Prednáška 4 - Regulárne výrazy, vlastnosti regulárnych jazykov

Ing. Viliam Hromada, PhD.

C-510
Ústav informatiky a matematiky
FEI STU

viliam.hromada@stuba.sk

Regulárne gramatiky - opakovanie

Pre pripomenutie: Gramatika $G = (N, T, P, S)$ sa nazýva regulárna, ak každé jej pravidlo spĺňa jeden z 2 tvarov:

1. $A \rightarrow xB, A \in N, B \in N, x \in T^+$
2. $A \rightarrow w, A \in N, w \in T^*$

Každý jazyk, ktorý je možné generovať regulárnou gramatikou sa nazýva **regulárny**.

Konečné automaty vs gramatiky

- Konečné automaty predstavujú jeden z tzv. **akceptačných spôsobov** špecifikácie jazyka.
- Gramatiky predstavujú **generatívny spôsob** špecifikácie jazyka.
- Preto je vhodná otázka: existuje nejaký súvis medzi **gramatikami** a **automatmi**?
- O vzťahu regulárnych gramatík a konečných automatov pojednávajú nasledovné 2 vety.

Regulárna gramatika $\Rightarrow$ KA

Veta

Nech $G = (N, T, P, S)$ *je regulárna gramatika. Potom existuje nedeterministický konečný automat* $M = (Q, T, \delta, q_0, F)$ *taký, že* $L(M) = L(G)$.

Veta

Nech $M = (Q, T, \delta, q_0, F)$ *je DKA. Potom existuje taká regulárna gramatika* $G = (N, T, P, S)$*, že* $L(G) = L(M)$.

- Teda každý jazyk, ku ktorému vieme nájsť regulárnu gramatiku, ktorá ho generuje, je zároveň akceptovateľný nejakým (deterministickým/nedeterministickým konečným automatom).
- A naopak, každý jazyk, ku ktorému existuje nejaký DKA/NKA, ktorý ho akceptuje, je generovateľný nejakou regulárnou gramatikou.

Pripomeňme si Chomského hierarchiu

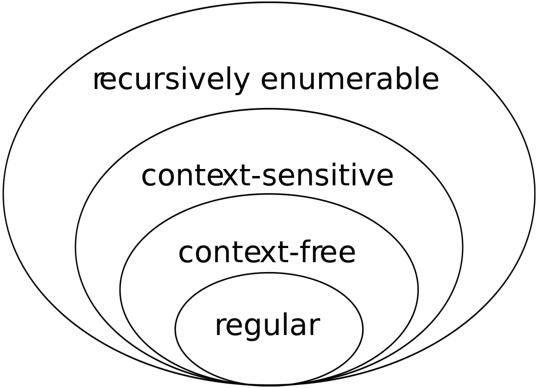
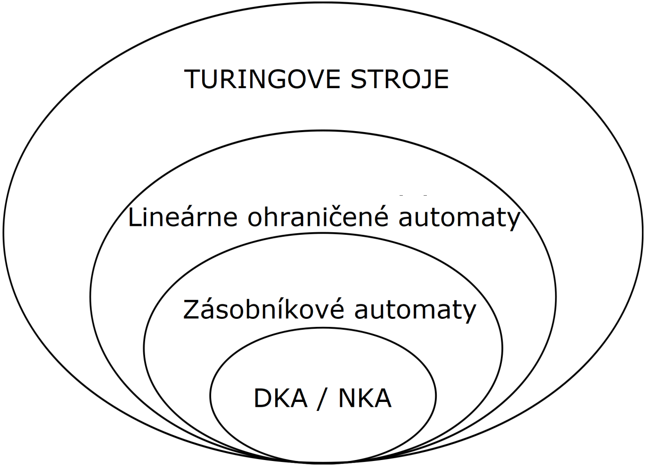


