Automaty a gramatiky (BI-AAG)

4. Regulární výrazy

Jan Holub

Katedra teoretické informatiky
Fakulta informačních technologií
ČVUT v Praze

Regulární výrazy

Definice

Regulární výraz V nad abecedou Σ je definován takto:

1. $\emptyset, \varepsilon, a$ jsou regulární výrazy pro všechna $a \in \Sigma$.
2. Jsou-li x, y regulární výrazy nad Σ, pak:

   (a) $(x + y)$ (sjednocení, alternativa),

   (b) $(x.y)$ (zřetězení),

   (c) $(x)^*$ (iterace)

   jsou regulární výrazy nad Σ. $\square$

Regulární výrazy

Definice

Hodnota $h(x)$ *regulárního výrazu* x je definována takto:

1. $h(\emptyset) = \emptyset, h(\varepsilon) = \{\varepsilon\}, h(a) = \{a\},$
2. $h(x + y) = h(x) \cup h(y),$
   $h(x.y) = h(x).h(y),$
   $h(x^*) = (h(x))^*.$ $\square$

Regulární výrazy — Axiomy

$A_1 : x + (y + z) = (x + y) + z$ (asociativnost sjednocení),
$A_2 : x + y = y + x$ (komutativnost sjednocení),
$A_3 : x + \emptyset = x$ ($\emptyset$ je nulový prvek pro sjednocení),
$A_4 : x + x = x$ (idempotence sjednocení),
$A_5 : x.(y.z) = (x.y).z$ (asociativnost zřetězení),
$A_6 : \varepsilon x = x\varepsilon = x$ (ε je jednotkový prvek pro zřetězení)
$A_7 : \emptyset x = x\emptyset = \emptyset$ ($\emptyset$ je nulový prvek pro zřetězení),
$A_8 : x.(y + z) = x.y + x.z$ (distributivnost zleva),
$A_9 : (x + y).z = x.z + y.z$ (distributivnost zprava),
$A_{10} : x^* = \varepsilon + x^*x$
$A_{11} : x^* = (\varepsilon + x)^*$
$A_{12} : x = x\alpha + \beta \Rightarrow x = \beta\alpha^*$ (řešení levé regulární rovnice),
$A_{13} : x = \alpha x + \beta \Rightarrow x = \alpha^*\beta$ (řešení pravé regulární rovnice).

Regulární rovnice

Příklad

$L = \{$ sudý počet jedniček následovaný příponou 010 $\}$.

$$(R)\;\; x = 11x + 010.$$

řešení: $x = (11)^*010$

$(11)^*010 = 11(11)^*010 + 010$ ($x = \alpha x + \beta \Rightarrow x = \alpha^*\beta$)

$(11)^*010 = (11(11)^* + \varepsilon)010$ ($xy + y = (x + \varepsilon)y$)

$(11)^*010 = (11)^*010$ ($xx^* + \varepsilon = x^*$)

Regulární rovnice

Definice

Standardní soustava regulárních rovnic má tvar:

$X_i = \alpha_{i0} + \alpha_{i1}X_1 + \alpha_{i2}X_2 + ... + \alpha_{in}X_n, 1 \leq i \leq n$, kde $X_1, X_2, ..., X_n$ jsou neznámé a α_{ij} jsou regulární výrazy nad abecedou Σ, která neobsahuje $X_1, X_2, ..., X_n$. $\square$

Regulární rovnice

Příklad

$$A = 1A + 1B$$
$$B = 0A + 0B + 0.$$

$A = 1^*1B$
$B = 01^*1B + 0B + 0.$
$B = (01^*1 + 0)B + 0$
$B = (01^*1 + 0)^*0 = (0(1^*1 + \varepsilon))^*0 = (01^*)^*0.$

Řešení:

$$A = 1^*1(01^*)^*0$$
$$B = (01^*)^*0.$$

Derivace regulárních výrazů

Definice

Derivace $\frac{d}{dx}$ regulárního výrazu V podle řetězce $x \in \Sigma^*$:

$\frac{dV}{dx} = V', h(V') = \{y : xy \in h(V)\}$

Derivace regulárních výrazů

Definice

Derivace $\frac{d}{dx}$ regulárního výrazu V podle řetězce $x \in \Sigma^*$:

