



Théorie des Langages

Pr. Youness Tabii
Youness.tabii@um5.ac.ma
ENSIAS
UM5 - Rabat

Plan

- Concepts de base
- Expression régulière
- Automate à état finis
- Déterminisation d'automate (NFA → DFA)
- Transformation d'une Expression régulière en automate
- Transformation DFA en Expression régulière
- Minimisation d'une Expression régulière
- Minimisation d'un automate

Concepts de base

Alphabet

- ❖ Un **Alphabet** est un ensemble fini de symboles indivisibles (lettres, séparateurs, chiffres,...).
- ❖ Nous désignerons par Σ ou A les alphabets dénotant le vocabulaire d'un langage.
 - ❖ Par exemple, le vocabulaire des chiffres est : $\Sigma = \{0,1,2,3,4,5,6,7,8,9\}$.
- ❖ Dans un langage, nous utilisons un **alphabet** pour construire des **mots** (**w**) : $w \subseteq \Sigma^*$ ou $w \subseteq A^*$.
 - ❖ Par exemple un nombre est un **mot** de chiffre. $\Sigma^* = \{..., 11, 12, 13, ...\}$.

Langage

- ❖ L'ensemble de tous les **mots** construits sur un **alphabet** sera noté par Σ^* .
- ❖ Un langage est un ensemble de **mots** construit sur un **alphabet** (Σ ou A).
- ❖ Tout langage défini sur un **alphabet** Σ (ou A) sera une partie de Σ^* (ou A^*).

Langage

- ❖ $|w|$ = longueur du **mot** w.
- ❖ Le **mot** de longueur zéro est noté ε . ε est le **mot** vide, il est neutre pour la concaténation des langages.
- ❖ Ordre partiel sur les mots :
 - ❖ a est préfixe de b si $\exists a'$, tel que $b = aa'$
 - ❖ a est suffixe de b si $\exists a'$, tel que $b = a'a$
 - ❖ a est **sous mot** de b si $b = b_0a_0b_1a_1 \dots b_na_n$ avec b_i et $a_i \in \Sigma^*$ pour $a = a_0a_1 \dots a_n$

Langage

Opérations sur les langages : Soit L_1 et L_2 deux langages sur Σ^* :

- ❖ **Union** : $L_1 \cup L_2 \{a \mid a \in L_1 \text{ ou } a \in L_2\}$
- ❖ **Intersection** : $L_1 \cap L_2 \{a \mid a \in L_1 \text{ et } a \in L_2\}$
- ❖ **Différence** : $L_1/L_2 \{a \mid a \in L_1 \text{ et } a \notin L_2\}$
- ❖ **Complément** : $\overline{L_2} \{a \in L_1 \mid a \notin L_2\}$
- ❖ **Concaténation** : $L_1L_2 \{ab \mid a \in L_1 \text{ et } b \in L_2\}$

Les Expressions Régulières

Définition

- ❖ Les **Expressions Régulières (ER)** sont des motifs destinés à décrire des chaînes ou des sections complètes de texte.
- ❖ Les motifs sont construits selon une syntaxe simple qui permet de décrire de manière abstraite les caractéristiques d'une chaîne de caractères.
- ❖ La syntaxe d'une **expression régulière** comporte des règles et des symboles spéciaux appelés méta-caractères.
- ❖ Le mot **Régulière** doit être compris dans le sens qui obéit aux règles.

Besoins d'ER

L'utilisation des **ER** peuvent comprendre :

- ✓ La **recherche d'éléments** dans un texte (ou un flux de données) : Il s'agit de reconnaître des éléments particuliers.
- ✓ **L'extraction d'information** : Il s'agit d'isoler des parties d'une chaîne de caractères.
- ✓ La **validation de données** : Il s'agit de vérifier la validité d'une information.
- ✓ La **substitution et le reformatage** : il s'agit de reconnaître un texte pour le réécrire en un autre.

Exemple

Exemple d'une ER de validation des variables en C :

Soit :

- **L** un ensemble fini de lettres,
- **S** un ensemble fini de séparateurs (ensemble valide du langage),
- **C** l'ensemble fini des entiers de 0 à 9.
- Alors l'**expression régulière des variables** valides en langage **C** qui vérifie qu'une
 - variable commence par une Lettre
 - puis peut comporter des Lettres, des séparateurs ou des chiffres

$$L(L+S+C)^*$$

Exemple

Exemple d'une ER de validation des variables en C :

$$L(L+S+C)^*$$

- Le **+** représente la sélection (disjonction), elle peut se noter **|**, **ou...**
- Le ***** signifie 0 ou plusieurs occurrences de la sous-expression.
- La juxtaposition des symboles du vocabulaire, ici la juxtaposition de **L** et **()** représente leur concaténation.
- Les parenthèses, **()** représentent les groupements des sous-expressions.

