AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Własności języków regularnych

Teoria automatów i języków formalnych

Dr inż. Janusz Majewski
Katedra Informatyki

Uzasadnienie lematu o pompowaniu

Jeśli jakiś język jest regularny, to jest on akceptowany przez deterministyczny automat skończony o pewnej określonej liczbie stanów. Załóżmy, że ta liczba stanów wynosi *k*. Rozważamy słowo należące do tego języka o długości *k* lub więcej symboli. Słowo to jest akceptowane przez nasz deterministyczny automat skończony posiadający *k* stanów. Aby było ono zaakceptowane, automat startując ze stanu początkowego musi przeczytać co najmniej *k* symboli i zatrzymać się w pewnym stanie końcowym akceptującym. Słowo wyznacza więc w grafie automatu ścieżkę końcową o liczbie krawędzi co najmniej *k*. Wobec tego liczba stanów (węzłów grafu), które znajdują się na ścieżce końcowej wyznaczonej przez to słowo w grafie automatu, wynosi co najmniej *k+1*. Ponieważ jednak automat posiada tylko *k* stanów, co najmniej jeden stan na ścieżce wyznaczonej przez słowo musi się powtórzyć (musimy dwukrotnie przejść przez co najmniej jeden węzeł grafu). Przechodząc dwukrotnie przez jakiś węzeł (stan) robimy pętlę. Moglibyśmy przejść przez tę pętlę więcej niż jeden raz – w rzeczywistości tyle razy, ile chcemy. Moglibyśmy też nie wchodzić w tę pętlę ani razu – i zawsze doszlibyśmy do stanu akceptującego. Właśnie pokazaliśmy w sposób uproszczony, że jeśli mamy wystarczająco długie słowo akceptowane przez automat skończony, to możemy znaleźć podłańcuch tego słowa (naszą „pętlę”) położony blisko początku tego słowa, który może być „napompowany”, tj. powtórzony tyle razy, ile chcemy, a wynikowe słowo będzie akceptowane przez ten automat skończony.

Lemat o pompowaniu języków regularnych

Niech L będzie językiem regularnym. Wtedy istnieje stała k, taka że jeśli w jest dowolnym słowem z L oraz $|w| \geq k$, to w możemy przedstawić w postaci $w = xuz$, gdzie $|xu| \leq k$, $|u| \geq 1$, oraz xu^iz należy do L dla każdego $i \geq 0$. Co więcej, k nie jest większe niż liczba stanów najmniejszego automatu skończonego akceptującego L.

Formalnie można to zapisać jak niżej:

Twierdzenie: Jeśli L jest językiem regularnym to:

$$(\exists\, k)\ (\ (w \in L\ \wedge\ |w| \geq k) \Rightarrow (w = xuz\ \wedge\ |xu| \leq k\ \wedge\ |u| \geq 1\ \wedge\ (\forall i \geq 0)\ (xu^iz \in L)\ )\ )$$

Warto zauważyć, że lemat o pompowaniu mówi, że jeśli język regularny zawiera długie słowo xuz, to zawiera nieskończony zbiór słów postaci xu^iz.

Przykład (1)

Przykład: Czy język $\{ 0^i 1^i \mid i \geq 0 \}$ jest językiem regularnym?

Nie; przypuśćmy dla dowodu nie wprost, że język ten jest regularny i niech k będzie stałą z lematu o rozrastaniu języków regularnych. Rozważamy słowo $w = 0^k 1^k = xuz$. Słowo w ma długość równą $2k > k$. Wówczas u może zawierać od jednej do maksymalnie k cyfr 0. Inne przypadki nie mogą być brane pod uwagę, gdyż pompowana część u musi być blisko początku słowa ($|xu| \leq k$). Rozważymy łańcuch xu^2z. Zawiera on co najmniej $k+1$ i co najwyżej $2k$ cyfr 0. Wówczas xu^2z nie należy do języka, gdyż liczba cyfr 1 pozostaje bez zmiany. Tak więc xu^2z w żadnym możliwym przypadku nie należy do naszego języka, język ten nie może być regularny.

Przykład (2)

Przykład: Czy język $\{ a^i b^j c^{i+j} \mid i \geq 1, j \geq 1 \}$ jest językiem regularnym?