1. $\frac{dV}{d\varepsilon} = V$
2. pro $a \in \Sigma$ platí:

   $\frac{d\varepsilon}{da} = \emptyset \quad \frac{d\emptyset}{da} = \emptyset$

   $$\frac{db}{da} = \left\langle \begin{array}{l} \emptyset, \text{ jestliže } a \neq b \\ \\ \varepsilon, \text{ jestliže } a = b \end{array} \right.$$

   $\frac{d(U+V)}{da} = \frac{dU}{da} + \frac{dV}{da}$

   $\frac{d(UV)}{da} = \frac{dU}{da}V + \{\frac{dV}{da} : \varepsilon \in h(U)\}$

   $\frac{d(V^*)}{da} = \frac{dV}{da}.V^*$
3. Pro $x = a_1 a_2 \ldots a_n, a_i \in \Sigma$ platí

   $\frac{dV}{dx} = \frac{d}{da_n}(\frac{d}{da_{n-1}}(\ldots \frac{d}{da_2}(\frac{dV}{da_1})\ldots))$

$\square$

Derivace regulárních výrazů

Příklad

Regulární výraz $y = (0+1)^*.1$.

$$
\begin{aligned}
\frac{dy}{d\varepsilon} &= (0+1)^*.1 \\
\frac{dy}{d1} &= \frac{d(0+1)^*}{d1}.1 + \frac{d1}{d1} \\
&= \frac{d(0+1)}{d1}.(0+1)^*.1 + \varepsilon \\
&= (\frac{d0}{d1} + \frac{d1}{d1})(0+1)^*.1 + \varepsilon \\
&= (\emptyset + \varepsilon).(0+1)^*.1 + \varepsilon \\
&= (0+1)^*.1 + \varepsilon \\
\frac{dy}{d0} &= \frac{d(0+1)^*}{d0}.1 + \frac{d1}{d0} \\
&= \frac{d(0+1)}{d0}.(0+1)^*.1 + \emptyset \\
&= (\varepsilon + \emptyset).(0+1)^*.1 + \emptyset \\
&= (0+1)^*.1
\end{aligned}
$$

Integrál regulárních výrazů

Definice

Integrál regulárního výrazu V podle řetězce $x \in \Sigma^*$ je definován takto:

$$h(\int V\ dx) = \{xy : y \in h(V)\}.$$

Pro integrování regulárních výrazů platí tato pravidla:

1. $\int V d\varepsilon = V$
2. pro $a \in \Sigma$ platí:
   $\int \varepsilon\ da = a,$
   $\int \emptyset\ da = \emptyset,$
   $\int b\ da = ab,$
   $\int (U + V)\ da = \int U\ da + \int V\ da,$
   $\int (U.V)\ da = aUV,$
   $\int V^*\ da = aV^*.$
3. pro $x = a_1 a_2 \cdots a_n \in \Sigma^*$ platí:
   $\int V\ dx = \int \cdots [\int (\int V\ da_n)\ da_{n-1}] \cdots da_1.$

Integrál regulárních výrazů

$$\frac{d}{dx}\int V\ dx = V,$$
$$\int \frac{dV}{dx}\ dx = V.\ \text{(?)}$$

Integrál zahrnující integrační konstantu Z:

$$\int V\ dx = xV + Z$$
$$\frac{dZ}{dx} = \emptyset$$

Integrál regulárních výrazů

Příklad

Regulární výraz $(0+1)^*.1$.

$$\int (0+1)^*.1 \; d1 = 1.(0+1)^*.1 + Z_1,$$
$$\int (0+1)^*.1 \; d0 = 0.(0+1)^*.1 + Z_0.$$

Úprava regulárních výrazů

Definice

(a) regulární výrazy x, y nazveme *identické* (označíme $x \equiv y$), jestliže x a y jsou úplně stejné řetězce symbolů.

(b) regulární výrazy x, y nazveme *ekvivalentní* (označíme $x = y$), jestliže mají stejnou hodnotu, $h(x) = h(y)$, tj. regulární množiny, které tyto výrazy popisují, jsou stejné.

