Langages formels : hauteur(s) d'étoile

Stéphane Le Roux leroux@lsv.fr

ENS Paris-Saclay

2023-2024

Expressions rationnelles (rappel)

Syntaxe : Soit Σ un alphabet fini et non-vide. Les expressions rationnelles sur Σ sont des mots sur l'alphabet $\underline{\Sigma} \cup \{(,), +, \cdot, ^*, \underline{\emptyset}\}$. Leur ensemble $\mathrm{ER}(\Sigma)$ est défini par récurrence structurelle.

- $\underline{\emptyset} \in \mathrm{ER}(\Sigma)$
- Pour tout $a \in \Sigma$, on pose $\underline{a} \in \mathrm{ER}(\Sigma)$
- Pour tout $E, F \in \mathrm{ER}(\Sigma)$ on pose
  - $(E + F) \in \mathrm{ER}(\Sigma)$
  - $(E \cdot F) \in \mathrm{ER}(\Sigma)$
  - $(E^*) \in \mathrm{ER}(\Sigma)$

Sémantique : À chaque expression rationnelle sur Σ, on associe un langage sur Σ, par récurrence structurelle.

- $\mathcal{L}(\underline{\emptyset}) := \emptyset$
- Pour tout $a \in \Sigma$, on pose $\mathcal{L}(\underline{a}) := \{a\}$
- Pour tout $E, F \in \mathrm{ER}(\Sigma)$ on pose
  - $\mathcal{L}(E + F) := \mathcal{L}(E) \cup \mathcal{L}(F)$
  - $\mathcal{L}(E \cdot F) := \mathcal{L}(E) \cdot \mathcal{L}(F)$
  - $\mathcal{L}(E^*) := \mathcal{L}(E)^*$

Expressions rationnelles avec mot vide explicite

Expressions rationelles, déf. alternative avec mot vide explicite, ER_ϵ

- Ajout d'un cas à la définition de la syntaxe par récurrence structurelle. $\mathrm{ER}_\epsilon ::= \underline{\emptyset} \mid \underline{\Sigma} \mid \mathrm{ER}_\epsilon + \mathrm{ER}_\epsilon \mid \mathrm{ER}_\epsilon \cdot \mathrm{ER}_\epsilon \mid \mathrm{ER}_\epsilon^* \mid \underline{\epsilon}$
- Ajout d'un cas à la définition de la sémantique par récurrence structurelle : $\mathcal{L}(\underline{\epsilon}) := \{\epsilon\}$.

Lemme

Un langage est définissable par expressions rationnelles ssi il l'est par expressions rationnelles avec mot vide explicite.

- $\mathrm{ER}(\Sigma) \subseteq \mathrm{ER}_\epsilon(\Sigma)$ et $\mathcal{L}[\mathrm{ER}(\Sigma)] \subseteq \mathcal{L}[\mathrm{ER}_\epsilon(\Sigma)]$
- Soit $f : \mathrm{ER}_\epsilon(\Sigma) \to \mathrm{ER}(\Sigma)$ définie par réc. struct. sur l'argument.
  - $f(\underline{\emptyset}) := \underline{\emptyset}$ et $f(\underline{a}) := \underline{a}$.
  - $f(E + F) := f(E) + f(F)$ et $f(E \cdot F) = f(E) \cdot f(F)$ et $f(E^*) = f(E)^*$
  - $f(\underline{\epsilon}) := \underline{\emptyset}^*$

On peut montrer par réc. struct. que f est bien à valeur dans $\mathrm{ER}(\Sigma)$ et que $\forall E \in \mathrm{ER}_\epsilon(\Sigma), \mathcal{L}(E) = \mathcal{L}(f(E))$. Donc $\mathcal{L}[\mathrm{ER}_\epsilon(\Sigma)] \subseteq \mathcal{L}[\mathrm{ER}(\Sigma)]$

Hauteur d'étoile

Hauteur d'étoile d'une expression rationnelle avec mot vide explicite

- $h(\underline{\emptyset}) := h(\underline{\epsilon}) := h(\underline{a}) := 0$ pour tout $a \in \Sigma$.
- $h(E + F) := h(E \cdot F) := \max(h(E), h(F))$
- $h(E^*) := 1 + h(E)$

Examples

- $\underline{a} + \underline{\epsilon}$
- $(\underline{a} \cdot (\underline{b} + \underline{\epsilon})) + (\underline{b} \cdot \underline{b} \cdot \underline{a})$
- $\underline{a}^* \cdot (\underline{b}^* + (\underline{a} \cdot \underline{b})^*)$
- $(\underline{a}^* \cdot \underline{b})^*$
- $(\underline{a}^*)^*$

Hauteur d'étoile d'un langage

La hauteur d'étoile d'un langage rationnel est la hauteur d'étoile minimale d'une expression rationnelle avec mot vide le définissant.