Nie; przypuśćmy dla dowodu nie wprost, że język ten jest regularny i niech k będzie stałą z lematu o rozrastaniu języków regularnych. Rozważamy słowo $w = a^k b^k c^{2k} = xuz$. Słowo w ma długość równą $4k > k$. Wówczas u może zawierać od jednej do maksymalnie k liter a. Inne przypadki nie mogą być brane pod uwagę, gdyż pompowana część u musi być blisko początku słowa ($|xu| \leq k$). Rozważymy łańcuch xu^2z. Zawiera on co najmniej $k+1$ i co najwyżej $2k$ liter a. Wówczas xu^2z nie należy do języka, gdyż liczba liter b pozostaje bez zmiany, zaś liter c jest zbyt mało. Tak więc xu^2z w żadnym możliwym przypadku nie należy do naszego języka, język ten nie może być regularny.

Definicja domknięcia funkcji przejścia automatu skończonego (niedeterministycznego, z ε-ruchami)

Domknięciem funkcji przejścia automatu skończonego $\delta: Q \times (\Sigma \cup \{\varepsilon\}) \mapsto 2^Q$ nazywamy funkcję:

$$\hat{\delta} : Q \times \Sigma^* \mapsto 2^Q$$

taką, że:

(1) $\hat{\delta}(q, \varepsilon) = \varepsilon\text{–}CLOSURE(q)$

(2) $\hat{\delta}(q, wa) = \varepsilon\text{–}CLOSURE(P)$ dla $q \in Q$, $w \in \Sigma^*$, $a \in \Sigma$ i gdzie

$$P = \bigcup_{r \in R} \delta(r, a), \quad R = \hat{\delta}(q, w)$$

Wartością funkcji $\delta(q,a)$, $a \in \Sigma$ jest zbiór stanów, do których automat może przejść startując ze stanu q przy wejściu będącym pojedynczym symbolem a. Wartością funkcji $\hat{\delta}(q, w)$ jest zbiór stanów, do których automat startując ze stanu q może przejść po przetworzeniu łańcucha w. Dlatego też – jeśli nie będzie to prowadzić do niejednoznaczności – będziemy dalej stosować symbol δ zarówno do oznaczenia funkcji przejścia, jak i domknięcia (uogólnienia) funkcji przejścia (opuszczając daszek).

Własności zamkniętości języków regularnych (1)

Twierdzenie: Klasa języków regularnych jest zamknięta ze względu na sumę teoriomnogościową, złożenie oraz domknięcie Kleene'go. Formalnie, jeśli L_1 i L_2 są językami regularnymi, to językami regularnymi są także $L_1 \cup L_2$, $L_1 L_2$ oraz L_1^*.

Uzasadnienie wynika natychmiast z definicji języka (zbioru) regularnego.

Twierdzenie: Klasa języków regularnych jest zamknięta ze względu operację dopełnienia. Formalnie, jeśli język L jest językiem regularnym, gdzie $L \subseteq \Sigma^*$, to język $\overline{L} = \Sigma^* - L$ też jest językiem regularnym.

Jeżeli język L jest językiem regularnym, to istnieje deterministyczny zupełny automat skończony A akceptujący ten język. Jeżeli w tym automacie stany akceptujące zmienimy na nieakceptujące i na odwrót, to otrzymamy automat akceptujący słowa nie należące do języka L i tylko takie słowa. Zatem automat A będzie akceptował dopełnienie języka L, a więc dopełnienie musi być językiem regularnym.

Własności zamkniętości języków regularnych (2)

Twierdzenie: Klasa języków regularnych jest zamknięta ze względu na iloczyn teoriomnogościowy. Formalnie, jeśli L_1 i L_2 są językami regularnymi, to językami regularnymi jest także $L_1 \cap L_2$.

Zamkniętość ze względu na iloczyn teoriomnogościowy wynika z zamkniętości ze względu na sumę teoriomnogościową oraz dopełnienie (por. prawa de Morgana).

