Formális nyelvek

Aszalós László, Mihálydeák Tamás

Számítógéptudományi Tanszék

December 6, 2017

Problémafelvetés

- Az informatikában nagyon gyakori feladat az elemzés és a generálás:
  - A különféle programok/számítógépek közti adatforgalom különféle szabványokat használ (HTML, XML, json, YAML, stb). Ezért a küldő oldalon generálni kell ilyen adatokat, a fogadó oldalon pedig elemezni.
  - A forrásprogramból a futtatható kód fordítás során készül el. A fordítás első része a lexikai, szintaktikai és szemantikai elemzés, melyet a kódgenerálás követ. (Interpreter esetén az utóbbi a végrehajtásra cserélődik.)
- Például űrlapok esetén:
  - A bevitt adat valóban egy email cím?
  - A jelszó tartalmaz nem betű karaktert?
  - Generált biztonságos jelszó felajánlása.

A formális nyelv definíciója

Definíció

- *Legyen Σ egy véges nemüres halmaz! A Σ elemeit betűknek, magát halmazt a halmazt ábécének nevezzük.*
- *Σ^+ a Σ betűiből álló $w = a_1 \dots a_n$ $(n \geq 1, a_i \in \Sigma)$ alakú véges hosszú sorozatok, a Σ feletti véges, nem üres szavak halmaza.*
- *A $w \in \Sigma^+$ szó hosszán a benne szereplő betűk számán értjük $(|a_1 \dots a_n| = n)$.*
- *Az üres szót (melynek nincs egy betűje sem, és a hossza definíció szerint 0) az ε jelöli.*
- *A $\Sigma^\star = \Sigma^+ \cup \{\varepsilon\}$ halmazt a Σ feletti szavak halmazának nevezzük.*
- *A Σ feletti szavak egy tetszőleges L halmazát a Σ ábécé feletti (formális) nyelvnek nevezzük $(L \subseteq \Sigma^\star)$.*
- *Az üres nyelvet, melynek nincs egy szava sem, a $L_\emptyset$ jelöli.*

Műveletek formális nyelvekkel

Definíció

Legyenek L, L_1 és L_2 a Σ ábécé feletti nyelvek! A következő műveleteket definiáljuk:

- *$L_1 \cup L_2 = \{x \mid x \in L_1$ vagy $x \in L_2\}$ (unió);*
- *$L_1 \cap L_2 = \{x \mid x \in L_1$ és $x \in L_2\}$ (metszet);*
- *$L_1 \circ L_2 = \{xy \mid x \in L_1$ és $y \in L_2\}$ konkatenáció (összefűzés)*
- *Hatványozás, azaz ismételt (önmagával vett) konkatenáció:*
  - *$L^0 = \{\varepsilon\}$*
  - *$L^{n+1} = L^n \circ L$*
- *Tranzitív lezárt (Kleene-csillag művelet): $L^\star = \cup_{i=0}^{\infty} L^i$, azaz*

$$L^\star = \{x_1 x_2 \dots x_k \mid k \geq 0 \text{ és } x_i \in L\}$$

Megjegyzés

- $L_\varepsilon^\star = \{\varepsilon\}$, *ahol* $L_\varepsilon = \{\varepsilon\}$.
- $L_\emptyset^\star = \{\varepsilon\}$, *mert definíció szerint* $L_\emptyset^0 = \{\varepsilon\}$

Példák

Legyen $\Sigma = \{a, b, c\}$, $L_1 = \{a, bb, ccc\}$ és $L_2 = \{ac, bc\}$. Ekkor

- $L_1 \cup L_2 = \{a, ac, bb, bc, ccc\}$;
- $L_1 \circ L_2 = \{aac, abc, bbac, bbbc, cccac, cccbc\}$;
- $L_1^\star = \{\varepsilon, a, bb, ccc, aa, abb, accc, bba, bbbb, bbccc, \dots\}$;
- Σ önmaga, mint ábécé feletti nyelv, a nyelv tranzitív lezártja, azaz $\Sigma^\star$ éppen a Σ ábécé felett szavak halmazát adja meg.

