Автомат Томпсона

Лучшая команда разработчиков по ТФЯ

2022 г.

Адгоритм Томпсона и НКА

Основные сведения

В информатике алгоритм построения Томпсона представляет собой метод преобразования регулярного выражения в эквивалентный недетерминированный конечный автомат (НКА). Этот НКА можно использовать для сопоставления строк с регулярным выражением. Регулярные выражения и недетерминированные конечные автоматы - это два представления формальных языков.

Недетерминированные КА

Определение

Недетерминированный конечный автомат (НКА) – это детерминированный конечный автомат (ДКА), который не выполняет следующие условия:

- любой его переход единственным образом определяется по текущему состоянию и входному символу;
- чтение входного символа требуется для каждого изменения состояния.

Недетерминированные КА

Определение

Недетерминированный конечный автомат (NFA) — это пятёрка $\mathscr{A} = \langle Q, \Sigma, q_0, F, \delta \rangle$, где:

- Q — множество состояний;
- Σ — алфавит терминалов;
- δ — множество правил перехода вида $\langle q_i, (a_i|\varepsilon), M_i \rangle$, где $q_i \in Q$, $a_i \in \Sigma$, $M_i \in 2^Q$;
- $q_0 \in Q$ — начальное состояние;
- $F \subseteq Q$ — множество конечных состояний.

Сокращаем: $\langle q_1, a, q_2 \rangle \in \delta \Leftrightarrow \langle q_1, a, M \rangle \in \delta \ \& \ q_2 \in M$.

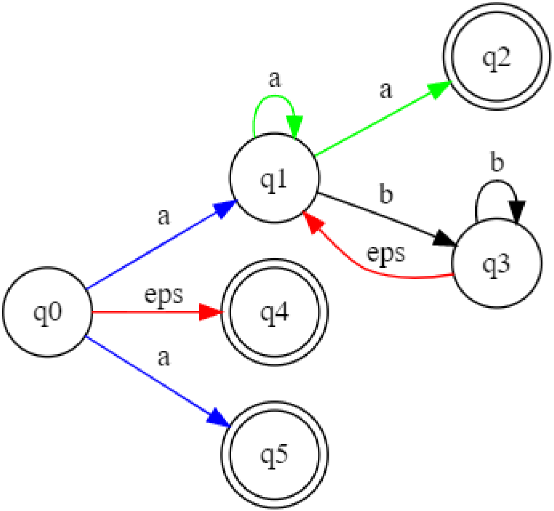
Недетерминированные КА

- $q \xrightarrow{\varepsilon} q' \Leftrightarrow (q = q') \vee \exists p_1, \dots, p_k(\langle q, \varepsilon, p_1\rangle \in \delta \;\&\; \langle p_k, \varepsilon, q'\rangle \in \delta \;\&\; \forall i, 1 \leqslant i < k \langle p_i, \varepsilon, p_{i+1}\rangle \in \delta)$.
- $q \xrightarrow{a} q' \Leftrightarrow \exists p, p'(q \xrightarrow{\varepsilon} p \;\&\; \langle p, a, p'\rangle \in \delta \;\&\; p' \xrightarrow{\varepsilon} q')$.
- $q \xrightarrow{a_1 \dots a_k} q' \Leftrightarrow \exists p_1, \dots, p_{k-1}(q \xrightarrow{a_1} p_1 \;\&\; p_{k-1} \xrightarrow{a_k} q' \;\&\; \forall i, 1 \leqslant i < k-1(p_i \xrightarrow{a_{i+1}} p_{i+1}))$.

Определение

Язык $\mathscr{L}$, распознаваемый НКА $\mathscr{A}$ — это множество слов $\left\{w \mid \exists q \in F(q_0 \xrightarrow{w} q)\right\}$.

Пример НКА



Конструкция автомата Томпсона

Алгоритм построения $\mathtt{Thompson}(r)$

Алгоритм работает рекурсивно, разбивая выражение на составляющие его подвыражения, из которых будет построен НКА с использованием набора правил. Точнее, из регулярного выражения r полученный автомат A с переходной функцией δ учитывает следующие свойства:

- A имеет ровно одно начальное состояние q_0, которое недоступно ни из какого другого состояния. То есть для любого состояния q и любой буквы a $\delta(q, a)$ не содержит q_0.
- A имеет ровно одно конечное состояние q_f, которое недоступно ни из какого другого состояния. То есть для любой буквы a, $\delta(q_f, a) = \emptyset$.
- Пусть c - число конкатенаций регулярного выражения r, а s — количество символов, не считая круглых скобок, то есть $|,^*$, a, ε. Тогда число состояний A равно $2s - c$ (линейно по размеру r).