Figure: https://commons.wikimedia.org/wiki/File:Chomsky-hierarchy.svg

Z hľadiska výpočtových zariadení



Jazyk, ktorý nie je regulárny

- Ako naznačuje hierarchia, existujú jazyky, ktoré nie sú regulárne - a teda sa nedajú generovať regulárnou gramatikou, resp. akceptovať konečným automatom.
- Učebnicové príklady takýchto jazykov, napríklad nad abecedou $\Sigma = \{a, b\}$ sú:
  - $L = \{a^n b^n \mid n \in \{0, 1, 2, ...\}\}$
  - $L = \{w \in \{a, b\}^* \mid \sharp_a(w) = \sharp_b(w)\}$
  - $L = \{ww^R \mid w \in \{a, b\}^*\}$
  - $L = \{ww \mid w \in \{a, b\}^*\}$

Regulárne jazyky

Ako už bolo spomínané, hoci sú regulárne jazyky pomerne obmedzené, používajú sa napríklad na:

- Popis ľubovoľnej konečnej množiny (napr. kľúčových slov programovacieho jazyka)
- Popis identifikátorov premenných, funkcií, atď.
- Popis číselných konštánt

Taktiež tzv. **regulárne výrazy** úzko súvisia s regulárnymi jazykmi...

Regulárne výrazy

Regulárny výraz je popri regulárnych gramatikách a konečných automatoch iný spôsob špecifikácie regulárnych jazykov.

Regulárne výrazy

Definícia

Nech Σ je abeceda. Potom:

1. *$\emptyset$ je regulárny výraz popisujúci prázdny jazyk,*
2. *ε je regulárny výraz popisujúci jazyk $\{\varepsilon\}$,*
3. *a, pre $a \in \Sigma$, je regulárny výraz popisujúci jazyk $\{a\}$.*
4. *Ak R_1 a R_2 sú regulárne výrazy popisujúce jazyky L_1 a L_2, potom $(R_1 \mid R_2)$ je regulárny výraz popisujúci jazyk $L_1 \cup L_2$ (zjednotenie).*
5. *Ak R_1 a R_2 sú regulárne výrazy popisujúce jazyky L_1 a L_2, potom $(R_1 R_2)$ je regulárny výraz popisujúci jazyk $L_1 L_2$ (zreťazenie).*
6. *Ak R je regulárny výraz popisujúci jazyk L, potom (R^*) je regulárny výraz popisujúci iteráciu L^* jazyka L.*
7. *iné regulárne výrazy ako tie, zostrojené podľa bodov 1-6, neexistujú.*

Príklad: Nech $\Sigma = \{0, 1, ..., 9, a, b, ..., z\}$. Potom možné regulárne výrazy:

- $(0 \mid 1) = \{0, 1\}$
- $(1(0^*)1) = \{11, 101, 1001, ...\}$
- $((10)(0 \mid 1)^*) = \{10w \mid w \in \{0, 1\}^*\}$
- $((begin) \mid (end)) = \{begin, end\}$
- $((0 \mid 1 \mid ... \mid 9)(0 \mid 1 \mid ... \mid 9)^*)$ - celočíselné konštanty s prípadnými bezvýznamnými nulami zľava, bez znamienka
- $((\varepsilon \mid + \mid -)(0 \mid ((1 \mid ... \mid 9)(0 \mid 1 \mid ... \mid 9)^*)))$ - celočíselné konštanty bez/so znamienkom, bez bezvýznamných núl zľava
- $((a \mid ... \mid z)^*(begin)(a \mid ... \mid z)^*)$ - všetky textové reťazce obsahujúce *begin* ako podreťazec

Regulárne výrazy $\Rightarrow$ NKA

Veta

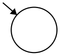
Nech R je regulárny výraz popisujúci jazyk L. Potom existuje nedeterministický konečný automat M taký, že $L(M) = L$.

Ukážeme si dôkaz tejto vety pomocou tzv. **Thompsonovej konštrukcie**.