Formális nyelvek megadása

- Felsorolással (mint az előző példában):
  - szigorúan véve ekkor csak véges sok szava lehet a nyelvünknek.
- Generálással (specifikus induktív definícióval):
  - generatív grammatikák (Chomsky féle nyelvosztályok).
- Automatával:
  - az automata eldönti, hogy egy szó eleme-e a nyelvünknek.
- A nyelv szavainak halmazát leíró kifejezések megadásával:
  - például a reguláris kifejezések rendszere.

Megjegyzés

A felsorolt lehetőségek különböző tulajdonsággal rendelkező nyelvek megadására alkalmasak.

Reguláris kifejezések

Feladat

- Legyen $\Sigma = \{0, 1\}$ egy ábécé. Adjuk meg (fejezzük ki) más (lehetőség szerint egyszerűbb) nyelvekkel azt az L nyelvet, amely olyan szavakból áll, amely tartalmazza a 010 szót!
- Megoldás a "szavak szintjén": Ha $w_1, w_2 \in \Sigma^\star$, akkor $w_1 010 w_2 \in L$.
  - De: hogyan állíthatjuk elő a w_1 és w_2 szavakat?
  - Tudjuk, hogy nem üres esetben w_1 és w_2 a 0 és az 1 karaktereket tartalmazza.
  - Legyen $L_0 = \{0\}$ és $L_1 = \{1\}$, ekkor $L_0 \cup L_1 = \{0, 1\}$, és $(L_0 \cup L_1)^\star$ éppen a w_1 és w_2 szavak lehetséges alakjait tartalmazza (már az üres szóval kiegészítve).
  - Ha $L_{010} = \{010\}$, akkor $L = (L_0 \cup L_1)^\star \circ L_{010} \circ (L_0 \cup L_1)^\star$.
  - L_0, L_1, L_{010} egyszavas nyelvek! (Az egyszavas nyelveket alapnyelveknek is nevezik.)
  - Másként leírva: $(0 + 1)^\star \cdot 010 \cdot (0 + 1)^\star$

Példák reguláris kifejezésekre

Példák

- Adjuk meg reguláris kifejezéssel azt a nyelvet a $\{0, 1\}$ ábécé felett, amely azon szavakból áll, amelyek tartalmazzák részszóként a 000 vagy az 111 szót!
- $(0 + 1)^\star \cdot (000 + 111) \cdot (0 + 1)^\star$
- Adjuk meg reguláris kifejezéssel azt a nyelvet a $\{0, 1\}$ ábécé felett, amely azon szavakból áll, amelyekben az 1-esek száma osztható 5-tel!
- $0^\star + (0^\star \cdot 1 \cdot 0^\star \cdot 1 \cdot 0^\star \cdot 1 \cdot 0^\star \cdot 1 \cdot 0^\star \cdot 1 \cdot 0^\star)^\star$

Reguláris kifejezések

Definíció

Legyen Σ egy tetszőleges ábécé úgy, hogy $\emptyset, \varepsilon, +, \cdot, ^\star, (,) \notin \Sigma$. A Σ ábécé feletti reguláris kifejezések $\mathcal{R}$ halmazát az alábbi induktív definíció adja meg:

1. $\emptyset \in \mathcal{R}$
2. $\varepsilon \in \mathcal{R}$
3. *Ha* $a \in \Sigma$, *akkor* $a \in \mathcal{R}$.
4. *Ha* $R_1 \in \mathcal{R}$ *és* $R_2 \in \mathcal{R}$, *akkor* $(R_1 + R_2) \in \mathcal{R}$.
5. *Ha* $R_1 \in \mathcal{R}$ *és* $R_2 \in \mathcal{R}$, *akkor* $(R_1 \cdot R_2) \in \mathcal{R}$.
6. *Ha* $R \in \mathcal{R}$, *akkor* $R^\star \in \mathcal{R}$.

