Université Paris Cité
Automates et Analyse Lexicale

UFR Informatique
Année 2023–2024

Handout 2

Expressions Rationnelles

1 Définition

L'ensemble des *expressions rationnelles* sur un alphabet Σ est défini par induction. C'est le plus petit ensemble ExpRat tel que

- $\epsilon \in$ ExpRat
- $\emptyset \in$ ExpRat
- pour tout $a \in \Sigma$: $a \in$ ExpRat
- si $r_1, r_2 \in$ ExpRat, alors $(r_1 + r_2) \in$ ExpRat
- si $r_1, r_2 \in$ ExpRat, alors $(r_1 r_2) \in$ ExpRat
- si $r \in$ ExpRat, alors $r^* \in$ ExpRat

Les anglophones les appellent *regular expressions.*

On associe à chaque expression rationnelle r un langage $\mathcal{L}(r)$, sa sémantique. Cette association est formalisée par une fonction définie par récurrence :

- $\mathcal{L}(\epsilon) = \{\epsilon\}$
- $\mathcal{L}(\emptyset) = \{\}$ (l'ensemble vide)
- si $a \in \Sigma$, alors $\mathcal{L}(a) = \{a\}$
- $\mathcal{L}((r_1 + r_2)) = \mathcal{L}(r_1) \cup \mathcal{L}(r_2)$
- $\mathcal{L}((r_1 r_2)) = \mathcal{L}(r_1) \cdot \mathcal{L}(r_2)$
- $\mathcal{L}(r^*) = (\mathcal{L}(r))^*$

Un langage L est *rationnel* quand il existe une expression rationnelle r telle que $L = \mathcal{L}(r)$. *Rat* est la classe de tous les langages rationnelles.

Deux expressions rationnelles r_1 et r_2 sont *équivalentes* quand $\mathcal{L}(r_1) = \mathcal{L}(r_2)$.

2 Raccourcis

Quand on écrit des expressions rationnelles on se permet d'omettre des parenthèses qui ne sont pas utiles, sachant que les deux opérateurs binaires $+$ et $\cdot$ sont associatifs, et que * lie plus fortement que $\cdot$, qui lie plus fortement que $+$.

Quelques raccourcis ("sucre syntaxique") fréquents :

- $r?$ est une abréviation pour $r + \epsilon$
- r^+ est une abréviation pour rr^*

On se permet des séquences de définitions d'expressions rationnelles, c-à-d on donne chaque fois un nom à une expression rationnelle, et quand on écrit une autre expression rationnelle dans la suite on peut se servir d'un nom qu'on a déjà défini. Par exemple

$$\begin{aligned} r_1 &= a(a+b)a \\ r_2 &= br_1^*b \end{aligned}$$

Attention on a dans ce type de définitions pas droit à des cycles.

3 Questions faciles et difficiles

3.1 Propriétés de clôture de la classe des langages rationnels

On a d'abord évidemment :

- La classe des langages rationnels est close sous union, c-à-d quand L_1 et L_2 sont rationnels, alors $L_1 \cup L_2$ est également rationnel.
- La classe des langages rationnels est close sous concaténation, c-à-d quand L_1 et L_2 sont rationnels, alors $L_1 \cdot L_2$ est également rationnel.
- La classe des langages rationnels est close sous l'étoile de Kleene, c-à-d quand L est rationnel, alors L^* l'est également.

Mais on peut imaginer d'autres propriétés de clôture qui sont beaucoup moins évidentes :

- Est-ce que la classe des langages rationnels est close sous intersection ?
- Est-ce que la classe des langages rationnels est close sous complément ?

Nous allons trouver des réponses à ces questions pendant ce semestre.

3.2 Décider des langages rationnels

C'est le problème : étant donné un mot w et une expression rationnelle r, est-ce que $w \in \mathcal{L}(r)$? On peut imaginer un algorithme naïf qui essaye toutes les factorisations possibles de w et puis de faire correspondre les facteurs aux parties de r, mais ça serait énormément inefficace. Comment faire mieux ?

L'outil `grep` que vous avez vu dans le cours IS1 peut en fait pas seulement chercher un mot donné dans un texte, il peut plus généralement chercher des mots qui sont spécifiés par une expression rationnelle. Et cela de façon très efficace. Nous allons voir dans les semaines à venir la technologie qui est à la base de `grep`.

Comment faire pour décider si deux expressions rationnelles sont équivalentes ? Réponse à voir dans quelques semaines dans ce cours.

3.3 Des langages qui ne sont pas rationnels ?

Est-ce que tous les langages sont rationnels ? On peut voir avec un peu de math que cela est impossible : Étant donné un alphabet Σ, l'ensemble ExpRat des expressions rationnelles sur Σ est dénombrable infini, tandis que la classe de tous les langages sur Σ est indénombrable. Il y a donc beaucoup plus de langages que de langages rationnels.

Mais à quoi ressemble un langage non-rationnel ? En général, comment peut-on montrer qu'un langage donné n'est *pas* rationnel ? C'est une des questions que nous allons étudier vers la fin du semestre.