Конструкция автомата Томпсона

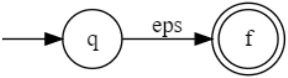
Алгоритм построения $\mathtt{Thompson}(r)$

Алгоритм работает рекурсивно, разбивая выражение на составляющие его подвыражения, из которых будет построен НКА с использованием набора правил. Точнее, из регулярного выражения r полученный автомат A с переходной функцией δ учитывает следующие свойства:

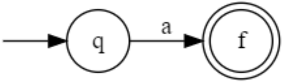
- Число переходов, выходящих из любого состояния, не более двух.
- Поскольку НКА из m состояний и не более e переходов из каждого состояния может соответствовать строке длиной n за время $O(emn)$, НКА Томпсона может выполнять сопоставление с образцом за линейное время, предполагая алфавит фиксированного размера.

Правила

$N(s)$ и $N(t)$ являются NFA подвыражений s и t соответственно. Пустое выражение ε преобразуется в

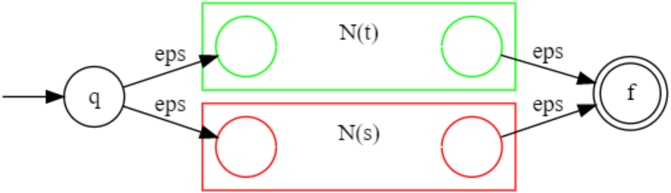


Символ a входного алфавита преобразуется в



Правила

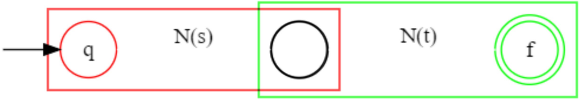
$\mathrm{N}(s)$ и $\mathrm{N}(t)$ являются NFA подвыражений s и t соответственно. Выражение объединения $s \mid t$ преобразуется в



Состояние q переходит через ε либо в начальное состояние $\mathrm{N}(s)$, либо $\mathrm{N}(t)$. Их конечные состояния становятся промежуточными состояниями всего НКА и сливаются через два ε-перехода в конечное состояние НКА.

Правила

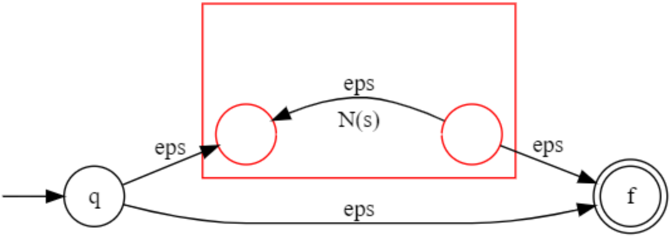
$N(s)$ и $N(t)$ являются NFA подвыражений s и t соответственно.
Выражение конкатенации st преобразуется в



Начальное состояние $N(s)$ является начальным состоянием всего НКА.
Конечное состояние $N(s)$ становится начальным состоянием $N(t)$.
Конечное состояние $N(t)$ является конечным состоянием всего НКА.

Правила

$N(s)$ и $N(t)$ являются NFA подвыражений s и t соответственно.
Выражение Клини Стар s^* преобразуется в



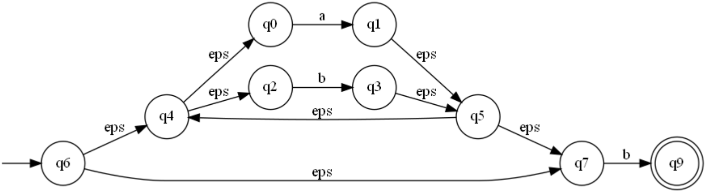
ε-переход соединяет начальное и конечное состояние НКА с промежуточным НКА $N(s)$. Другой ε-переход от внутреннего конечного к внутреннему начальному состоянию $N(s)$ допускает повторение выражения s в соответствии с оператором *.
Заключенное в скобки выражение (выражения) преобразуется в само $N(s)$.

Пример автомата Томпсона

Исходное регулярное выражение:

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

Автомат Томпсона:



Свойства автомата Томпсона

- Единственное начальное состояние
- Единственное конечное состояние
- Не больше двух переходов из каждого состояния