Formális nyelvek - 7. előadás

Csuhaj Varjú Erzsébet

Algoritmusok és Alkalmazásaik Tanszék
Informatikai Kar
Eötvös Loránd Tudományegyetem
H-1117 Budapest
Pázmány Péter sétány 1/c

E-mail: csuhaj@inf.elte.hu

Reguláris kifejezések

Motiváció

Ismeretes, hogy **minden véges nyelv reguláris**. Tudjuk továbbá, hogy az $\mathcal{L}_3$ **nyelvosztály** (a reguláris nyelvek osztálya) **zárt az unió, a konkatenáció és az iteráció lezártja** műveletekre nézve.

Következésképpen, kiindulva véges számú véges nyelvből és az előzőekben felsorolt, ún. reguláris műveleteket véges sokszor alkalmazva reguláris nyelvet kapunk.

Kérdés az, hogy vajon ezzel az **eljárással minden reguláris nyelvet elő tudunk-e állítani**, azaz, ez a módszer elégséges-e az $\mathcal{L}_3$ nyelvosztály leírására?

Reguláris kifejezések

Definíció

Legyenek V és $V' = \{\varepsilon, \cdot, +, *, (,)\}$ diszjunkt ábécék. A $V \cup V'$ ábécé feletti reguláris kifejezéseket rekurzív módon a következőképpen definiáljuk:

1. ε reguláris kifejezés V felett.
2. Minden $a \in V$ reguláris kifejezés V felett.
3. Ha R reguláris kifejezés V felett, akkor $(R)^*$ is reguláris kifejezés V felett.
4. Ha Q és R reguláris kifejezések V felett, akkor $(Q)\cdot(R)$ és $(Q)+(R)$ is reguláris kifejezések V felett.

A $*$, $\cdot$ és $+$ szimbólumok rendre az iteráció lezártjára, a konkatenációra és az unióra utalnak, azt jelölik.

A reguláris kifejezések által jelölt nyelv

Minden reguláris kifejezés jelöl (meghatároz) valamely reguláris nyelvet.

A $V \cup V'$ ábécé felett megadott R reguláris kifejezés által jelölt nyelvet L_R-el jelöljük és a következőképpen definiáljuk:

- $L_\varepsilon = \{\varepsilon\}$,
- $L_a = \{a\}$, minden $a \in V$-re,
- Továbbá minden R, Q reguláris kifejezésre $V \cup V'$ felett $L_{(R+Q)} = L_R \cup L_Q$, $L_{(R \cdot Q)} = L_R L_Q$, valamint $L_{(R)^*} = (L_R)^*$.

Például a az $\{a\}$ nyelvet, $a + b$ az $\{a\} \cup \{b\} = \{a, b\}$ nyelvet és $a \cdot b$ az $\{a\}\{b\} = \{ab\}$ nyelvet jelöli.

Példák

Legyen $V = \{a, b\}$. Az alábbi reguláris kifejezések mellett az általuk jelölt nyelv található.

Megjegyzés:

A zárójelek egyrésze elhagyható, ha a műveleteken precedenciát definiálunk. A szokásos sorrend $*, \cdot, +$.

- a^* ugyanaz, mint $(a)^*$ és az $\{a\}^*$ nyelvet jelöli.
- $(a + b)^*$ ugyanaz, mint $((a) + (b))^*$ és az $\{a, b\}^*$ nyelvet jelöli.
- $a^* \cdot b$ ugyanaz, mint $((a)^*) \cdot (b)$ és az $\{a\}^*b$ nyelvet jelöli.
- $b + a \cdot b^*$ ugyanaz, mint $(b) + ((a) \cdot (b)^*)$ és a $\{b\} \cup \{a\}\{b\}^*$ nyelvet jelöli.
- $(a + b) \cdot a^*$ ugyanaz, mint $((a) + (b)) \cdot ((a)^*)$ és az $\{a, b\}\{a\}^*$ nyelvet jelöli.

Egyenlőségek reguláris kifejezésekre

Könnyen látható, hogy $\{a,b\}\{a\}^* = \{a\}\{a\}^* \cup \{b\}\{a\}^*$. Így

$$(a+b)\cdot a^* = a\cdot a^* + b\cdot a^*,$$

azaz, a két reguláris kifejezés ugyanazt a nyelvet jelöli.

Legyenek P, Q, R reguláris kifejezések. Akkor P, Q és R helyébe reguláris kifejezéseket írva fennállnak az alábbi egyenlőségek.

$$P+(Q+R) = (P+Q)+R \qquad P\cdot(Q\cdot R) = (P\cdot Q)\cdot R$$

$$P+Q = Q+P \qquad P\cdot(Q+R) = P\cdot Q + P\cdot R$$

$$(P+Q)\cdot R = P\cdot R + Q\cdot R \qquad P^* = \varepsilon + P\cdot P^*$$

$$\varepsilon\cdot P = P\cdot\varepsilon = P \qquad P^* = (\varepsilon+P)^*$$

Egyenlőségek reguláris kifejezésekre - folytatás

Ha a fenti egyenlőségekben a P, Q, R reguláris kifejezéseket reguláris kifejezésekkel helyettesítjük, reguláris kifejezéseket kapuk.

Azonban sem az előbbi egyenlőségekből, sem egyenlőségek más véges halmazából nem kaphatjuk meg az összes reguláris kifejezést kizárólag helyettesítés segítségével.