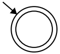
Dôkaz - $\emptyset, \varepsilon, a$

Stačí ukázať, že pre prvých 6 bodov z definície vieme vždy zostrojiť príslušný automat:

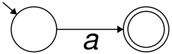
- Výraz $\emptyset$:



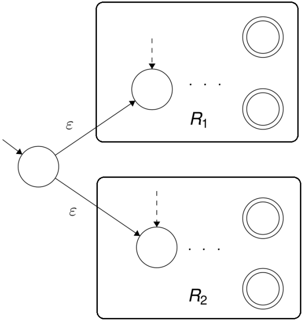
- Výraz ε:



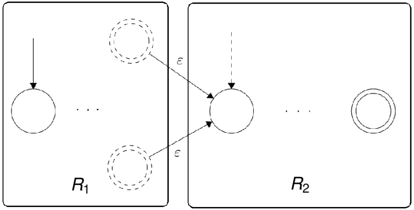
- Výraz a:



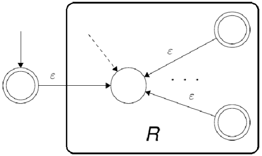
Dôkaz - $R_1 \mid R_2, R_1 R_2, R^*$



$R_1 \mid R_2$



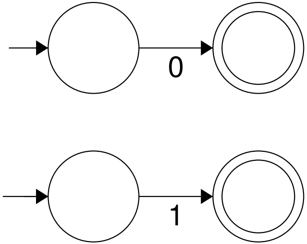
$R_1 R_2$



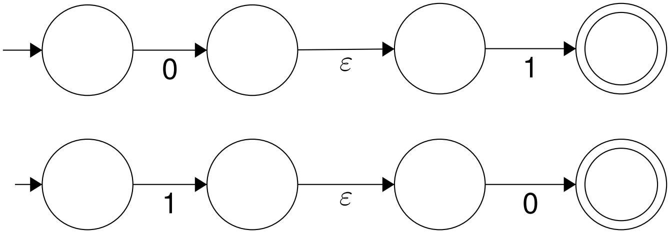
R^*

Príklad

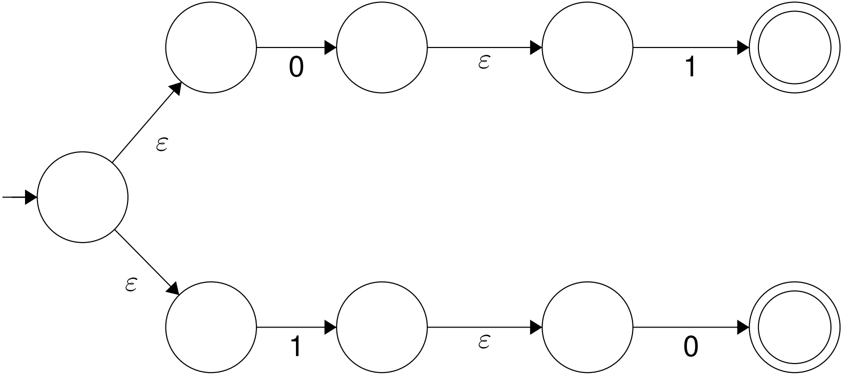
Nájdite DKA, ktorý akceptuje jazyk popísaný regulárnym výrazom: $(01|10)^*(\varepsilon|0|1)$. Najprv skonštruujeme NKA, ktorý rozpoznáva daný regulárny výraz... Začneme s elementárnymi KA pre 0 a 1.



