Leçon 909 Langages rationnels - Exemples et Applications

(I) Formalisme syntaxique

1.1) Définition - Langage rationnel

définition 1 (Alphabet, mot, langage) [CAR], p 13
- Un alphabet Σ est un ensemble fini dont les éléments sont appelés lettres ou symboles
- Un mot est une suite finie d'éléments de Σ. Le mot vide est noté ε. Σ^* est l'ensemble des mots, appelé monoïde libre sur Σ
- Un langage L est une partie de Σ^*

Exemple: Mots dont la longueur est un nombre premier.

Définition 2 (Opérations rationnelles) [CAR], p 15
Les opérations rationnelles sur les langages sont :
- L'union de deux langages rationnels L et L' est noté $L+L'$.
- Le produit de deux langages rationnels L et L' est $LL' = \{uv / u\in L, v\in L'\}$
- Soit $L \subset A^*$. on définit l'étoile de L par :

$$L^0=\{\varepsilon\},\quad L^{i+1}=LL^i,\quad L^* = \bigcup_{i\geq 0} L^i$$

Définition 3 (Langages rationnels) [CAR], p 33
La classe $\mathcal{R}$ des langages rationnels sur Σ est la plus petite famille de langages tels que
(i) $\emptyset \in \mathcal{R}$ et $\{a\}\in\mathcal{R}$ pour tout $a\in\Sigma$
(ii) $\mathcal{R}$ est close par les opérations rationnelles

Exemples : $\{\varepsilon\} = \emptyset^*$, $\Sigma^* \in \mathcal{R}$

1.2) Expressions régulières

Définition 4 (Expression régulière) [WOL], p 10
La classe $\mathcal{E}$ des expressions rationnelles sur Σ est la plus petite famille d'expressions telles que
(i) $\emptyset\in\mathcal{E}$ et $a\in\mathcal{E}$ pour tout $a\in\Sigma$
(ii) Si $(E,E')\in\mathcal{E}^2$, alors $((E+E'),(EE'),E^*)\in\mathcal{E}^3$

Remarque : Par abus de notation, s'il n'y a pas ambiguïté, on omettra les parenthèses, et si $A=\{a_1,\dots,a_n\}$, on notera A par $a_1+\dots+a_n$.

Exemples : $(a+b^*)$ noté $a+b^*$, $(ab)^*a$, A^*, a^*, $(a+aa)^*$ sont des expressions rationnelles

Propriété 5 Les expressions rationnelles sont non ambigües. On peut donc définir des fonctions inductives sur les expressions rationnelles

Définition 6 Le langage $L(E)$ associé à une expression rationnelle est défini par :
(1) $L(\emptyset)=\emptyset$, $L(\varepsilon)=\{\varepsilon\}$, $L(a)=\{a\}$ pour tout $a\in\Sigma$
(2) $L((E_1+E_2)) = L(E_1)+L(E_2)$
(3) $L((E_1E_2)) = L(E_1)L(E_2)$
(4) $L(E_1^*) = L(E_1)^*$

Propriété 7 L'application $L : \mathcal{E} \to \mathcal{R}$, $E \mapsto L(E)$ est surjective.

Contre exemple pour l'injectivité : $L((a+aa)^*) = L(a^*)$

1.3) Équation sur les langages [CAR], 40

On sera amenés à résoudre des équations linéaires sur les langages. Pour cela, on introduit des outils

Lemme 8 : Lemme d'Arden
Soient K, L deux langages et l'équation $X=KX+L$, X désignant un langage
1) Si $\varepsilon\notin K$, l'unique solution est $X = K^*L$
2) Si $\varepsilon\in K$, les solutions sont de la forme $X=K^*(L+Y)$, $Y\subseteq A^*$

Elimination de Gauss
Pour résoudre un système d'équations $\{X_p = \sum_{q\in Q} A X_q + Y, p\in Q\}$ on procède par élimination successive des langages X_p en utilisant le lemme d'Arden.

