X5I0050 - Langages et automates
Langages rationnels

D. Béchet & T. Sadiki

Université de Nantes & Université Internationale de Rabat

17 septembre 2014

Introduction

Langages rationnels (appelés aussi langages réguliers)
Parmi tous les langages constitués par les parties d'un monoïde libre A^* engendrés par un alphabet A, on distingue une classe particulière : la classe **Rat(A^*)** des langages rationnels sur A

Intérêts

- Caractérisation précise du langage
- Une méthode simple permet de décider pour chacun de ses langages si un mot lui appartient ou non : les automates finis
- Présent dans de très nombreux langages de programmation et utilitaires pour :
  - Tester l'appartenance d'une sous-chaîne de caractères à un langage rationnel
  - Diviser une chaîne de caractères en sous-chaîne de caractères / *tokens*
  - Effectuer des remplacements de sous-chaînes de caractères par d'autres chaînes de caractères

Plan Chapitre 2

1. Langages rationnels
2. Expressions rationnelles
3. Définitions rationnelles
4. Systèmes d'équations rationnelles
5. Lemme de l'étoile

Définition d'un langage rationnel

Définition 2.1 - Langage rationnel

Soit A un alphabet. Les **langages rationnels** sur A sont les éléments de la classe **Rat(A^*)** définie récursivement de la façon suivante : **Rat(A^*)** est le plus petit sous-ensemble de $\mathcal{P}(A^*)$ tel que :

1. $\emptyset \in Rat(A^*)$
2. $\{\varepsilon\} \in Rat(A^*)$
3. $\forall a \in A^*, \{a\} \in Rat(A^*)$
4. **Union** Si $L_1 \in Rat(A^*)$ et $L_2 \in Rat(A^*)$ alors $L_1 \cup L_2 \in Rat(A^*)$
5. **Produit** Si $L_1 \in Rat(A^*)$ et $L_2 \in Rat(A^*)$ alors $L_1 \times L_2 \in Rat(A^*)$
6. **Fermeture** Si $L_1 \in Rat(A^*)$ alors $L_1^* \in Rat(A^*)$

Théorème des parties finies

Théorème 2.1 - Théorème des parties finies

Toute partie finie L de A^* est dans $Rat(A^*)$

Démonstration : très simplement, un langage fini est l'union des langages contenant un des mots du langage

Exemples de langages rationnels

1. Soit l'alphabet $A = \{0, 1\}$, le langage décrivant les symboles de la table ASCII en binaire isur 8 bits est-il un langage rationnel ?
2. Soit l'alphabet $A = \{0, 1\}$, le langage décrivant les chaînes de caractères (en Pascal ou C par exemple) en binaire (sur 8 bits) est-il un langage rationnel ?

Exemples de langages rationnels - Réponses

1. Soit l'alphabet $A = \{0, 1\}$, le langage décrivant les symboles de la table ASCII en binaire sur 8 bits est-il un langage rationnel ?
   $\Longrightarrow$ Oui car c'est un langage fini (128 mots possibles)
2. Soit l'alphabet $A = \{0, 1\}$, le langage décrivant les chaînes de caractères (en Pascal ou C par exemple) en binaire (sur 8 bits) est-il un langage rationnel ?
   $\Longrightarrow$ Oui. En Pascal, les chaînes ont au plus 255 caractères et donc le langage est fini. En C, une chaîne est une suite de caractères en binaire (sauf le caractère de code binaire 00000000) terminée par un caractère de code binaire 00000000. La réponse est plus complexe si l'on s'intéresse au chaînes valides codées en UTF-8, UTF-16 ou UTF-32

Définition d'une expression rationnelle

Définition 2.2 - Expression rationnelle

Les **expressions rationnelles** sur A décrivent les **langages rationnels**.
Elles sont définies de la façon suivante :

