Teorie jazyků a automatů— 6. přednáška

Marie Demlová
http://math.feld.cvut.cz/en/people/demlova

4. 11. 2024

Regulární výrazy

Regulární výraz reprezentující jazyk $L(M)$.

Ke stavovému diagramu přidáme

- nový počáteční stav *start* a nový koncový stav *fin*,
- hranu ze *start* do počátečního stavu automatu a hrany z koncových stavů do *fin*, všechny označené ε.

S takto vzniklým orientovaným grafem provádíme úpravy:

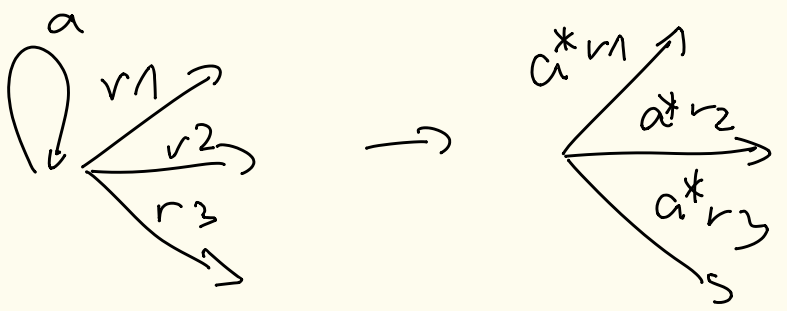
a) Odstranění paralelních hran.

b) Odstranění smyčky.

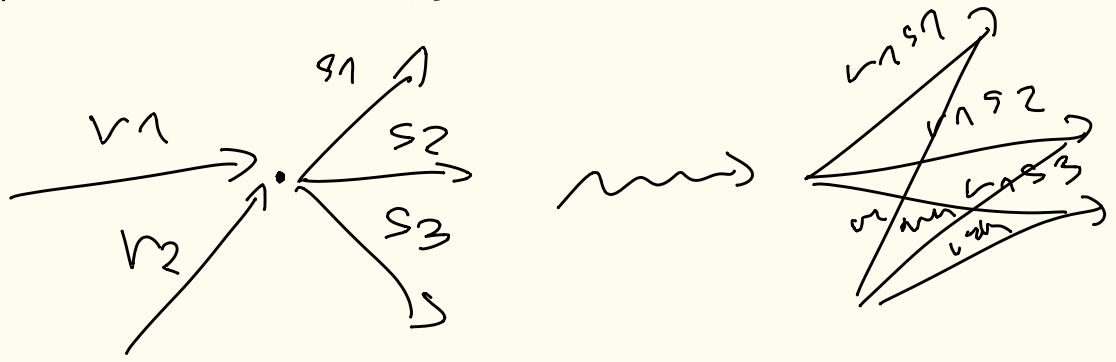
c) Odstranění vrcholu q.

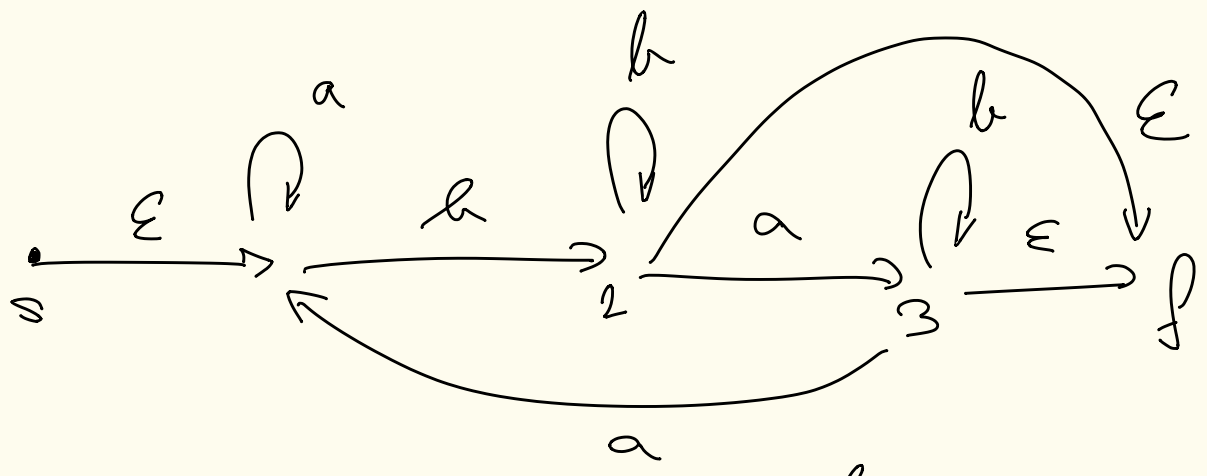
Končíme, když zbyde *start* a *fin* a jediná hrana. Ta je ohodnocena hledaným regulárním výrazem.

