Substitúcie

Peter Kostolányi

12. februára 2025

1 Definícia

Pod substitúciou sa v teórii formálnych jazykov chápe „zovšeobecnenie“ homomorfizmu, kde obrazom slov už nie sú slová, ale jazyky.

Definícia 1. Nech Σ, Γ sú abecedy. Zobrazenie $\tau\colon \Sigma^* \to 2^{\Gamma^*}$ nazveme *substitúciou*, ak $\tau(\varepsilon) = \{\varepsilon\}$ a pre všetky $u, v \in \Sigma^*$ je

$$\tau(uv) = \tau(u)\tau(v).$$

Poznámka 1. Požiadavka $\tau(\varepsilon) = \{\varepsilon\}$ je v uvedenej definícii skutočne podstatná, pretože *nevyplýva* z vlastnosti $\tau(uv) = \tau(u)\tau(v)$ pre všetky $u, v \in \Sigma^*$, čím sa substitúcie odlišujú od homomorfizmov. Nájdenie vhodného protipríkladu je jednou z úloh určených na nasledujúce cvičenie.

Poznámka 2. Ak $h\colon \Sigma^* \to \Gamma^*$ je homomorfizmus, zobrazenie $\tau\colon \Sigma^* \to 2^{\Gamma^*}$ definované pre všetky $w \in \Sigma^*$ ako $\tau(w) = \{h(w)\}$ je zjavne substitúcia. Aj keď teda homomorfizmus z formálneho hľadiska nie je substitúciou, možno substitúcie chápať ako zovšeobecnenie homomorfizmov.

Poznámka 3. Definíciu substitúcie možno interpretovať aj tak, že ide o homomorfizmus monoidov $(\Sigma^*, \cdot, \varepsilon)$ a $(2^{\Gamma^*}, \cdot, \{\varepsilon\})$.

Z nasledujúceho tvrdenia vyplýva, že podobne ako homomorfizmy sú aj substitúcie jednoznačne určené obrazmi jednotlivých písmen – každé zobrazenie zo Σ do 2^{Γ^*} teda *jednoznačne* určuje substitúciu zo Σ^* do 2^{Γ^*}. V nasledujúcom budeme túto skutočnosť využívať bez toho, aby sme na to explicitne upozorňovali.

Tvrdenie 1. *Nech Σ, Γ sú abecedy a $f\colon \Sigma \to 2^{\Gamma^*}$ je zobrazenie. Potom existuje* práve jedna *substitúcia $\tau\colon \Sigma^* \to 2^{\Gamma^*}$ taká, že pre všetky $c \in \Sigma$ je $\tau(c) = f(c)$.*

Dôkaz. Definujme zobrazenie $\tau\colon \Sigma^* \to 2^{\Gamma^*}$ pre všetky $k \in \mathbb{N}$ a všetky $a_1, \ldots, a_k \in \Sigma$ predpisom $\tau(a_1 \ldots a_k) = f(a_1) \ldots f(a_k)$. Pre všetky $c \in \Sigma$ potom $\tau(c) = f(c)$.

Dokážeme, že zobrazenie τ je substitúcia. Rovnosť $\tau(\varepsilon) = \{\varepsilon\}$ vyplýva bezprostredne z definície τ. Nech teraz $u, v \in \Sigma^*$, pričom $u = a_1 \ldots a_m$ a $v = b_1 \ldots b_n$ pre nejaké $a_1, \ldots, a_m, b_1, \ldots, b_n \in \Sigma$. Potom

$$\tau(uv) = \tau(a_1 \ldots a_m b_1 \ldots b_n) = f(a_1) \ldots f(a_m) f(b_1) \ldots f(b_m) = \tau(u)\tau(v).$$

Zostáva dokázať jedinečnosť substitúcie τ. Nech $\tau'\colon \Sigma^* \to 2^{\Gamma^*}$ je substitúcia taká, že pre všetky $c \in \Sigma$ je $\tau'(c) = f(c)$. Pre všetky $k \in \mathbb{N}$ a $a_1, \ldots, a_k \in \Sigma$ potom

$$\tau'(a_1 \ldots a_k) = \tau'(a_1) \ldots \tau'(a_k),$$

čo možno dokázať matematickou indukciou: pre $k = 0$ a $k = 1$ je tvrdenie triviálne – a ak tvrdenie platí pre $k = s$, tak aj pre $k = s + 1$ dostávame

$$\tau'(a_1 \ldots a_s a_{s+1}) = \tau'(a_1 \ldots a_s)\tau'(a_{s+1}) = \tau'(a_1) \ldots \tau'(a_s)\tau'(a_{s+1}).$$

Preto

$$\tau'(a_1 \ldots a_k) = \tau'(a_1) \ldots \tau'(a_k) = f(a_1) \ldots f(a_k) = \tau(a_1 \ldots a_k).$$

Keďže sú $k \in \mathbb{N}$ a $a_1, \ldots, a_k \in \Sigma$ ľubovoľné, nutne $\tau' = \tau$. □

Podobne ako v prípade homomorfizmov možno definíciu substitúcie aditívne rozšíriť aj na jazyky.

