



AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Zbiory łańcuchów, języki

Języki formalne i automaty

Dr inż. Janusz Majewski
Katedra Informatyki



Wprowadzenie

- Dowiedzieliśmy się, że symbol jest pojęciem pierwotnym, niedefiniowanym
- Zdefiniowaliśmy alfabet jako skończony i niepusty zbiór symboli
- Wiemy, że z symboli alfabetu można budować łańcuchy (słowa, napisy)
- Obecnie będziemy rozważać zbiory łańcuchów, które później nazwiemy językami (formalnymi)



Składanie zbiorów łańcuchów (1)

Niech U_1 i U_2 będą zbiorami łańcuchów nad alfabetami odpowiednio Σ_1 i Σ_2 . Złożeniem U_1U_2 tych zbiorów jest zbiór zawierający łańcuchy postaci x_1x_2 , gdzie $x_1 \in U_1$, zaś $x_2 \in U_2$.

$$U_1U_2 = \{ x_1x_2 \mid x_1 \in U_1, x_2 \in U_2 \}$$

Złożenie zbiorów łańcuchów jest:

- nieprzemienne (na ogół $U_1U_2 \neq U_2U_1$)
- łączne ($U_1U_2U_3 = (U_1U_2)U_3 = U_1(U_2U_3)$)
- posiada element neutralny $\{\varepsilon\}$ będący zbiorem jednoelementowym, którego jedynym elementem jest łańcuch pusty ($U\{\varepsilon\} = \{\varepsilon\}U = U$)



Składanie zbiorów łańcuchów (2)

Przykład składania zbiorów łańcuchów:

$$\Sigma_1 = \{a, b\}$$

$$\Sigma_2 = \{a, b, n, r\}$$

$$U_1 = \{a, ba\}$$

$$U_2 = \{rnaba, rab\}$$

$$U_1U_2 = \{arnaba, barnaba, arab, barab\}$$

$$U_2U_1 = \{rnabaa, rnababa, raba, rabba\}$$



Notacja potęgowa

Notacja „potęgowa”

Niech U będzie zbiorem łańcuchów. Wtedy zbiory będące wynikiem kolejnego składania zbioru U z samym sobą oznaczamy w uproszczeniu:

$$U^0 = \{\varepsilon\}$$

$$U^1 = U$$

$$U^2 = UU$$

$$U^3 = UUU, \dots, \text{ itd.}$$

Definiujemy dalej:

$$U^* = U^0 \cup U^1 \cup U^2 \cup U^3 \cup \dots$$

$$U^+ = U^1 \cup U^2 \cup U^3 \cup \dots$$



Zbiór słownikowy (1)

Zbiór wszystkich łańcuchów nad alfabetem

Zbiór wszystkich łańcuchów nad alfabetem Σ oznaczamy Σ^* . Zbiór wszystkich niepustych łańcuchów nad alfabetem Σ oznaczamy Σ^+ .

Przykład:

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

$$\Sigma^* = \{ \varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, \dots \}$$

$$\Sigma^+ = \{ 0, 1, 00, 01, 10, 11, 000, 001, \dots \}$$



Zbiór słownikowy (2)

Formalnie stosując notację „potęgowa” i wykorzystując definicję złożenia zbiorów łańcuchów mamy dla alfabetu Σ definicję rekurencyjną:

$$\Sigma^0 = \{\varepsilon\}$$

$$\Sigma^1 = \Sigma = \{x \mid x \text{ jest słowem nad } \Sigma, |x| = 1\}$$

$$\Sigma^2 = \Sigma^1 \Sigma$$

.....

$$\Sigma^n = \Sigma^{n-1} \Sigma = \{x \mid x \text{ jest słowem nad } \Sigma, |x| = n\}$$

$$\Sigma^* = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \dots$$

$$\Sigma^+ = \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \dots$$

Zachodzi:

$$\Sigma^+ = \Sigma \Sigma^*$$



Uporządkowanie zbioru słownikowego (1)

Niech \leq będzie relacją liniowego porządku na zbiorze (alfabecie) Σ .