II Automates finis

2.1) Définitions des automates

[CAR], p35-36

Définition 9 (Automate fini)

Un automate fini est un quintuplet $A = (Q, \Sigma, \delta, I, F)$ où Q est fini, $I \subset Q$, $F \subset Q$, $\delta \subset Q \times \Sigma \times Q$.

Les éléments de Q sont les états, ceux de I sont les états initiaux, ceux de F les états finaux, ceux de δ les transitions.

Définition 10 (chemin, chemin acceptant, mot accepté, langage accepté)

- Un chemin est une suite finie $q_0 \xrightarrow{a_1} q_1 \cdots \xrightarrow{a_n} q_n$ telle que $\forall i \in [\![1,n]\!]$, $(q_{i-1}, a_i, q_i) \in \delta$. On note $q_0 \xrightarrow{a_1 \cdots a_n} q_n$
- Un chemin est acceptant si $q_0 \in I$, $q_n \in F$
- Un mot est accepté par A s'il est l'étiquette d'un chemin acceptant.
- Le langage accepté par A est l'ensemble des mots acceptés par A
- Un langage est reconnaissable s'il existe un état qui l'accepte.

Exemple : (cf Annexe) : recherche de mot dans un texte : abaa

Définition 11 (Automate émondé) Un automate est émondé si par tout état passe au moins un chemin acceptant.

Exemple : cf Annexe

Définition 12 (Automate normalisé) A est normalisé si $I = \{i\}$, $F = \{f\}$ et $\forall a \in \Sigma, q \in Q$, $(q, a, i) \notin \delta$ et $(f, a, q) \notin \delta$

Propriété 13 Pour tout automate A, il existe un automate normalisé A' acceptant le même langage. (A et A' sont équivalents)

Exemple : cf Annexe

Définition 14 (Automate déterministe) A est déterministe si $I = \{i\}$ et $(p, a, q) \in \delta$, $(p, a, q') \in \delta \Rightarrow q = q'$

Exemple : cf Annexe

2.2) Les langages reconnaissables sont rationnels

[CAR], p75

Définition 15 (Grammaires linéaires à gauche)

- Une grammaire linéaire à gauche $\mathcal{G}$ est un quadruplet (Σ, V, P, S) où Σ et V sont des alphabets finis et disjoints, appelés terminaux et variables, et P est une partie finie de $V \times (\Sigma V \cup \{\varepsilon\})$ appelé règles. $S \in V$ est le symbole de départ.
- (Dérivation) u se dérive en v si il existe $\alpha \in \Sigma^*$, $X \in V$ et $w \in \Sigma V \cup \{\varepsilon\}$ tels que
$$u = \alpha X, \quad v = \alpha w, \quad (X, w) \in P.$$
On note $u \to v$.
On note $u \to^* v$ s'il existe une suite finie de dérivations passant de u à v.
- (Langage engendré) $\hat{L}_{\mathcal{G}} = \{v \in \Sigma^*(V \cup \{\varepsilon\}),\ S \to^* v\}$
$$L_{\mathcal{G}} = \hat{L}_{\mathcal{G}} \cap \Sigma^*.$$

Définition 16 Soit $A = (Q, \Sigma, \delta, I, F)$, $p \in Q$

[SAK], p107

$$L_p = \{f \in \Sigma^* \mid \exists t \in T,\ p \xrightarrow[A]{f} t\}$$

- On a $L_p = \bigcup_{(x,q) \mid (p,x,q) \in \delta} x L_q + \delta_{p,F}$, $\delta_{p,F} = \varepsilon$ si $p \in F$, $\emptyset$ sinon

et $L(A) = \bigcup_{p \in I} L_p$

Proposition 17 Les langages reconnaissables sont engendrés par une grammaire linéaire à gauche, et sont rationnels

2.3) Les langages rationnels sont reconnaissables

Lemme 18 Un langage L est reconnaissable si et seulement si $L \setminus \{\varepsilon\}$ est reconnaissable

- On peut construire par induction un automate normalisé reconnaissant $L \setminus \{\varepsilon\}$ défini par une expression rationnelle (cf Annexe)

Théorème de Kleene : Les langages rationnels et les langages reconnaissables sont les mêmes

III) Application du Théorème de Kleene

Le théorème de Kleene donne des conditions de rationnalité des langages.