Definícia 2. Nech Σ, Γ sú abecedy, $\tau\colon \Sigma^* \to 2^{\Gamma^*}$ je substitúcia a $L \subseteq \Sigma^*$ je jazyk. *Obraz jazyka L pri zobrazení substitúciou τ* je jazyk

$$\tau(L) = \bigcup_{w \in L} \tau(w).$$

Príklad 1. Nech $\Sigma = \{a, b\}$ a substitúcia $\tau\colon \Sigma^* \to 2^{\Sigma^*}$ je daná ako $\tau(a) = \{\varepsilon, b\}$ a $\tau(b) = a^*$. Potom pre $L = \{ab, bb\}$ je

$$\tau(L) = \tau(ab) \cup \tau(bb) = (a^* \cup ba^*) \cup a^*a^* = a^* \cup ba^*.$$

Je zrejmé, že substitúcia je veľmi silná operácia – jazyk $\tau(a)$ napríklad ani len nemusí byť rekurzívne vyčísliteľný. Z tohto dôvodu sa v súvislosti s konkrétnymi triedami jazykov budeme väčšinou zaoberať substitúciami, ktorých obor hodnôt je určitým spôsobom obmedzený. To odôvodňuje zavedenie pojmu $\mathscr{L}$-substitúcie pre triedu jazykov $\mathscr{L}$.

Definícia 3. Nech Σ, Γ sú abecedy, $\tau\colon \Sigma^* \to 2^{\Gamma^*}$ je substitúcia a $\mathscr{L}$ je trieda jazykov. Hovoríme, že τ je $\mathscr{L}$-substitúcia, ak pre všetky $c \in \Sigma$ je $\tau(c) \in \mathscr{L}$.

Poznámka 4. Z uvedenej definície ešte pre $\mathscr{L}$-substitúciu $\tau\colon \Sigma^* \to 2^{\Gamma^*}$ vo všeobecnosti *nevyplýva* $\tau(w) \in \mathscr{L}$ pre všetky $w \in \Sigma^*$. Táto implikácia je však zjavne pravdivá v prípade, že je trieda $\mathscr{L}$ uzavretá na zreťazenie (a teda platí aj pre všetky triedy jazykov Chomského hierarchie).

V podobnom duchu tiež definujeme *regulárnu substitúciu* ako $\mathscr{R}$-substitúciu, *bezkontextovú substitúciu* ako $\mathscr{L}_{CF}$-substitúciu a podobnú terminológiu používame aj pre ďalšie známe triedy jazykov.

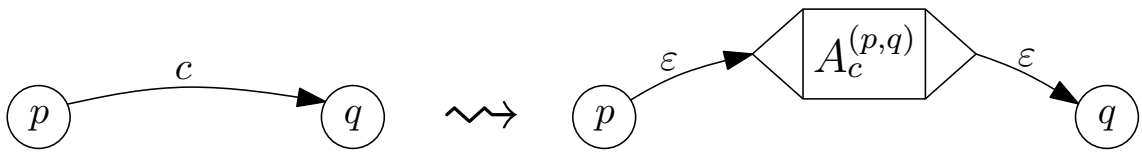
2 Uzavretosť tried $\mathscr{R}$ a $\mathscr{L}_{CF}$ na substitúciu

Z predchádzajúcich úvah okrem iného vyplýva, že nemá zmysel očakávať uzavretosť nejakej zmysluplnej a netriviálnej triedy jazykov na všeobecnú substitúciu. Aj preto budeme hovoriť, že trieda $\mathscr{L}$ je *uzavretá na substitúciu*, ak je uzavretá na $\mathscr{L}$-substitúciu. V nasledujúcom dokážeme, že triedy $\mathscr{R}$ a $\mathscr{L}_{CF}$ sú uzavreté na substitúciu.

Veta 1. *Trieda $\mathscr{R}$ je uzavretá na substitúciu.*

Dôkaz. Nech Σ, Γ sú abecedy, $L \subseteq \Sigma^*$ je regulárny jazyk akceptovaný deterministickým konečným automatom $A = (K, \Sigma, \delta, q_0, F)$ a $\tau\colon \Sigma^* \to 2^{\Gamma^*}$ je regulárna substitúcia. Zostrojíme nedeterministický konečný automat A' taký, že $L(A') = \tau(L)$.