Przykład: $\Sigma = \{a, b, c\}$. Uporządkowanie „alfabetyczne”: $a < b$; $a < c$; $b < c$

Zdefiniujemy relację \leq_s określoną na Σ^* . Powiemy, że

$$x = a_1 a_2 \dots a_m \text{ jest w relacji } \prec_s \text{ z } y = b_1 b_2 \dots b_n \text{ (} x \prec_s y \text{)}$$

$x, y \in \Sigma^*$; $a_i, b_j \in \Sigma$ dla $1 \leq i \leq m$, $1 \leq j \leq n$, gdy spełniony jest jeden z dwóch poniższych warunków:

- $m < n$,
- $m = n$ oraz $a_i < b_i$ dla pewnego $i \leq m = n$ oraz $a_j = b_j$ dla wszystkich $1 \leq j < i$.

Relacja \prec_s jest quasi-porządkiem, który w zbiorze Σ^* definiuje

porządek \leq_s nazywamy porządkiem standardowym. (Σ^*, \leq_s) jest zbiorem dobrze uporządkowanym.

Przykład: $\Sigma = \{a, b\}$. Porządek alfabetu: $a < b$. Zbiór słownikowy Σ^* w porządku standardowym:

$$\Sigma^* = \{\varepsilon, a, b, aa, ab, ba, bb, aaa, aab, aba, abb, baa, bab, \dots\}$$



Uporządkowanie zbioru słownikowego (2)

Niech \leq będzie relacją liniowego porządku na zbiorze (alfabecie) Σ .

Zdefiniujemy relację \leq_L określoną na Σ^* . Powiemy, że $x = a_1a_2\dots a_m$ jest w relacji \prec_L z $y = b_1b_2\dots b_n$ ($x \prec_L y$; $x, y \in \Sigma^*$; $a_i, b_j \in \Sigma$ dla $1 \leq i \leq m$, $1 \leq j \leq n$, $k = \min(m, n)$), gdy spełniony jest jeden z dwóch poniższych warunków:

(1) $a_i \prec b_i$ dla pewnego $i \leq k$ oraz $a_j = b_j$ dla wszystkich $1 \leq j < i$

(2) $m < n$ oraz $a_i = b_i$ dla wszystkich $1 \leq i \leq m = k$.

Relacja \prec_L jest quasi-porządkiem, który w zbiorze Σ^* definiuje porządek \leq_L , który nazywamy porządkiem leksykograficznym. (Σ^*, \leq_L) jest zbiorem liniowo uporządkowanym. Nie jest on jednak zbiorem dobrze uporządkowanym.



Uporządkowanie zbioru słownikowego (3)

Przykład: $\Sigma = \{a, b\}$. Porządek alfabetu: $a \prec b$. Kilka pierwszych elementów zbioru słownikowego Σ^* w porządku leksykograficznym:

$$\Sigma^* = \{\epsilon, a, aa, aaa, aaaa, aaaaa, \dots\}$$

Dlaczego uporządkowanie leksykograficzne nie jest porządkiem dobrym? Weźmy nieskończony podzbiór zbioru Σ^* : $\{b, ab, aab, aaab, aaaab, \dots\}$. Elementy tego podzbioru są uporządkowane malejąco:

$$b \succ_L ab \succ_L aab \succ_L aaab \succ_L aaaab \succ_L \dots$$

Oczywiście podzbiór ten nie ma elementu najmniejszego, uporządkowanie leksykograficzne nie jest więc porządkiem dobrym. Porządek leksykograficzny w nieskończonym zbiorze Σ^* jest bardzo skomplikowany i trudno go sobie wyobrazić.



Uporządkowanie zbioru słownikowego (4)

Porządek leksykograficzny jest jednak powszechnie wykorzystywany do skończonych zbiorów łańcuchów, np. do określania kolejności słów w encyklopediach, słownikach, leksykonach. Wówczas quasi-porządek \prec jest powszechnie przyjętym uporządkowaniem liter w alfabecie pewnego języka naturalnego.