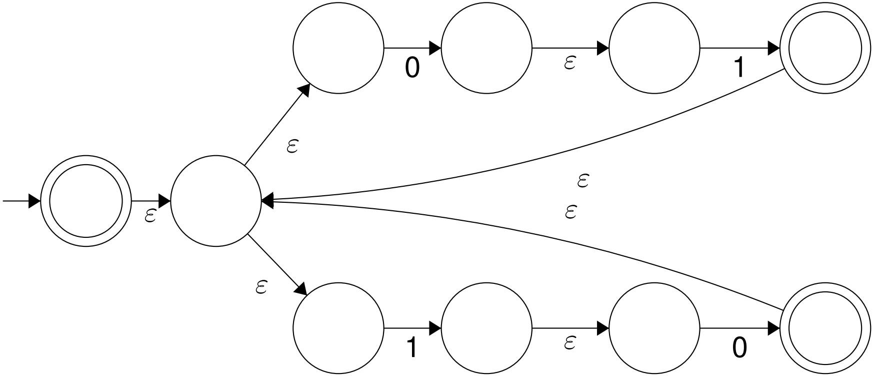
Z nich zreťazením dostávame 2 NKA - prvý pre 01 a druhý pre 10:



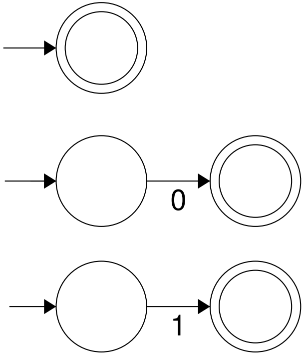
Zjednotením NKA pre 01 a NKA pre 10 vznikne NKA pre (01|10):



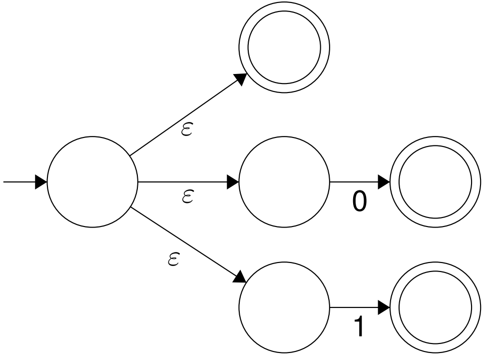
Iteráciou NKA pre (01|10) vznikne NKA pre $(01|10)^*$:



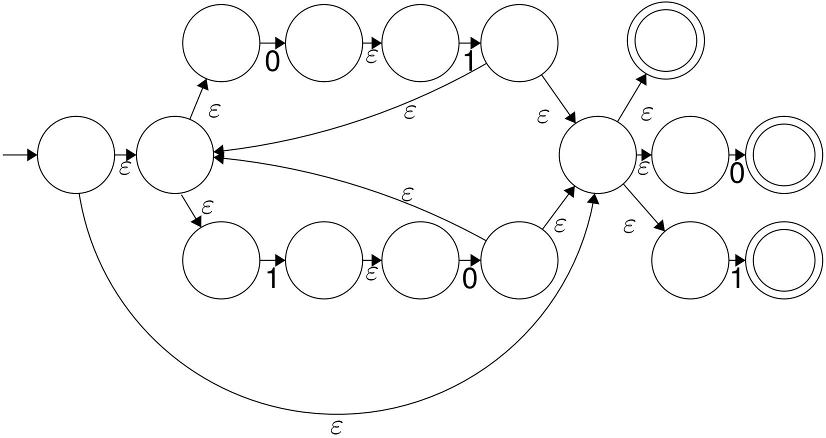
Ďalej potrebujeme zostrojiť NKA pre $(\varepsilon|0|1)$. Tri elementárne NKA, pre $\varepsilon, 0, 1$ sú nasledovné:



