

Języki, automaty i obliczenia

Wykład 1: Języki regularne

Sławomir Lasota

Uniwersytet Warszawski

25 lutego 2015

1 Słowa, języki

2 Języki regularne

- A – niepusty skończony zbiór, zwany *alfabetem*.
- *Słowo* nad A to dowolny skończony ciąg elementów zbioru A . Formalnie, słowo to funkcja $w : \{1 \dots n\} \rightarrow A$.
- *Słowo puste* ε to pusty ciąg, albo pusta funkcja $\emptyset \rightarrow A$.

Przykład

Niech $A = \{a, b\}$. Palindrom $w = abba$, $|w| = 4$.

Notacja

A^* - zbiór wszystkich słów nad A .

A^+ - zbiór wszystkich niepustych słów nad A .

A^n - zbiór wszystkich słów długości n nad A .

Przykład

$$w_0 = b$$

$$w_1 = a$$

$$w_2 = ab$$

$$w_3 = aba$$

$$w_4 = abaab$$

$$w_5 = abaababa$$

$$w_6 = abaababaabaab$$

$$w_7 = abaababaabaababaababa$$

$$w_8 = \dots$$

- *Konkatenacja* dwóch słów

$$w = a_1 \dots a_n \in A^n, v = b_1 \dots b_m \in A^m,$$

oznaczana $w \cdot v$, to słowo $a_1 \dots a_n b_1 \dots b_m$ długości $n + m$.

- Element neutralny:

$$\varepsilon \cdot w = w \cdot \varepsilon = w$$

- Łączność:

$$(w \cdot v) \cdot u = w \cdot (v \cdot u)$$

- Monoid słów $(A^*, \cdot, \varepsilon)$, czyli monoid wolny nad A .

Notacja

Będziemy pomijać symbol konkatenacji i pisać wv zamiast $w \cdot v$.

- Jeśli $u = wv$, to w nazywamy *prefiksem* u , a v *sufiksem* u .
- Relacja prefiksu (sufiksu) jest porządkiem częściowym w A^* .

- Język nad alfabetem A to dowolny podzbiór zbioru A^* , $L \subseteq A^*$.

Przykład

$$A = \{a, b\}$$

$$L = \{w_n : n \in \mathbb{N}\} = \text{Fib.}$$

$$A = \{a\}$$

$$L = \{w \in A^* : |w| \text{ jest liczbą pierwszą}\}.$$

$$A = \{0, 1\}$$

$$L = \{\text{bin}(n) : n \text{ jest liczbą pierwszą}\}.$$

- Każdy język jest przeliczalny, ale języków jest nieprzeliczalnie wiele.

- Konkatenacja języków:

$$LM = \{wv : w \in L, v \in M\}$$

- Konkatenacja jest *rozdzielna* względem sumy (ale nie względem przecięcia)

$$L(M \cup N) = LM \cup LN$$

Pytanie

Łączność? Element neutralny?

Notacja

Dla $n \geq 0$, potęga $L^n = \underbrace{L \dots L}_{n \text{ razy}}$. W szczególności $L^0 = \{\varepsilon\}$.

Definicja indukcyjna:

$$L^0 = \{\varepsilon\}, \quad L^{n+1} = L^n L$$

A^n to szczególny przypadek, jeśli utożsamimy A ze zbiorem wszystkich słów długości 1.

Iloraz lewostronny i prawostronny:

$$M^{-1}L = \{w : \exists v \in M. vw \in L\} \quad LM^{-1} = \{w : \exists v \in M. wv \in L\}$$

Uwaga: M^{-1} nic nie oznacza.

Pytanie

$$\{a\}\{a\}^{-1}L = ?$$

Pytanie

Czy z $LM = N$ wynika $L^{-1}N = M$?

Operacja *iteracji* (gwiazdki):

$$L^* = \bigcup_{n \in \mathbb{N}} L^n$$

Przykład

$A = \{a\}$	$\{aa\}^*$ słowa długości parzystej	$(aa)^*$
$A = \{a, b\}$	$\{a, ab\}^*$ słowa nie zaczynające się od b , w których litery b nie leżą nigdy obok siebie.	$(a + ab)^*$
$A = \{a, \neg a, b, c\}$	$(\{a\}\{b\})^* \{\neg a\}\{c\}$	$(ab)^* \neg a c$

A^* to szczególny przypadek, jeśli utożsamimy A ze zbiorem wszystkich słów długości 1.