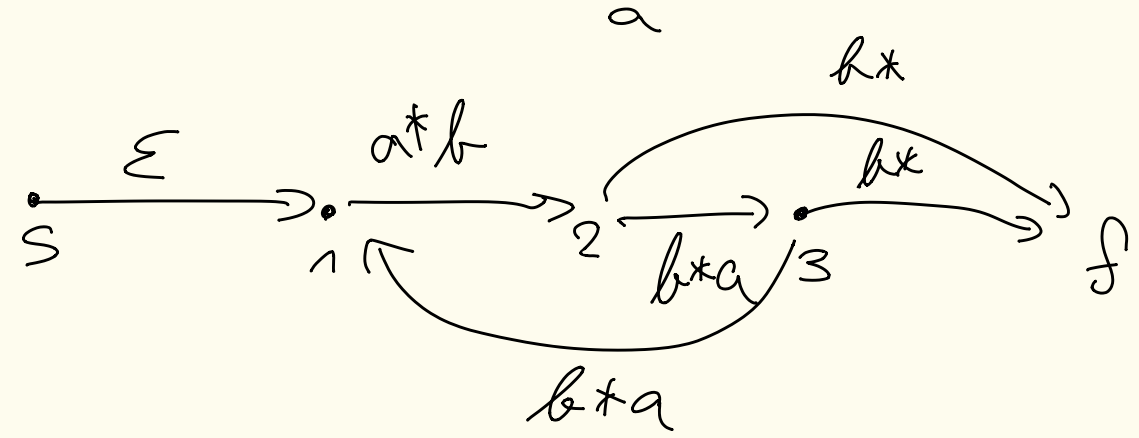
odstranění smyček



odstranění vrcholu







$$a^*b(b^*ab^*aa^*b)^*(b^*+b^*ab^*)$$

Regulární výrazy

Regulární výraz reprezentující jazyk $L(M)$ – příklad.

Příklad. Vytvořte regulární výraz, který reprezentuje jazyk přijímaný

	a	b
$\rightarrow 1$	1	2
$\leftarrow 2$	3	2
$\leftarrow 3$	1	3

Regulární výrazy

Zápis regulárních výrazů v praxi

- Regulární výraz . (tečka) reprezentuje libovolný znak.
- Pro zápis obecnější množiny znaků se používá konstrukce s hranatými závorkami.
- Znak | („svislítko“, v UNIXu roura) značí alternativu, tedy operaci, která se v algebraickém zápisu značí $+$.
- Hvězdička (Kleeneho operace) se zapisuje v řádce, nikoli jako exponent.
- Znak $+$ označuje jedno nebo více opakování předcházejícího. Tedy $\mathbf{r}+ = \mathbf{r}\mathbf{r}*$.

Regulární výrazy

- Znak ? označuje žádný nebo jeden výskyt předchozího. Tedy **r**? označuje to, co by se v algebraickém zápise vyjádřilo jako $(\varepsilon + \mathbf{r})$.
- Pro obecnější počty opakování předchozího znaku nebo výrazu v závorkách se používá konstrukce se složenými závorkami.

  Příklady:

  - $\mathbf{r}\{3\}$ značí přesně 3 opakování,
  - $\mathbf{r}\{3,\}$ značí 3 nebo více opakování,
  - $\mathbf{r}\{3,5\}$ značí 3 až 5 opakování.

Regulární výrazy

- Znak „stříška“ na začátku výrazu značí začátek textu nebo řádky.
- Znak „dolar“ na konci výrazu značí konec textu nebo řádky.
- Některé znaky mají v praktických regulárních výrazech speciální význam popsaný výše. Potřebujeme-li některý z těchto znaků použít ve výrazu jako obyčejný znak, je třeba před tento znak zapsat obrácené lomítko.
- Některé speciální znaky se bohužel v praxi používají nejednotným způsobem.

Pro zevrubnější poučení doporučuji text Pavla Satrapy `http://www.nti.tul.cz/~satrapa/docs/regvyr/`, kde je i tabulka odlišností syntaxe regulárních výrazů v různých implementacích.

Další uzávěrové vlastnosti třídy regulárních jazyků

Homomorfismus

Máme dvě abecedy Σ, Γ a zobrazení $h\colon \Sigma \to \Gamma^\star$. h rozšíříme na zobrazení $\Sigma^\star$ do $\Gamma^\star$ takto:

- $h(\varepsilon) = \varepsilon$,
- $h(u\,a) = h(u)\,h(a)$.

Obraz jazyka L **v** h je

$$h(L) = \{h(w) \mid w \in L\}.$$

Další uzávěrové vlastnosti třídy regulárních jazyků

Substituce.