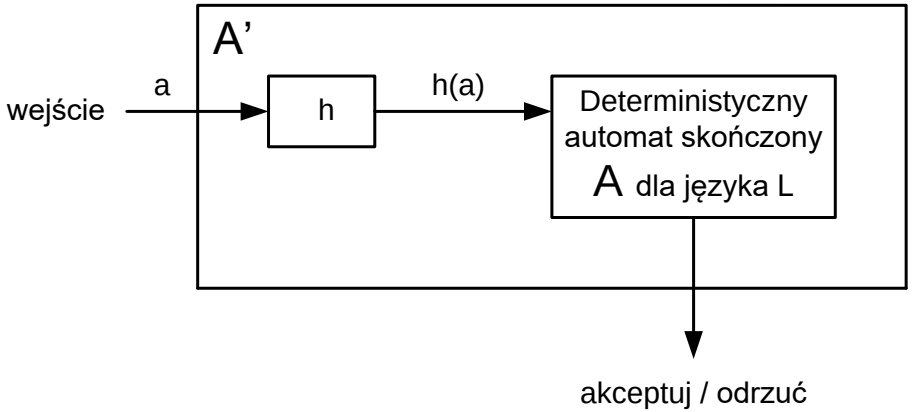
$$L_1 \cap L_2 = \overline{\overline{L_1} \cup \overline{L_2}}$$

Twierdzenie: Klasa języków regularnych jest zamknięta ze względu na operację podstawienia (w tym także homomorfizmu).

Niech $L \subseteq \Sigma^*$ będzie językiem regularnym oraz dla każdego $a \in \Sigma$ niech $L_a \subseteq \Gamma^*$ będzie językiem regularnym. Niech $f: \Sigma \mapsto 2^{\Gamma^*}$ będzie podstawieniem określonym jako $f(a)=L_a$. Zastępując każde wystąpienie symbolu a w wyrażeniu regularnym reprezentującym L wyrażeniem regularnym reprezentującym L_a, otrzymamy nowe wyrażenie regularne. Aby dowieść, że reprezentuje ono $f(L)$, zauważamy, że podstawienie sumy teoriomnogościowej, złożenia i domknięcia jest odpowiednio sumą teoriomnogościową, złożeniem i domknięciem podstawień. Tak więc $f(L)$ jest regularny.

Własności zamkniętości języków regularnych (3)

Twierdzenie: Klasa języków regularnych jest zamknięta ze względu na przeciwobrazy homomorficzne.



Załóżmy, że A = <Q, Σ, δ, q_0, F> jest deterministycznym automatem skończonym akceptującym język regularny L, zaś $h: \Gamma \mapsto \Sigma^*$ jest homomorfizmem. Konstruujemy deterministyczny automat skończony $A' = <Q, \Gamma, \delta', q_0, F>$, gdzie $\delta'(q, a) = \delta(q, h(a))$ dla każdego $q \in Q$, $a \in \Gamma$. Wówczas można pokazać, że $\delta'(q_0, x) = \delta(q_0, h(x))$. Zatem A' akceptuje x wtedy i tylko wtedy, gdy A akceptuje $h(x)$. Inaczej mówiąc język regularny akceptowany przez A' to $L(A') = h^{-1}(L(A))$, czyli klasa języków regularnych jest zamknięta ze względu na przeciwobrazy homomorficzne.

Własności zamkniętości języków regularnych (4)

Twierdzenie: Klasa języków regularnych jest zamknięta ze względu na ilorazy (dzielenie przez dowolne zbiory).

Załóżmy, że A = <Q, Σ, δ, q_0, F> jest automatem skończonym akceptującym język regularny L, zaś M jest dowolnym językiem. Iloraz L/M jest akceptowany przez automat skończony $A' = <Q, \Sigma, \delta, q_0, F'>$, który zachowuje się jak A z tym jednym wyjątkiem, że stanami końcowymi w A' są wszystkie stany q automatu A, dla których istnieje $y \in M$ takie, że $\delta(q, y) \in F$. Przy tym założeniu $\delta(q_0, x) \in F'$ wtedy i tylko wtedy, gdy istnieje $y \in M$ takie, że $\delta(q_0, xy) \in F$. Tym samym A' akceptuje L/M, więc L/M jest językiem regularnym.