Még egy további szabályra van szükségünk, nevezetesen, ha

$P = R + P \cdot Q$ és $\varepsilon \notin Q$, akkor $P = R \cdot Q^*$.

Egyenlőségek reguláris kifejezésekre - folytatás

A teljesség biztosítása céljából még hozzáadjuk az $\emptyset$ szimbólumot a reguláris kifejezések halmazához, amely az üres nyelvet jelöli.

Ebben az esetben nincs szükségünk a ε szimbólumra, mivel $\emptyset^* = \{\varepsilon\}$.

Így, a definícióban helyettesíthetjük a ε szimbólumot az $\emptyset$ szimbólummal.

Ekkor helyettesítjük ε-t a megelőző axióma rendszerben $(\emptyset)^*$-gal és még egy további axiómát tekintünk:

$$\emptyset \cdot P = P \cdot \emptyset = \emptyset.$$

Az egyenlőségek, valamint a helyettesítés és a fenti feltételes egyenlőség elégséges ahhoz, hogy levezessünk minden érvényes egyenlőséget reguláris kifejezések között.

Reguláris kifejezések versus reguláris nyelvek

Tétel

Minden reguláris kifejezés egy reguláris (3-típusú) nyelvet jelöl, és megfordítva, minden reguláris nyelvhez megadható egy, ezen nyelvet jelölő reguláris kifejezés.

Bizonyításvázlat

1) Az állítás első fele a megelőző diszkusszióból következik.

2) Megmutatjuk, hogy minden L reguláris nyelvhez, amelyet a $G = (N, T, P, S)$ normálformában adott reguláris grammatika generál, meg tudunk konstruálni egy reguláris kifejezést, amely az L nyelvet jelöli.

Legyen $N = \{A_1, \ldots, A_n\}$, $n \geq 1$, $S = A_1$. (G minden szabálya vagy $A_i \to aA_j$ vagy $A_i \to \varepsilon$ alakú, ahol $a \in T$, $1 \leq i, j \leq n$.)

Azt mondjuk, hogy az $A_i \Longrightarrow^* uA_j$ ($u \in T^*$) levezetés **érinti** az A_m nemterminálist, ha A_m előfordul valamely közbülső mondatformában A_i és uA_j között a levezetésben.

Az $A_i \Longrightarrow^* uA_j$ levezetést k-**megszorítottnak** nevezzük, ha $0 \leq m \leq k$ teljesül minden A_m nemterminálisra, amely a levezetésben előfordul.

Bizonyításvázlat - folytatás

Definiáljuk a következő halmazokat:

$E_{i,j}^k = \{u \in T^* \mid$ létezik $A_i \Longrightarrow^* uA_j$ k-megszorított levezetés $\}$.

k**-szerinti indukcióval** bizonyítjuk, hogy az $E_{i,j}^k$ nyelvhez létezik őt jelölő reguláris kifejezés i, j, k-ra, ahol $0 \leq i, j, k \leq n$.

Bázis: $i \neq j$ esetén az $E_{i,j}^0$ halmaz vagy üres vagy T-beli betűkből áll. ($a \in E_{i,j}^0$, akkor és csak akkor, ha $A_i \to aA_j \in F$.) Ha $i = j$, akkor $E_{i,i}^0$ tartalmazza ε-t és nulla vagy több elemét T-nek, így $E_{i,j}^0$ reguláris kifejezéssel jelölhető.

Indukciós lépés: tegyük fel, hogy rögzített k-ra, $0 < k \leq n$, az $E_{i,j}^{k-1}$ nyelvek mindegyike jelölhető reguláris kifejezéssel. Akkor minden i, j, k-ra fennáll, hogy

$$E_{i,j}^k = E_{i,j}^{k-1} + E_{i,k}^{k-1} \cdot (E_{k,k}^{k-1})^* \cdot E_{k,j}^{k-1}.$$

Ekkor az indukciós feltevés alapján $E_{i,j}^k$ szintén jelölhető reguláris kifejezéssel.

Legyen I_ε azon i indexek halmaza, amelyekre $A_i \to \varepsilon$. Akkor $L(G) = \cup_{i \in I_\varepsilon} E_{1,i}^n$, azaz, L reguláris kifejezéssel jelölhető.

Helyettesítés

Definíció

Legyen V egy ábécé, valamint legyen minden $a \in V$-re V_a ábécé és $s(a) \subseteq V_a^*$. Minden $u = a_1 a_2 \ldots a_n \in V^*$ szóra definiáljuk az u szó s helyettesítését a következőképpen:

$$s(u) = s(a_1)s(a_2)\ldots s(a_n).$$

Legyen továbbá $s(\varepsilon) = \varepsilon$. Az s helyettesítés kiterjeszthető bármely $L \subseteq V^*$ nyelvre a következő módon: $s(L) = \{w \mid w \in s(u), u \in L\}$.

Reguláris nyelvek zártsága a helyettesítésre nézve

Reguláris kifejezéseket használva könnyen látható, hogy az $\mathcal{L}_3$ nyelvosztály zárt a helyettesítésre nézve. A reguláris kifejezések halmaza nyilvánvalóan zárt a kifejezés minden betűjének valamely reguláris kifejezéssel való helyettesítésére nézve. (Lásd a megelőző diszkussziót).

Megjegyzés: A helyettesítés a homomorfizmus általánosítása.