Langages Formels

Licence Informatique – ENS Cachan
Partiel du 29 mars 2012
durée 3 heures

Document autorisé : polycopié du cours.
Toutes les réponses devront être correctement justifiées.
Le chiffre en regard d'une question est une indication sur sa difficulté ou sa longueur.

1 Langages commutatifs

Soit $\Sigma = \{a_1, \ldots, a_k\}$ un alphabet de taille $k \geq 1$. *L'image de Parikh* d'un mot $v \in \Sigma^*$ est le vecteur d'entiers $\psi(v) = (|v|_{a_1}, \ldots, |v|_{a_k}) \in \mathbb{N}^k$. La définition s'étend aux langages $L \subseteq \Sigma^*$ par $\psi(L) = \{\psi(v) \mid v \in L\}$. Par exemple, si $k = 3$ alors $\psi(a_1a_2a_1) = (2, 1, 0)$ et $\psi((a_1a_3)^*) = \{(n, 0, n) \mid n \geq 0\}$.

L'application de Parikh $\psi : \Sigma^* \to \mathbb{N}^k$ est un morphisme du monoïde libre $(\Sigma^*, \cdot, \varepsilon)$ dans le monoïde commutatif $(\mathbb{N}^k, +, \bar{0})$, où $\bar{0} = (0, \ldots, 0)$ et l'addition est composante par composante. Pour $a \in \mathbb{N}$ et $u = (i_1, \ldots, i_k) \in \mathbb{N}^k$, on utilisera aussi le produit externe $au = (ai_1, \ldots, ai_k)$. Pour $u_1, \ldots, u_m \in \mathbb{N}^k$, on note $\langle u_1, \ldots, u_m\rangle$ le sous-monoïde de $\mathbb{N}^k$ engendré par $u_1, \ldots, u_m$:

$$\langle u_1, \ldots, u_m\rangle = \{a_1u_1 + \cdots + a_mu_m \mid a_1, \ldots, a_m \in \mathbb{N}\}.$$

Par exemple, $\psi((a_1a_3)^*) = \langle(1, 0, 1)\rangle$. Un sous-ensemble de $\mathbb{N}^k$ est *linéaire* s'il est de la forme

$$u_0 + \langle u_1, \ldots, u_m\rangle = \{u_0 + a_1u_1 + \cdots + a_mu_m \mid a_1, \ldots, a_m \in \mathbb{N}\}$$

avec $u_0, u_1, \ldots, u_m \in \mathbb{N}^k$. Finalement, un sous-ensemble de $\mathbb{N}^k$ est *semi-linéaire* si c'est une union finie d'ensembles linéaires.

[1] **a)** Montrer que tout ensemble semi-linéaire est l'image de Parikh d'un langage rationnel.

[1] **b)** Montrer que la somme de deux ensembles linéaires est encore linéaire.
Montrer que la somme de deux ensembles semi-linéaires est encore semi-linéaire.

Le sous-monoïde engendré par un sous-ensemble $K \subseteq \mathbb{N}^k$ est

$$\langle K\rangle = \{u_1 + \cdots + u_m \mid m \geq 0 \text{ et } u_1, \ldots, u_m \in K\}.$$

[5] **c)** Montrer que le sous-monoïde engendré par un ensemble *linéaire* est semi-linéaire mais pas forcément linéaire.
Montrer que $\langle K \cup K'\rangle = \langle K\rangle + \langle K'\rangle$.
Montrer que le sous-monoïde engendré par un ensemble *semi-linéaire* est semi-linéaire.
En déduire que l'image de Parikh d'un langage rationnel est semi-linéaire.

[1] **d)** Montrer qu'il existe des langages non algébriques dont l'image de Parikh est semi-linéaire.

Nous allons dans la suite prouver le théorème de Parikh : l'image de Parikh d'un langage algébrique est semi-linéaire.

Soit $G = (\Sigma, V, P, S)$ une grammaire algébrique. Pour un arbre de dérivation s de G, on note $\mathrm{rac}(s)$ le symbole à sa racine et $\mathrm{Fr}(s) \in (\Sigma \cup V)^*$ sa *frontière*. On note $h(s)$ la hauteur de l'arbre s (la hauteur d'un arbre trivial réduit à sa racine est 0). Finalement, on note $\mathrm{var}(s)$ l'ensemble des variables de V qui apparaissent dans l'arbre s.

Une *pompe* est un arbre de dérivation s de G qui est non trivial (n'est pas réduit à la racine) et dont la frontière est dans $\Sigma^* \cdot \mathrm{rac}(s) \cdot \Sigma^*$. Une pompe s de racine x correspond donc à une dérivation (non triviale) $x \xrightarrow{+} uxv$ de G avec $u, v \in \Sigma^*$.

On note $r \lhd t$ si r et t sont deux arbres de dérivation de G et que t peut s'obtenir à partir de r par insertion d'une pompe s à un nœud de r étiqueté par $\mathrm{rac}(s)$. En termes de dérivations, si la pompe s correspond à $x \xrightarrow{+} uxv$ alors il existe deux dérivations $\mathrm{rac}(r) \xrightarrow{*} \alpha x \gamma$ et $x \xrightarrow{*} \beta$ telles que r correspond à la dérivation $\mathrm{rac}(r) \xrightarrow{*} \alpha x \gamma \xrightarrow{*} \alpha\beta\gamma = \mathrm{Fr}(r)$ et t correspond à la dérivation $\mathrm{rac}(t) = \mathrm{rac}(r) \xrightarrow{*} \alpha x \gamma \xrightarrow{+} \alpha uxv\gamma \xrightarrow{*} \alpha u\beta v\gamma = \mathrm{Fr}(t)$.