Hauteur d'étoile nulle

$\mathrm{ER}_{0,\epsilon}(\Sigma)$

- Retrait du cas de l'étoile à la définition de $\mathrm{ER}_{\epsilon}(\Sigma)$:
  $\mathcal{E} ::= \underline{\emptyset} \mid \underline{\Sigma} \mid \mathcal{E} + \mathcal{E} \mid \mathcal{E} \cdot \mathcal{E} \mid \underline{\epsilon}$
  où $\mathcal{E}$ est un raccourci pour $\mathrm{ER}_{0,\epsilon}(\Sigma)$.
- Retrait du cas de l'étoile à la définition de la sémantique par récurrence structurelle.

Lemme

Soit $E \in \mathrm{ER}_{\epsilon}$. Alors $E \in \mathrm{ER}_{0,\epsilon}$ ssi $h(E) = 0$.

- Mq $\forall E \in \mathrm{ER}_{0,\epsilon}, h(E) = 0$ par récurrence sur E.
  - $h(\underline{\emptyset}) := h(\underline{\epsilon}) := h(\underline{a}) := 0$ pour tout $a \in \Sigma$.
  - $h(E + F) := h(E \cdot F) := \max(h(E), h(F)) = 0$ par HR.
- Mq $\forall E \in \mathrm{ER}_{\epsilon}, h(E) = 0 \Rightarrow E \in \mathrm{ER}_{0,\epsilon}$
  - $\underline{\emptyset}, \underline{a} \in \mathrm{ER}_{0,\epsilon}$, donc l'implication est vraie quelque soit sa prémisse.
  - Si $h(E + F) = 0$, alors $h(E) = h(F) = 0$, d'où $E, F \in \mathrm{ER}_{0,\epsilon}$ par HR, et $E + F \in \mathrm{ER}_{0,\epsilon}$ par définition de $\mathrm{ER}_{0,\epsilon}$. Pareil pour $E \cdot F$.
  - $h(E^*) \neq 0$, donc $h(E^*) = 0 \Rightarrow E^* \in \mathrm{ER}_{0,\epsilon}$.

Hauteur d'étoile nulle (II)

Théorème

Soit Σ un ensemble fini et $L \in \Sigma^*$. Les trois assertions suivantes sont équivalentes.

1. L est fini.
2. $L \in \mathcal{L}[\mathrm{ER}_{0,\epsilon}(\Sigma)]$
3. L est de hauteur d'étoile nulle.

Montrons $1 \Leftrightarrow 2$.

- $1 \Rightarrow 2$. Posons $E_\epsilon := \underline{e}$ et $\forall ua \in \Sigma^+, E_{ua} := (E_u \cdot \underline{a})$. On a $\forall u \in \Sigma^*, \mathcal{L}(E_u) = \{u\}$, prouvable par récurrence sur u. Soit $L = \{u^1, \ldots, u^n\} \subseteq \Sigma^*$ fini. Si $L = \emptyset$ alors $L = \mathcal{L}(\underline{\emptyset})$ avec $\underline{\emptyset} \in \mathrm{ER}_{0,\epsilon}(\Sigma)$. Sinon, soit $E = (\ldots(E_{u^1} + E_{u^2})\ldots) + E_{u^{n-1}}) + E_{u^n})$. Alors $E \in \mathrm{ER}_{0,\epsilon}(\Sigma)$ et $L = \mathcal{L}(E)$.
- $2 \Rightarrow 1$.Mq $\forall E \in \mathrm{ER}_{0,\epsilon}(\Sigma), |\mathcal{L}(E)| \in \mathbb{N}$, par réc. struct. sur E.
  - $|\mathcal{L}(\underline{\emptyset})| = 0$ et $|\mathcal{L}(\underline{\epsilon}))| = |\mathcal{L}(\underline{a})| = 1$.
  - $|\mathcal{L}(E+F)| \leq |\mathcal{L}(E)| + |\mathcal{L}(F)|$ et $\mathcal{L}(E \cdot F) \leq |\mathcal{L}(E)||\mathcal{L}(F)|$ sont finis si $|\mathcal{L}(E)|$ et $|\mathcal{L}(F)|$ le sont.