Keďže je substitúcia τ regulárna, pre každé $c \in \Sigma$ existuje nedeterministický konečný automat A_c v „prasiatkovom“ normálnom tvare taký, že $L(A_c) = \tau(c)$. Automat A' zostrojíme z automatu A tak, že každý prechod na písmeno c nahradíme samostatnou kópiou automatu A_c tak, ako je to znázornené na obrázku 1. Každá kópia automatu A_c musí byť označená počiatočným a koncovým stavom prechodu, ktorý nahrádza – inak by mohlo dôjsť k „pomiešaniu stavov“ jednotlivých kópií automatu A_c.



Obr. 1: Každý prechod automatu A na písmeno c nahradíme samostatnou kópiou automatu A_c.

Predpokladajme, že pre všetky $c \in \Sigma$ je $A_c = (K[A_c], \Gamma, \delta[A_c], q_0[A_c], \{q_F[A_c]\})$, pričom pre každé dve rôzne $a, b \in \Sigma$ sú množiny stavov $K[A_a]$ a $K[A_b]$ disjunktné a pre všetky $c \in \Sigma$ sú disjunktné množiny K a $K[A_c] \times K^2$. Konštrukciu automatu A' potom môžeme zapísať nasledovne: $A' = (K', \Sigma', \delta', q_0', F')$, kde

$$K' = K \cup \bigcup_{c \in \Sigma} \left(K[A_c] \times K^2 \right),$$

$\Sigma' = \Gamma$, $q_0' = q_0$ a $F' = F$. Pre všetky stavy $p \in K$ ďalej

$$\delta'(p, \varepsilon) = \{(q_0[A_c], p, \delta(p, c)) \mid c \in \Sigma\},$$

čo zodpovedá prechodom na ε vedúcim zo stavov z množiny K do počiatočných stavov kópií automatov A_c. Pre všetky $p, r \in K$, $c \in \Sigma$, $q \in K[A_c] - \{q_F[A_c]\}$ a $z \in \Gamma \cup \{\varepsilon\}$ položíme

$$\delta'((q, p, r), z) = \{(q', p, r) \mid q' \in \delta[A_c](q, z)\},$$

čo zodpovedá prechodom v rámci jednotlivých kópií automatov A_c. Pre všetky $p, r \in K$ a $c \in \Sigma$ napokon definujeme prechod

$$\delta'((q_F[A_c], p, r), \varepsilon) = \{r\}$$

vedúci z akceptačného stavu danej kópie automatu A_c do príslušného stavu z K. □

Veta 2. *Trieda $\mathscr{L}_{CF}$ je uzavretá na substitúciu.*

Dôkaz. Nech L je bezkontextový jazyk generovaný bezkontextovou gramatikou $G = (N, T, P, \sigma)$. Nech τ je bezkontextová substitúcia na T^*. Gramatiku G' generujúcu jazyk $L(G') = \tau(L)$ možno získať nahradením každého terminálu $c \in T$ počiatočným neterminálom bezkontextovej gramatiky pre $\tau(c)$ (za predpokladu disjunktnosti jednotlivých množín neterminálov a neexistencie symbolu, ktorý je súčasne neterminálom jednej gramatiky a terminálom inej gramatiky).

Nech teda pre každé $c \in T$ je $G_c = (N_c, T_c, P_c, \sigma_c)$ bezkontextová gramatika taká, že $L(G_c) = \tau(c)$. Predpokladajme navyše, že

$$\begin{aligned} N_c \cap N_d &= \emptyset \qquad \forall c, d \in T \text{ také, že } c \neq d, \\ N_c \cap N &= \emptyset \qquad \forall c \in T, \\ \left(N \cup \bigcup_{c \in T} N_c \right) \cap \left(\bigcup_{c \in T} T_c \right) &= \emptyset. \end{aligned}$$

Uvažujme homomorfizmus h na $(N \cup T)^*$ taký, že pre všetky $\xi \in N$ je $h(\xi) = \xi$ a pre všetky $c \in T$ je $h(c) = \sigma_c$. Gramatiku $G' = (N', T', P', \sigma')$ potom môžeme zostrojiť nasledovne:

$$\begin{aligned} N' &= N \cup \bigcup_{c \in T} N_c, \\ T' &= \bigcup_{c \in T} T_c, \\ P' &= \{\xi \to h(w) \mid \xi \in N;\ w \in T^*;\ \xi \to w \in P\} \cup \bigcup_{c \in T} P_c, \\ \sigma' &= \sigma. \end{aligned}$$

□