Langages rationnels : Exemples et applications

Les instances d'un problème donné peuvent être représentées par une chaîne de caractères :
- → nombres (séquences de bits)
- → image (suites des intensités de chaque point)
- → son (suite de nombres représentant les tonalités)

Ainsi, beaucoup de problèmes se réduisent à décider l'appartenance d'un élément à un ensemble. Si l'on peut décrire l'ensemble à l'aide d'une expression rationnelle, la théorie des langages rationnels intervient pour résoudre le problème. Celle-ci est très liée à la théorie des automates finis.

(I) Généralités [Ro3, Wo?, HOP]

1) Langage sur un alphabet

Définition 1 : Un alphabet Σ est un ensemble non vide fini de symboles, appelés lettres.

exemple : $A = \{0,1\}$; $B = \{a,b,c,e\}$ sont des alphabets.

Définition 2 : Un mot sur Σ est une séquence finie de lettres de Σ

exemple :
- ε = mot vide et un mot sur tout alphabet
- $w_1 = 001$; $w_2 = 101$ sont des mots sur A.
- $w_1' = abce$; $w_2' = aec$ sont des mots sur B.

Notations :
- $|w|$ est la longueur du mot, ie. le nombre de symboles qui le compose. ($|\varepsilon| = 0$)
- Σ^* est l'ensemble des mots sur Σ.
- $w_1 w_2 = w_1 \cdot w_2$ est la concaténation de w_1 et w_2 : les symboles de w_1 suivis de ceux de w_2.

Définition 3 : Un langage L sur Σ est un ensemble de mots sur Σ. C'est une partie de Σ^*.

2) Langages et expressions rationnelles

Définition 4 : On définit les expressions rationnelles de manière inductive :

- Cas de base : $\emptyset$, ε, $a \in \Sigma$ sont des expressions rationnelles
- Induction : Soient $\mathcal{E}_1, \mathcal{E}_2$ deux expressions rationnelles.
  - $\mathcal{E}_1 + \mathcal{E}_2 = \mathcal{E}_1 \cup \mathcal{E}_2$ est une expression rationnelle : clôture par union
  - $\mathcal{E}_1 \cdot \mathcal{E}_2 = \{w_1 w_2 / (w_1, w_2) \in \mathcal{E}_1 \times \mathcal{E}_2\}$ est une expression rationnelle : clôture par concaténation
  - $\mathcal{E}_1^*$ est une expression rationnelle : clôture par passage à l'étoile.

Définition 5 : On définit par le même procédé d'induction le langage associé à une expression rationnelle $\mathcal{E}$ noté $\mathcal{L}(\mathcal{E})$:

- Cas de base : $\mathcal{L}(\emptyset) = \emptyset$; $\mathcal{L}(\varepsilon) = \{\varepsilon\}$, $\mathcal{L}(a) = \{a / a \in \Sigma\}$
- Induction : Si $\mathcal{E}_1$ et $\mathcal{E}_2$ sont deux expressions rationnelles, alors $\mathcal{L}(\mathcal{E}_1 + \mathcal{E}_2) = \mathcal{L}(\mathcal{E}_1) \cup \mathcal{L}(\mathcal{E}_2)$, $\mathcal{L}(\mathcal{E}_1 \cdot \mathcal{E}_2) = \mathcal{L}(\mathcal{E}_1) \cdot \mathcal{L}(\mathcal{E}_2)$, $\mathcal{L}(\mathcal{E}_1^*) = \mathcal{L}(\mathcal{E}_1)^*$.

exemples :
- $\mathcal{L}\left((0+1)\cdot[(0+1)\cdot(0+1)]^*\right)$ est l'ensemble des mots de longueur impaire sur A.
- $\mathcal{L}\left(((b\cdot c)^* + (c\cdot b)^*) + (b\cdot(c\cdot b)^*) + (c\cdot(b\cdot c)^*)\right)$ est l'ensemble des mots où b et c sont alternés. ■

Définition 6 : Un langage L est dit rationnel s'il existe une expression rationnelle $\mathcal{E}$ telle que $L = \mathcal{L}(\mathcal{E})$. On notera $Rat(\Sigma)$ l'ensemble des langages rationnels sur Σ.

Remarques :
- $Rat(\Sigma)$ est la plus petite classe de langages stable par $\cup, \cdot, ^*$ et contenant $\emptyset, \{\varepsilon\}, \{a\}$ pour $a \in \Sigma$.
- Plusieurs expressions peuvent décrire un même langage :
  $\mathcal{L}((ab)^* + (ba)^* + a(ba)^* + b(ab)^*) = \mathcal{L}((\varepsilon+b)(ab)^*(\varepsilon+a))$
  ce qui motive la définition suivante.

Définition 7 : $\mathcal{E}_1$ et $\mathcal{E}_2$ sont dites équivalentes si elles sont associées au même langage : $\mathcal{L}(\mathcal{E}_1) = \mathcal{L}(\mathcal{E}_2)$.