Máme dvě abecedy Σ, Γ a zobrazení $\sigma : \Sigma \to \mathcal{P}(\Gamma^\star)$. σ rozšíříme na zobrazení $\Sigma^\star$ do $\mathcal{P}(\Gamma^\star)$ takto:

- $\sigma(\varepsilon) = \{\varepsilon\}$,
- $\sigma(u\,a) = \sigma(u)\,\sigma(a)$.

Obraz jazyka L **v** σ je

$$\sigma(L) = \bigcup \{\sigma(w) \mid w \in L\}.$$

Další uzávěrové vlastnosti třídy regulárních jazyků

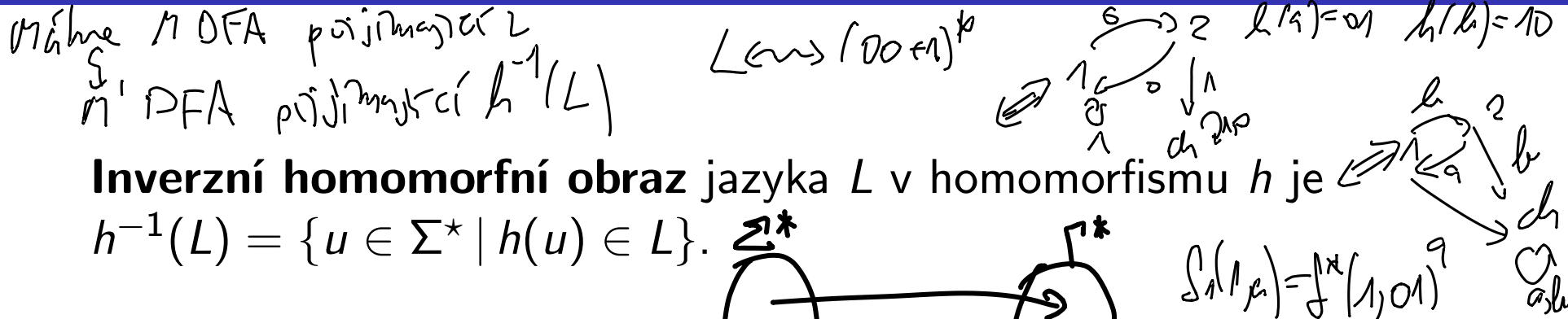
Věta.

Je dána substituce σ z $\Sigma^\star$ do jazyků nad abecedou Γ. Jestliže každý jazyk $\sigma(a)$ je regulární a jestliže L je regulární jazyk nad abecedou Σ, pak jazyk $\sigma(L)$ je regulární jazyk nad abecedou Γ.

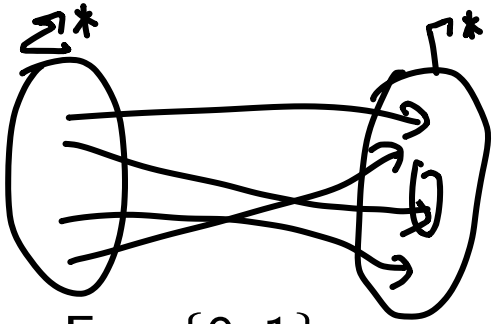
Důsledek.

Je-li h homomorfismus a L regulární jazyk, pak jazyk $h(L)$ je také regulární jazyk.

Další uzávěrové vlastnosti třídy regulárních jazyků



Inverzní homomorfní obraz jazyka L v homomorfismu h je $h^{-1}(L) = \{u \in \Sigma^* \mid h(u) \in L\}$.



Příklad.

Uvažujme jazyk L nad abecedou $\Gamma = \{0, 1\}$ popsaný regulárním výrazem $(00 + 1)^\star$ a homomorfismus h určený $h(a) = 01$ a $h(b) = 10$.

Pak $h^{-1}(L)$ je reprezentován regulárním výrazem $(ba)^\star$.

Další uzávěrové vlastnosti třídy regulárních jazyků

Věta.

Jestliže h je homomorfismus, $h\colon \Sigma^\star \to \Gamma^\star$ a L je regulární jazyk nad abecedou Γ, pak inversní obraz $h^{-1}(L)$ je také regulární (nad Σ).

Algoritmická řešitelnost úloh pro regulární jazyky

1. Pro daný konečný automat M (ať deterministický nebo nedeterministický) a slovo $w \in \Sigma^\star$ rozhodnout, zda $w \in L(M)$.

2. Pro daný konečný automat M rozhodnout, zda $L(M) = \Sigma^\star$.

3. Pro dva konečné automaty M_1 a M_2 rozhodnout, zda $L(M_1) = L(M_2)$.

Algoritmická řešitelnost úloh pro regulární jazyky

Tvrzení.