Ponieważ dla zbioru słownikowego Σ^* można określić porządek liniowy (np. leksykograficzny) lub porządek dobry (np. standardowy), można więc wszystkie elementy zbioru słownikowego ułożyć w ciąg i ponumerować. Świadczy to o równoliczności zbioru słownikowego ze zbiorem liczb naturalnych.



Definicja języka

Definicja języka

Niech Σ będzie alfabetem, Σ^* - zbiorem wszystkich łańcuchów nad alfabetem Σ .

Dowolny podzbiór L zbioru Σ^* nazywamy językiem L nad alfabetem Σ .

$$L \subseteq \Sigma^*$$

Przykłady:

$L_0 = \emptyset$ - język pusty

$L_1 = \{\varepsilon\}$ - język zawierający tylko słowo puste

$L_2 = \Sigma^*$ - język zawierający wszystkie słowa nad alfabetem Σ

$L_3 = \{\varepsilon, 0, 01, 001\}$ - język zawierający skończoną liczbę słów

$L_4 = \{0, 01, 011, 0111, \dots\} = \{01^n \mid n \geq 0\}$ - język nieskończony



Operacje na językach (1)

Niech L , L_1 i L_2 będą językami odpowiednio nad alfabetami Σ , Σ_1 i Σ_2 .

$$L \subseteq \Sigma^* \quad L_1 \subseteq \Sigma_1^* \quad L_2 \subseteq \Sigma_2^*$$

Najczęściej wykorzystuje się następujące operacje na językach:

Suma teoriomnogościowa

$$L_1 \cup L_2 = \{ x \mid x \in L_1 \vee x \in L_2 \}$$

Złożenie języków

$$L_1 L_2 = \{ x_1 x_2 \mid x_1 \in L_1 \wedge x_2 \in L_2 \}$$

Domknięcie Kleene'a (gwiazdka Kleene'a) L^*

$$L^0 = \{\varepsilon\}$$

$$L^1 = L$$

$$L^2 = L^1 L$$

.....

$$L^n = L^{n-1} L$$

$$L^* = L^0 \cup L^1 \cup L^2 \cup L^3 \cup \dots$$



Operacje na językach (2)

Rozpatruje się także operacje przecięcia (iloczynu teoriomnogościowego), dopełnienia i inne

Przecięcie (iloczyn teoriomnogościowy)

$$L_1 \cap L_2 = \{ x \mid x \in L_1 \wedge x \in L_2 \}$$

Dopełnienie języka L względem Σ^*

$$\bar{L} = \Sigma^* - L$$



Przedrostki, przyrostki (1)

Niech $z \in L \subseteq \Sigma^*$ będzie słowem z języka L .

Przedstawimy z w postaci:

$$z = xy \quad x, y \in \Sigma^*$$

x nazywamy przedrostkiem (prefiksem) słowa z , zaś y nazywamy przyrostkiem (sufiksem) słowa z .

x nazywamy przedrostkiem właściwym słowa $z \Leftrightarrow y \neq \varepsilon$.

y nazywamy przyrostkiem właściwym słowa $z \Leftrightarrow x \neq \varepsilon$.

Przykład:

Rozważamy słowo $abbb$

Przedrostki tego słowa to: $\varepsilon, a, ab, abb, abbb$

Przedrostki właściwe tego słowa to: ε, a, ab, abb



Przedrostki, przyrostki (2)

Język L ma własność przedrostkową, jeśli żaden przedrostek właściwy słowa tego języka nie jest identyczny z żadnym słowem tego języka.

Język L ma własność przyrostkową, jeśli żaden przyrostek właściwy słowa tego języka nie jest identyczny z żadnym słowem tego języka.

Przykład :

$$L = \{0^n1 \mid n \geq 0\} = \{1, 01, 001, 0001, \dots\}$$

L nie posiada własności przyrostkowej, gdyż np. słowo 0001 ma przyrostek właściwy 01 będący słowem tego języka.

L posiada własność przedrostkową, gdyż wszystkie przedrostki właściwe słów tego języka mają postać $\{0^n \mid n \geq 0\}$, i żaden z nich nie jest identyczny z żadnym słowem tego języka.