1. $\emptyset$ est une expression rationnelle qui décrit le langage rationnel $\emptyset$
2. ε est une expression rationnelle qui décrit le langage rationnel $\{\varepsilon\}$
3. $\forall a \in A$, **a** est une expression rationnelle qui décrit le langage rationnel $\{a\}$
4. Si l_1 et l_2 sont des expressions rationnelles qui décrivent L_1 et $L_2 \in Rat(A^*)$ alors :
   1. $(\mathbf{l_1}|\mathbf{l_2})$ est une expression rationnelle qui décrit le langage rationnel $L_1 \cup L_2 \in Rat(A^*)$
   2. $(\mathbf{l_1}.\mathbf{l_2})$ est une expression rationnelle qui décrit le langage rationnel $L_1 \times L_2 \in Rat(A^*)$
   3. $\mathbf{l_1^*}$ est une expression rationnelle qui décrit le langage rationnel $L_1^* \in Rat(A^*)$

Notations et propriétés des expressions rationnelles

Préséance : (pour éviter les parenthèses inutiles)

$* > . > |$

Par exemple : $a.b^*|c|d^* \equiv ((a.(b^*))|(c|(d^*)))$

Notations

Le . est le plus souvent omis

Comme $l.\epsilon = \epsilon.l = l$ et $l.\emptyset = \emptyset.l = \emptyset$, ϵ et $\emptyset$ ne sont en général pas utilisés

Ajout de l'opérateur $+$: $a^+ = (a.a^*)$

Ajout de l'opérateur $?$: $a^? = (a|\varepsilon)$

Propriétés

Mêmes propriétés que sur les langages (associativité, distributivité, idempotence, élément neutre/absorbant, ...)

Théorème des expressions rationnelles

Théorème 2.2 - Théorème des expressions rationnelles

Tout langage dénoté par une expression rationnelle est un langage rationnel et tout langage rationnel est dénoté par une expression rationnelle

Démonstration : très simple en comparant la manière dont les expressions rationnelles et les langages rationnels sont construits. La seule "difficulté" consiste à s'apercevoir que le langage rationnel $\{a\}$ avec $a \in A^*$ peut être obtenu par la concaténation des langages rationnels constitués des symboles de a (ou bien du mot vide)

Expression rationnelle standard et équivalence entre expressions rationnelles

Définition 2.3 - Expression rationnelle standard

Une **expression rationnelle** est dite **standard** si et seulement si les seuls opérateurs utilisés sont les opérateurs ., | et *

Définition 2.4 - Équivalence entre expressions rationnelles

Deux **expressions rationnelles** w_1 et w_2 sont dites **équivalentes** (ou par abus de langage "égales"), noté "$\mathbf{w_1 \equiv w_2}$" (ou "$w_1 = w_2$"), si elles décrivent le même langage rationnel.

Propriétés des expressions rationnelles

Si l_1 et l_2 des expressions rationnelles :

- **Union**
  - commutativité : $(l_1|l_2) = (l_2|l_1)$ - indempotence : $(l|l) = l$
  - élément neutre : $(l|\emptyset) = (\emptyset|l) = l$
  - associativité : $(l_1|(l_2|l_3)) = ((l_1|l_2)|l_3)$
- **Mise à l'étoile**
  - idempotence : $l^{**} = l^*$ - élément neutre : $\emptyset^* = \varepsilon$
- **Mise à l'étoile et union**
  - absorption : $l^* = l|l^*$
- **Concaténation**
  - élément neutre : $(l.\varepsilon) = (\varepsilon.l) = l$ - absorption : $(l.\emptyset) = (\emptyset.l) = \emptyset$
  - associativité : $(l_1.(l_2.l_3)) = ((l_1.l_2).l_3)$
- **Concaténation et union**
  - distributivité à droite (resp. à gauche) de la concaténation par rapport à l'union : $(l_1|l_2)l_3 = (l_1l_3|l_2l_3)$; $l_1(l_2l_3) = (l_1l_2|l_1l_3)$

Facteur itérant et mot primitif

Définition 2.5 - Facteur itérant

Un **facteur itérant** est une sous-chaîne non vide pouvant être "étoilée"
Autrement dit, v est facteur itérant de $l \in L$ si

$$l = uvw,\ |v| \neq 0,\ \forall i \geq 0,\ uv^i w \in L\ (uv^* w \subset L)$$

Définition 2.6 - mot primitif

Un **mot primitif** est un mot l tel que si $l = v^n$ alors $n = 1$, c'est-à-dire s'il n'est pas puissance d'un autre mot que lui-même.