Niestety, podany szkic dowodu nie jest konstruktywny, gdyż nie podaje efektywnego algorytmu wyznaczania stanów końcowych automatu A'. Efektywny algorytm istnieje dla M będącego językiem regularnym, ale nie dla M dowolnego.

Przykład – idealne przetasowanie (1)

Idealne przetasowanie języków L_A i L_B definiujemy następująco:

Niech $L_A, L_B \subseteq \Sigma^*$

$$\text{Perfect\_Shuffle}(L_A, L_B) = \{ w \mid w = a_1b_1a_2b_2\ldots a_kb_k,$$
$$a_1a_2\ldots a_k \in L_A,\ b_1b_2\ldots b_k \in L_B, \text{ dla } a_i, b_i \in \Sigma \}$$

Czy klasa języków regularnych jest zamknięta ze względu na idealne przetasowanie?

Przykład – idealne przetasowanie (2)

Tak, weźmy bowiem:

$\Delta = \Sigma \times \Sigma$

Elementami alfabetu Δ są dwójki (a,b) takie że $a,b \in \Sigma$.

Rozważymy trzy homomorfizmy:

left: $\Delta \rightarrow \Sigma^*$ taki że left((a,b)) = a

right: $\Delta \rightarrow \Sigma^*$ taki że right((a,b)) = b

unpair: $\Delta \rightarrow \Sigma^*$ taki że unpair((a,b)) = ab

Widzimy, że:

$L_1 = left^{-1}(L_A) = \{(a_1,b_1)\,(a_2,b_2)\ldots(a_n,b_n) \mid (a_i,b_i) \in \Delta,\ a_1a_2\ldots a_n \in L_A\}$

$L_2 = right^{-1}(L_B) = \{(a_1,b_1)\,(a_2,b_2)\ldots(a_n,b_n) \mid (a_i,b_i) \in \Delta,\ b_1b_2\ldots b_n \in L_B\}$

Przykład – idealne przetasowanie (3)

Mamy dalej:

$$L_3 = L_1 \cap L_2 = \{ (a_1,b_1)\,(a_2,b_2)\ldots(a_n,b_n) \mid (a_i,b_i) \in \Delta,\ a_1a_2\ldots a_n \in L_A,\ b_1b_2\ldots b_n \in L_B \}$$

oraz

$$\text{unpair}(L_3) = \text{Perfect-Shuffle}(L_A, L_B) = \{ w \mid w = a_1b_1a_2b_2\ldots a_kb_k,\ a_1a_2\ldots a_k \in L_A,\ b_1b_2\ldots b_k \in L_B, \text{ dla } a_i, b_i \in \Sigma \}$$

Jeśli L_A i L_B są językami regularnymi, to z uwagi na zamkniętość klasy języków regularnych ze względu na przeciwobrazy homomorficzne, przecięcie i homomorfizmy, Perfect-Shuffle(L_A, L_B) jest także językiem regularnym, więc klasa języków regularnych jest zamknięta ze względu na idealne przetasowanie.

Inny przykład (1)

Rozważymy język:

$L = \{ a^i b^j c^k \mid i,j,k \geq 0 \text{ oraz jeśli } i = 1 \text{ to } j = k \}$

Ten język nie jest regularny, bowiem:

$L \cap \mathbf{ab^*c^*} = \{ab^n c^n \mid n \geq 0 \}$

Zdefiniujmy teraz homomorfizm $h: \{a,b,c\}^* \rightarrow \{0,1\}^*$

$h(a) = \varepsilon; \quad h(b) = 0; \quad h(c) = 1$

Wtedy

$h(L) = \{ 0^n 1^n \mid n \geq 0 \}$

Język h(L) nie jest regularny, jak wykazano wcześniej na wykładzie. Skoro h(L) nie jest regularny, to także L nie jest regularny, gdyż h(L) został otrzymany z L poprzez przecięcie i homomorfizm, które to są operacjami zachowującymi regularność.

Innny przykład (2)

Ale każde odpowiednio długie słowo języka $L = \{ a^i b^j c^k \mid i,j,k \geq 0$ oraz jeśli $i = 1$ to $j = k\}$ da się napompować. Weźmy bowiem stałą $k = 3$. Rozważamy dowolne słowo $w \in L$ takie że $|w| \geq 3$. Pokażemy, że każde takie w da się przedstawić w postaci $w = xuy$ i da się napompować.

