Chapitre un

Langages rationnels

MPI/MPI*, lycée Faidherbe

Résumé

Dans ce chapitre, après avoir défini les notions formelles de mots et de langages, on introduit un formalisme efficace permettant de décrire certains langages que l'on peut être amené à reconnaître. La caractéristique de ces langages est la possibilité de les décrire par des motifs (patterns en anglais), c'est-à-dire par des formules qu'on appelle, dans ce contexte, expressions régulières. On conclut par l'étude d'un autre type de langages dont on verra plus tard qu'ils sont en fait rationnels.

I Mots et langages

I.1 Définitions

On va ici formaliser la notion de chaînes de caractères.

Définition 1 : mot

Un *alphabet* est un ensemble fini, ses éléments sont appelés *lettres*.
Un *mot* sur un alphabet Σ est une suite finie d'éléments de Σ .

- Dans les exemples on considérera souvent un alphabet à 2 lettres $\Sigma = \{a, b\}$.
- Un mot sur Σ est noté par la concaténation de ses lettres : $w = u_1 u_2 \cdots u_n$ avec $u_i \in \Sigma$.

Définition 2 : longueur

Le nombre de lettres d'un mot w est sa longueur, notée $|w|$.
L'ensemble des mots de longueur p se note Σ^p.
Le nombre d'occurrences d'une lettre a dans un mot w est noté $|w|_a$.

On a donc $|w| = \sum_{x \in \Sigma} |w|_x$.

Notations

La notion de suite finie contient le cas d'une suite vide.

- On note ε ou 1_Σ (ou parfois Λ) le mot vide ; sa longueur est nulle.
- L'ensemble des mots sur Σ , y compris le mot vide, est noté Σ^*.
- $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$ est l'ensemble des mots de longueur au moins 1.
- Σ^p est l'ensemble des mots de longueur p ; en particulier $\Sigma^0 = \{\varepsilon\}$.
- $\Sigma^+ = \bigcup_{p \in \mathbb{N}^*} \Sigma^p$ et $\Sigma^* = \bigcup_{p \in \mathbb{N}} \Sigma^p$.

Définition 3 : langage

Un langage est un ensemble de mots (qui peut contenir ε) , c'est une partie de Σ^*.
Les *langages élémentaires* sur un alphabet Σ sont :

- l'ensemble vide, $\emptyset$,
- le langage réduit au mot vide, $\{\varepsilon\}$,
- les singletons $\{x\}$ pour toute lettre $x \in \Sigma$.

I.2 Produit

Définition 4 : produit de mots

Le *produit* de deux mots est le mot obtenu par la concaténation de leurs lettres :
si $u = x_1 \dots x_n$ et $v = y_1 \dots y_m$ alors $w = u.v = z_1 \cdots z_{m+n}$
avec $z_i = x_i$ pour $1 \leqslant i \leqslant n$ et $z_i = y_{i-n}$ pour $n+1 \leqslant i \leqslant n+m$.
Si $w = u.v$ on dit que u est un *préfixe* de w et v est un *suffixe* de w.
Si $w = u.v.u'$ on dit que v est un *facteur* de w.
On définit $u^0 = \varepsilon$ et, par récurrence, $u^{n+1} = u^n.u$.

On a $u.\varepsilon = \varepsilon.u = u$ et$u^1 = u$
ε et u sont des préfixes, suffixes et facteurs de u.
Attention à ne pas confondre $L.L$ et $\{w.w \ ; \ w \in L\}$.

Théorème 1 : propriétés du produit

1. Pour $p \leqslant |u|$ u admet un unique préfixe (resp. suffixe) de longueur p.
2. $(u.v).w = u.(v.w)$ la loi est associative, on omettra les parenthèses dans les produits.
3. $|u.v| = |u| + |v|$.
4. Si $u.v = \varepsilon$ alors $u = v = \varepsilon$.
5. $u^n.u^m = u^{n+m}$.
6. Si $u.v = u.v'$ alors $v = v'$ (simplification à gauche).
7. Si $u.v = u'.v$ alors $u = u'$ (simplification à droite).

Le produit est donc une loi interne dans Σ^*, associative et admettant un élément neutre, ε, cela confère à $(\Sigma^*, .)$ une structure de *monoïde*, comme $(\mathbb{N}, +)$ et $(A, \times)$ si $(A, +, \times)$ est un anneau.

Définition 5 : produit de langages

Le produit des langages L_1 et L_2 est le langage $L_1.L_2$ défini par

$$L_1.L_2 = \{u_1.u_2 \ ; \ u_1 \in L_1, \ u_2 \in L_2\}$$

On pose $L^0 = \{\varepsilon\}$ et, par récurrence $L^{n+1} = L^n.L$.

$L^1 = L$ et, pour $n \geqslant 2$, $L^n = \{u_1.u_2.\cdots.u_n \ ; \ u_i \in L\}$.
On notera que la définition est cohérente pour $L = \Sigma$: Σ^p au sens du produit est bien l'ensemble des mots de longueur p.