(c) regulární výrazy x, y nazveme *podobné*, jestliže se dají na sebe převést pomocí následujících identit:

$$x + x = x$$
$$x + y = y + x$$
$$(x + y) + z = x + (y + z)$$
$$x + \emptyset = x$$
$$x.\emptyset = \emptyset.x = \emptyset$$
$$x.\varepsilon = \varepsilon.x = x$$

Úprava regulárních výrazů

Věty v axiomatické teorii regulárních výrazů

$V_1: \quad \emptyset^* = \varepsilon$

$V_2: \quad x^* + x = x^*$

$V_3: \quad (x^*)^* = x^*$

$V_4: \quad (x + y)^* = (x^* y^*)^*$

$V_5: \quad x^* y = y + x^* x y$

$V_6: \quad x^* y = y + x x^* y$

$V_7: \quad x^* y = (x^n)^*.(y + xy + x^2 y + ... + x^{n-1} y)$

$V_8:$ Jestliže $\varepsilon \in h(x)$, pak $x x^* = x^*$

$V_9: \quad (xy)^* x = x(yx)^*$

$V_{10}: \quad (x + y)^* = (x^* + y^*)^*$

Úprava regulárních výrazů

Příklad

Regulární výrazy

$x = \varepsilon + 1^*(011)^*(1^*(011)^*)^*$

$y = (1 + 011)^*$

Jsou ekvivalentní?

$$\begin{aligned} x &= \varepsilon + 1^*(011)^*(1^*(011)^*)^* \\ &= (1^*(011)^*)^* \\ &= (1 + 011)^* \\ &= y. \end{aligned}$$

$$\varepsilon + xx^* = x^*$$

$$(x^*y^*)^* = (x + y)^*$$

Úprava regulárních výrazů

Příklad

Je výraz $(\varepsilon + \emptyset)(0 + 1)^*1 + (0 + 1)^*\emptyset$ podobný výrazu $(0 + 1)^*1$?

$$
\begin{aligned}
&(\varepsilon + \emptyset)(0 + 1)^*1 + (0 + 1)^*\emptyset = && x.\emptyset = \emptyset \\
&= (\varepsilon + \emptyset)(0 + 1)^*1 + \emptyset = && x + \emptyset = x \\
&= (\varepsilon + \emptyset)(0 + 1)^*1 = && x + \emptyset = x \\
&= \varepsilon(0 + 1)^*1 = && \varepsilon.x = x \\
&= (0 + 1)^*1. && \square
\end{aligned}
$$