Reguláris kifejezések által leírt/felismert nyelv

Definíció

Legyen $\mathcal{R}$ a Σ ábécé feletti reguláris kifejezések halmaza és $R \in \mathcal{R}$. A R reguláris kifejezés által leírt/felismert nyelvet az alábbi induktív definíció adja meg:

1. *Ha* $R = \emptyset$, *akkor* $L_R = L_\emptyset$.
2. *Ha* $R = \varepsilon$, *akkor* $L_R = L_\varepsilon = \{\varepsilon\}$.
3. *Ha* $R = a, a \in \Sigma$, *akkor* $L_R = \{a\}$.
4. *Ha* $R = (R_1 + R_2)$, *akkor* $L_R = L_{R_1} \cup L_{R_2}$.
5. *Ha* $R = (R_1 \cdot R_2)$, *akkor* $L_R = L_{R_1} \circ L_{R_2}$.
6. *Ha* $R = R_1^\star$, *akkor* $L_R = L_{R_1}^\star$.

Reguláris nyelv

Definíció

A Σ ábécé feletti L nyelv reguláris, ha létezik egy olyan Σ ábécé feletti R reguláris kifejezés, hogy $L_R = L$

Tétel

Ha a Σ ábécé feletti L_1 és L_2 nyelvek regulárisak, akkor az $L_1 \cup L_2$, $L_1 \cap L_2$, $L_1 \backslash L_2$, $\overline{L_1} (= \Sigma^\star \backslash L_1)$, $L_1 \circ L_2$ és $L_1^\star$ nyelvek is regulárisak.

Véges automaták

- A legegyszerűbb automaták: szemléletesen csak olvasni tud és nincs memóriája.
- Állapottere véges (mást nem is nagyon tudunk róla mondani).

Definíció

A véges automata egy rendezett ötös: $A = \langle Q, \Sigma, \delta, q_0, F \rangle$, ahol

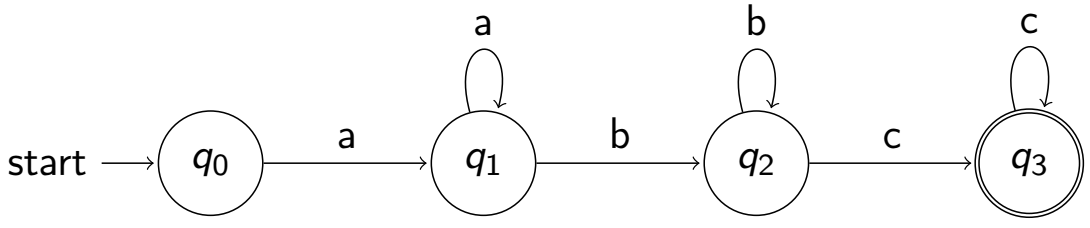
- *Q (belső) állapotok véges halmaza,*
- *Σ (külső) véges ábécé (az elemzendő jelsorozat ábécéje),*
- *$\delta : Q \times \Sigma \to Q$ átmenetfüggvény,*
- *$q_0 \in Q$ kezdőállapot és*
- *$F \subseteq Q$ az elfogadó állapotok halmaza.*

Megjegyzés.

- Az átmenetfüggvény azt mondja meg, hogy egy adott állapotban egy adott karakter beolvasásának hatására milyen új állapotot vesz fel az automata. Ha egy karaktert elolvastunk, akkor a következő karakter válik érvényessé (az olvasófej alatt a szalag egy karakternyit elmozdul).
- Az automata az elemzett jelsorozatot elfogadhatja, ha
  - a jelsorozatot végig olvasta és
  - az utolsó karakter elolvasása után elfogadó állapotban van.