I.3 Ordre lexicographique

On peut définir un d'ordre partiel dans Σ^* par u est un préfixe de v ; pour définir un ordre total, on doit, par exemple, pouvoir comparer deux lettres.

Définition 6 : ordre lexicographique

Si Σ est un ensemble fini muni d'une relation d'ordre $\preccurlyeq$ on définit une relation d'ordre total sur Σ^* par $u \leqslant v$ si et seulement si u est un préfixe de v ou il existe $w, u', v' \in \Sigma^*$ et $a, b \in \Sigma$ tels que $u = w.a.u'$, $v = w.b.v'$ et $a \prec b$.

Théorème 2 : ordre total

L'ordre lexicographique est une relation d'ordre total.

II Langages rationnels

Un mot $u = x_1x_2\cdots x_n$ est le produit des mots réduits à une lettre : $u = x_1 \cdot x_2 \cdots x_n$. Ainsi un langage réduit à un mot est un produit de langages élémentaires : $\{u\} = \{x_1\} \cdot \{x_2\} \cdots \{x_n\}$.
On en déduit qu'un langage fini est une union de produit de langages élémentaires.
Inversement tout langage obtenu par une suite d'unions et de produits à partir de langages élémentaires est fini.
On ajoute une opération.

Définition 7 : étoile de Kleene

L'étoile d'un langage L est $L^* = \bigcup_{n\in\mathbb{N}} L^n$.

On définit aussi $L^+ = \bigcup_{n\in\mathbb{N}^*} L^n$

Théorème 3 Propriétés de l'étoile

1. L^* est l'ensemble des produits de mots de L.
2. Si $L = \{w\}$ alors $L^* = \{w^n \ ; \ n \in \mathbb{N}\}$.
3. $\emptyset^* = \{\varepsilon\}^* = \{\varepsilon\}$.
4. Si L contient un mot non vide alors L^* est infini.

Les langages rationnels sont les langages qu'on peut construire à partir des langages élémentaires par un nombre fini d'unions, de produits et d'étoiles de Kleene. On va donner deux caractérisations plus rigoureuses de ces langages.
La première méthode consiste à stratifier l'ensemble des langages rationnels.
Pour un alphabet donné Σ , on note $\mathcal{R}_0(\Sigma) = \big\{\emptyset, \{\varepsilon\}\big\} \cup \bigcup_{a\in\Sigma}\{a\}$ l'ensemble des langages élémentaires sur Σ et on définit par récurrence

$$\mathcal{R}_{n+1}(\Sigma) = \big\{L_1 \cup L_2 \ ; \ L_1, L_2 \in \mathcal{R}_n(\Sigma)\big\} \cup \ \big\{L_1.L_2 \ ; \ L_1, L_2 \in \mathcal{R}_n(\Sigma)\big\} \cup \ \big\{L^* \ ; \ L \in \mathcal{R}_n(\Sigma)\big\}$$

Remarque : si $\{\varepsilon\} \in \mathcal{R}_n(\Sigma)$ alors $\{\varepsilon\} = \{\varepsilon\}.\{\varepsilon\} \in \mathcal{R}_{n+1}(\Sigma)$, or $\{\varepsilon\} \in \mathcal{R}_0(\Sigma)$;
on en déduit que $\{\varepsilon\} \in \mathcal{R}_n(\Sigma)$ pour tout n.
Ainsi, si $L \in \mathcal{R}_n(\Sigma)$ alors $L = L.\{\varepsilon\} \in \mathcal{R}_{n+1}(\Sigma)$; on en déduit $\mathcal{R}_n(\Sigma) \subset \mathcal{R}_{n+1}(\Sigma)$

Définition 8 : langages rationnels

L est rationnel sur l'alphabet Σ s'il existe $n \in \mathbb{N}$ tel que $L \in \mathcal{R}_n(\Sigma)$.

Autrement dit, l'ensemble $\mathrm{Rat}(\Sigma)$ les langages rationnels sur Σ est $\mathrm{Rat}(\Sigma) = \bigcup_{n\in\mathbb{N}} \mathcal{R}_n(\Sigma)$.

Théorème 4 Caractérisation

Rat(Σ) est le plus petit sous ensemble de $\mathcal{P}(\Sigma^*)$ qui contient les langages élémentaires pour Σ et qui est stable par produit, union et étoile.

"plus petit" signifie ici que Rat(Σ) vérifie les propriétés et que si une partie de $\mathcal{P}(\Sigma^*)$ vérifie les propriétés alors il contient $\mathcal{P}(\Sigma^*)$. On peut aussi exprimer cette définition sous la forme $\mathcal{P}(\Sigma^*)$ est l'intersection des parties de $\mathcal{P}(\Sigma^*)$ qui contiennent les langages élémentaires pour Σ et qui sont stables par produit, union et étoile.

Une conséquence utile de la définition est un mode de démonstration.

