

REGULÁRNÍ VÝRAZY A RACIONÁLNÍ MNOŽINY

Nechť $(M, \cdot, 1_M)$ je monoid, tedy množina M s binární asociativní operací a neutrálním prvkem (jednotkou).

Racionální podmnožiny monoidu. Řekneme, že systém $\mathcal{F}$ podmnožin monoidu M je *racionálně uzavřený*, pokud pro libovolné dvě množiny $A, B \in \mathcal{F}$ je také

- $A \cup B \in \mathcal{F}$,
- $A \cdot B \in \mathcal{F}$,
- $\langle A \rangle \in \mathcal{F}$.

Součin množin přitom definujeme $A \cdot B = \{a \cdot b \mid a \in A, b \in B\}$ a uzávěrem $\langle A \rangle$ množiny A míníme nejmenší podmonoid M obsahující A .

Racionálním uzávěrem nějakého systému $\mathcal{F}$ podmnožin monoidu M rozumíme nejmenší racionálně uzavřený systém obsahující $\mathcal{F}$.

Podmnožina monoidu M se nazývá *racionální*, pokud leží v racionálním uzávěru konečných podmnožin M , který značíme $\text{Rat}M$. Je to tedy nejmenší třída, která obsahuje prázdnou množinu, všechny jednoprvkové množiny a je uzavřená na součin, sjednocení a uzávěr.

Regulární výrazy. *Regulárním výrazem* nad monoidem M je každé správně utvořené slovo nad M a operačními symboly $\{1, 0, +, \cdot, *\}$, kde 1 a 0 jsou konstanty, $+$ a $\cdot$ jsou binární symboly a $*$ je unární symbol.

Regulární výrazy jsou tedy dány následující indukční definicí.

- $1, 0$ a $m \in M$ jsou regulární výrazy (nazývané *atomické*);
- jsou-li r a s regulární výrazy, je i
 - $(r \cdot s)$
 - $(r + s)$
 - (r^*)

regulární výraz.

Symbol $*$ se nazývá *Kleenova hvězda*. Pokud přijmeme obvyklou přednost symbolu násobení před symbolem sčítání a přednost Kleenovy hvězdy před symboly sčítání a násobení, můžeme vynechávat některé závorky. Symbol násobení také obvykle nahrazujeme prostým zřetězením. Výrazy $r \cdot r^*$ a $r^* \cdot r$ se obvykle zkracují na r^+ (tzv. *Kleenovo plus*).

Rozšířený regulární výraz je definován přidáním unárního symbolu $\bar{}$ a pravidla

- je-li r regulární výraz, je i $\bar{r}$ regulární výraz.

Dále nás bude zajímat výhradně monoid Σ^* , což je volný monoid nad abecedou Σ s operací zřetězení a prázdným slovem ε jako jednotkou. Prvky tohoto monoidu se nazývají *slova* a jeho podmnožiny *jazyky*.

Souvislost.

Pozorování. Každou racionální podmnožinu monoidu Σ^* lze definovat regulárním výrazem nad Σ . A naopak, každý regulární výraz nad Σ definuje nějakou racionální podmnožinu Σ^* .

Souvislost je dána následujícími přirozenými pravidly.

- Regulárnímu výrazu 0 odpovídá prázdná množina;
- regulárnímu výrazu 1 odpovídá množina $\{\varepsilon\}$, kde ε je prázdné slovo;

- regulárnímu výrazu $a \in \Sigma$ odpovídá množina $\{a\}$;
- regulárnímu výrazu $(r \cdot s)$, kde r je regulární výraz odpovídající množině A a s je regulární výraz odpovídající množině B , odpovídá množina $A \cdot B$;
- regulárnímu výrazu $(r + s)$, kde r je regulární výraz odpovídající množině A a s je regulární výraz odpovídající množině B , odpovídá množina $A \cup B$;
- regulárnímu výrazu r^* , kde r je regulární výraz odpovídající množině A , odpovídá množina $\langle A \rangle$.

