Automate des occurrences [Cormen p 915-921]

Soit Σ un alphabet $< \infty$, $t \in \Sigma^*$ un texte et $m \in \Sigma^*$ un mot que l'on souhaite rechercher dans t.

notation + def On note $\mathcal{P}$ l'ensemble des préfixes de m

De plus $\forall u \in \Sigma^*$, on note $\sigma_m(u)$ = le plus long suffixe de u qui soit préfixe de m.

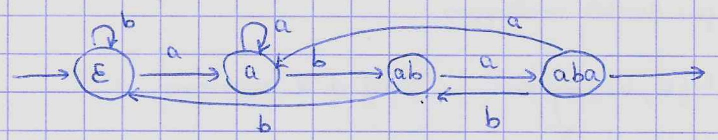
def Automate des occurrences

états

on considère l'automate $A_m = (P, \Sigma, \delta, \{\varepsilon\}, \{m\})$

avec $\forall p \in P, \forall a \in \Sigma, \delta(p,a) = \sigma_m(pa)$

Ex Pour le mot $m = aba$, on a A_m =



interprétations :

état = jusque là on a reconnu ce préfixe

transition soit on grandit le préfixe reconnu,

soit on rebrousse chemin mais pas + que nécessaire

Simuler la recherche avec t = aabbabab

Th A_m reconnaît le langage $L = \Sigma^* m$. De plus, il est minimal pour ce langage L

démo • On étend comme d'habitude la fonction de transition δ sur les lettres en $\bar{\delta}$ sur les mots tq $\forall p \in P, \forall a \in \Sigma, \forall w \in \Sigma^*$, $\begin{cases} \bar{\delta}(p,a) = \delta(p,a) \\ \bar{\delta}(p,aw) = \bar{\delta}(\delta(p,a), w) \end{cases}$

On s'autorise à écrire δ à la place de $\bar{\delta}$

• On montre que $\forall p \in P, \forall w \in \Sigma^*$, $\delta(p,w)$ = plus long suffixe de pw qui est préfixe de m

$= \sigma_m(pw)$

$p_1 \cdots p_k$ (k peut ê = 0)

Soit $p \in P$. Montrons / rec sur $n = |w| \in \mathbb{N}^*$, H_n : " $\forall w \in \Sigma^*$ de taille n, $\delta(p,w) = \sigma_m(pw)$ "

H_1 est OK par déf de δ

Soit $n \geq 1$ tq H_n et $w \in \Sigma^*$ de taille $n+1$. En particulier $\exists a \in \Sigma, \tilde{w} \in \Sigma^*, w = a\tilde{w}$

On a $\delta(p, w) = \delta(p, a\tilde{w}) = \delta(\delta(p,a), \tilde{w})$ /déf

$= \delta(\sigma_m(pa), \tilde{w})$

$= \sigma_m(\sigma_m(pa)\tilde{w})$ par HR (car $|\tilde{w}| = n$)

à montrer

$= \sigma_m(pa\tilde{w}) = \sigma_m(pw)$

Il suffit pour conclure de mq $\sigma_m(\sigma_m(pa)\tilde{w}) = \sigma_m(pa\tilde{w})$

Deux cas se présentent : → Si $\sigma_m(pa) \neq \varepsilon$ alors $\exists\, i \in [\![1, k]\!]$ tq $p_i \dots p_k a = \sigma_m(pa)$.

En particulier, $\forall\, j < i$, $p_j \dots p_k a$ n'est PAS préfixe de m
(sinon χ avec + long suff qui est préf)

On a donc bien $\sigma_m(\sigma_m(pa)\tilde{w})$

$= \sigma_m(p_i \dots p_k a\tilde{w}) = \sigma_m(pa\tilde{w})$

(rien ne sert de chercher $\sigma_m(pa\tilde{w})$ plus à gauche que p_i car ces mots st pas préf de m)

→ Si $\sigma_m(pa) \overset{(*)}{=} \varepsilon$ alors $\sigma_m(\sigma_m(pa)\tilde{w}) = \sigma_m(\tilde{w})$

À nouveau, comme aucun suffixe de pa n'est préfixe de m $(*)$, rien ne sert de chercher $\sigma_m(pa\tilde{w})$ à gauche de $\tilde{w}$ donc $\sigma_m(pa\tilde{w}) = \sigma_m(\tilde{w})$ aussi

Cela conclut la réc

- Montrons que $\mathcal{L}(A_m) = \Sigma^* m$ par double inclusion.

  $\subset$ Si $w \in \mathcal{L}(A_m)$ alors $\delta(\varepsilon, w) = m$ / déf d'ê reconnu
  $\delta(\varepsilon, w) = \sigma_m(w)$ / point précédent.

  Donc m est suffixe de w donc $w \in \Sigma^* m$

  $\supset$ Si $w \in \Sigma^* m$, $\sigma_m(w) = m$ / déf σ_m
  $\sigma_m(w) = \delta(\varepsilon, w)$ / pt ↺ donc $w \in \mathcal{L}(A_m)$

démo + propre : $\left\{\begin{array}{l} P \longrightarrow \{u^{-1}L\} \\ p \longmapsto p^{-1}L \end{array}\right.$ est injective
car si $p = m_1 \dots m_k \neq p' = m_1 \dots m_{k'}$ $(k < k')$,
$m_{k+1} \dots m_\ell \in p^{-1}L$ et $\notin p'^{-1}L$ car trop petit. Dc $|P| \leq$ nb résiduels !

- Montrons que A_m est minimal pour $\Sigma^* m = L$.

  A_m est complet, det et reconnaît $\Sigma^* m$: il suffit de mq il a le bon nb d'états cad que $|P| =$ nb de résiduels de $\Sigma^* m$. Notons $m = m_1 \dots m_\ell$. Soit $u \in \Sigma^*$ et calculons $u^{-1}L$

  En fait la valeur de $u^{-1}L$ ne dépend que de $\sigma_m(u)$ Ds tous les cas $u^{-1}L = \{$mots reconnus à partir de l'état $\sigma_m(u)$ dans $A_m\}$

  Si $\sigma_m(u) = \varepsilon$, $u^{-1}L = \Sigma^* m$ et $\forall\, k \in [\![1, \ell]\!]$, si $\sigma_m(u) = m_1 \dots m_k$, $u^{-1}L = \{m_{k+1} \dots m_\ell\} \cup \Sigma^* m$
(aucun suffixe de u préfixe de m)
$= \{$mots reconnus à partir de l'état $m_1 \dots m_k$ ds $A_m\}$

  Ces résiduels sont 2 à 2 $\neq$: ce sont donc les résiduels de L

  On en déduit que $|$résiduels de $L| =$ nb de préfixes de $m = |P|$

  Donc A_m est minimal

et 2 à 2 $\neq$ car m reconnu ds ε
$m_2 \dots m_k$ rec ds m_1, $m_3 \dots m_k$ rec ds $m_1 m_2$
(mais rec ds les autres) ≠ car de longueur ≠

(19) Une fois l'automate construit, en passant t ds A_m, on a 1 algo en $O(|t|)$ pour la recherche de m ds t. ⚠ toutefois on ne prend pas ici en compte le temps de calcul de l'automate A_m (= le calcul de ses transitions) mais y a 1 algo qui le fait en $|\Sigma||m|$