Dolne ograniczenie na rozmiar wyrażeń regularnych równoważnych automatowi skończonemu

Filip Murlak

31 marca 2011

Niniejszy tekst jest w całości oparty na artykule Andrzeja Ehrenfeuchta i Paula Zeigera [1]. Naszym celem jest udowodnienie następującego dolnego ograniczenia:

Twierdzenie 1. *Dla każdego n istnieje taki automat deterministyczny o n stanach, że każde wyrażenie regularne równoważne temu automatowi ma długość co najmniej 2^n.*

Dla $n \in \mathbb{N}$ zdefiniujmy automat $\mathcal{A}_n$ o stanach $Q_n = \{0, 1, \ldots, n-1\}$, nad alfabetem

$$\Sigma_n = \{a_{ij} \mid i, j \in Q_n\}$$

o przejściach $i \xrightarrow{a_{ij}} j$. Wszystkie stany automatu są jednocześnie początkowe i akceptujące. Odrzucenie słowa przez automat zdarza się jedynie wtedy, gdy na tym słowie nie istnieje żaden bieg, tzn. gdy brakuje przejścia odpowiadającego kolejnej literze. Nietrudno pokazać następującą własność.

Lemat 1.

$$L(\mathcal{A}_n) = \{a_{i_1 i_2} a_{i_2 i_3} a_{i_3 i_4} \ldots a_{i_{k-1} i_k} \mid k \in \mathbb{N},\ \ i_1, i_2, \ldots, i_k \in Q_n\}.$$

Zauważmy , że $L(\mathcal{A}_n)$ jest zamknięty na podsłowa.

Powiemy, że wyrażenie regularne α *pokrywa* słowo w, ozn. $w \sqsubseteq \alpha$, o ile istnieją takie u, v, że $uwv \in L(\alpha)$. Będziemy pisali również $\beta \sqsubseteq \alpha$, gdy $w \sqsubseteq \alpha$ dla każdego $w \in L(\beta)$.

Lemat 2. *Dla każdego słowa w i wyrażenia α, jeśli $w^k \sqsubseteq \alpha$ dla pewnego $k > 2|\alpha|$, to $w^* \sqsubseteq \alpha$.*

Dowód. Dla wyrażenia α można zbudować równoważny automat A o liczbie stanów co najwyżej $2|\alpha|$. Przypuśćmy, że $vw^k v' \in L(\alpha)$. Zatem istnieje bieg akceptujący $q_0 q_1 \ldots q_m$ automatu A na $vw^k v' \in L(\alpha)$, gdzie $m = |v| + k|w| + |v'|$. Rozważmy stany $q_{|v|+i|w|}$ dla $i = 0, 1, \ldots, k$. Skoro $k > 2|\alpha|$, to $q_{|v|+i|w|} = q_{|v|+j|w|}$ dla pewnych $i < j$. Używając pompowania widzimy, że automat zaakceptuje każde słowo postaci $vw^{k+\ell(j-i)}v'$, co daje tezę lematu. □

Zdefiniujmy

$$k(\alpha, w) = \sup\{k \mid w^k \sqsubseteq \alpha\}.$$

Na mocy Lematu 2, $k(\alpha, w) \leq 2|\alpha|$ lub $k(\alpha, w) = \infty$.

Lemat 3. *Jeśli $k(\alpha, w) = \infty$, to istnieje takie β^*, podwyrażenie α, że $k(\beta^*, w) = \infty$ i $k(\beta, w) < \infty$.*

Dowód. Łatwo zauważyć, że jeśli γ_0 jest podwyrażeniem γ_1, to $k(\gamma_0, w) \leq k(\gamma_1, w)$. Możemy zatem wybrać minimalne podwyrażenie γ wyrażenia α, takie że $k(\gamma, w) = \infty$. Dla uzasadnienia tezy lematu wystarczy pokazać, że γ jest postaci β^*. Wyrażenie γ nie może być postaci $\beta_1 + \beta_2$, bo $k(\beta_1 + \beta_2, w) = \max(k(\beta_1, w), k(\beta_2, w))$. Podobnie, zauważając, że $k(\beta_1\beta_2, w) \leq k(\beta_1, w) + k(\beta_2, w) + 1$, wykluczamy $\gamma = \beta_1\beta_2$. Stąd γ musi być postaci β^*. □

Twierdzenie wynika wprost z poniższego faktu.

Lemat 4. *Dla każdego n istnieje takie słowo $w \in L(\mathcal{A}_n)$ odpowiadające ścieżce ze stanu 0 do stanu 0, że jeśli $w \sqsubseteq \alpha$, to $|\alpha| \geq 2^{n-1}$.*

Dowód. Dla $n = 1$ wystarczy wystarczy wziąć a_{00}: najkrótsze wyrażenie pokrywające a_{00} musi być niepuste.