Jediné, co zbývá ošetřit jsou jednoprvkové množiny $\{u\}$, kde $u = a_1 a_2 \cdots a_k$, $a_i \in \Sigma$ a $k \geq 2$. Taková množina je zjevně definována regulárním výrazem $a_1 a_2 \cdots a_k$.

Pro rozšířený regulární výraz navíc platí, že

- regulárnímu výrazu $\bar{r}$, kde r je regulární výraz odpovídající množině A , odpovídá množina $\Sigma^* \setminus A$ (tedy doplněk A v Σ^*).

Skutečnost, že třída racionálních jazyků (tedy podmnožin Σ^*) je uzavřená na komplement plyne z Kleenovy věty. Uvedená korespondence je důvodem, proč se racionální jazyky zpravidla označují jako *regulární jazyky*.

KONEČNÉ AUTOMATY

Konečný automat $\mathcal{A}$ je pětice $(Q, \Sigma, \delta, q_0, F)$, kde Q je konečná množina stavů, Σ je konečná abeceda, $\delta \subseteq Q \times \Sigma \times Q$ je *přechodová relace*, $q_0 \in Q$ je *počáteční stav* a $F \subseteq Q$ je množina *přijímajících stavů*.

Smyslem přechodové relace je přiřadit danému okamžitému popisu automatu, tj. aktuálnímu stavu a aktuálnímu čtenému znaku, následující stav. Protože automat není obecně deterministický, je obrazem tohoto přiřazení množina stavů. Můžeme tedy psát $\delta(q, s) = P$, kde P je množina všech stavů q' , pro které platí $(q, s, q') \in \delta$. Ještě obecněji můžeme pro libovolné $R \in Q$ psát $\delta(R, s) = P$, kde P je množina stavů p takových, že existuje $r \in R$ takové, že $(r, s, p) \in \delta$. Symbolem δ^* značíme „obal“ relace δ , tedy relaci $\delta^* \subseteq Q \times \Sigma^* \times Q$, která obsahuje δ , pro každé $q \in Q$ obsahuje (q, ε, q) a pro všechna $u, v \in \Sigma^*$ platí

$$(q, u, q') \in \delta^* \ \& \ (q', v, q'') \in \delta^* \implies (q, uv, q'') \in \delta^*.$$

Zápis $\delta^*(R, s) = P$, budeme zkracovat jako $R \cdot s = P$, přičemž pro jednoprvkové množiny $\{q\}$ píšeme prostě q . Definici δ^* lze stručně zapsat pravidlem

$$R \cdot su = (R \cdot s) \cdot u.$$

Pokud platí, že pro každé $(q, s) \in Q \times \Sigma$ existuje nejvýše jedno q' takové, že $(q, s, q') \in \delta$, řekneme, že automat $\mathcal{A}$ je *deterministický*. Pokud existuje právě jedno takové q' , je automat *úplný deterministický*.

Jazyk *přijímaný* automatem $\mathcal{A}$ je definován jako

$$L(\mathcal{A}) = \{w \in \Sigma^* \mid q_0 \cdot w \cap F \neq \emptyset\}.$$

Jinak řečeno, $L(\mathcal{A})$ je množina slov, pro které existuje cesta automatem končící v přijímajícím stavu. Pokud dva automaty přijímají stejný jazyk, řekneme, že jsou *ekvivalentní*.

Věta. *Každý jazyk přijímaný konečným automatem je přijímán také úplným deterministickým konečným automatem.*

Důkaz. Nechť je dán obecně nedeterministický automat $\mathcal{A} = (Q, \Sigma, \delta, q_0, F)$. Definujme $\mathcal{B} = (\mathcal{P}(Q), \Sigma, \delta', q_0, F)$, kde $\mathcal{P}(Q)$ je potenční množina Q , $\delta' : \mathcal{P}(Q) \times \Sigma \rightarrow$

$\mathcal{P}(Q)$ je definována výše přijatou konvencí a F' je množina množin, které obsahují alespoň jeden přijímající stav.

