

# Automaty z wagami, wykład 1

Paweł Laskoś-Grabowski

7 października 2008

Wiemy, że języki regularne można definiować za pomocą wyrażeń regularnych skonstruowanych z liter, operacji konkatenacji, sumy i gwiazdki Kleenego. Wysokość gwiazdkową (*starheight*) wyrażenia definiujemy jako najwyższy poziom zagnieżdżenia gwiazdek w wyrażeniu, np.  $\mathbf{Sh}((a^*bc)^*) = 2$ . Wysokość gwiazdkowa języka to minimalna wysokość gwiazdkowa wśród wyrażeń opisujących ten język.

Złożoność problemu, by na podstawie danego DFA  $\mathcal{A}$  określić  $\mathbf{Sh}(L(\mathcal{A}))$  była nieznana do 1986, obecnie wiadomo, że jest w **2EXPSPACE**.

Można również zdefiniować rozszerzone wyrażenia regularne — korzystające również z operacji dopełnienia (a więc i przekroju). Nie wiadomo, czy istnieje język taki, że jego wysokość gwiazdkowa w sensie wyrażeń rozszerzonych jest większa od 1.

Udowodnimy, że hierarchia **Sh** dla (zwykłych) wyrażeń jest nieskończona, a nawet ścisła.

**Twierdzenie 1.** Dla dowolnej naturalnej  $n$  istnieje język regularny  $L$  taki, że  $\mathbf{Sh}(L) = n$ .

Zdefiniujemy indukcyjnie następujący ciąg wyrażeń regularnych oraz słów:

$$\begin{aligned}\alpha_1 &= (ab)^* \\ \alpha_i &= \left( a^{2^{i-1}} \alpha_{i-1} b^{2^{i-1}} \alpha_{i-1} \right)^* \\ w(1, j) &= ab \\ w(i, j) &= a^{2^{i-1}} w(i-1, j)^j b^{2^{i-1}} w(i-1, j)^j\end{aligned}$$

Oczywiście  $w(i, j) \in L(\alpha_i)$  oraz  $\mathbf{Sh}(L(\alpha_i)) \leq i$ .

Niech  $K_i$  będzie zbiorem języków  $L$  takich, że

- (1)  $\exists t \in \mathbb{Z} \quad \forall w \in L \quad \#_a(w) = \#_b(w) + t$
- (2)  $\forall j_0 \in \mathbb{N} \quad \exists j > j_0 \quad \exists w \in L \quad w(i, j)^j$  jest podsłowem  $w$
- (3)  $L$  ma minimalną **Sh** spośród języków spełniających (1, 2).

$K_i$  jest niepusta, bo  $L(\alpha_i)$  spełnia warunki (1, 2). **Sh** języków z tej klasy to  $d_i \leq i$ . Rozważmy zatem język  $L \in K_i$ , dla którego najlepsze (w sensie **Sh**) wyrażenie ma postać

$$\alpha = \sum \gamma_0^* \gamma_1 \gamma_2^* \gamma_3 \dots \gamma_{2n-1} \gamma_{2n}^*$$

gdzie suma jest skończenie indeksowana, a  $\mathbf{Sh}(\gamma_j) < d_i$ .

**Lemat 1.** Każde  $\gamma_j$  spełnia warunek (1), żadne  $\gamma_j$  nie spełnia warunku (2).

*Dowód.* Zauważmy, że całe wyrażenie  $\beta = \gamma_0^* \dots \gamma_{2n}^*$  spełnia warunek (1). Dla parzystych  $\gamma$  zachodzi  $\#_a = \#_b$  — w przeciwnym wypadku słowa z inną ilością „powtórzeń” słowa dopasowującego się do takiej  $\gamma$  miałyby różne różnice  $\#_a - \#_b$ . Podobnie gdyby do  $\gamma_j$  dla  $j$  nieparzystego dopasowywały się słowa  $w', w''$  o różnych wartościach tej różnicy, to słowa  $w_1 w' w_2, w_1 w'' w_2$  dopasowujące się do  $\beta$  również miałyby różne wartości tej różnicy.

Teraz ponieważ  $\mathbf{Sh}(\gamma_j) < d_i$ , to gdyby  $\gamma_j$  spełniało warunek (2), otrzymalibyśmy sprzeczność z warunkiem (3) należenia do  $K_i$ .  $\square$

**Lemat 2.** W sumie  $\alpha$  można wybrać taki składnik  $\beta$  i indeks  $0 \leq k \leq n$ , że podwyrażenie  $\gamma_{2k}^*$  w  $\beta$  spełnia warunek (2).