Uporządkowanie słów należących do języka

- Zbiór słownikowy można uważać za zbiór liniowo lub dobrze uporządkowany, np. poprzez porządek leksykograficzny (Σ^*, \leq_L) lub standardowy (Σ^*, \leq_S) .
- W taki sam sposób można uporządkować słowa dowolnego języka $L \subseteq \Sigma^*$ (określając relację \leq na alfabecie Σ oraz redukując relację \leq_S lub \leq_L określoną na Σ^* do L). Mówimy wówczas o leksykograficznym lub standardowym porządku słów danego języka.
- Przykładem porządku leksykograficznego \leq_L dla skończonych zbiorów (języków) może być uporządkowanie słów w encyklopediach, słownikach, leksykonach – wówczas \leq jest powszechnie przyjętym uporządkowaniem liter w alfabecie pewnego języka naturalnego.



Moc zbioru wszystkich języków (1)

Lemat: Zbiór \mathcal{B} wszystkich nieskończonych ciągów zerojedynkowych jest nieprzeliczalny.

Założmy dla dowodu nie wprost, że zbiór wszystkich nieskończonych łańcuchów zerojedynkowych jest przeliczalny. Można więc te łańcuchy wypisać i ponumerować, na przykład tak:

numer	łańcuch
1	01100010010100...
2	10001001001110...
3	01110101010010...
4	11101100111011...
...	...

Skonstruujemy łańcuch x różny od wszystkich wypisanych łańcuchów. Jeśli n -ty łańcuch ma na n -tej pozycji zero, to x będzie miał na n -tej pozycji jedynekę i na odwrót, jeśli n -ty łańcuch ma na n -tej pozycji jedynekę, to x będzie miał na n -tej pozycji zero. U nas $x = 1101...$

Łańcuch x jest różny co najmniej na jednej pozycji od każdego z wypisanych łańcuchów, wobec tego jest różny od każdego z wszystkich łańcuchów. Doszliśmy do sprzeczności. Zbioru wszystkich nieskończonych łańcuchów zerojedynkowych nie da się ponumerować, jest to więc zbiór nieprzeliczalny. (Jest to metoda diagonalizacji Cantora).



Moc zbioru wszystkich języków (2)

Twierdzenie: Zbiór $\mathcal{L} = 2^{\Sigma^*}$ wszystkich języków (wszystkich podzbiorów zbioru Σ^* - zbioru słownikowego nad danym alfabetem Σ) jest nieprzeliczalny.

Pokażemy, że \mathcal{L} jest nieprzeliczalny konstruując bijekcję między \mathcal{B} (zbiorem wszystkich nieskończonych łańcuchów zerojedynkowych) i \mathcal{L} dowodzącą, że oba te zbiory są tej samej mocy. Ponumerujmy słowa z Σ^* (wiadomo, że można, np. stosując porządek standardowy): $\Sigma^* = \{s_1, s_2, s_3, \dots\}$. Każdy język $L \in \mathcal{L} = 2^{\Sigma^*}$ odpowiada unikalnemu ciągowi z \mathcal{B} - i-ty bit tego ciągu jest równy jeden wtedy i tylko wtedy, gdy $s_i \in L$, w przeciwnym przypadku bit ten jest równy zero. Taki ciąg nazywamy ciągiem charakterystycznym χ_L języka L .

Przykład: $\Sigma = \{a, b\}$

$\Sigma^* = \{ \varepsilon, a, b, aa, ab, ba, bb, aaa, aab, aba, abb, baa, \dots \}$

$L = \{ a, aa, ab, bb, aaa, aab, abb, \dots \}$

$\chi_L = 0 \ 1 \ 0 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1 \ 0 \ 1 \ 0 \ \dots$

Odwzorowanie $f: \mathcal{L} \rightarrow \mathcal{B}$, gdzie $f(L) = \chi_L$ jest bijekcją, zatem, ponieważ \mathcal{B} jest nieprzeliczalny, to \mathcal{L} także jest nieprzeliczalny.