Définition rationnelle

Définition 2.7 - Définition rationnelle

Si A est un alphabet de symboles de base, une **définition rationnelle** est une suite de n définitions ($n \in \mathbb{N}$) de la forme

$$d_i \rightarrow r_i \text{ pour } i \in \{1, \ldots, n\}$$

où chaque d_i est un nom distinct, n'appartenant pas à $(A \cup \{|, ^*, ^+, ^?, \varepsilon, \emptyset, (,)\})^*$, et chaque r_i une expression rationnelle sur l'alphabet augmenté $A \cup \{d_1, \ldots, d_{i-1}\}$.

Attention : les definitions ne sont pas récurives

Exemple de définition rationnelle

1. Donner l'alphabet et les définitions rationnelles permettant de décrire le langage dont les mots sont les nombres décimaux

Exemple de définition rationnelle - réponse

1. Donner l'alphabet et les définitions rationnelles permettant de décrire le langage dont les mots sont les nombres décimaux

$\Longrightarrow$

L'alphabet : $\mathcal{A} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, `, `, '\}$

Les définitions rationnelles :
$$\begin{cases} D & \rightarrow & (0|1|2|3|4|5|6|7|8|9) \\ DNZ & \rightarrow & (1|2|3|4|5|6|7|8|9) \\ E & \rightarrow & 0|DNZ \cdot D^* \\ N & \rightarrow & E|E, D^+ \end{cases}$$

Systèmes d'équations rationnelles

Définition 2.8 - Systèmes d'équations rationnelles

Les **systèmes d'équations rationnelles** sont des ensembles d'équations dont les coefficients sont des expressions rationnelles.

Définition 2.9 - Systèmes d'équations rationnelles en forme standard

Un ensemble d'**équations rationnelles** d'indéterminés $D = \{X_1, \ldots, X_n\}$ est en **forme standard** si $\forall X_i \in D$, il y a une équation de la forme $X_i = a_{i_0} \mid a_{i_1} X_1 \mid \cdots \mid a_{i_n} X_n$ avec a_{i_j} des expressions rationnelles sur A telles que $A \cap D = \emptyset$.

Remarques : les équations en forme standard sont linéaires à droites. On a des propriétés similaires avec les systèmes d'équations linéaires à gauche

Lemme d'Arden

Lemme 2.1 - Lemme d'Arden

Soient K et L deux langages sur A^* ($K \subseteq A^*$ et $L \subseteq A^*$),
Si $\varepsilon \notin K$, alors :

- K^*L est l'unique solution de l'équation $X = KX \mid L$
- LK^* est l'unique solution de l'équation $X = XK \mid L$

Si $\varepsilon \in K$, alors :

- Pour $X = KX \mid L$: A^* est solution et K^*L la plus petite solution
- Pour $X = XK \mid L$: A^* est solution et LK^* la plus petite solution

Identités remarquables :

- $(E \mid F)^* = E^*(FE^*)^*$ (e1)
- $(E \mid F)^* = (E^*F)^*E^*$ (e1')
- $(EF)^* = \epsilon \mid E(FE)^*F$ (e2)

Lemme d'Arden - Démonstration

K^*L est l'unique/plus petite solution de l'équation $X = KX \mid L$

Démonstration

1. K^*L est une solution de l'équation $X = KX \mid L$:
   Il faut démontrer l'inclusion du langage de gauche L dans le langage de droite $KX \mid L$ et réciproquement
2. K^*L est la plus petite solution de l'équation $X = KX \mid L$:
   On démontre par récurrence sur la taille des mots de K^*L que si M est une solution de l'équation alors tout mot de K^*L doit lui appartenir
3. Si $\varepsilon \notin K$, la solution K^*L est unique :
   Ceci se démontre par récurrence sur la taille des mots d'une solution M de l'équation en utilisant le fait que l'équation permet de décomposer un mot de M soit en un mot de L, soit en un mot $k \cdot m$ avec $k \in K$ et $m \in M$