Úprava regulárních výrazů

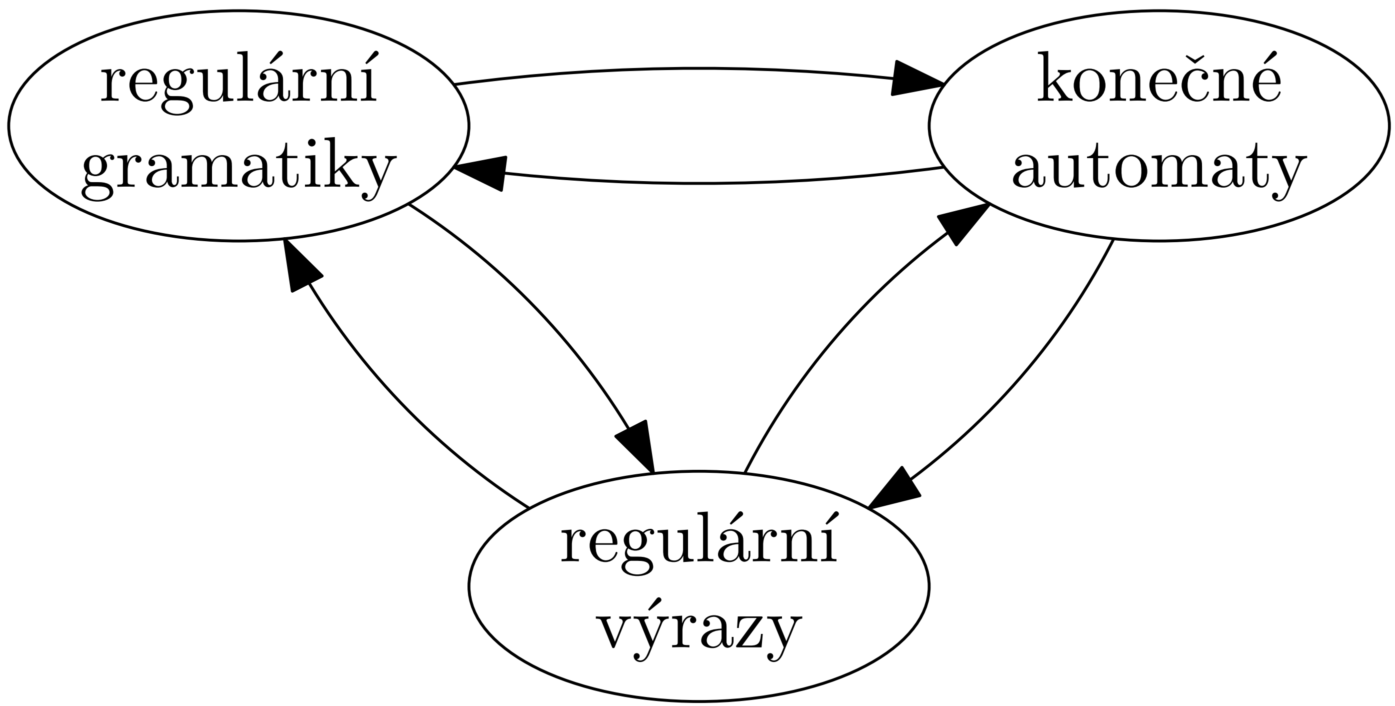
Příklad

Ekvivalentní RV, které se na sebe nedají převést:

- $B = (10^*1 + 1)^*(10^*1 + 0)$
- $B' = 0^*(1(0 + 10^*1)^*(100^* + 1) + 0)$

Vztahy mezi formálními systémy pro r.j.

Vztahy mezi formálními systémy pro popis regulárních jazyků



Vztah mezi reg. gr. a koneč. automaty

Algoritmus Konstrukce NKA pro danou pravou reg. gr.
Vstup: Pravá regulární gramatika $G = (N, T, P, S)$.
Výstup: NKA $M = (Q, T, \delta, q_0, F)$ takový, že $L(G) = L(M)$.
Metoda:

1. Množina vstupních symbolů automatu M je rovna T.
2. Množina stavů $Q = N \cup \{A\}, A \notin N$.
3. Zobrazení δ:
   je-li $B \to aC \in P$, pak $\delta(B, a)$ obsahuje C,
   je-li $B \to a \in P$, pak $\delta(B, a)$ obsahuje A.
4. $q_0 = S$.
5. $F = \{S, A\}$, jestliže $S \to \varepsilon \in P$,
   $F = \{A\}$, jestliže $S \to \varepsilon \notin P$.

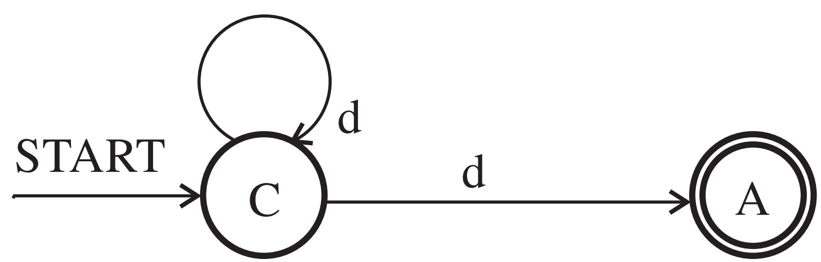
Vztah mezi reg. gr. a koneč. automaty

Příklad

$G = (\{C\}, \{d\}, \{C \to d \mid dC\}, C)$.

