Języki, automaty i obliczenia

Wykład 1: Języki regularne

Sławomir Lasota

Uniwersytet Warszawski

2 marca 2016

Plan

1. Słowa, języki
2. Języki regularne

Słowa

- A – niepusty skończony zbiór, zwany *alfabetem*.
- *Słowo* nad A to dowolny skończony ciąg elementów zbioru A. Formalnie, słowo to funkcja $w : \{1 \dots n\} \to A$.
- *Słowo puste* ε to pusty ciąg, albo pusta funkcja $\emptyset \to A$.
- Długość słowa w, np. $|aba| = 3$.

Przykład

$A = \{a, b\}$

$w_0 = b$ $\quad w_1 = a$ $\quad w_2 = ab$ $\quad w_3 = aba$ $\quad w_4 = abaab$ $\quad w_5 = abaababa$ $\quad w_6 = abaababaabaab$ $\quad w_7 = abaababaabaababaababa$ $\quad w_8 = \dots$

Notacja

A^* - zbiór wszystkich słów nad A.
A^+ - zbiór wszystkich niepustych słów nad A.
A^n - zbiór wszystkich słów długości n nad A.

Konkatenacja słów

- *Konkatenacja* dwóch słów

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

$$w \cdot v = a_1 \dots a_n b_1 \dots b_m, \quad |w \cdot v| = n + m.$$

- 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

$A = \{a, b\}$	$L = \{w_n : n \in \mathbb{N}\} = \text{Fib}.$
$A = \{a\}$	$L = \{w \in A^* : \lvert w \rvert \text{ jest liczbą pierwszą}\}.$
$A = \{0, 1\}$	$L = \{\text{bin}(n) : n \text{ jest liczbą pierwszą}\}.$

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

Konkatenacja języków

- 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:

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

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

Operacja odwrotna do konkatenacji?

Iloraz lewostronny i prawostronny:

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

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

Przykład

$$\{a^n : n \in \mathbb{N}\}^{-1}\{a^n b^n : n \in \mathbb{N}\} = \{a^m b^n : m, n \in \mathbb{N}, m \leq n\}$$

Pytanie

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

Iteracja

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 lezą nigdy obok siebie	$(a + ab)^*$
$A = \{a, \neg a, b, c\}$	$(\{a\}\{b\})^*\{\neg a\}\{c\}$	$(a\,b)^*\neg a\,c$

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

Pytanie

$\emptyset^* = ?$

Problemy decyzyjne

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 $w \in L$?

Homomorfizmy

- *Homomorfizm* to funkcja $h : A^* \to B^*$, która zachowuje konkatenację:

$$h(wv) = h(w)h(v) \qquad \text{(w szczególności } h(\varepsilon) = \varepsilon\text{)}.$$

  Innymi słowy, to homomorfizm monoidów.

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

$$A \to B^*.$$

Przykład

$A = \{a, b\}$	$h(\text{abaab}) = \text{cccccc}$
$B = \{c, d\}$	$\vec{h}(\text{Fib}) = ?$
h wyznaczony przez $(a \mapsto cc, b \mapsto \varepsilon)$	$\vec{h}^{-1}(\{d\}^*) = \{b\}^*$

Podstawienia

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

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

- rozszerzamy h do wszystkich słów:

$$\begin{aligned} \widehat{h}(\varepsilon) &= \{\varepsilon\} \\ \widehat{h}(aw) &= h(a)\,\widehat{h}(w) \end{aligned}$$

- rozszerzamy $\widehat{h}$ do języków:

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

Przykład

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

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

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

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

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

$\widehat{h}(w) = \{2x \oplus 4x \oplus y \oplus 2x, \ \ldots\}$

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

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

Plan

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 $L \cup M \in \mathcal{R}$,
- jeśli $L \in \mathcal{R}$ i $M \in \mathcal{R}$, to $LM \in \mathcal{R}$,
- jeśli $L \in \mathcal{R}$, to $L^* \in \mathcal{R}$.

Pytanie

Który z warunków jest nadmiarowy?

