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.

НКА

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.

НКА

Определение

НКА формально представляется как 5-кортеж $(Q, \Sigma, \Delta, q_0, F)$, состоящий из:

- конечного множества состояний Q.
- конечного множества входных символов Σ.
- функции переходов $\Delta : Q \times \Sigma \to P(Q)$.
- начального состояния $q_0 \in Q$.
- множества состояний F распознаваемых как конечные состояния $F \subseteq Q$.

Здесь $P(Q)$ означает степень множества Q.

Конструкция автомата Томпсона

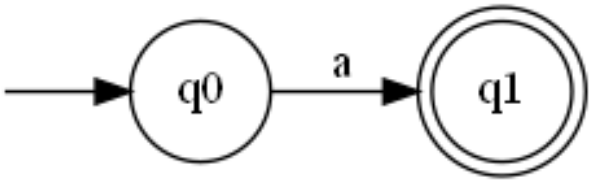
Алгоритм построения `Thompson(r)`

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:

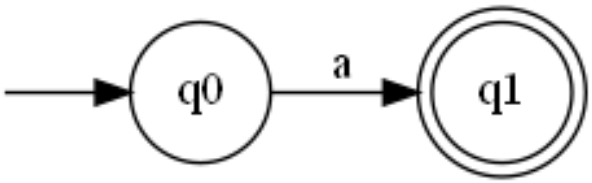
- A имеет ровно одно начальное состояние q_0, которое недоступно ни из какого другого состояния. То есть для любого состояния q и любой буквы a $\Delta(q, a)$ не содержит q_0.
- A имеет ровно одно конечное состояние q_f, которое недоступно ни из какого другого состояния. То есть для любой буквы , $\Delta(q_f, a) = \emptyset$.
- 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).
- Число переходов, выходящих из любого состояния, не более двух.
- Поскольку НКА из m состояний и не более e переходов из каждого

Правила

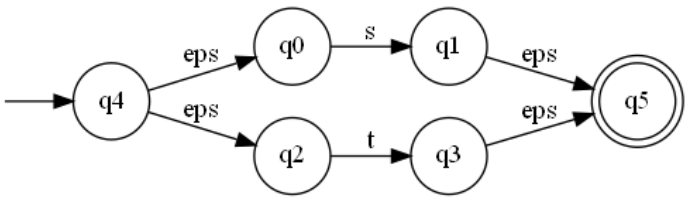
Пустое выражение ε преобразуется в



Символ a входного алфавита преобразуется в



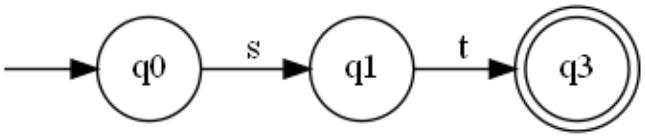
Выражение объединения $s \mid t$ преобразуется в



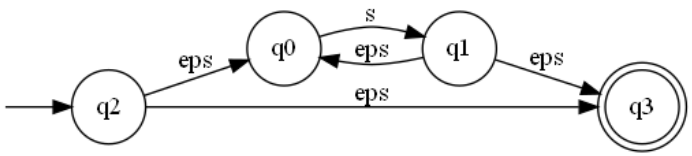
Состояние q_4 переходит через ε либо в начальное состояние $N(s)$, либо $N(t)$. Их конечные состояния становятся промежуточными состояниями всего НКА и сливаются через два ε-перехода в конечное состояние НКА.

Правила

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.
Выражение Клини Стар s* преобразуется в



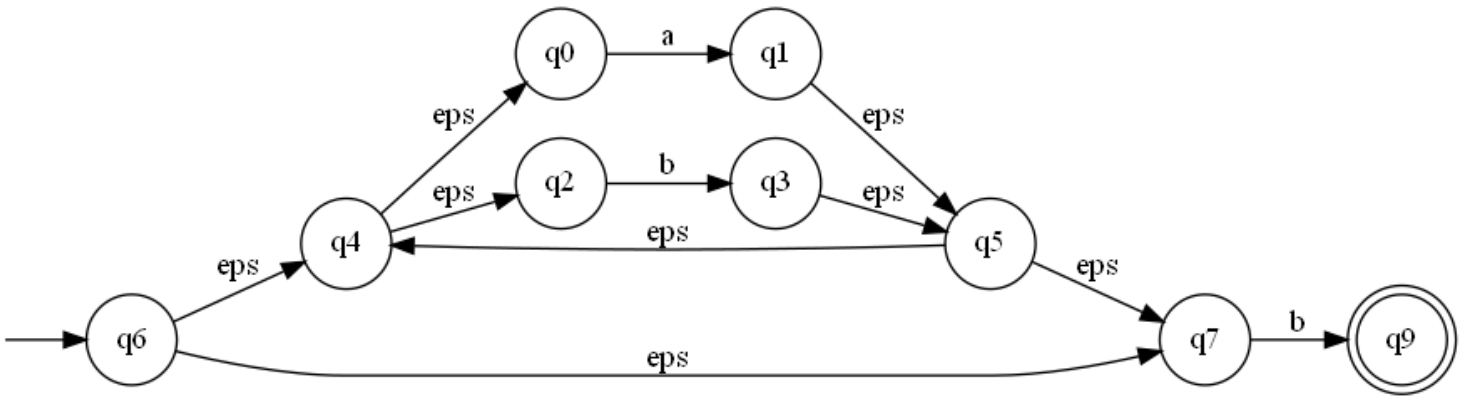
ε-переход соединяет начальное и конечное состояние НКА с промежуточным НКА N(s). Другой ε-переход от внутреннего конечного к внутреннему начальному состоянию N(s) допускает повторение выражения s в соответствии с оператором *.
Заключенное в скобки выражение (выражения) преобразуется в само N(s).

Пример автомата Томпсона

Исходное регулярное выражение:

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

Автомат Томпсона:



Свойства автомата Томпсона

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