Proposition 8 : L'ensemble des sous mots d'un langage L quelconque est rationnel

Proposition 9 : Il existe des langages non rationnels

Problèmes :
- Peut-on déterminer si deux expressions rationnelles sont équivalentes ?
- Pour w mot sur Σ et L langage sur Σ, peut-on déterminer si $w \in L$?

II Lien avec les automates finis [Car ; Sak ; Hop]

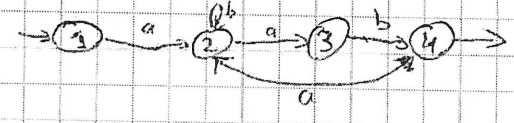
Les automates finis donnent une autre définition des langages rationnels et sont un outil parfois plus facile à manier pour certaines preuves.

1) Préliminaires

Définition 10 : Un automate fini $\mathcal{A}$ sur Σ est la donnée d'un quintuplet $(Q, \Sigma, \delta, I, F)$, avec Q ensemble fini, $I, F \subseteq Q$, $\delta \subseteq Q \times \Sigma \times Q$. Les éléments de Q sont appelés les états, ceux de I les états initiaux, ceux de F les états finaux. Les éléments (q, a, q') de δ sont les transitions de l'automate.

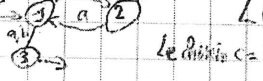
Remarques :
- Généralement, on prendra $Q = [\![1, n]\!]$, $n \in \mathbb{N}^*$ car le nom des états n'importe pas pour l'étude de l'automate
- On représente un automate $\mathcal{A}$ par un graphe avec arêtes étiquetées et des étiquettes sur les états initiaux et finaux.

Exemple : $\mathcal{A} = (\{1,2,3,4\}; \{a,b\}; \{(1,a,2),(2,b,2),(2,a,3),(3,b,4),(4,a,2)\}, \{1\}, \{4\})$ est représenté par :



Définition 11 :
- Un chemin dans $\mathcal{A}$ est une suite de transitions consécutives $(q_0, a_1, q_1)(q_1, a_2, q_2)\dots(q_{n-1}, a_n, q_n)$ notée $q_0 \xrightarrow{a_1 \dots a_n} q_n$ où q_0 et q_n sont les états de départ et d'arrivée.
- Un chemin est dit acceptant si $q_0 \in I$ et $q_n \in F$.
- Si $q_0 \xrightarrow{a_1 \dots a_n} q_n$ est un chemin, on dit que $a_1 \dots a_n$ est l'étiquette du chemin.

Définition 12 : Un mot est dit accepté par $\mathcal{A}$ si il est l'étiquette d'un chemin acceptant dans $\mathcal{A}$. On note $L(\mathcal{A})$ l'ensemble des mots acceptés par $\mathcal{A}$.

Exemple : $\mathcal{A} =$  $L(\mathcal{A}) = a(bb^*a)^*(a+b)$.

Le chemin $c = (a,a,3)(3,b,2)(2,a,3)(3,b,3)$ est acceptant.

Définition 13 : Un automate est dit émondé si par tout état passe un chemin acceptant

Définition 14 : Un automate est dit normalisé si $|I| = |F| = 1$ et qu'il n'y a pas de transitions de la forme (q, a, i) ou (f, a, q) si $I = \{i\}$ et $F = \{f\}$.

Proposition 15 : Pour tout automate $\mathcal{A}$, il existe un automate normalisé $\mathcal{A}'$ tel que $L(\mathcal{A}) = L(\mathcal{A}') \setminus \{\varepsilon\}$

2) Langages rationnels et reconnaissables

Lemme 16 : (Arden) L'équation $X = AX + B$ d'inconnue X avec A, B deux langages sur Σ fixé admet pour solution :
- Si $\varepsilon \notin A$, $X = A^*B$
- Si $\varepsilon \in A$, $\exists$ langage $P \subset \Sigma^*$, $X = A^*(B + P)$

Théorème 17 : (Kleene) [DEV]

Un langage L est rationnel si et seulement si il est reconnaissable par automate fini, c'est-à-dire qu'il existe $\mathcal{A}$ tel que $L = L(\mathcal{A})$.

Lemme de l'étoile 18 : Soit $L \in \mathrm{Rat}(\Sigma)$, alors $\exists N \in \mathbb{N}^*, \forall \beta \in \Sigma^*$

$$(|\beta| \geq N) \Rightarrow (\exists (u,v,w) \in (\Sigma^*)^3,\ v \neq \varepsilon,\ \beta = uvw \text{ et } |uv| \leq N \text{ et } uv^*w \subseteq L)$$

Lemme de l'étoile par blocs 19 : Soit $L \in \mathrm{Rat}(A)$, alors $\exists N \in \mathbb{N}^*, \forall f \in L$, pour toute factorisation $f = uv_1 \dots v_N w$, $v_i \neq \varepsilon$, $\exists 0 \leq j < k \leq N$,

$$uv_j \dots v_{k-1} (v_{j+1} \dots v_k)^* v_{k+1} \dots v_N w \subseteq L.$$

Applications :
- $\{a^n b^n / n \in \mathbb{N}\}$ n'est pas rationnel
- L'ensemble des palindromes n'est pas rationnel

Cela donne une preuve à la proposition 9 qui est difficile à montrer sans les automates

- $\mathrm{Rat}(\Sigma)$ est stable par $\cap, {}^c, \setminus$.