Je dán deterministický konečný automat M s n stavy. Pak

1. Jazyk $L(M)$ je neprázdný právě tehdy, když M přijímá slovo w délky $|w| < n$.

2. Jazyk $L(M)$ je nekonečný právě tehdy, když M přijímá slovo v délky $n \leq |v| < 2n$.

Algoritmická řešitelnost úloh pro regulární jazyky

Jiný pohled na DFA.

Dvousměrné automaty.

Gramatiky

Gramatika je uspořádaná čtveřice $\mathcal{G} = (N, \Sigma, S, P)$, kde

- N je konečná množina tzv. *neterminálů*;
- Σ je konečná neprázdná množina tzv. *terminálů*, platí $N \cap \Sigma = \emptyset$;
- $S \in N$ je *startovací symbol*;
- P je konečná množina pravidel typu $\alpha \to \beta$, kde α a β jsou slova nad $N \cup \Sigma$ taková, že α obsahuje alespoň jeden neterminál.

Gramatiky

Přímé odvození.

Dána gramatika $\mathcal{G} = (N, \Sigma, S, P)$. Řekneme, že δ se přímo odvodí (též přímo přepíše) z γ, jestliže existuje v P pravidlo $\alpha \to \beta$ a slova $\varphi, \psi \in (N \cup \Sigma)^\star$ taková, že $\gamma = \varphi\,\alpha\,\psi$ a $\delta = \varphi\,\beta\,\psi$.

Tento fakt zapisujeme $\gamma \Rightarrow_{\mathcal{G}} \delta$.

Gramatiky

Odvození, derivace.

Dána gramatika $\mathcal{G} = (N, \Sigma, S, P)$. Řekneme, že δ se odvodí z γ, jestliže

- buď $\gamma = \delta$,
- nebo existuje posloupnost přímých odvození

$$\gamma = \gamma_1 \Rightarrow_{\mathcal{G}} \gamma_2 \Rightarrow_{\mathcal{G}} \ldots \Rightarrow_{\mathcal{G}} \gamma_k = \delta.$$

Značíme $\gamma \Rightarrow^{\star}_{\mathcal{G}} \delta$ a konečná posloupnost je derivace.

Gramatiky

Jazyk generovaný gramatikou.

Slovo $w \in \Sigma^\star$ je generováno gramatikou $\mathcal{G}$, jestliže existuje derivace $S \Rightarrow_{\mathcal{G}}^\star w$.

Jazyk $L(\mathcal{G})$ generovaný gramatikou $\mathcal{G}$ je množina všech slov generovaných $\mathcal{G}$, tj.

$$L(\mathcal{G}) = \{w \mid S \Rightarrow_{\mathcal{G}}^\star w\}.$$

Gramatiky

Konvence.

- Neterminály značíme obvykle velkými písmeny $A, B, X, Y, \ldots$.
- Terminály značíme obvykle malými písmeny ze začátku abecedy, například $a, b, c, d, \ldots$.
- Slova z $(N \cup \Sigma)^\star$ obvykle značíme řeckými písmeny α, β, $\ldots$.
- Terminální slova, tj. slova z $\Sigma^\star$, značíme malými písmeny z konce abecedy $u, w, x, y, \ldots$.

Gramatiky

Chomského hierarchie.

- Gramatiky typu 0 jsou nejobecnější. Jazyky generované gramatikami typu 0 jsou *jazyky typu 0*.

- Gramatiky typu 1, též kontextové gramatiky, jsou gramatiky s pravidly tvaru

$$\alpha A \beta \to \alpha \gamma \beta,$$

kde $\alpha, \beta, \gamma \in (N \cup \Sigma)^\star$, A je neterminál a $\gamma \neq \varepsilon$. Jedinou výjimku tvoří pravidlo $S \to \varepsilon$, pak se ale S nevyskytuje na pravé straně žádného pravidla.

Jazyky generované gramatikami typu 1 jsou *jazyky typu 1*, též *kontextové jazyky*.

Gramatiky

Chomského hierarchie — II.

- Gramatiky typu 2, též bezkontextové gramatiky, (CF gramatiky) jsou gramatiky s pravidly tvaru

$$A \to \gamma,$$

kde $\gamma \in (N \cup \Sigma)^\star$ a A je neterminál.

Jazyky generované gramatikami typu 2 jsou *bezkontextové jazyky* nebo *jazyky typu 2*.

- Gramatiky typu 3 neboli regulární gramatiky (též pravé lineární gramatiky) jsou gramatiky s pravidly tvaru

$$A \to w\,B,\ A \to w,$$

kde A, B jsou neterminály a w je terminální slovo.

Jazyky generované gramatikami typu 3 jsou *regulární jazyky* nebo *jazyky typu 3*.