Je snadné ověřit, že $\mathcal{B}$ je úplný deterministický a $L(\mathcal{A}) = L(\mathcal{B})$. $\square$

Pro zjednodušení některých úvah je užitečné povolit tzv. ε -přechody, tj. dovolit, aby δ obsahovala prvky (q, ε, q') . Všimněte si, že pro každý konečný automat lze zkonstruovat jemu ekvivalentní, který ε -přechody neobsahuje.

Věta (Kleene). *Jazyk je regulární právě když je přijímán konečným automatem.*

Důkaz. „ $\Rightarrow$ “ Jednoprvkové množiny i prázdná množina jsou jistě přijímány konečným automatem. Chceme ukázat, že třída jazyků přijímaných konečným automatem je uzavřená na součin, sjednocení a uzávěr. Automat přijímající součin jazyků $L(\mathcal{A})$ a $L(\mathcal{B})$ vznikne přidáním ε -přechodu mezi přijímající stavy automatu $\mathcal{A}$ a počáteční stav automatu $\mathcal{B}$. Automat přijímající sjednocení lze získat z direktního součinu obou automatů vhodnou volbou přijímajících stavů (nebo také dvěma ε -přechody z nového vstupního stavu do vstupních stavů $\mathcal{A}$ a $\mathcal{B}$). Automat přijímající uzávěr $L(\mathcal{A})$ vznikne přidáním ε -přechodu mezi přijímající stavy a počáteční stav automatu $\mathcal{A}$, případně přidáním nového přijímajícího počátečního stavu s ε přechodem do původního počátečního stavu, pokud tento nebyl přijímající, což zajistí přijetí prázdného slova.

„ $\Leftarrow$ “ Nechť $\mathcal{A} = (Q, \Sigma, q_0, \delta, F)$ je úplný deterministický konečný automat se stavy $Q = \{q_0, q_1, \dots, q_n\}$. Označme $L_{r,s}^{(m)}$ jazyk obsahující slova vedoucí ze stavu q_r do stavu q_s a neprocházející skrz žádný stav q_k , $k \geq m$. Přesněji, slovo u leží v $L_{r,s}^{(m)}$, právě když

- $\delta^*(q_r, u) = q_s$ a
- pokud slovo $v \neq u$ je neprázdným prefixem u a $\delta^*(q_r, v) = q_k$, pak $k < m$.

Ukážeme, že $L_{r,s}^{(m)}$ je regulární jazyk pro všechna $r, s \in \{0, 1, \dots, n\}$ a všechna $m \in \{0, 1, \dots, n+1\}$. Postupujme indukcí podle m . Je-li $m = 0$, pak je jazyk $L_{r,s}^{(m)}$ roven $\{a \in \Sigma \mid \delta(q_r, a) = q_s\}$, a je tedy regulární.

Přepokládejme nyní, že $L_{r,s}^{(m)}$ jsou regulární pro dané $m \leq n$ a všechna r, s . Každé slovo $u \in L_{r,s}^{(m+1)}$, které neleží v $u \in L_{r,s}^{(m)}$, lze faktorizovat jako $u_1 u_2 \dots u_\ell$ tak, že

- $\delta^*(q_r, u_1) = q_m$ a $u_1 \in L_{r,m}^{(m)}$;
- $\delta^*(q_m, u_i) = q_m$ a $u_i \in L_{m,m}^{(m)}$ pro všechna $i = 2, 3, \dots, \ell - 1$;
- $\delta^*(q_m, u_\ell) = q_s$ a $u_\ell \in L_{m,s}^{(m)}$.

Z toho plyne, že

$$L_{r,s}^{(m+1)} = L_{r,s}^{(m)} \cup \left(L_{r,m}^{(m)} \cdot \langle L_{m,m}^{(m)} \rangle \cdot L_{m,s}^{(m)} \right).$$

Tím je dokázáno, že $L_{r,s}^{(m+1)}$ je regulární jazyk, a díky indukci je regulární $L_{r,s}^{(n+1)}$. Nyní si stačí všimnout, že jazyk přijímaný automatem $\mathcal{A}$ je roven

$$L(\mathcal{A}) = \bigcup_{q_f \in F} L_{q_0, q_f}^{(n+1)}.$$

$\square$