Proposition 20 : L'appartenance d'un mot à un langage rationnel est décidable.

Corollaire 21 : L'équivalence de deux expressions rationnelles est décidable

3) Quotients et automates minimaux [Car]

Définition 22 : Soit $L \subseteq \Sigma^*$ un langage sur Σ. Le quotient à gauche de L par un mot $u \in \Sigma^*$ est le langage $u^{-1}L = \{v \in \Sigma^* / uv \in L\}$.
Le quotient à gauche de L par $K \subseteq \Sigma^*$ est $K^{-1}L = \bigcup_{u \in K} u^{-1}L$.

Proposition 23 : ~~$\forall L, \forall u \in \Sigma^*$~~, pour tout mot u, v, et w et tous langages K et L, on a les relations suivantes :

- $w^{-1}(K \cup L) = w^{-1}K \cup w^{-1}L$
- $a^{-1}(KL) = (a^{-1}K)L + \varepsilon(K)a^{-1}L$.
- $w^{-1}(KL) = (w^{-1}K)L + \sum_{uv=w} \varepsilon(u^{-1}K)v^{-1}L$
- $a^{-1}(L^*) = (a^{-1}L)L^*$
- $w^{-1}L^* = \sum_{uv=w} \varepsilon(u^{-1}L^*)(v^{-1}L)L^*$
- $(uv)^{-1}L = v^{-1}(u^{-1}L)$

($\varepsilon(L) = \varepsilon \in L$)

Définition 24 : Un automate $\mathcal{A} = (Q, \Sigma, \delta, I, F)$ est déterministe si :

- $|I| = 1$
- $(p,a,q), (p,a,q') \in \delta \Rightarrow q = q'$

Lemme 25 : Soit $\mathcal{A} = (Q, \Sigma, \delta, \{i\}, F)$ un automate déterministe tel que $L(\mathcal{A}) = L$.

Pour tout chemin $i \xrightarrow{u} q$, on a $u^{-1}L = \{v / q \xrightarrow{v} f \text{ avec } f \in F\}$

Définition 26 : Soit $L \in Rat(\Sigma)$. L'automate minimal de L est un automate $\mathcal{A}_L = (Q, \Sigma, \delta, I, F)$ où $Q = \{a^{-1}L / u \in \Sigma^*\}$, $I = \{L\}$, $F = \{u^{-1}L / u \in L\}$
$\delta = \{u^{-1}L \xrightarrow{a} (ua)^{-1}L / (q,a) \in Q \times A\}$.

Lemme 27 : $L(\mathcal{A}_L) = L$

Corollaire 28 : $(L \in Rat(\Sigma)) \Leftrightarrow$ (L admet un nombre fini de quotient à gauche)

III. Applications des langages rationnels [BBC], ~~[Lot]~~, [Car]

1) Recherche de motifs

On veut rechercher un mot $u \in \Sigma^*$ dans un texte. Généralement, on veut reconnaître $\Sigma^* u \Gamma^*$

→ biologie : recherches de motifs dans une chaine de nucléotides.
→ informatique : traitement de texte, recherche en remplacement d'une chaine de caractères par une autre.

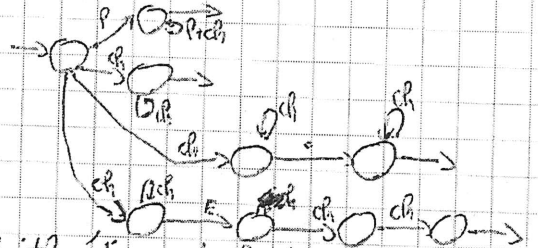
Résolution du problème par l'algorithme de Knuth-Morris-Pratt.

2) Analyse lexicale

Il s'agit de la première étape d'un compilateur avant l'analyse syntaxique. On décompose le programme source du compilateur en unités lexicales, c'est-à-dire que l'on regroupe les caractères qui codent le même élément du programme (entier, mot du langage : begin, else, ...).

exemple : On veut reconnaître $P(ch + P)^* + ch.ch^* + (ch.ch^*.ch^*.ch^* + ch.ch^*.ch.ch)$
où $ch \in \{a, b, \ldots, z\}$ et $P \in \{0, \ldots, 9\}$.

Il suffit de manipuler :



3) Arithmétique de Presburger. [DEV] [Car]

On se place dans le cadre de la logique du premier ordre sur $\mathbb{N}$.
On considère les opérations $+$ et $=$ uniquement. Appelle ceci l'arithmétique de Presburger.

Définition 29 : On dit que une théorie logique est décidable si le problème de savoir si une formule close est vrai est décidable.

Théorème 30 (Presburger) : La théorie au premier ordre des entiers munie de l'addition est décidable.

Réf : [BBC] Beauquier, Bretel, Christienne / [Car] Carton / [Lot] Lot / [Per] Perrin / [Sak] Sakarovitch
[Hop] Hopcroft - Ullman / [Roz] Rozenberg - Salomaa