Pytanie

$\emptyset^* = ?$

Języki będziemy utożsamiać z *zadaniami obliczeniowymi*, albo *problemami decyzyjnymi*.

Przykład

Dane wejściowe: graf skierowany G

Wynik: rozstrzygnąć, czy G ma cykl Hamiltona?

Graf można opisać jako słowo nad $A = \{0, 1, \#\}$, np. 110#101#001

Język $L \subseteq A^*$ można utożsamić z następującym zadaniem obliczeniowym:

Dane wejściowe: $w \in A^*$

Wynik: rozstrzygnąć, czy $w \in L$?

- *Homomorfizm* to funkcja $h : A^* \rightarrow B^*$ która zachowuje słowo puste i konkatencję:

$$h(\varepsilon) = \varepsilon \quad h(wv) = h(w)h(v).$$

Innymi słowy, to homomorfizm monoidów.

- Homomorfizm jest jednoznacznie wyznaczony przez swoje wartości dla słów jednoliterowych, czyli przez funkcję

$$A \rightarrow B^*.$$

Przykład

$$A = \{a, b\}$$

$$B = \{c, d\}$$

h wyznaczony przez $(a \mapsto cc, b \mapsto \varepsilon)$

$$h(abaab) = cccccc$$

$$\vec{h}(\text{Fib}) = ?$$

$$\vec{h}^{-1}(\{d\}^*) = \{b\}^*$$

- Funkcja $\hat{h} : \mathcal{P}(A^*) \rightarrow \mathcal{P}(B^*)$ wyznaczona jednoznacznie przez

$$h : A \rightarrow \mathcal{P}(B^*)$$

- rozszerzamy h do wszystkich słów:

$$\hat{h}(\varepsilon) = \{\varepsilon\}$$

$$\hat{h}(aw) = h(a) \hat{h}(w)$$

- rozszerzamy $\hat{h}$ do języków:

$$\hat{h}(L) = \bigcup_{w \in L} \hat{h}(w)$$

Przykład

$$A = \{a, b, x, y, \oplus\}$$

$$B = \{2, 3, 4, x, y\}$$

$$w = ax \oplus ax \oplus by \oplus ax$$

$$L = ((ax + by) \oplus)^*(ax + by)$$

$$h : a \mapsto \{2, 4\}, b \mapsto \{\varepsilon, 3\}$$

$$x \mapsto \{x\}, y \mapsto \{y\}, \oplus \mapsto \{\oplus\}$$

$$\hat{h}(w) = \{2x \oplus 4x \oplus y \oplus 2x, \dots\}$$

$$\hat{h}(L) = ?$$

1 Słowa, języki

2 Języki regularne

Języki regularne nad A to najmniejszy zbiór języków $\mathcal{R}$ spełniający poniższe warunki:

- $\emptyset \in \mathcal{R}$,
- $\{\varepsilon\} \in \mathcal{R}$,
- $\{a\} \in \mathcal{R}$, dla każdego $a \in A$,
- jeśli $L \in \mathcal{R}$ i $M \in \mathcal{R}$, to $LM \in \mathcal{R}$ oraz $L \cup M \in \mathcal{R}$,
- jeśli $L \in \mathcal{R}$, to $L^* \in \mathcal{R}$.

Pytanie

Który z warunków jest nadmiarowy?

Przykład

$$A = \{a, b\}$$

$$\begin{aligned} \{a, ab\}^* &= (\{a\} \cup \{a\}\{b\})^* \\ &\{b\}\{a\}^*\{b\} \cup \{\varepsilon\} \end{aligned}$$

$$\begin{aligned} (a + ab)^* \\ ba^*b + \varepsilon \end{aligned}$$

Konwencja notacyjna

- $+$ zamiast $\cup$
- nie używamy „wąsów” $\{, \}$
- $*$ wiąże silniej niż konkatenacja, a ta silniej niż $+$

Przykład

$\{a\}$	a
$\{a\}\{b\}\{c\}$	abc
$\{a\} \cup \{\varepsilon\}$	$a + \varepsilon$
$(\{b\} \cup \{\varepsilon\})(\{a\} \cup \{a\}\{b\})^*$	$(b + \varepsilon)(a + ab)^*$