- Jeśli $i \neq 2$ to dzielimy słowo w tak: $x = \varepsilon$, u jest pierwszym symbolem słowa w, zaś y jest resztą słowa w. Mamy, $w=xuy$, $|xu|<3$, $|u|>0$. Obserwujemy, że łańcuch $xu^p y$, gdzie $p \neq 1$ ma tę własność, że liczba liter a jest różna od 1, więc $xu^p y \in L$ dla każdego p.
- Jeśli $i = 2$ to dzielimy słowo w tak: $x = aa$, u jest pierwszym symbolem występującym po części x, zaś y jest resztą słowa w. Mamy, $w=xuy$, $|xu| \leq 3$, $|u|>0$. i teraz łańcuch $xu^p y$ dla każdego p posiada dwie litery a na początku, więc $xu^p y \in L$ dla każdego p.

Inny przykład (3)

Lemat o pompowaniu mówi, że każdy język regularny spełnia warunki lematu (jego tezę), ale nie mówi, że tylko języki regularne spełniają te warunki. W niniejszym przykładzie pokazaliśmy, że język L nie będący językiem regularnym też spełnia tezę lematu o pompowaniu języków regularnych.

Twierdzenie Myhilla-Nerode'a (1)

Twierdzenie:

Następujące warunki są równoważne:

(1) Język $L \subseteq \Sigma^*$ jest akceptowany przez pewien deterministyczny zupełny automat skończony.

(2) Język L jest sumą teoriomnogościową pewnych klas abstrakcji pewnej prawostronnie niezmienniczej relacji równoważności o indeksie skończonym.

(3) Relacja R_L indukowana przez język L jest relacją o indeksie skończonym.

Twierdzenie Myhilla-Nerode'a (2)

(1) $\Rightarrow$ (2)

Niech $A = <Q, \Sigma, F, q_0, \delta>$ będzie deterministycznym zupełnym automatem skończonym akceptującym język L.

Zdefiniujemy relację $R_A \subseteq \Sigma^* x \Sigma^*$ taką, że $xR_Ay \equiv \delta(q_0,x)=\delta(q_0,y)$. Relacja R_A jest relacją prawostronnie niezmienniczą, gdyż dla dowolnych x i y, jeśli xR_Ay, tzn. jeśli $\delta(q_0,x)=\delta(q_0,y)$, to dla dowolnego słowa z zachodzi:

$$\delta(q_0,xz)=\delta(\delta(q_0,x),z)=\delta(\delta(q_0,y),z)=\delta(q_0,yz)$$

Relacja R_A jest też relacją równoważności (proszę sprawdzić). Relacja R_A jest relacją o indeksie skończonym, ponieważ indeks relacji (liczba klas równoważności) nie przekracza liczby stanów automatu A. Wynika to z faktu, że dowolne dwa słowa, dla których przetwarzanie kończy się w tym samym stanie, są ze sobą w relacji R_A. Wynika z tego, że zbiór wszystkich słów, dla których przetwarzanie kończy się w pewnym stanie, zawiera się w pewnej klasie abstrakcji relacji R_A. Tak określone zbiory słów stanowią podział zbioru wszystkich słów nad alfabetem Σ. Co więcej zbiory te są klasami abstrakcji, gdyż dowolne dwa słowa, dla których przetwarzanie przez automat kończy się w różnych stanach, nie są ze sobą w relacji R_A.

Oczywiście język L jest sumą zbiorów tych słów, dla których przetwarzanie kończy się w stanach akceptujących automatu A.