Algorithme de résolution d'un système d'équations rationnelles

Soit un système de n équations rationnelles (linéaires à droite) en forme standard notées $X_1, \ldots, X_n$. Cette méthode se déroule en trois phases :

1. Écrire, les équations du système sous la forme $X_i = x_i X_i \mid Y_i$ où x_i est une expression rationnelle sur A et Y_i est une expression rationnelle de la forme $y_0 \mid y_1 X_1 \mid \ldots \mid y_{i-1} X_{i-1} \mid y_{i+1} X_{i+1} \mid \ldots \mid y_n X_n$ avec y_i des expressions rationnelles sur A.
2. Pour toutes les équations X_i de X_1 à X_n, sachant que selon le lemme d'Arden, l'unique solution (ou la plus petite si $\varepsilon \in x_i$) pour $X_i = x_i X_i \mid Y_i$ est $x_i^* Y_i$, remplacer l'équation X_i par $X_i = x_i^* Y_i$ et dans toutes les équations $X_{i+1} \ldots X_n$ remplacer la variable X_i par $x_i^* Y_i$
3. Pour toutes les variables X_i de X_n à X_1, remplacer X_i par sa valeur dans toutes les équations de X_{i-1} à X_1.

Exemples de systèmes d'équations rationnelles

Soient a et b deux expressions rationnelles. Montrer que :

1. $a(a \mid ba)^* = (a \mid ab)^* a$
2. $(a \mid b)^+ = a^+ \mid a^*(ba^*)^+$

Exemples de systèmes d'équations rationnelles - Solution

Soient a et b deux expressions rationnelles. Montrer que :

1. $a(a \mid ba)^* = (a \mid ab)^* a$
   $\Longrightarrow$ Les deux expressions rationnelles sont solutions de l'équation $X = (a|ab)X|a$ qui n'accepte qu'une seule solution (car $\varepsilon \notin a|ab$)
2. $(a \mid b)^+ = a^+ \mid a^*(ba^*)^+$
   $\Longrightarrow$ Les deux expressions rationnelles sont solutions de l'équation $X = (a|b)X|a|b$ qui n'a qu'une seule solution (car $\varepsilon \notin a|b$)

Lemme de l'étoile

Lemme 2.2 - Lemme de l'étoile

Soit L un langage rationnel sur un alphabet A. Alors il existe un entier naturel n tel que pour tout mot z de L vérifiant $|z| > n$, il existe $u, v, w \in A^*$ tels que $z = uvw$, $v \neq \varepsilon$, $|uv| \leq n$ et pour tout $i \in N$, $uv^i w \in L$.

Autre nom : Lemme de la pompe, Lemme d'itération, Lemme d'Ogden pour les rationnels

Utile pour démontrer qu'un langage n'est pas rationnel

$Rat(A^*) \subset Etoile(A^*) \subset P(A^*)$

Exemples d'utilisation du lemme de l'étoile

1. Monter que $\{a^n b^p : n < p\}$ n'est pas rationnel
2. $\{a^p : p \textit{ premier}\}$ est-il rationnel ?

Exemples d'utilisation du lemme de l'étoile - Réponse du (1)

1. Monter que $\{a^n b^p : n < p\}$ n'est pas rationnel
   $\Longrightarrow$ Démonstration par l'absurde. Si le langage était rationnel, alors il vérifirait la propriété du lemme de l'étoile : $\exists n \in \mathbb{N}$ tel que ... Soit le mot $a^n b^{n+1}$ de ce langage. Il peut être décomposé en uvw suivant le lemme de l'étoile avec $|uv| \leq n$ et $|v| > 0$. Donc v (de même que u) ne contient que des a. De plus v n'est pas vide. Donc le mot $uv^{|w|}w$ contient plus ou autant de a que de $b \rightarrow$ Contradiction $\Rightarrow$ le langage n'est pas rationnel
2. $\{a^p : p \textit{ premier}\}$ est-il rationnel ?