Utożsamiamy:

- literę a z językiem zawierającym jedno jednoliterowe słowo $\{a\}$
- słowo ε z językiem $\{\varepsilon\}$

- Łączność i przemienność:

$$L + (M + N) = (L + M) + N$$

$$L(MN) = (LM)N$$

$$L + M = M + L$$

ale nie:

$$LM = ML$$

- Idempotentność:

$$L + L = L$$

- Elementy neutralne i zerowe:

$$L\varepsilon = L$$

$$L\emptyset = \emptyset$$

$$L + \emptyset = L$$

- Prawo rozdzielności:

$$L(M + N) = LM + LN$$

- Prawa dla gwiazdki:

$$(L^*)^* = L^*$$

$$\emptyset^* = \varepsilon$$

$$\varepsilon^* = \varepsilon$$

$$(L^* M^*)^* = (L + M)^*$$

$$L(ML)^* = (LM)^* L$$

$$(L + \varepsilon)^* = L^*$$

- Fałszywki:

$$(L + M)^* = L^* + M^*$$

$$L + \varepsilon = L$$

$X \subseteq \mathbb{N}$ jest *prawie okresowy* jeśli

$$\exists o \in \mathbb{N}. \exists n_0 \in \mathbb{N}. \forall n > n_0. n \in X \iff (n + o) \in X.$$

Twierdzenie

Gdy A jest jednoelementowy, język $L \subseteq A^*$ jest regularny wtw gdy

$\{|w| : w \in L\}$ jest prawie okresowy.

Twierdzenie

Gdy A jest jednoelementowy i $L \subseteq A^*$, język L^* jest regularny.

Dlaczego języki regularne?

- prosta klasa, która zawiera wiele naturalnych języków
- wiele równoważnych definicji (np. *rozszerzone wyrażenia regularne*)
- klasa zamknięta na wiele naturalnych operacji
- zastosowania praktyczne, np.:
 - wyszukiwanie wzorca (grep)
 - analiza leksykalna

Wyrażenia regularne w Unixie:

.	$a + b + \dots$
[a b c d] [a-d]	$a + b + c + d$
[:digit:]	$0 + 1 + \dots + 9$
_ _	$_ + _$
e?	$e + \varepsilon$
e+	$e^* e$
e{n}	$\underbrace{e \dots e}_{n \text{ razy}}$
...	

Analiza leksykalna:

while	return(WHILE);
[A-Za-z] [A-Za-z0-9]*	return(ID);
>=	return(GE);
...	

Twierdzenie

Klasa języków regularnych jest zamknięta na

- *dopełnienie: jeśli L jest regularny to dopełnienie L też*
- *przecięcia: jeśli L i M są regularne to $L \cap M$ też*
- *obrazy homomorficzne: jeśli L jest regularny to $\vec{h}(L)$ też*
- *podstawienia: jeśli L jest regularny to $\hat{h}(L)$ też*
- *przeciwbrazy homomorficzne: jeśli L jest regularny to $\vec{h}^{-1}(L)$ też*
- ...

Dowód dla obrazów homomorficznych:

$$\begin{aligned}\vec{h}(LM) &= \vec{h}(L)\vec{h}(M) \\ \vec{h}(L \cup M) &= \vec{h}(L) \cup \vec{h}(M) \\ \vec{h}(L^*) &= (\vec{h}(L))^*\end{aligned}$$

- Dodajemy operację dopełnienia $\bar{L} = A^* - L$.
- Rozszerzone wyrażenia regularne są równoważne wyrażeniom regularnym, ale bardziej zwarte.

Przykład

$$A = \{0, 1, \#\} \quad v_n = \#\text{bin}(0)\#\text{bin}(1)\#\dots\#\text{bin}(2^n - 1)\#$$

Dopełnienie języka $\{v_n\}$ opisuje wyrażenie regularne rozmiaru $\text{poly}(n)$.

Najkrótsze wyrażenie regularne opisujące $\{v_n\}$ jest wykładnicze względem n .
Dlaczego?

Jaki to język?

$$(aa + bb + (ab + ba)(aa + bb)^*(ab + ba))^*$$

