Formální jazyky a automaty

Petr Šimeček

Úvod

Formální jazyky a automaty jsou základním kamenem teoretické informatiky. Na počátku se zmíníme o Chomského klasifikaci gramatik, nástroje, který lze aplikovat třeba i v lingvistice či analýze gregoriánských chorálů.

Hlavní oblastí našeho zájmu budou konečné automaty. Hádejte, kdo obstarává to, že pod seznamem, atlasem, altavistou ... najdete vše tak rychle?

Na závěr se zmíníme o Turingových strojích a dokážeme, že počítače nikdy nebudou předvídat budoucnost, ani nikdy nespočítají všechny možné reálné funkce.

Jazyky

Definice.

- *Libovolná neprázdná množina V je abecedou. Její prvky nazýváme znaky nebo symboly.*
- *Slovem nad abecedou V myslíme konečnou posloupnost znaků z V, značíme $w = a_1 a_2 \cdots a_n$.*
- *Délka slova w je délka w jako posloupnosti, značíme $|w| = n$.*
- *Prázdné slovo (posloupnost délky 0) značíme ε, $|\varepsilon| = 0$. Symbolem V^* značíme množinu všech slov nad abecedou V, symbolem V^+ množinu $V \setminus \{\varepsilon\}$.*

Definice. *Jazykem L nad abecedou V rozumíme libovolnou (ne nutně konečnou!) množinu slov, tedy $L \subseteq V^*$.*

Definice. *Nechť $u = a_1 a_2 \ldots a_m$, $v = b_1 b_2 \ldots b_n$ jsou slova nad V. Definujme operaci zřetězení $V^* \times V^* \to V^*$ tak, že $uv = a_1 a_2 \ldots a_m b_1 b_2 \ldots b_n$ (napíšeme u,v za sebe).*

Gramatiky

Definice. (Gramatika) *Gramatika je čtveřice $G = (N, \Sigma, P, S)$ s následujícím významem:*

- *N je množina neterminálů, obvykle značíme $A, B, \ldots$*

- Σ *je množina terminálů, obvykle značíme* $a, b, \ldots$; *platí* $N \cap \Sigma = \emptyset$, *dále zavedeme značení* $V = N \cup \Sigma$.
- P *je množina pravidel (pravidlem rozumíme dvojici slov z nichž první obsahuje alespoň jeden neterminál, značíme* $\alpha \to \beta$*).*
- S *je startovní symbol,* $S \in N$.

Definice. *Budeme nyní definovat odvození v* G*:*
- *v jednom kroku:* $u \Rightarrow_G v$, *pokud existuje pravidlo, které* u *přepíše na* v.
- *ve více krocích:* analogicky $\Rightarrow_G^*$.
- *ve více krocích, ale v minimálně jednom:* $\Rightarrow_G^+$.

Definice.
- *Jazyk generovaný gramatikou* G *je množina* $L(G) = \{w \in \Sigma^*; S \Rightarrow_G^* w\}$ *slov z terminálů odvoditelných z* S.
- *Gramatiky* G_1 *a* G_2 *jsou ekvivalentní, pokud* $L(G_1) = L(G_2)$.

Definice. (Chomského hierarchie gramatik)
- *Gramatika typu* 0 *je libovolná gramatika.*
- *Gramatika typu* 1 *(kontextová) je gramatika, pro jejíž všechna pravidla* $\alpha \to \beta$ *platí* $|\alpha| \leq |\beta|$.
- *Gramatika typu* 2 *(bezkontextová) je gramatika, jejíž všechna pravidla jsou tvaru* $A \to \beta$, *kde* A *je neterminál,* $\beta \in V^+$.
- *Gramatika typu* 3 *(regulární) je gramatika, jejíž všechna pravidla jsou tvaru* $A \to \beta$, *kde* A *je neterminál,* $\beta \in V^+$ *je buď* a *nebo* aB.

Deterministické konečné automaty

Definice. *Konečný automat (KA)* M *nad abecedou* Σ *je pětice* (K,Σ,δ,q_0,F), *kde*
- K *je konečná neprázdná množina stavů.*
- Σ *je abeceda KA.*
- δ *je přechodová funkce:* $K \times \Sigma \to K$.
- q_0 *je* $q_0 \in K$, *počáteční stav.*
- F *je* $F \subset K$, *množina koncových stavů.*

Definice. *Přechodovou funkci* δ *lze induktivně rozšířit na zobecněnou přechodovou funkci* $\delta^* : K \times \Sigma^* \to K$ *předpisem:*
- $\delta^*(q, a) := \delta(q, a)$ *pro* $a \in \Sigma$.
- $\delta^*(q, xa) := \delta(\delta^*(q, x), a)$ *pro* $x \in \Sigma^+$, $a \in \Sigma$.