3.1) Lemme de l'Étoile

[SAK], p78

Lemme 19 (Lemme de l'étoile)

Si L est rationnel, $\exists N \in \mathbb{N}$ tel que pour tout $f \in L$, si $f = g_1 h g_2$ et $|h| \geq N$, alors il existe une factorisation

$h = uvw$, $v \neq \varepsilon$ et $g_1 u v^* w g_2 \subset L$.

Contre exemple: Ce lemme ne donne pas une CNS : [SAK], p80

$$L = \{(aab)^n (abb)^n \mid n \in \mathbb{N}\} \cup A^*(aaa + bbb)A^*$$

Lemme 20 (Lemme de l'étoile par blocs)

Si L est rationnel, $\exists N \in \mathbb{N}$ tel que pour tout $f \in L$, pour toute factorisation $f = u v_1 v_2 \dots v_N w$, $|v_i| \geq 1$, alors il existe $0 \leq j < k \leq N$ tel que

$$u v_1 \dots v_j (v_{j+1} \dots v_k)^* v_{k+1} \dots v_N w \subset L$$

Contre exemple Ce lemme ne donne pas une CNS

$$L = \{a^{k_1} b a^{k_2} b \dots a^{k_n} b,\ \exists i,\ i \neq k_i\}$$

3.2) Quotients à gauche et minimisation

[CAR], 45, 46

Définition 21 (Quotient à gauche)

$L \subset A^*$, le quotient à gauche de L par $u \in A^*$ est $u^{-1}L = \{v \in A^*, uv \in L\}$

Proposition 22 Un langage est rationnel si et seulement si il a un nombre fini de quotients à gauche

Définition 23 (Automate minimal)

Soit L un langage rationnel. L'automate minimal de L est $A_L = (Q, \Sigma, E, \{i\}, F)$, où $\cdot\ Q = \{u^{-1}L, u \in \Sigma^*\}$

$\cdot\ i = L$

$\cdot\ E = \{u^{-1}L \to^{a} (ua)^{-1}L, u \in \Sigma^*, a \in \Sigma\}$ $\cdot\ F = \{u^{-1}L, u \in L\}$

Définition 24 Congruence de Nerode

$A = (Q, \Sigma, \delta, \{i\}, F)$ un automate déterministe complet. Alors $q \sim q' \Leftrightarrow (\forall w \in \Sigma^*, (q.w \in F \Leftrightarrow q'.w \in F))$

Propriété 25 Si A reconnaît L, l'automate minimal de L est égal à $A/\sim = (Q/\sim, \Sigma, \delta', \{[i]\}, \{[f], f \in F\})$

$\delta' = \{[q] \xrightarrow{a} [q.a],\ q \in Q, a \in \Sigma\}$.

IV) Applications des langages rationnels

4.1) Recherche de mot clans un texte

[BBC]

Algorithme de Knuth-Morris-Pratt (DVP)

4.2) Problème de séparation par Automate

[FB]

Le problème de séparation par automates est NP-complet (DVP)

Références

[CAR] Olivier Carton, Langages formels.

[WOL] Pierre Wolper, Introduction à la Calculabilité

[SAK] Jacques Sakarovitch, Éléments de théorie des automates

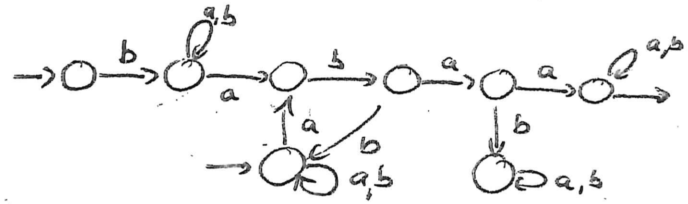
[BBC] Beauquier, Berstel, Chrétienne, Éléments d'Algorithmique..