*Dowód.* Przypuśćmy nie wprost, że nie ma takiego  $k$ . Istnieje zatem takie  $j_0$ , że dla  $j > j_0$  słowo  $w(i, j)^j$  nie dopasowuje się do żadnego z  $\gamma_0^*, \gamma_1, \gamma_2^* \dots$ . Rozważmy DFA rozpoznający język  $\beta$  o  $j_1$  stanach. Niech  $j_2 > \max(j_0, j_1)$ , wtedy  $w(i, j_2)^{j_2}$  jest pod słowem pewnego słowa z  $L(\beta)$ . Zatem dla tego słowa na ścieżce przejść rozważanego DFA musi istnieć cykl, a wtedy istnieje liczba  $0 < m < j_2$  taka, że dla  $t \in \mathbb{N}$  i pewnych  $w_1, w_2$  jest  $w_1 w(i, j_2)^{j_2+mt} w_2 \in L(\beta)$ .

Możemy wziąć takiej postaci  $j > (2n+1)j_2$  i przyjrzeć się słowu  $w_1 w(i, j_2)^j w_2$ . Nie jest możliwe, aby każda kopia słowa  $w(i, j_2)^{j_2}$  zawierała w sobie granicę między  $\gamma_l, \gamma_{l+1}$  — granic tych jest  $2n$ , kopii słowa —  $2n+1$ . Zatem któraś trafia do pewnego  $\gamma_i$  lub  $\gamma_i^*$ , sprzeczność.

Z lematu 1 wynika dodatkowo, że nieparzysty  $\gamma$  nie może zawierać takiego słowa, więc musi ono znajdować się w którymś  $\gamma_{2k}$ .  $\square$

**Lemat 3.** W sumie  $\alpha$  można wybrać taki składnik  $\beta$  i indeks  $0 \leq k \leq n$ , że podwyrażenie  $\gamma_{2k}$  w  $\beta$  spełnia warunek (2'), tzn. (2) z  $i$  zastąpionym przez  $i-1$ .

*Dowód.* Rozważmy  $\gamma_{2k}$  wybrane w lemacie 2. Wiemy, że  $\gamma_{2k}^*$  spełnia (2), przypuśćmy nie wprost, że  $\gamma_{2k}$  nie spełnia (2'). Istnieje zatem  $j_0$  takie, że dla  $j > j_0$   $w(i-1, j)^j$  nie jest pod słowem żadnego słowa z  $L(\gamma_{2k})$ . Istnieje też  $j_1 > j_0$  takie, że  $w(i, j_1)^{j_1}$  jest pod słowem pewnego  $w \in L(\gamma_{2k}^*)$ .

Oznaczmy  $w = w_1 w_2 \dots w_m, w_j \in L(\gamma_{2k})$  i zastanówmy się, jak jeden egzemplarz  $w(i, j_1) = \underbrace{a^{2^{i-2}} w(i-1, j_1)^{j_1}}_u \underbrace{a^{2^{i-2}} w(i-1, j_1)^{j_1}}_v$  może być umieszczony w czynnikach  $w$ .

Niech  $w_p, w_q$  oznaczają czynniki, w których zaczynają się odpowiednio słowa  $u, v$ , oraz  $w_q = xy$  gdzie  $x$  jest sufiksem  $u$ .

$p < q$ , ponieważ w przeciwnym przypadku całe  $w(i-1, j_1)^{j_1}$  znajduje się w jednym czynniku dopasowującym się do  $\gamma_{2k}$ . Poczyńmy następujące spostrzeżenia:

*Spostrzeżenie 1.* W każdym sufiksie  $w(i-1, j_1)^{j_1}$  jest nie mniej  $b$  niż  $a$ .

*Spostrzeżenie 2.* W każdym prefiksie  $v$  jest więcej  $b$  niż  $a$ .

W całym  $w_q$  jest tyle samo  $a$ , co  $b$ , a  $x$  jest sufiksem  $w(i-1, j_1)^{j_1}$ . W  $y$  musi być zatem nie mniej  $a$  niż  $b$ . Więc albo  $y$  zawiera całe  $v$  (a dalej jeszcze więcej liter  $a$ ), co znów daje sprzeczność, albo  $y$  jest puste. Ten sam argument stosujemy teraz do  $w_{q+1}$ , które zaczyna się tam, gdzie  $v$ .  $\square$

*Dowód twierdzenia 1.* Pokażemy, że  $d_i \geq i$ , co razem ze znaną już nierównością przeciwną dowiedzie twierdzenia. Baza indukcji ( $i = 1$ ) jest trywialna.

Rozważmy zatem  $L \in K_i$  opisany przez wyrażenie  $\alpha$  jak powyżej i wybierzmy  $\gamma_{2k}$  z lematów 2, 3. Wiemy, że  $\mathbf{Sh}(\gamma_{2k}) < d_i$ , a zatem  $\mathbf{Sh}(L(\gamma_{2k})) < d_i$ . Wyrażenie to spełnia warunki **(1, 2')**, a więc warunki **(1, 2)** należenia do  $K_{i-1}$ , zatem dla  $L' \in K_{i-1}$  jest  $\mathbf{Sh}(L') < d_i$ . Oznacza to, że  $d_{i-1} \leq d_i - 1$ , co przy założeniu indukcyjnym  $d_{i-1} \geq i - 1$  daje tezę.  $\square$