Hauteur d'étoile arbitraire

Théorème

Pour tout $n \in \mathbb{N}$ il existe un langages rationnel de hauteur d'étoile n.

Preuve

En TD ?

Expressions rationnelles avec mot vide et complément

Expressions rationnelles avec mot vide et complément, $\mathrm{ER}_{\epsilon,c}$

Ci-dessous, on écrit $\mathcal{E}$ ou lieu de $\mathrm{ER}_{\epsilon,c}$.

- $\mathcal{E} ::= \underline{\emptyset} \mid \underline{\Sigma} \mid \mathcal{E} + \mathcal{E} \mid \mathcal{E} \cdot \mathcal{E} \mid \mathcal{E}^* \mid \underline{\epsilon} \mid \mathcal{E}^c$
- $\mathcal{L}(E^c) := \Sigma^* \setminus \mathcal{L}(E)$.

Lemme

$\mathcal{L}[\mathrm{ER}] = \mathcal{L}[\mathrm{ER}_{\epsilon,c}]$

Preuve

- $\mathrm{ER} \subseteq \mathrm{ER}_{\epsilon,c}$ et $\mathcal{L}[\mathrm{ER}] \subseteq \mathcal{L}[\mathrm{ER}_{\epsilon,c}]$
- Pas de moyen simple de traduire $E \in \mathrm{ER}_{\epsilon,c}$ en $f(E) \in \mathrm{ER}$. Pour tout $E \in \mathrm{ER}_{\epsilon,c}$, construisons un ($\epsilon$-)AFN $\mathcal{A}_E$ telle que $\mathcal{L}(\mathcal{A}_E) = \mathcal{L}(E)$. Par récurrence sur E.
  - Pour les cas autres que E^c, on invoque les constructions vues en cours.
  - Pour E^c, on déterminise $\mathcal{A}_E$ avant d'inverser les états acceptants.

Langages sans étoile

Expressions sans étoile et avec complément

$\mathrm{ER}_{0,c}$ est défini par $\mathcal{E} ::= \underline{\emptyset} \mid \underline{\Sigma} \mid \mathcal{E} + \mathcal{E} \mid \mathcal{E} \cdot \mathcal{E} \mid \mathcal{E}^c$

Langage sans étoile

Un langage est dit sans étoile s'il est définissable par une expression de $\mathrm{ER}_{0,c}$

- $\Sigma^* = \mathcal{L}(\underline{\emptyset}^c)$
- $\Sigma^* a \Sigma^* bb \Sigma^* = \mathcal{L}(\underline{\emptyset}^c \cdot \underline{a} \cdot \underline{\emptyset}^c \cdot \underline{b} \cdot \underline{b} \cdot \underline{\emptyset}^c)$
- $\{\epsilon\} = \Sigma^* \setminus \cup_{a \in \Sigma} a \Sigma^* = \mathcal{L}((+_{a \in \Sigma} \underline{a} \cdot \underline{\emptyset}^c)^c)$
- $(ab)^+ = (a\Sigma^* \cap \Sigma^* b) \setminus \Sigma^*(aa + bb)\Sigma^*$ (pour $a \neq b$)
- $(ab)^* = (ab)^+ + \epsilon$
- $(aa)^*$ n'est pas sans étoile. Et $(aa)^+$?

Autres expressions rationnelles

- Rappel : $\mathrm{ER}_{\epsilon,c}$ est défini par $\mathcal{E} ::= \underline{\emptyset} \mid \underline{\Sigma} \mid \mathcal{E}+\mathcal{E} \mid \mathcal{E}\cdot\mathcal{E} \mid \mathcal{E}^* \mid \underline{\epsilon} \mid \mathcal{E}^c$.
- Rappel : $\mathrm{ER}_{0,c}$ est défini par $\mathcal{E} ::= \underline{\emptyset} \mid \underline{\Sigma} \mid \mathcal{E}+\mathcal{E} \mid \mathcal{E}\cdot\mathcal{E} \mid \mathcal{E}^c$

Expressions sans étoile, avec mot vide et complément

$\mathrm{ER}_{0,\epsilon,c}$ est défini par $\mathcal{E} ::= \underline{\emptyset} \mid \underline{\Sigma} \mid \mathcal{E}+\mathcal{E} \mid \mathcal{E}\cdot\mathcal{E} \mid \underline{\epsilon} \mid \mathcal{E}^c$

Lemme

$\mathcal{L}[\mathrm{ER}_{0,\epsilon,c}] = \mathcal{L}[\mathrm{ER}_{0,c}]$

Preuve par double inclusion.