Załóżmy, że istnieje słowo $w \in L(\mathcal{A}_{n-1})$ o takich własnościach. Zauważmy, że $w \in L(\mathcal{A}_n)$ i że ścieżka odpowiadająca temu słowu zaczyna się i kończy w stanie 0, i nie przechodzi przez stan $n-1$. Dla $k = 0, 1, 2, \ldots, n-1$ zdefiniujmy w_k jako słowo otrzymane z w przez zamianę każdej litery a_{ij} na literę $a_{i+k,\,j+k}$ (operacje na indeksach interpretujemy modulo n). Mamy wtedy $w_0 = w$. Nie trudno zauważyć, że $w_k \in L(\mathcal{A}_n)$ i ścieżka odpowiadająca w_k nie przechodzi przez stan $k-1$. Pokażemy, że słowo

$$u = (w_0)^{2^n} a_{01} (w_1)^{2^n} a_{12} (w_2)^{2^n} a_{23} \ldots (w_{n-1})^{2^n} a_{n-1,\,0}$$

spełnia założenia dla n.

Weźmy wyrażenie α pokrywające u. Z definicji u, $k(\alpha, w_i) \geq 2^n$ dla wszystkich i. Na mocy Lematu 2, jeśli dla pewnego i wartość $k(\alpha, w_i)$ jest skończona, to $2|\alpha| > k(\alpha, w_i) \geq 2^n$, co daje tezę.

Załóżmy zatem, że $k(\alpha, w_i) = \infty$ dla wszystkich i. na mocy Lematu 3, istnieją podwyrażenia α_i^*, dla których $k(\alpha_i^*, w_i) = \infty$ i $k(\alpha_i, w_i) < \infty$. Z założenia indukcyjnego, $|\alpha_i| \geq 2^{n-2}$. (Istotnie, gdyby istniało krótsze wyrażenie pokrywające w_i, to zamieniając litery otrzymalibyśmy krótsze wyrażenie pokrywające $w_0 = w$.) Niech β^* będzie najkrótszym z tych wyrażeń.

Z Lematu 1 wynika, że istnieje takie j, że każde słowo z β^* jest postaci $a_{jk} \ldots a_{\ell j}$. Inaczej mówiąc, słowa zgodne z β^* odpowiadają ścieżkom zaczynającym i kończącym się w pewnym ustalonym stanie j. Niech $\gamma = \alpha_{j+1}$ (ścieżka odpowiadająca w_{j+1} omija stan j).

Jeśli wyrażenia β i γ są rozłączne, to mamy $|\alpha| > |\beta| + |\gamma| \geq 2^{n-2} + 2^{n-2}$, co daje tezę. Przypuśćmy przeciwnie. Wtedy jedno z wyrażeń musi być podwyrażeniem drugiego. Skoro β jest najkrótsze, to β jest podwyrażeniem γ (być może niewłaściwym). Niech γ_0 będzie wyrażeniem powstałym przez zastąpienie β^* przez ε. Pokażemy, że $w_{j+1} \sqsubseteq \gamma_0$.

Mamy $w_j^* \sqsubseteq \gamma$. Skoro obliczenie odpowiadające w_{j+1} nie przechodzi przez stan j, to nie zawiera żadnej litery postaci $a_{j\ell}$ ani $a_{\ell j}$. Zatem jeśli nie zachodzi $w_{j+1}^* \sqsubseteq \gamma_0$, to musi zachodzić $w_{j+1}^* \sqsubseteq \beta$: w_{j+1}^* musi się zmieścić między dwoma kolejnymi odwiedzinami w stanie j, a każde obliczenie na słowie z $L(\beta)$ zaczyna się i kończy w j. Ale skoro β^* jest podwyrażeniem γ^*, to β jest podwyrażeniem γ. Stąd $w_{j+1}^* \sqsubseteq \gamma = \alpha_{j+i}$, co jest sprzeczne z definicją α_{j+1}.

Mamy zatem $w_{j+1} \sqsubseteq \gamma_0$ i na podstawie założenia indukcyjnego wnioskujemy, że $|\gamma_0| \geq 2^{n-2}$. Zatem $|\alpha| \geq |\gamma| \geq |\gamma_0| + |\beta| \geq 2^{n-2} + 2^{n-2}$. □

Literatura

[1] Andrzej Ehrenfeucht, Paul Zeiger, *Complexity measures for regular expressions*, Proc. STOC 1974:75-79.