Les expressions régulières

Définitions

Introduction

Comment décrire un langage ??
Étant donné un mot, appartient il à un langage donné ??
Nous allons parler de la théorie des langages, en particulier nous décrivons les expressions régulières, et par conséquent les langages réguliers

Les langages

Définitions

On appelle **alphabet** *un ensemble fini non vide A de symboles (lettres de 1 ou plusieurs caractères).*
On appelle **mot** *toute séquence finie d'éléments de A.*
On note ε le **mot vide**.
On note A^* l'ensemble infini contenant tous les mots possibles sur A.
On note A^+ l'ensemble des mots non vides que l'on peut former sur A, c'est à dire $A^+ = A^* - \{\varepsilon\}$
On note $|m|$ la longueur du mot m, c'est à dire le nombre de symboles de A composant le mot.
On note A^n l'ensemble des mots de A^* de longueur n. Remarque : $A^* = \bigcup_{n=0}^{\infty} A^n$

Exemples

Soit l'alphabet $A = \{a, b, c\}$. $aaba$, $bbbacbb$, c et ε sont des mots de A^*, de longueurs respectives 4, 7, 1 et 0.
Soit l'alphabet $A = \{aa, b, c\}$. aba n'est pas un mot de A^*. $baab$, caa, bc, $aaaa$ sont des mots de A^* de longueurs respectives 3, 2, 2 et 2.

Notation

On note . l'opérateur de **concaténation** de deux mots : si $u = u_1 \dots u_n$ (avec $u_i \in A$) et $v = v_1 \dots v_p$ (avec $v_i \in A$), alors la concaténation de u et v est le mot $u.v = u_1 \dots u_n v_1 \dots v_p$
Remarque : un mot de n lettres est en fait la concaténation de n mots d'une seule lettre.

Les langages

Propriété

$|u.v| = |u| + |v|$
$(u.v).w = u.(v.w)$ *(associativité)*
ε *est l'élément neutre pour* . : $u.\varepsilon = \varepsilon.u = u$

Remarque : nous écrirons désormais uv pour $u.v$

Définition

On appelle **langage** *sur un alphabet A tout sous-ensemble de A^*.*

Exemples

Soit l'alphabet $A = \{a, b, c\}$
Soit L_1 l'ensemble des mots de A^* ayant autant de a que de b. L_1 est le langage infini $\{\varepsilon, c, ccc, \ldots, ab, ba, \ldots, abccc, acbcc, accbc, \ldots, aabb, abab, abba, baab, \ldots, accbccbccccca, \ldots, bbccccaccbcabcccacac, \ldots\}$
Soit L_2 l'ensemble de tous les mots de A^* ayant exactement 4 a. L_2 est le langage infini $\{aaaa, aaaac, aaaca, \ldots, aabaa, \ldots, caaaba, \ldots, abcabbbaacc, \ldots\}$

Opération sur les langages

Opérations sur les langages

union : $L_1 \bigcup L_2 = \{w \text{ tq } w \in L_1 \text{ ou } w \in L_2\}$
intersection : $L_1 \bigcap L_2 = \{w \text{ tq } w \in L_1 \text{ et } w \in L_2\}$
concaténation : $L_1 L_2 = \{w = w_1 w_2 \text{ tq } w_1 \in L_1 \text{ et } w_2 \in L_2\}$
puissance : $L^n = \{w = w_1 \ldots w_n \text{ tq } w_i \in L \text{ pour tout } i \in \{1, \ldots, n\}\}$
étoile : $L^* = \cup_{n \geq 0} L^n$

Les langages réguliers

Problème

étant donné un langage, comment décrire tous les mots acceptables ? Comment décrire un langage?
Il existe plusieurs types de langage (classification), certains étant plus facile à décrire que d'autres. On s'intéresse ici aux **langages réguliers**.

Définitions

Un **langage régulier** L *sur un alphabet* A *est défini récursivement de la manière suivante :*

- $\{\varepsilon\}$ *est un langage régulier sur* A
- *Si* a *est une lettre de* A, $\{a\}$ *est un langage régulier sur* A
- *Si* R *est un langage régulier sur* A, *alors* R^n *et* R^* *sont des langages réguliers sur* A
- *Si* R_1 *et* R_2 *sont des langages réguliers sur* A, *alors* $R_1 \bigcup R_2$ *et* $R_1 R_2$ *sont des langages réguliers*

Les langages réguliers se décrivent très facilement par une **expression régulière.**

Les langages réguliers

Définitions

Les expressions régulières (E.R.) sur un alphabet A et les langages qu'elles décrivent sont définis récursivement de la manière suivante :

- *ε est une E.R. qui décrit le langage $\{\varepsilon\}$*
- *Si $a \in A$, alors a est une E.R. qui décrit $\{a\}$*
- *Si r est une E.R. qui décrit le langage R, alors $(r)^*$ est une E.R. décrivant R^**
- *Si r est une E.R. qui décrit le langage R, alors $(r)^+$ est une E.R. décrivant R^+*
- *Si r et s sont des E.R. qui décrivent respectivement les langages R et S, alors $(r)|(s)$ est une E.R. décrivant $R \bigcup S$*
- *Si r et s sont des E.R. qui décrivent respectivement les langages R et S, alors $(r)(s)$ est une E.R. décrivant RS*
- *Il n'y a pas d'autres expressions régulières*

Remarques

on conviendra des priorités décroissantes suivantes : $*$, concaténation, | C'est à dire par exemple que $ab^*|c = ((a)((b)^*))|(c)$
En outre, la concaténation est distributive par rapport à | : $r(s|t) = rs|rt$ et $(s|t)r = sr|tr$.

Les langages réguliers

Exemples

- $(a|b)^* = (b|a)^*$ dénote l'ensemble de tous les mots formés de a et de b, ou le mot vide.
- $(a)|((b)^*(c)) = a|b^*c$ est soit le mot a, soit les mots formés de 0 ou plusieurs b suivi d'un c. C'est à dire $\{a, c, bc, bbc, bbbc, bbbbc, \ldots\}$
- $(a^*|b^*)^* = (a|b)^* = ((\varepsilon|a)b^*)^*$ décrit tous les mots sur $A = \{a, b\}$ ou encore A^*
- $(a|b)^*abb(a|b)^*$ dénote l'ensemble des mots sur $\{a, b\}$ ayant le facteur abb
- $b^*ab^*ab^*ab^*$ dénote l'ensemble des mots sur $\{a, b\}$ ayant exactement 3 a
- $(abbc|baba)^+aa(cc|bb)* = \{abbcaa, \ldots, babaabbababaaa, \ldots, abbcabbcaaccbbbb, \ldots\}$

Remarques

$(a|b)^*a(a|b)^*$, qui décrit les mots sur $\{a, b\}$ ayant au moins un a est **ambiguë**. Car, par exemple, le mot $abaab$ "colle" à l'expression régulière de plusieurs manières :

$abaab = \varepsilon \,.\, a \,.\, baab$ avec $\varepsilon \in (a|b)^*$, et $baab \in (a|b)^*$

$abaab = ab \,.\, a \,.\, ab$ avec $ab \in (a|b)^*$, et $ab \in (a|b)^*$

$abaab = aba \,.\, a \,.\, b$ avec $aba \in (a|b)^*$, et $b \in (a|b)^*$

Par contre, l'e.r. $b^*a(a|b)^*$ décrit le **même langage** et **n'est pas** ambiguë.

$abaab = \varepsilon \,.\, a \,.\, baab$ avec $\varepsilon \in b^*$, et $baab \in (a|b)^*$