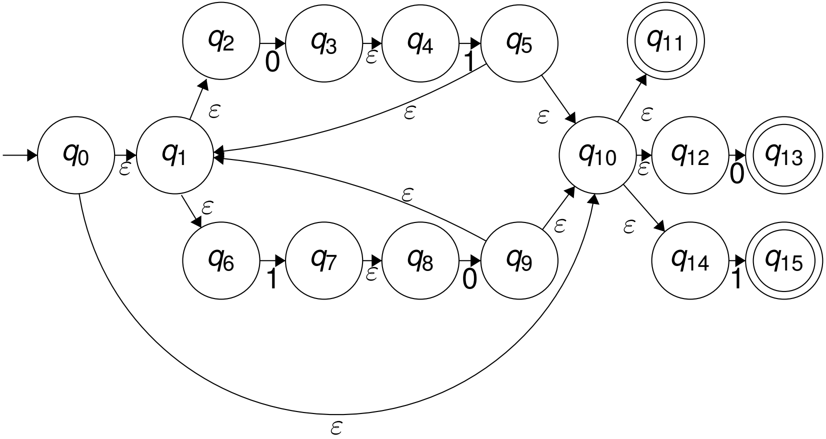
Ich zjednotením dostávame NKA pre $\varepsilon|0|1$:



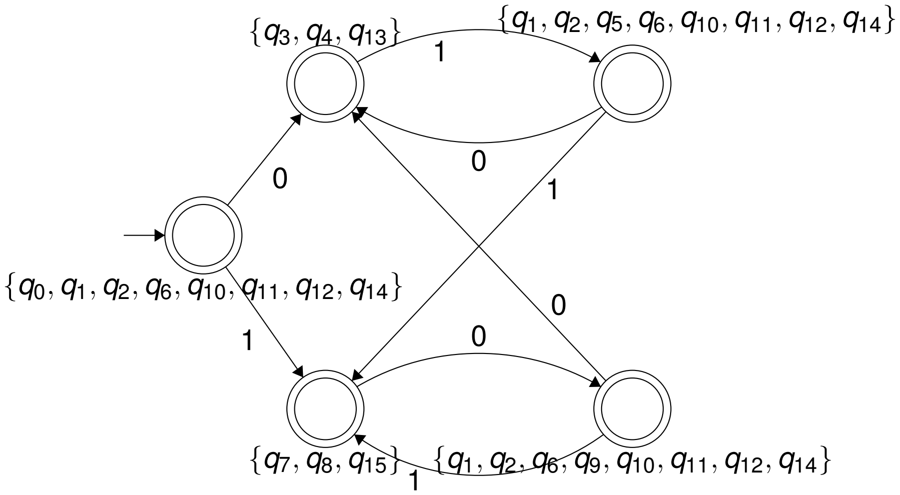
Na záver z NKA pre $(01|10)^*$ a z NKA pre $\varepsilon|0|1$ ich zreťazením dostávame výsledný NKA pre $(01|10)^*(\varepsilon|0|1)$



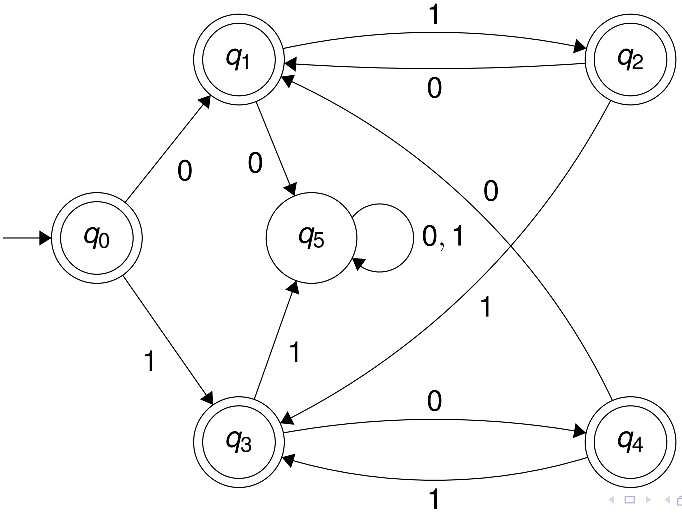
Následne automat podrobíme determinizácii. Najprv však pomenujme stavy (teraz je v princípe jedno, ako).