Twierdzenie Myhilla-Nerode'a (3)

(2)$\Rightarrow$(3)

Niech ρ będzie prawostronnie niezmienniczą relacją równoważności o indeksie skończonym. Niech język L będzie sumą pewnych klas abstrakcji relacji ρ. Pokażemy, że każda klasa abstrakcji relacji ρ jest zawarta w pewnej klasie abstrakcji relacji R_L indukowanej przez język L. Niech $x,y \in \Sigma^*$ i niech $x \rho y$ (co oczywiście oznacza, że x i y należą do tej samej klasy abstrakcji relacji ρ). Wobec tego $(\forall z \in \Sigma^*)\ xz\ \rho\ yz$ – na mocy prawostronnej niezmienniczości relacji ρ. To znaczy, że dla dowolnego $z \in \Sigma^*$ oba słowa xz i yz należą do jakiejś jednej klasy abstrakcji. Ponieważ język L jest sumą teoriomnogościową pewnych klas abstrakcji relacji ρ, to dla danego słowa z albo oba słowa xz i yz należą do języka L, albo oba słowa nie należą do języka L. To wyczerpuje definicję relacji R_L indukowanej przez język L. Ostatecznie dla dowolnych słów x i y, jeżeli pozostają w relacji ρ, to pozostają też w relacji R_L. Czyli każda klasa abstrakcji relacji ρ zawiera się w pewnej klasie abstrakcji relacji R_L, co oznacza, że indeks relacji R_L jest nie większy niż indeks relacji ρ, czyli skończony.

Twierdzenie Myhilla-Nerode'a (4)

(3)$\Rightarrow$(1)

Niech L będzie językiem nad alfabetem Σ, niech R_L będzie relacją o indeksie skończonym indukowaną przez język L. Następujący automat będzie akceptował język L:

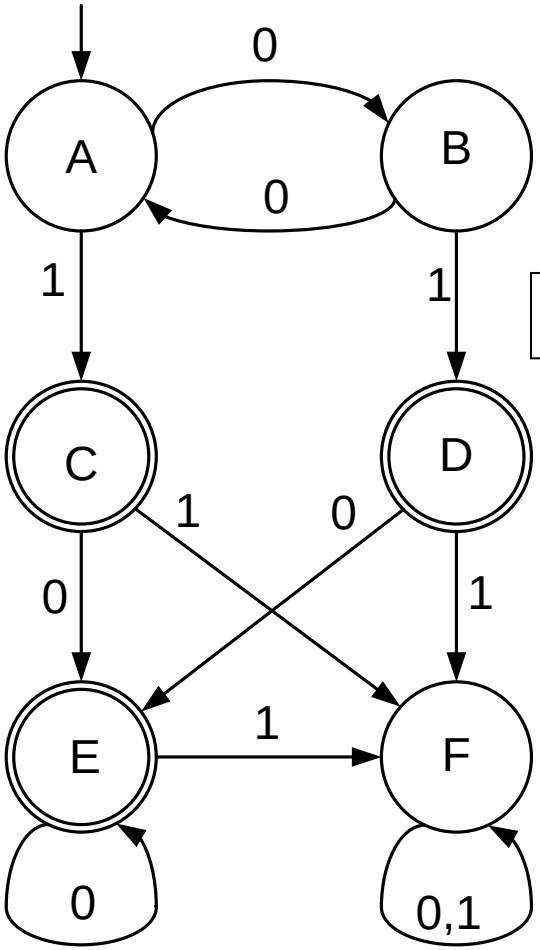
$A = <Q, \Sigma, F, q_0, \delta>$

gdzie:

- zbiór stanów Q jest zbiorem indeksowanym klasami abstrakcji relacji R_L,
- stanem początkowym q_0 jest stan indeksowany klasą abstrakcji zawierającą słowo puste $q_{[\varepsilon]}$,
- stanami akceptującymi są stany indeksowane klasami abstrakcji zawartymi w języku L,
- funkcja przejścia jest zdefiniowana następująco: $\delta(q_{[w]},a)=q_{[wa]}$, gdzie $w \in \Sigma^*$, $a \in \Sigma$ oraz $[w]$ i $[wa]$ są klasami abstrakcji o reprezentantach w i wa.

Tak skonstruowany automat A jest deterministycznym zupełnym i minimalnym automatem skończonym akceptującym język L.

Przykład – język **0*10***



⇒

Relacja R_A:

K_A = **(00)***

K_B = **(00)*0**

K_C = **(00)*1**

K_D = **(00)*01**

K_E = **0*100***

K_F = **0*10*1(0|1)***

⇒

Relacja R_L:

K_1 = **0***

K_2 = **0*10***

K_3 = **0*10*1(0|1)***

⇓