- On a $\mathrm{ER}_{0,c} \subseteq \mathrm{ER}_{0,\epsilon,c}$ et $\mathcal{L}[\mathrm{ER}_{0,c}] \subseteq \mathcal{L}[\mathrm{ER}_{0,\epsilon,c}]$.
- On a $\forall E \in \mathrm{ER}_{0,\epsilon,c} \exists f(E) \in \mathrm{ER}_{0,c}, \mathcal{L}(f(E)) = \mathcal{L}(E)$. En effet, on remplace $\underline{\epsilon}$ par $(+_{a\in\Sigma}\underline{\emptyset}^c \cdot \underline{a} \cdot \underline{\emptyset}^c)^c$ par réc. sur $E \in \mathrm{ER}_{0,\epsilon,c}$. D'où $\mathcal{L}[\mathrm{ER}_{0,\epsilon,c}] \subseteq \mathcal{L}[\mathrm{ER}_{0,c}]$.

Hauteur d'étoile généralisée

Hauteur d'étoile dans $\mathrm{ER}_{\epsilon,c}$

- $H(\underline{\emptyset}) := H(\underline{\epsilon}) := H(\underline{a}) := 0$ pour tout $a \in \Sigma$.
- $H(E + F) := H(E \cdot F) := \max(H(E), H(F))$
- $H(E^*) := 1 + H(E)$
- $H(E^c) := H(E)$

- $(\underline{a} \cdot (\underline{b} + \underline{\epsilon})^c) + (\underline{b} \cdot \underline{b} \cdot \underline{a})$
- $((\underline{a}^* \cdot \underline{b})^c)^* + (((\underline{a}^c)^c)^c)^c$

Hauteur d'étoile généralisée d'un langage

La hauteur d'étoile généralisée d'un langage rat. est la hauteur d'étoile minimale d'une expression rat. avec ϵ et complément le définissant.

- Lemme : $H(L) \leq h(L)$
- On ne sait pas (encore) s'il existe des langages rationnels de hauteur d'étoile généralisée 2 ou plus.

Langages sans étoile (II)

- Rappel dans $\mathrm{ER}_{\epsilon,c}$:
  - $H(\underline{\emptyset}) := H(\underline{\epsilon}) := H(\underline{a}) := 0$ pour tout $a \in \Sigma$.
  - $H(E+F) := H(E \cdot F) := \max(H(E), H(F))$
  - $H(E^*) := 1 + H(E)$
  - $H(E^c) := H(E)$
- Rappel : $\mathrm{ER}_{0,c}$ est défini par $\mathcal{E} ::= \underline{\emptyset} \mid \underline{\Sigma} \mid \mathcal{E} + \mathcal{E} \mid \mathcal{E} \cdot \mathcal{E} \mid \mathcal{E}^c$

Lemme

Un langage rationnel est sans étoile ssi il est de hauteur d'étoile généralisée nulle.

Preuve

$L \subseteq \Sigma^*$ est sans étoile ssi $\exists E \in \mathrm{ER}_{0,c}$ telle que $L = \mathcal{L}(E)$, ssi $\exists E \in \mathrm{ER}_{0,\epsilon,c}$ telle que $L = \mathcal{L}(E)$ (car $\mathcal{L}[\mathrm{ER}_{0,c}] = \mathcal{L}[\mathrm{ER}_{0,\epsilon,c}]$)

- $\mathrm{ER}_{0,c} \subseteq \mathrm{ER}_{\epsilon,c}$. On a $\forall E \in \mathrm{ER}_{0,c}, H(E) = 0$, par récurrence facile sur E.
- On a $\forall E \in \mathrm{ER}_{\epsilon,c}, H(E) = 0 \Rightarrow E \in \mathrm{ER}_{0,\epsilon,c}$, par réc. facile sur E.

Avant-goût du théorème de Schützenberger

Théorème de Schützenberger, 1965

Soit L un langage rationnel. Les trois assertions suivantes sont équivalentes.

1. L est sans étoile
2. Le monoïde syntaxique de L est fini et apériodique
3. L est reconnu par un monoïde fini et apériodique

Retour sur les monoïdes

Monoïde apériodique

Un groupe $(G, \cdot_G, e_G)$ est dit contenu dans un monoïde $(M, \cdot_M, e_M)$ si

- $G \subseteq M$
- $\forall x, y \in G, x \cdot_G y = x \cdot_M y$

Si de plus $e_G = e_M$, on parle de sous-groupe d'un monoïde.
Un monoïde est dit apériodique s'il ne contient que des groupes triviaux, i.e. réduits à un élément.

