mid \mathtt{a}_7\mathtt{b}_8)^*$.

Линеаризованное выражение

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

В описании данных множеств содержится исчерпывающая информация о языке $\mathscr{L}(\mathtt{r}_{\mathsf{Lin}})$.

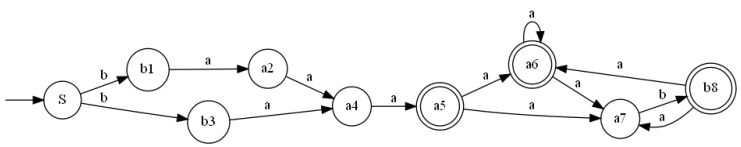
Конструкция автомата Глушкова

Алгоритм построения $\texttt{Glushkov}(\mathsf{r})$

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

Пример автомата Глушкова

Исходное регулярное выражение:
`(ba | b)aa(a | ab)*`
Линеаризованное регулярное выражение: `(ba | b)aa(a | ab)*`
Автомат 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.