Ekvivalentný deterministický automat vyzerá nasledovne:



Po preznačení stavov a doplnení na úplný automat (pridáme pascu - stav q_5), dostávame:



Zhrnutie

Pre regulárne jazyky teda platí:

- Dajú sa popísať pomocou regulárnej gramatiky
- Dajú sa popísať pomocou (deterministického) konečného automatu
- Dajú sa popísať pomocou regulárneho výrazu.

Uzáverové vlastnosti regulárnych jazykov

Veta

Nech L_1 a L_2 sú regulárne jazyky. Potom zjednotenie jazykov $L_1 \cup L_2$ je tiež regulárny jazyk.

Veta

Nech L_1 a L_2 sú regulárne jazyky. Potom zreťazenie jazykov $L_1 L_2$ je tiež regulárny jazyk.

Veta

Nech L je regulárny jazyk. Potom iterácia jazyka L^* je tiež regulárny jazyk.

Všetky tri hore uvedené vety priamo vyplývajú z definície regulárnych výrazov, resp. z konštrukcie NKA, ktorý príslušné výrazy akceptuje.

Veta

Nech L je regulárny jazyk nad abecedou T. Potom doplnok jazyka L^C je regulárny jazyk.

KA akceptujúci L^C získame, ak v **úplnom** automate akceptujúcom L vymeníme akceptujúce stavy za neakceptujúce a naopak.

Veta

Nech L_1 a L_2 sú regulárne jazyky. Potom prienik jazykov $L_1 \cap L_2$ je regulárny jazyk.

Dôkaz: Nech $M_1 = (Q_1, \Sigma_1, \delta_1, q_{01}, F_1)$ a $M_2 = (Q_2, \Sigma_2, \delta_2, q_{02}, F_2)$ sú KA také, že $L(M_1) = L_1, L(M_2) = L_2$. Potom konečný automat M, akceptujúci $L(M) = L_1 \cap L_2$ je pätica $M = (Q, \Sigma, \delta, q_0, F)$, kde:

- $Q = \{(q_1, q_2) \mid q_1 \in Q_1 \wedge q_2 \in Q_2\}, Q \subseteq Q_1 \times Q_2$
- $\Sigma = \Sigma_1 \cap \Sigma_2$
- $\delta((q_1, q_2), a) = (\delta_1(q_1, a), \delta_2(q_2, a))$
- $q_0 = (q_{01}, q_{02})$
- $F = \{(q_1, q_2) \mid q_1 \in F_1 \wedge q_2 \in F_2\}$

Konečný automat je teda kombináciou M_1 a M_2 a „sleduje“, akými cestami sa môžu uberať oba automaty a akceptuje slovo len vtedy, keď by ho akceptoval M_1 a súčasne aj M_2 (teda logicky patrí do $L_1 \cap L_2$).

Príklad: Nech $M_1 = (\{q_{01}, q_{11}\}, \{0, 1\}, \delta_1, q_{01}, \{q_{11}\})$ a $M_2 = (\{q_{02}, q_{12}, q_{22}\}, \{0, 1, 2\}, \delta_2, q_{02}, \{q_{22}\})$, kde prechodové tabuľky:

δ_1	0	1
q_{01}	q_{11}	-
q_{11}	q_{11}	q_{11}

δ_2	0	1	2
q_{02}	q_{12}	q_{22}	q_{22}
q_{12}	q_{12}	q_{22}	q_{12}
q_{22}	q_{12}	q_{22}	q_{12}

Riešenie:

δ	0	1
(q_{01}, q_{02})	(q_{11}, q_{12})	-
(q_{11}, q_{12})	(q_{11}, q_{12})	(q_{11}, q_{22})
(q_{11}, q_{22})	(q_{11}, q_{12})	(q_{11}, q_{22})

Použitá literatúra

Dedera, Ľ: Počítačové jazyky a ich spracovanie.

Linz, P.: An Introduction to Formal Languages and Automata.

Molnár, Ľ.: Gramatiky a jazyky.