Exemple

Exemple de construction (mot d'un langage) :

- ❖ L'**ER** $a+b$ dénote le langage $\{a, b\}$ pour un alphabet Σ constitué des caractères a et b .
- ❖ L'**ER** $(a+b)(a+b)$ dénote le langage $\{aa, ab, ba, bb\}$, soit le langage de toutes les chaînes de longueur 2 sur l'alphabet Σ .
- ❖ L'**ER** a^* dénote le langage des chaînes de $0, 1$ ou plusieurs a , soit $\{\varepsilon, a, aa, aaa, \dots\}$.

Remarque : Il est à noter que ε représente l'élément vide.

ER - Langage

- ❖ Un Langage rationnel est un langage reconnue par une **expression régulière (expression rationnelle)**.
- ❖ Un langage $L \subseteq \Sigma^*$ est rationnel s'il existe une **ER** telle que $L = L(\text{ER})$.
- ❖ A et B, deux **ER**, sont équivalentes si leurs deux langages rationnels sont équivalents : $L(A) = L(B)$.
- ❖ Pour montrer que deux langages rationnels sont équivalents il faut prouver que $L_1 \subseteq L_2$ et que $L_2 \subseteq L_1$

ER - Langage

Soit Σ un alphabet. Une expression régulière **ER** sur Σ est une formule qui définit un langage $L(\text{ER})$ sur Σ , de la manière suivante :

- ❖ Si ϵ est une **ER** qui définit le langage $L(\epsilon)$ ou $\{\epsilon\}$
- ❖ Si $a \in \Sigma$, alors a est une **ER** qui définit langage $L(a)$ ou $\{a\}$

- ❖ Soient a et b deux **ER**, définissant les langages $L(a)$ et $L(b)$:
 - ❖ a^* est une **ER** définissant le langage $(L(a))^*$
 - ❖ ab est une **ER** définissant le langage $L(a)L(b)$
 - ❖ $(a+b)$ est une **ER** définissant le langage $L(a) \cup L(b)$

ER

❖ Commutativité de l'union (mais pas de la concaténation)

- ❖ $\alpha \beta \leftrightarrow \beta \alpha$
- ❖ $\alpha | \beta = \beta | \alpha$

❖ associativité

- ❖ $\alpha (\beta \gamma) = (\alpha \beta) \gamma$
- ❖ $\alpha | (\beta | \gamma) = (\alpha | \beta) | \gamma$

❖ distributivité

- ❖ $\alpha (\beta | \gamma) = \alpha \beta | \alpha \gamma$

❖ Écritures

- ❖ $\alpha | \beta = \alpha + \beta$
- ❖ Fermeture positive : l'écriture $\alpha^+ = \alpha \alpha^*$ (répétition un nombre de fois au moins égal à 1)

❖ Autres propriétés

- ❖ $w^0 = \varepsilon$
- ❖ $w^* = \cup_{\{i=0..\infty\}} w^i = w^0 | \cup_{\{i=1..\infty\}} w^i = \varepsilon | \cup_{\{i=1..\infty\}} w^i = \varepsilon | w^+$
- ❖ $|w_1.w_2| = |w_1| + |w_2|$
- ❖ $\alpha | \emptyset = \emptyset | \alpha = \alpha$ (élément neutre par rapport à l'union)
- ❖ $(\emptyset \alpha) = (\alpha \emptyset) = \emptyset$ (\emptyset est l'élément absorbant pour la concaténation)
- ❖ $(\varepsilon \alpha) = (\alpha \varepsilon) = \alpha$ (ε est l'élément neutre pour la concaténation)
- ❖ $(\alpha | \alpha) = \alpha$
- ❖ $\emptyset^* = \varepsilon$ ($\emptyset^0 = \varepsilon$ comme $0^0 = 1$)
- ❖ $\varepsilon | \alpha \alpha^* = \varepsilon | \alpha^+ = \alpha^*$ (analogie $0 \sim \emptyset$ et $1 \sim \varepsilon$) par rapport à la concaténation et la multiplication