[FB] Floyd, Beigel, The language of machines.

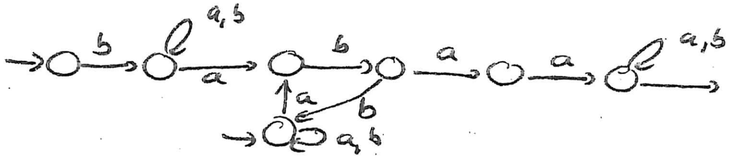
[HU] Hopcroft Ullmann, Introduction to automata theory.

Annexe

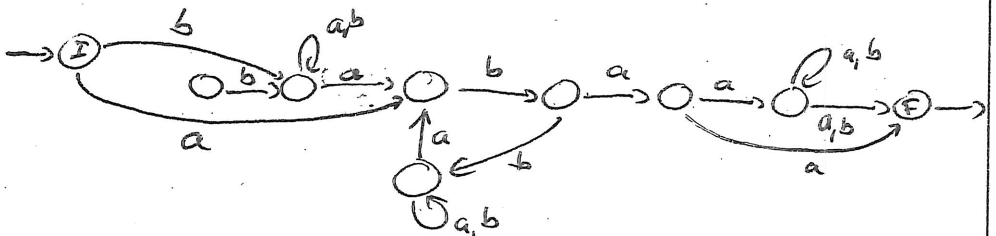
recherche de mot : abaa



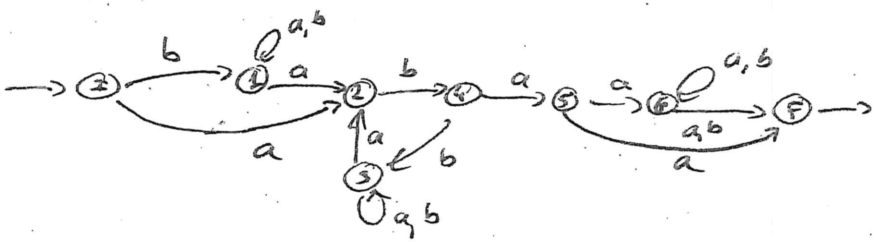
automate émondé



automate normalisé

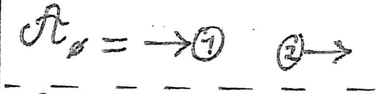


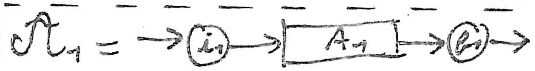
émondé



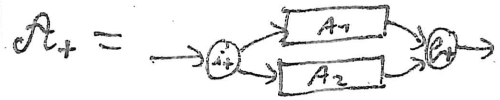
théorème de Kleene

 $\mathcal{L}(\mathcal{A}) = \{a\}$

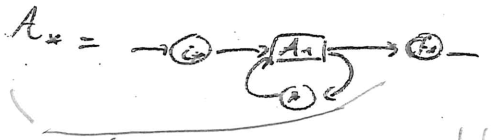
 $\mathcal{L}(\mathcal{A}_\emptyset) = \emptyset$

 $L_1 = \mathcal{L}(\mathcal{A}_1)$, $\varepsilon \notin L_1$

 $L_2 = \mathcal{L}(\mathcal{A}_2)$, $\varepsilon \notin L_2$

 $\mathcal{L}(\mathcal{A}_+) = L_1 + L_2$

 $\mathcal{L}(\mathcal{A}_.) = L_1 L_2$

 $\mathcal{L}(\mathcal{A}_*) = L_1^* \setminus \{\varepsilon\}$

Algo de Thomson ; Autre algo : Glushkov