Si $r \lhd t$ par insertion de la pompe s, on dit que s est contenue dans t. De plus le nombre de nœuds de r est strictement inférieur à celui de t puisque la pompe s est un arbre non trivial.

Une pompe *minimale* est une pompe s qui est minimale parmi les pompes pour la relation $\lhd$: si $r \lhd s$ alors r n'est pas une pompe.

[1] **e)** Montrer que si $r \lhd s$ et s est une pompe alors $\mathrm{Fr}(r) \in \Sigma^* \cdot \mathrm{rac}(r) \cdot \Sigma^*$.
Montrer que si $r \lhd s$ et s est une pompe minimale alors r est l'arbre trivial réduit à $\mathrm{rac}(s)$.

[1] **f)** Montrer que toute pompe contient une pompe minimale, i.e., si s est une pompe alors il existe un arbre de dérivation r et une pompe minimale s' tels que s s'obtient à partir de r par insertion de s'.

[4] **g)** Montrer que si s est une pompe minimale, alors $h(s) \leq 2|V|$.
Montrer qu'une grammaire contient un nombre fini de pompes minimales.
Donner une grammaire G et une pompe minimale s de G telle que $h(s) = 2|V|$.

On définit maintenant une relation d'ordre partiel sur les arbres de dérivation : $r \leq t$ si t peut s'obtenir à partir de r par une suite d'insertions de pompes minimales s telles que $\mathrm{var}(s) \subseteq \mathrm{var}(r)$. Donc si $r \leq t$ alors $\mathrm{var}(r) = \mathrm{var}(t)$ (ce qui n'est pas forcément le cas si $r \lhd t$).

[2] **h)** Soit r un arbre de dérivation de G tel que $\mathrm{Fr}(r) \in \Sigma^*$. Montrer que l'image de Parikh $\psi(\{\mathrm{Fr}(t) \mid r \leq t\})$ est un ensemble linéaire.

[5] **i)** Soit t un arbre de dérivation de G tel que $\mathrm{Fr}(t) \in \Sigma^*$. Montrer que si t contient une branche ayant au moins $|V|+1$ occurrences d'une même variable $x \in V$ alors t n'est pas $<$-minimal, i.e., il existe $r \neq t$ tel que $r \leq t$.
Montrer qu'une grammaire G admet un nombre fini d'arbres de dérivation qui sont $<$-minimaux et dont la frontière est dans Σ^*.

[1] **j)** Montrer que l'image de Parikh d'un langage algébrique est semi-linéaire.

2 Automate déterministe descendant d'arbres

Soit $A_p = \{d_1, \dots, d_p\}$ un alphabet de directions et Σ un alphabet disjoint de A_p. Soit $t : A_p^* \to \Sigma$ un arbre. Si $v \in \mathrm{dom}(t)$ est un nœud de t, on note $\mathrm{ar}_t(v)$ l'arité du nœud v dans l'arbre t. Soit $\Gamma = \Sigma \times \{0, \dots, p\} \times (A_p \cup \{\perp\})$. Le chemin d'un arbre $t : A_p^* \to \Sigma$ associé à une feuille $u = u_1 u_2 \cdots u_{n-1} u_n \in \mathrm{feuille}(t)$ est le mot de Γ^+

$$c(t,u) = (t(\varepsilon), \mathrm{ar}_t(\varepsilon), u_1)(t(u_1), \mathrm{ar}_t(u_1), u_2) \cdots (t(u_1 \dots u_{n-1}), \mathrm{ar}_t(u_1 \dots u_{n-1}), u_n)(t(u), 0, \perp)\,.$$

On note $\mathcal{C}(t) = \{c(t,u) \mid u \in \mathrm{feuille}(t)\} \subseteq \Gamma^+$ l'ensemble des chemins de t. Si $L \subseteq T_p(\Sigma)$ est un langage d'arbres, on note $\mathcal{C}(L) = \bigcup_{t \in L} C(t) \subseteq \Gamma^+$ l'ensemble des chemins des arbres de L.

[2] **a)** Montrer que si $L \subseteq T_p(\Sigma)$ est un langage d'arbres reconnaissable alors $\mathcal{C}(L) \subseteq \Gamma^+$ est un langage reconnaissable de mots.

Si $L \subseteq T_p(\Sigma)$ est un langage d'arbres, on définit la *clôture par chemins* de L comme l'ensemble $\mathcal{CC}(L) \subseteq T_p(\Sigma)$ des arbres t tels que $\mathcal{C}(t) \subseteq \mathcal{C}(L)$.

[2] **b)** Montrer que si $L \subseteq T_p(\Sigma)$ est un langage d'arbres reconnaissable alors $\mathcal{CC}(L)$ aussi.

[3] **c)** Montrer qu'un langage d'arbres reconnaissable $L \subseteq T_p(\Sigma)$ vérifie $L = \mathcal{CC}(L)$ si et seulement si il peut être reconnu par un automate déterministe descendant.

[1] **d)** Peut-on décider si un langage d'arbres reconnaissable $L \subseteq T_p(\Sigma)$ est reconnaissable par un automate déterministe descendant ?