Exercices

- ❖ Ecrire les expression régulières pour les langages suivants avec l'alphabet {a,b}
1. Langage accepte les mots de tailles au moins 2
 2. Langage accepte les mots de taille 2 au maximum

Solution

1. $L_1 = \{ab, ba, aaa, aab, bba, \dots\}$

➤ $ER_1 = (a+b)(a+b)(a+b)^*$

2. $L_2 = \{\varepsilon, a, b, aa, ab, ba, bb\}$

➤ $ER_2 = (\varepsilon+a+b)(\varepsilon+a+b)$

Langage régulier: Lemme de l'étoile

- Si un langage L sur un alphabet Σ ($L \subseteq \Sigma^*$) est **régulier**, si
 - $\exists n \in \mathbb{N}^*$ (qui dépend de L) tel que
 - $\forall u \in L$ avec $|u| \geq n$,
 - il existe une décomposition de **u** en trois parties **$u = fgh$** telle que :
 - $g \neq \varepsilon$
 - $|fg| \leq n$
 - $\forall k \geq 0, fg^k h \in L$

Langage régulier: Lemme de l'étoile

- ❖ Quelque soit le mot $|u| \geq n$,
- ❖ il existe un sous mot qui se répète.
- ❖ Et le schéma de répétition fg^* doit être commun à tous les mots du langage à partir d'une taille donnée n ($n+1$ qui est par la même occasion le nombre d'états qui reconnaît le mot fg)
- ❖ Attention, la décomposition doit toujours trouver un sous mot g (en boucle) non loin du début du mot $|fg| \leq n$ et ce quelque soit la taille du mot $|u| \geq n$!!

Langage régulier: Lemme de l'étoile

- ❖ Le lemme de l'étoile ne permet pas de démontrer qu'un langage est régulier (car il n'y a pas d'équivalence entre le fait de permettre une décomposition et la régularité du langage)
- ❖ Le lemme de l'étoile est souvent utilisé pour **démontrer qu'un langage donné est irrégulier** (par absurdité)
- ❖ Le lemme de l'étoile ne permet pas de démontrer pour tous les langages irréguliers qu'ils sont irréguliers (car des langages sont irréguliers pourtant ils permettent une décomposition)
 - ❖ $a^i b^j \ i \geq j$

Langage irrégulier

Exemple

- ❖ Montrer que le langage $\{a^n b^p : n < p\}$ est irrégulier:
 - ❖ Prenons un entier N suffisamment grand.
 - ❖ Pour tout mot z du langage tel que $|z| \geq N$,
 - ❖ il existe u, v, w de A^* tels que $z = uvw$, $|v| > 0$,
 $|uv| \leq N$ alors, normalement, pour tout entier i le mot $uv^i w$ est dans $\{a^n b^p : n < p\}$

Langage irrégulier

Exemple

- ❖ Montrer que le langage $\{a^n b^p : n < p\}$ est irrégulier:
- ❖ Prenons le mot $z=a^N b^{N+k}$ avec $k \geq 1$
 - ❖ Si $z=uvw$ alors $u=a^p$, $v=a^q$ avec $q>0$ et $w=a^{N-(p+q)} b^{N+k}$
 - ❖ Donc $z = a^p a^q a^{N-(p+q)} b^{N+k}$
 - ❖ $uv^i w = a^p (a^q)^i a^{N-(p+q)} b^{N+k} = a^N (a^q)^{i-1} b^{N+k} = a^{N+q(i-1)} b^{N+k}$
- ❖ Or pour tout $i > 1+k/q$
 - ❖ $uv^i w$ n'est plus dans $\{a^n b^p : n < p\}$
 - ❖ car alors $N+(i-1)q > N+k$!
 - ❖ Donc il ne vérifie pas le lemme de l'étoile et donc il **n'est pas régulier !**

Langage irrégulier

Exemple

- ❖ Application
 - ❖ Cas n=5 et p= 6, n<p
 - ❖ Le mot : Z = aaaaabbbbbbb
 - ❖ Prennons : **U**=a, **V**=aaaa, **W**=bbbbbb, i=2, k=1
 - ❖ Z=**UV²W**=aaaaaaa**abbbbbbb**
- ❖ Dans ce cas n=9 et p=6 → n>p,
 - ❖ Mot n'appartient pas au langage
- ❖ **{aⁿb^p : n < p} est irrégulière**