Przykład

$A = \{a, b\}$

$\{a, ab\} = \{a\} \cup \{a\}\{b\}$	$a + ab$
$\{b\}\{a\}^*\{b\} \cup \{\varepsilon\}$	$b\,a^*\,b + \varepsilon$
$\{a\}^*(\{b\}\{a\}^*\{b\}\{a\}^*)^*$	$a^*(b\,a^*\,b\,a^*)^*$

Wyrażenia regularne

Wyrażenie regularne to wyrażenie zbudowane z:

$$\emptyset \qquad \varepsilon \qquad a \qquad \_+\_ \qquad \_\cdot\_ \qquad \_^*$$

Konwencja notacyjna

- $+$ zamiast $\cup$
- nie używamy „wąsów" $\{\ \}$, czyli a oznacza język zawierający jedno jednoliterowe słowo $\{a\}$
- ε oznacza język $\{\varepsilon\}$
- * wiąże silniej niż konkatenacja, a ta silniej niż $+$

Przykład

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

Tożsamości

- Łączność i przemienność:

$$\begin{aligned} L + (M + N) &= (L + M) + N \\ L(MN) &= (LM)N \\ L + M &= M + L \end{aligned}$$

  ale nie:

$$LM = ML$$

- Idempotentność:

$$L + L = L$$

- Elementy neutralne i zerowe:

$$\begin{aligned} L\varepsilon &= \varepsilon L = L \\ L\emptyset &= \emptyset L = \emptyset \\ L + \emptyset &= \emptyset + L = L \end{aligned}$$

Tożsamości (c.d.)

- Prawo rozdzielności:

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

- Prawa dla gwiazdki:

$$\begin{aligned}(L^*)^* &= L^*\\ \emptyset^* &= \varepsilon\\ \varepsilon^* &= \varepsilon\\ (L^*M^*)^* &= (L+M)^*\\ L(ML)^* &= (LM)^*L\\ (L+\varepsilon)^* &= L^*\end{aligned}$$

- Fałszywki:

$$\begin{aligned}(L+M)^* &= L^*+M^*\\ L+\varepsilon &= L\end{aligned}$$

Alfabet jednoliterowy

$X \subseteq \mathbb{N}$ 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

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

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

Zastosowania

Wyrażenia regularne w Unixie:

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

Analiza leksykalna:

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

Operacje zachowujące regularność

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*
- *przeciwobrazy homomorficzne: jeśli L jest regularny to $\vec{h}^{-1}(L)$ też*
- *podstawienia regularne: jeśli L jest regularny to $\widehat{h}(L)$ też*
- *...*

Dowód dla obrazów homomorficznych:

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

Równoważna definicja: rozszerzone wyrażenia regularne

- Dodajemy operację dopełnienia $\overline{L} = A^* - L$.

Przykład

$$A = \{a, b\} \qquad L = \overline{a^* b\, a^* b\, a^*}$$

- Rozszerzone wyrażenia regularne są równoważne wyrażeniom regularnym, ale bardziej zwięzłe.

Przykład

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

- Dopełnienie języka $\{v_n\}$ można opisać wyrażeniem regularnym rozmiaru poly(n).
- Najkrótsze wyrażenie regularne opisujące $\{v_n\}$ jest wykładnicze względem n. Dlaczego?

Głębokość gwiazdkowa (*ang. star height*)

Głębokość gwiazdkowa wyrażenia regularnego e:

$$e \;=\; (a\,b^*\,c)^* \qquad \mathrm{gg}(e) \;=\; ?$$

Głębokość gwiazdkowa języka regularnego L:

$$\mathrm{gg}(L) \;=\; \min\{\,\mathrm{gg}(e) \;:\; \text{wyrażenie } e \;\; \text{definiuje } L\,\}$$

Pytanie (Eggan 1963)

- Czy języki regularne mają ograniczoną głębokość gwiazdkową?
- Czy dla danego języka regularnego można obliczyć jego głębokość gwiazdkową?

Twierdzenie (Hashiguchi 1988)

Głębokość gwiazdkowa języka regularnego jest obliczalna.

Obydwa pytania są otwarte dla rozszerzonych wyrażeń regularnych!

Co w następnym odcinku: automaty skończone

Jaki to język?

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