"Poêle à frire"

Soit M un monoïde fini. Pour tout $x \in M$, soient $l_x \geq 0$ et $p_x > 0$ tels que $x^{l_x+p_x} = x^{l_x}$ et $l_x + p_x$ minimal pour cette propriété.

- Donc $1, x, \ldots, x^{l_x+p_x-1}$ sont distincts deux à deux.
- Si $l \geq 0$ et $p > 0$ tel que $x^l = x^{l+p}$, alors $l_x \leq l$ et $p_x \leq p$.

Caractérisation des groupes et monoïdes apériodiques

Proposition

Soit M un monoïde fini. Les assertions suivantes sont équivalentes.

1. M est un groupe.
2. Pour tout $x \in M$, on a $l_x = 0$ (donc $x^{p_x} = 1$).
3. $\exists k > 0, \forall x \in M,\ x^k = 1$

Lemme

$G_x := \{x^{l_x}, \ldots, x^{l_x + p_x - 1}\}$ est un groupe.

Lemme

Soit M un monoïde fini. Les assertions suivantes sont équivalentes.

1. M est un monoïde apériodique.
2. Pour tout $x \in M$, on a $p_x = 1$ (et donc $x^{l_x+1} = x^{l_x}$)
3. $\exists k \in \mathbb{N}, \forall x \in M,\ x^{k+1} = x^k$

Théorème de Schützenberger

Example

Soit $\Sigma := \{a\}$ et $L := (aa)^+$. Alors $(\Sigma^*/\approx_L) = \{[\epsilon],[aa],[a]\}$. De plus $\{[aa],[a]\}$ est un groupe, donc $\Sigma^*/\approx_L$ n'est pas apériodique.

Théorème de Schützenberger, 1965

Soit L un langage rationnel. Les trois assertions suivantes sont équivalentes.

1. L est sans étoile
2. Le monoïde syntaxique de L est fini et apériodique
3. L est reconnu par un monoïde fini et apériodique

Preuve de $2 \Rightarrow 3$

Le monoïde syntaxique reconnaît L (via la surjection canonique)

Preuve du théorème de Schützenberger, $3 \Rightarrow 2$

Rappel : Si un monoïde M reconnaît $L \subseteq \Sigma^*$, alors $\Sigma^*/\approx_L$ est isomorphe à un quotient d'un sous-monoïde de M

Lemme

1. Tout sous-monoïde d'un monoïde apériodique est apériodique.
2. Tout quotient d'un monoïde apériodique est apériodique.

Preuve

1. Les groupes inclus dans le sous-monoïde sont aussi inclus dans le monoïde de départ.
2. Soit k tel que $\forall x \in M, x^{k+1} = x_k$. Soit $\sim$ une congruence sur M. Alors $[x]_\sim^{k+1} = [x^{k+1}]_\sim = [x^k]_\sim = [x]_\sim^k$.

Corollaire, $3 \Rightarrow 2$

Si M reconnaissant L est apériodique, alors $\Sigma^*/\approx_L$ l'est aussi

Preuve du théorème de Schützenberger, $1 \Rightarrow 2$

Lemme

Le monoïde syntaxique d'un langage rationnel L est apériodique ssi il existe $k \in \mathbb{N}$ tel que $\forall x, y, z \in \Sigma^*, xy^k z \in L \Leftrightarrow xy^{k+1} z \in L$

Pour tout langage L de monoïde syntaxique apériodique soit $i(L)$ le plus petit entier k vérifiant la formule du lemme ci-dessus.

Lemme

- Le langage vide et les $\{a\}$ ont des monoïdes syntaxiques apériodiques.
- Les langages rationnels de monoïdes syntaxiques apériodiques sont clos par union, concaténation, complémentaire.
- plus précisément :
  - $i(\emptyset) = 0$ et $i(\{a\}) = 2$.
  - $i(\Sigma^* \setminus L) = i(L)$
  - $i(K \cup L) \leq \max(i(K), i(L))$.
  - $i(KL) \leq i(K) + i(L)$

$1 \Rightarrow 2$: Si L est un langage rationnel sans étoile, son monoïde syntaxique est apériodique.

Preuve du théorème de Schützenberger, $2 \Rightarrow 1$

Mq si le monoïde syntaxique de L est fini et apériodique, alors L est sans étoile

Pas dans ce cours.