TD d'algorithmique numéro 9 - Recherche de motif

Olivier Bodini et Natacha Portier

lundi 19 novembre 2001

Question 1 — ⋆ Knuth Morris Pratt ⋆

On veut éviter le calcul de la fonction de transition de l'automate fini, de manière à améliorer l'efficacité de la recherche de motif.

On part de l'observation suivante: lorsque l'on cherche le motif *abacd* dans le texte *dfababacabacd*, au moment où l'on se rend compte que le décalage "3" ne convient pas (c'est à dire au moment d'examiner le quatrième caractère du motif, cf Figure 1), plutôt que de reprendre l'inspection avec le décalage suivant, on exploite la connaissance que l'on a du motif pour passer directement à un décalage plus "plausible".

d	f	a	b	a	b	a	c	a	b	a	c	d

a	b	a	c	d	⇝

Figure 1: Recherche de motif

Par exemple, sur la Figure 1, comme le deuxième caractère du motif est différent du premier, on peut sauter directement le décalage "4", car on sait qu'il ne conviendra pas.

Fonction préfixe En se basant sur les observations qui viennent d'être faites, proposer une manière d'associer à tout motif P de longueur m une fonction π (que l'on appellera la *fonction préfixe*), de $\{1; 2; \ldots ; m\}$ dans $\{0; 1; \ldots ; m-1\}$, qui "implémente" le saut de décalage esquissé ci-dessus. L'idée sera d'utiliser cette fonction dans une version modifiée de l'algorithme qui exécute l'automate pour la recherche de motif, de manière à en améliorer les performances (la fonction π pouvant être précalculée, avant que la recherche de motif ne commence).

Question 1,1 — ⋆ Définition ⋆

Écrire une procédure qui calcule la fonction π.

Question 1,2 — ⋆ Correction ⋆

On pose $\pi^*[q] = \{q, \pi[q], \pi^2[q], \ldots, \pi^t[q]\}$, la séquence s'arrêtant lorsque l'on atteint $\pi^t[q] = 0$.

1. Montrer que pour $q = 1, 2, \ldots, m$:

$$\pi^*[q] \; = \; \{k : P_k \sqsupset P_q\} \, .$$

2. Montrer que pour $q = 1, 2, \ldots, m$, si $\pi[q] > 0$ alors $\pi[q] - 1 \in \pi^*[q-1]$.
   Pour $q = 2, 3, \ldots, m$ on définit le sous ensemble $E_{q-1} \subseteq \pi^*[q-1]$ par

$$E_{q-1} = \{k \mid k \in \pi^*[q-1] \text{ et } P[k+1] = P[q]$$

3. Montrer que pour $q = 2, 3, \ldots, m$:

$$\pi[q] = \begin{cases} 0 & \text{si } E_{q-1} = \emptyset \\ 1 + \max\{k \in E_{q-1}\} & \text{si } E_{q-1} \neq \emptyset \end{cases}$$

4. Utiliser les raisonnements que l'on vient d'effectuer pour montrer la correction de la procédure calculant π.

Question 1,3 ⋆ Complexité ⋆

Comment évaluer la complexité de la procédure qui calcule π?

Question 1,4 ⋆ Reconnaissance de motifs ⋆

Écrire l'algorithme de recherche de motifs modifié pour utiliser les informations exploitées dans la définition π. En évaluer la complexité.

Question 1,5 ⋆ Correction du tout ⋆

Comment se justifie la correction de l'algorithme résultant ?

Question 2 ⋆ Boyer-Moore ⋆

On recherche un motif P de longueur m écrit dans un texte T avec l'alphabet Σ. L'algorithme de Boyer-Moore est efficace lorsque le motif est long et l'alphabet petit. Nous recherchons le motif de gauche à droite dans le texte, en comparant le motif avec le texte de *droite à gauche* et en faisant des sauts lorsque la comparaison échoue. Nous allons utiliser deux heuristiques pour faire des sauts de la taille la plus grande possible.

Question 2,1 ⋆ Exemple ⋆

Rechercher *r*éminiscence dans ... curieux indice pour passer par le chemin allant de la bonne recette ...

Question 2,2

Donner un algorithme qui calcule la fonction λ qui à une lettre de l'alphabet associe l'endroit le plus à droite du motif où elle apparait. Donner sa complexité. À quoi peut-elle servir ?

Question 2,3

On définit la relation $Q \sim R$ pour deux chaînes Q et R lorsque l'une est suffixe de l'autre. Les chaînes Q et R sont alors appelées chaînes similaires. Montrer que si Q est un suffixe de R et si S est un suffixe de R alors Q et S sont similaires.

Question 2,4

Pour tout entier j compris entre 0 et m on définit la fonction du bon suffixe γ :

$$\gamma[j] = m - \max\{k | 0 \leq k < m \text{ et } P[j+1, ..., m] \sim P_k\}$$

Quel est l'intéret de cette fonction ?

Question 2,5

Montrer que $\gamma[j] \leq m - \pi[m]$ pour tout j.

Question 2,6

Montrer que $\gamma[j] = m - \max(\{\pi[m]\} \cup \{k | \pi[m] < k < m \text{ et } P[j+1, ..., m] \sqsupset P_k\}$

Question 2,7

On définit P' comme l'image miroir du motif P et π' comme la fonction préfixe correspondant. Montrer que si k est la plus grande valeur posible telle que $P[j+1, ..., m] \sqsupset P_k$ alors

$$\pi'[l] = m - j$$

où $l = (m-k) + (m-j)$. Montrer que

$$\gamma[j] = \min(\{m - \pi[m]\} \cup \{l - \pi'[l] | 1 \leq l \leq m \text{ et } j = m - \pi'[l]\})$$

Question 2,8

Ecrire une fonction qui calcule γ en $\mathcal{O}(m)$.

Question 2,9

Donner un algorithme de recherche de motif qui utilise les deux décalages précédents. Donner sa complexité.