$M = (\{C, A\}, \{d\}, \delta, C, \{A\})$, kde δ:

δ	d
C	$\{C, A\}$
A	



Vztah mezi reg. gr. a koneč. automaty

Příklad

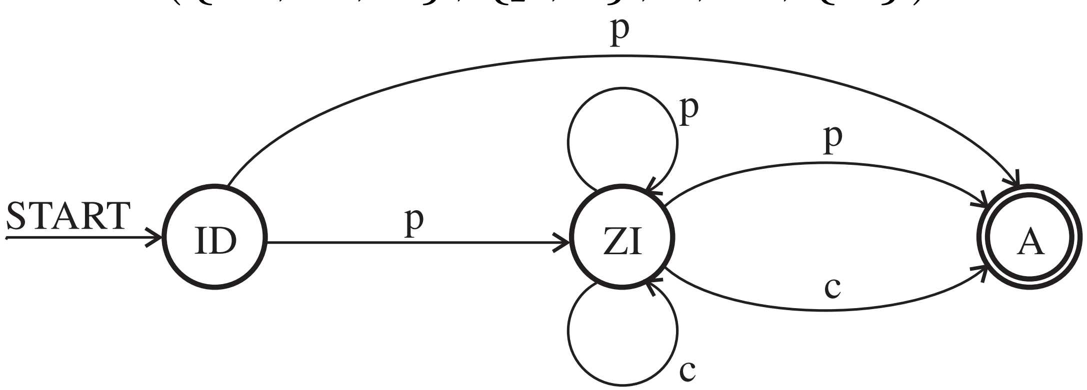
$G = (\{ID, ZI\}, \{p, c\}, P, ID)$, kde P obsahuje pravidla:

$ID \rightarrow p\, ZI \mid p$

$ZI \rightarrow p\, ZI \mid c\, ZI \mid p \mid c.$

(Gramatika generuje identifikátory podle obvyklé definice (p – písmeno, c – číslice).)

$M = (\{ID, ZI, A\}, \{p, c\}, \delta, ID, \{A\})$, kde δ



δ	p	c
ID	$\{ZI, A\}$	
ZI	$\{ZI, A\}$	$\{ZI, A\}$
A		

Vztah mezi reg. gr. a koneč. automaty

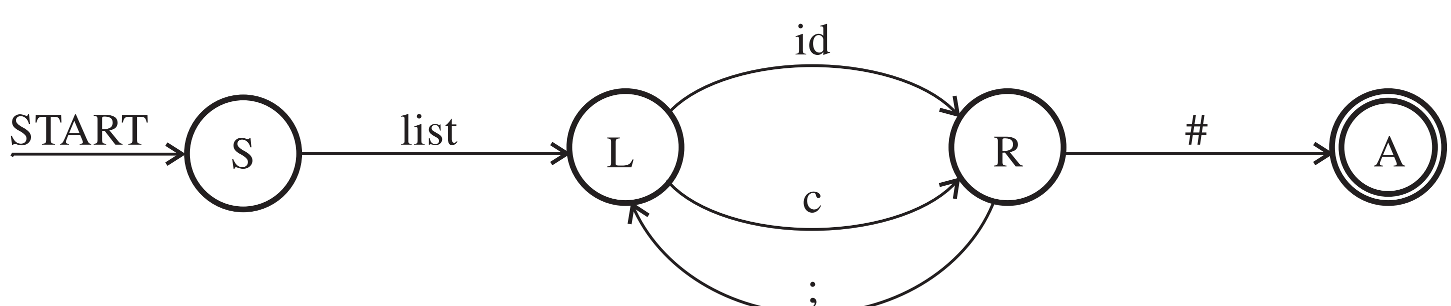
Příklad

Jazyk, který obsahuje řetězce tvaru: $list\ \ id; c; id; id; \ldots; c; c; id\#$

$G = (\{S, L, R\}, \{list, id, c, ;\,, \#\}, P, S)$, kde P:
$S \rightarrow list\, L$, $L \rightarrow id\, R \mid c\, R$, $R \rightarrow ;\ L \mid \#$

$M = (\{S, L, R, A\}, \{list, id, c, \#, ;\}, \delta, S, \{A\})$, kde δ:

δ	$list$	id	c	$;$	$\#$
S	L				
L		R	R		
R				L	A
A					



Vztah mezi reg. gr. a koneč. automaty

Algoritmus Konstrukce NKA pro levou regulární gramatiku.
Vstup: Levá regulární gramatika $G = (N, T, P, S)$.
Výstup: NKA $M = (Q, T, \delta, q_0, F)$ takový, že $L(G) = L(M)$.
Metoda:

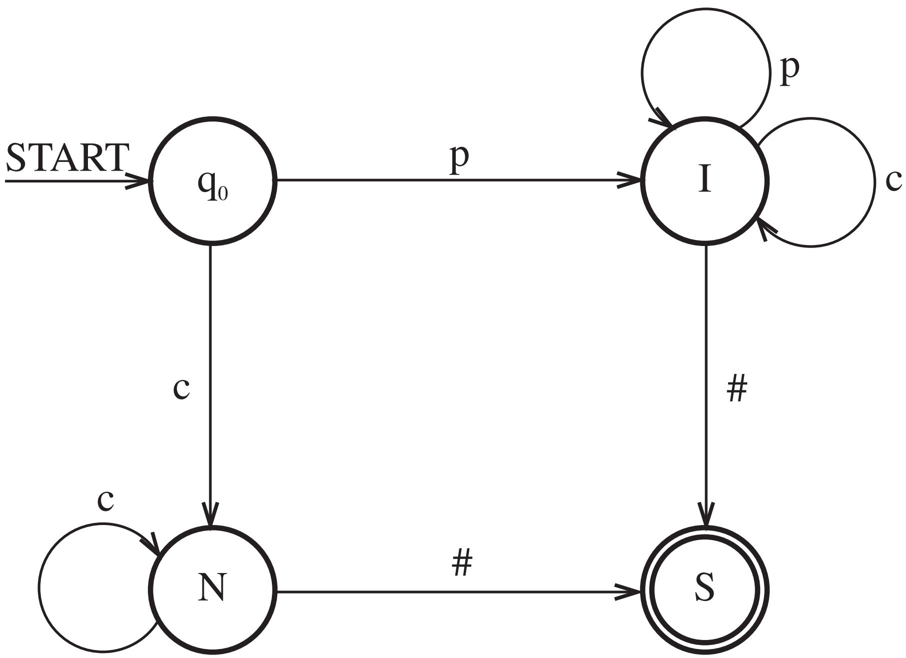
1. Množina vstupních symbolů automatu M je rovna T.
2. Množina stavů $Q = N \cup \{q_0\}$.
3. Zobrazení δ:
   Je-li $A \to Ba \in P$, pak $\delta(B, a)$ obsahuje A,
   je-li $A \to a \in P$, pak $\delta(q_0, a)$ obsahuje A.
4. $q_0 \in Q$ je počáteční stav automatu M.
5. Je-li $S \to \varepsilon \in P$, pak $F = \{S, q_0\}$,
   v opačném případě $F = \{S\}$.

Vztah mezi reg. gr. a koneč. automaty

Příklad

Levá regulární gramatika $G = (\{S, I, N\}, \{c, p, \#\}, P, S)$, kde P:
$S \rightarrow I\# \mid N\#$, $I \rightarrow p \mid Ip \mid Ic$, $N \rightarrow c \mid Nc$.

NKA $M = (\{S, I, N, q_0\}, \{c, p, \#\}, \delta, q_0, \{S\})$, kde δ:



	c	p	$\#$
S			
I	I	I	S
N	N		S
q_0	N	I	

Vztah mezi koneč. automaty a reg. gr.

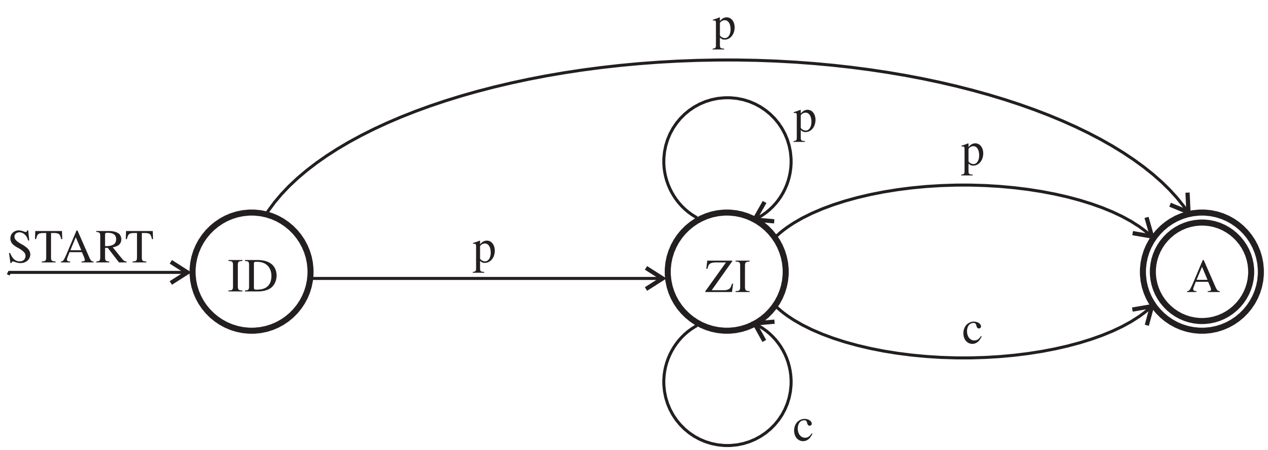
Algoritmus Konstrukce pravé reg. gramatiky pro daný NKA
Vstup: NKA $M = (Q, T, \delta, q_0, F)$.
Výstup: Pravá reg. gramatika $G = (N, T, P, S)$, $L(M) = L(G)$.
Metoda:

1. $N = Q$.
2. P sestrojíme takto:
   (a) Jestliže $\delta(B, a)$ obsahuje C, pak $B \to aC$ dáme do P.
   (b) Jestliže $\delta(B, a)$ obsahuje C a $C \in F$, pak $B \to a$ dáme do P.
   (c) Symbol $S = q_0$.
   (d) Jestliže $q_0 \in F$ a S není na pravé straně, pak $S \to \varepsilon$ dáme do P a počáteční symbol je S.
   (e) Jestliže $q_0 \in F$ a S je na pravé straně, pak přidáme do P pravidla tvaru $S' \to \alpha$, kde α jsou pravé strany pravidel tvaru $S \to \alpha,\ S'$ je počáteční symbol a pravidlo $S' \to \varepsilon$ přidáme do P.

Vztah mezi koneč. automaty a reg. gr.

Příklad

Sestrojíme pravou regulární gramatiku pro NKA:



Výsledná regulární gramatika $G = (\{\mathit{ID}, \mathit{ZI}, \mathit{A}\}, \{p, c\}, P, \mathit{ID})$, kde P:

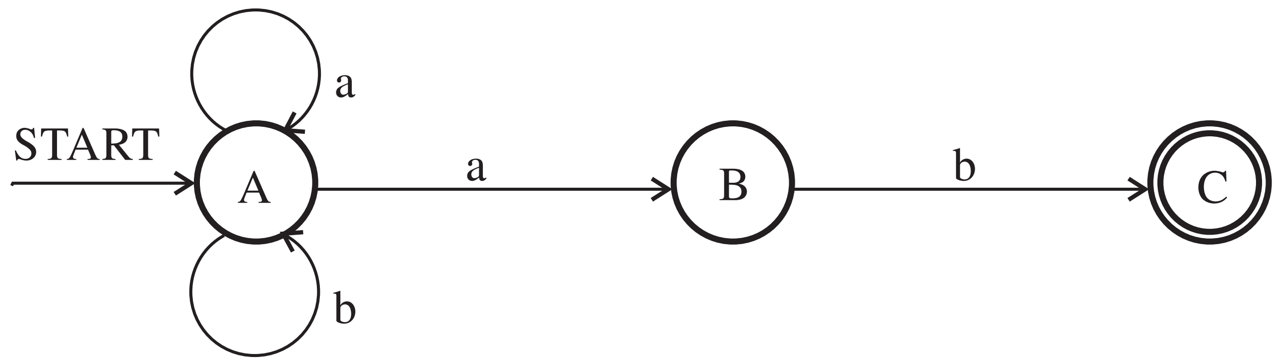
$\mathit{ID} \rightarrow p\mathit{ZI} \mid p \mid p\mathit{A}$

$\mathit{ZI} \rightarrow p\mathit{ZI} \mid p\mathit{A} \mid c\mathit{ZI} \mid c\mathit{A} \mid p \mid c$.

Vztah mezi koneč. automaty a reg. gr.

Příklad

Sestrojíme pravou regulární gramatiku pro NKA:



$G = (\{A, B, C\}, \{a, b\}, P, A)$, kde P:

$A \rightarrow aA \mid aB \mid bA$

$B \rightarrow bC \mid b$.