Théorème 5 Induction structurelle

Si une propriété P portant sur des langages est telle que

- $P(\emptyset)$, $P(\{\varepsilon\})$ et $P(\{a\})$ pour $a \in \Sigma$ sont vérifiées
- la vérité de $P(L_1)$ et $P(L_2)$ implique celle de $P(L_1 \cup L_2)$, $P(L_1.L_2)$ et $P(L_1^*)$

alors $P(L)$ est vraie pour tout langage rationnel.

III Langages réguliers

Définition 9 : Expression régulière

Σ est un alphabet ne contenant pas les lettres "$\varnothing$ " ,"ϵ", "*", "|", ".", "(" et ")". Une expression régulière sur Σ est une expression de la forme

- $\varnothing$
- ϵ
- $\mathtt{a}$ avec $a \in \Sigma$
- $(\mathtt{r}_1|\mathtt{r}_2)$ avec $\mathtt{r}_1$ et $\mathtt{r}_2$ expressions régulières
- $(\mathtt{r}_1 \cdot \mathtt{r}_2)$ avec $\mathtt{r}_1$ et $\mathtt{r}_2$ expressions régulières
- $(\mathtt{r}*)$ avec $\mathtt{r}$ expression régulière

Simplifications : pour éviter la multiplication des parenthèses, on peut

- éliminer la dernières parenthèses,
- considérer $*$ prioritaire devant les autres opérations donc remplacer $(\mathtt{r}*)$ par $\mathtt{r}*$,
- omettre . : remplacer $\mathtt{r}_1 \cdot \mathtt{r}_2$ par $\mathtt{r}_1\mathtt{r}_2$
- considérer . prioritaire devant | donc remplacer $(\mathtt{r}_1\mathtt{r}_2)$ par $\mathtt{r}_1\mathtt{r}_2$ quand le résultat est utilisé dans une expression avec |
- utiliser la priorité à gauche : remplacer $(\mathtt{r}_1\mathtt{r}2)\mathtt{r}_3$ par $\mathtt{r}_1\mathtt{r}_2\mathtt{r}_3$ et $(\mathtt{r}_1|\mathtt{r}2)|r_3$ par $\mathtt{r}_1|\mathtt{r}_2|\mathtt{r}_3$

Par exemple on peut écrire $(((\mathtt{r}_1 \cdot \mathtt{r}_2)|(\mathtt{r}_3*))|r_4)$ sous la forme $\mathtt{r}_1\mathtt{r}_2|\mathtt{r}_3*|\mathtt{r}_4$.
On ne devrait par remplacer $\mathtt{r}_1|(\mathtt{r}_2|\mathtt{r}_3)$ par $\mathtt{r}_1|\mathtt{r}_2|\mathtt{r}_3$ pour l'instant.

Définition 10 : Langage régulier

À toute expression régulière e sur Σ on peut associer un langage par la construction inductive :

- $L[\texttt{∅}] = \emptyset$,
- $L[\texttt{ϵ}] = \{\varepsilon\}$,
- $L[\texttt{a}] = \{a\}$ pour $a \in \Sigma$,
- $L[\texttt{r}_1\texttt{r}_2)] = L[\texttt{r1}] \cup L[\texttt{r}_2]$,
- $L[(\texttt{r}_1.\texttt{r}_2)] = L[\texttt{r}_1].L[\texttt{r}_2]$,
- $L[\texttt{r}*] = \big(L[\texttt{r}]\big)^*$.

On dit que $\texttt{r}$ dénote $L[\texttt{r}]$. Un langage dénoté par une expression réguliers est dit *régulier*.

Théorème 6 Première équivalence

Les langages rationnels sont les langages réguliers.

Il faut noter que le langage dénoté par une expression régulière est unique mais qu'il peut exister plusieurs expressions régulières qui dénotent un langage donné.
Par exemple le langage des mots sur $\Sigma = \{a, b\}$ contenant au moins un b peut être dénoté par `(a|b)*b(a|b)*` ou par `a*b(a|b)*`.

Définition 11 : Expressions régulières équivalentes

Deux expressions régulières sont équivalentes si elles dénotent le même langage.

IV Langages locaux

Nous allons, dans cette partie, donner une autre construction de langages. Elle consiste à ajouter, pas-à-pas, des lettres aux mots en fonction uniquement de la lettre précédente.

Définition 12 : Langage local

Un langage L sur l'alphabet Σ est local
s'il existe $P \subset \Sigma$, $S \subset \Sigma$ et $F \subset \Sigma^2$
tels que $u = u_1 u_2 \cdots u_n \in L \setminus \{\varepsilon\}$ est équivalent à
$u_1 \in P$, $u_n \in S$ et $u_i u_{i+1} \in F$ pour tout $i \in \{1, 2, \ldots, n-1\}$.

L'appartenance de ε à L est facultative ; en plus des 3 ensembles il faut donc un quatrième déterminant pour indiquer si on a $\varepsilon \in L$.

Définition 13 : Expressions linéaires

Une expression régulière est linéaire si toutes les lettres qui la composent sont distinctes.

Théorème 7

Si r est une expression régulière linéaire alors $L[\texttt{r}]$ est un langage local.