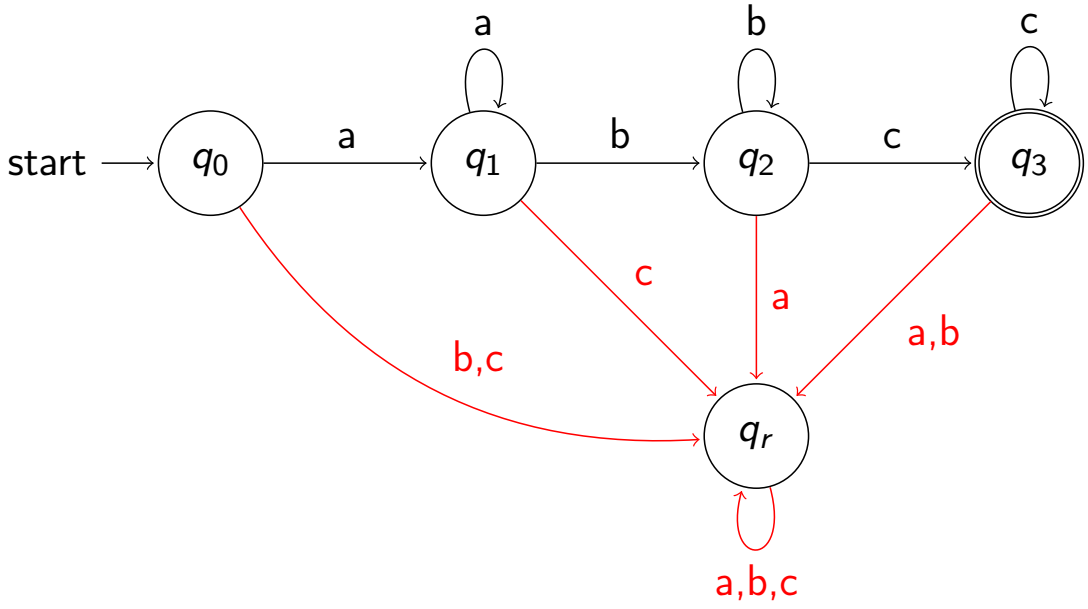
Legyen $A = \langle Q, \Sigma, \delta, q_0, F \rangle$, ahol

- $Q = \{q_0, q_1, q_2, q_3, q_r\}$ (belső) állapotok véges halmaza;
- $\Sigma = \{a, b, c\}$ (külső) véges ábécé (az elemzendő jelsorozat ábécéje);
- $\delta(q_0, a) = q_1, \delta(q_1, a) = q_1, \delta(q_1, b) = q_2, \delta(q_2, b) = q_2, \delta(q_2, c) = q_3, \delta(q_3, c) = q_3$ és δ értéke q_r egyébként (az átmenetfüggvény);
- $q_0 \in Q$ a kezdőállapot;
- $F = \{q_3\}$, azaz q_3 az elfogadó állapot.



$a^i b^j c^k$, ahol $i, j, k > 0$

- Definíció szerint egy állapotból az ábécé minden betűjéhez kell rendelni egy állapotot (δ teljes függvény). A rajz egyszerűbbé tétele érdekében a hiányzó nyilak egy nem jelölt *rontott* állapotba mutatnak, ahonnan minden inputra ugyanide jutunk vissza. Az ábrán a kiegészítés pirossal lett kiemelve.



- Grafikus ábrázolásban a belső állapotokat körökkel, az átmenetfüggvényt nyilakkal és a rájuk írt külső ábécé betűivel jelöljük. A kezdőállapot egy bemenő, felirat nélküli nyíl jelzi, míg a végállapotokat duplán rajzolt körök.

Elfogadó automata

Definíció

$A = \langle Q, \Sigma, \delta, q_0, F \rangle$ automata és $w = a_1 a_2 \dots a_n \in \Sigma^\star$ szó esetén azt mondjuk, hogy az automata elfogadja (felismeri) a w szót, ha

$$\delta(\dots(\delta(\delta(q_0, a_1), a_2)\dots), a_n) \in F.$$

Definíció

Legyen $A = \langle Q, \Sigma, \delta, q_0, F \rangle$ egy automata! Ha $L = \{w \in \Sigma^\star | A \text{ elfogadja } w\text{-t}\}$, akkor azt mondjuk, hogy az A elfogadja (felismeri) az L nyelvet. Jelölése $L = L_A$

Tétel

A véges automaták a reguláris nyelveket fogadják el, azaz ha L reguláris nyelv, akkor van olyan A automata, hogy $L = L_A$