Místo δ^* *lze psát pouze* δ *(na symbolech se rovnají).*

Definice.
- *Slovo w je přijímané (akceptované, rozpoznávané) KA, pokud $\delta(q_0, w) \in F$.*
- *Jazyk rozpoznávaný KA M je množina $L(M) = \{w \in \Sigma^*; w$ přijímané $M\}$.*
- *KA M_1 a M_2 jsou ekvivalentní, pokud $L(M_1) = L(M_2)$.*

Lemma. (pumpovací) *Nechť L je jazyk rozpoznávaný KA. Pak existuje číslo p takové, že každé slovo $w \in L$, $|w| \geq p$ lze rozložit takto: $w = xyz$, $|y| \leq p$ a všechna $w_i = xy^iz$ jsou také z jazyka L.*

Definice.
- *Nechť $\sim$ je ekvivalence, řekneme, že $\sim$ je pravá kongruence, pokud z $x \sim y$ plyne $xz \sim yz$ pro každou trojici x, y, z.*
- *Index ekvivalence je počet tříd odpovídajícího rozkladu.*

Věta. (Nerodova) *Následující tři tvrzení jsou ekvivalentní:*
- *Jazyk $L \subseteq \Sigma^*$ je akceptovaný nějakým KA.*
- *Jazyk L je sjednocením některých tříd rozkladu pravé kongruence s konečným indexem.*
- *Nechť $\sim$ je definovaná na Σ^* předpisem: $x \sim y$ právě tehdy, když pro každé $r \in \Sigma^*$ platí $xr \in L \Leftrightarrow yr \in L$. Pak takto definovaná $\sim$ má konečný index.*

Definice. *KA s nejmenším počtem stavů, který rozpoznává daný jazyk L, se nazývá minimální automat pro L.*

Věta. *Minimální automat je dán jednoznačně až na izomorfismus.*

Poznámka. Minimální automat dostaneme tak, že odstraníme „nedosažitelné stavy“ a „ztotožníme“ stavy „ekvivalentní“.

Nedeterministické konečné automaty

Místo KA budeme dále psát DKA, abychom odlišili od NKA.

Definice. *Nedeterministický konečný automat (NKA) M nad abecedou Σ je pětice (K,Σ,δ,q_0,F), kde*
- *K je konečná neprázdná množina stavů.*
- *Σ je abeceda KA.*
- *δ je přechodová funkce: $K \times \Sigma \to 2^K$ (Pozor!).*
- *q_0 je $q_0 \in K$, počáteční stav.*
- *F je $F \subset K$, množina koncových stavů.*

Definice. *Přechodovou funkci δ lze opět induktivně rozšířit na zobecněnou přechodovou funkci $\delta^* : K \times \Sigma^* \to K$ předpisem:*

- $\delta^*(q, \varepsilon) := \{q\}$.
- $\delta^*(q, xa) := \bigcup_{p \in \delta^*(q,x)} \delta(p, a)$ *pro* $x \in \Sigma^+$, $a \in \Sigma$.

Navíc pro NKA rozšiřujeme obor δ na $2^K \times \Sigma^*$: $\delta^+(P, x) = \bigcup_{p \in P} \delta^*(p, x)$.

Definice.

- *Slovo w je přijímané (akceptované, rozpoznávané) NKA, pokud $\delta(q_0, w) \cap F \neq \emptyset$.*
- *Jazyk rozpoznávaný NKA M je množina $L(M) = \{w \in \Sigma^*; w$ přijímané $M\}$.*
- *NKA M_1 a M_2 jsou ekvivalentní, pokud $L(M_1) = L(M_2)$.*

Věta. *Nechť L je jazyk rozpoznávaný NKA. Potom existuje DKA, jenž rozpoznává L.*

Regulární gramatiky a KA

Věta. *Nechť G je regulární gramatika. Potom existuje KA M takový, že $L(M) = L(G)$.*

Věta. *Nechť M je konečný automat. Potom existuje gramatika G taková, že $L(G) = L(M)$.*

Důsledek. *Jazyk L je generovaný nějakou gramatikou G právě tehdy když existuje nějaký KA, který jej rozpoznává.*

Věta. *Třída regulárních jazyků je uzavřená vzhledem k operaci*

(a) *sjednocení.*
(b) *doplňku.*
(c) *průniku.*
(d) *zřetězení.*
(e) *uzávěru.*

Poznámka. Všechny konečné jazyky jsou zjevně regulární.

Turingovy stroje

Definice. *Turingův stroj (TS) je systém* $T = (Q,\Sigma,\Gamma,\varepsilon,\delta,q_0,F)$*, kde*
- Q *je množina stavů.*
- Σ *je vstupní abeceda,* $\Sigma \subset \Gamma \setminus \{\varepsilon\}$.
- Γ *je pásková abeceda.*
- ε *je prázdný symbol (tzn. na pásce nic není),* $\varepsilon \in \Gamma$.
- δ *je přechodová fce* $Q \times \Gamma \to Q \times \Gamma \times \{L, R\}$.
- q_0 *je počáteční stav.*
- F *je množina koncových stavů.*

Poznámka. Turingův stoj je systém s konečně stavovou řídící jednotkou a pamětí ve formě pásky rozdělené na políčka. Na políčku může být zapsán symbol jisté abecedy nebo může být prázdné. Páska je alespoň v jednom směru nekonečná. TS v každém kroku změní obsah políčka na kterém je, posune se doleva či doprava a případně změní svůj stav. Změny závisí na stavu a obsahu políčka, na kterém se TS nachází.

Věta. (Turingova) *Problém zastavení Turingova stroje je nerozhodnutelný.*

Literatura

- Křetínský a kol., *Formální jazyky a automaty I,II* (k dostání na FI MU, knihkupectví u Marečka).
- `www.fi.muni.cz/zkusto` — zápisky studentů z FI MU.
- Chytil, *Automaty a gramatiky.*