Compilation

Avant propos - Théorie des langages

Nicolas Delestre

INSA
ROUEN NORMANDIE

Définitions

- La théorie des langages est une branche de l'informatique qui s'intéresse aux langages formels
- Un langage formel est un ensemble (fini ou infini) de mots
- Un mot est une suite fini de symboles issus d'un alphabet. Le mot vide (noté ϵ) est le mot contenant aucun symbole
- Un alphabet (souvent noté A ou Σ) contient un nombre fini de symboles

Objectifs

- Être capable de représenter un langage
- Être capable de savoir si un mot appartient (ou pas) à un langage et quelle est la complexité de l'algorithme pour le vérifier

Exemples de langage

- les naturels représentés en binaire (l'alphabet est composé des deux symboles 0 et 1, $\Sigma = \{0, 1\}$)

$$L = \{0, 1, 10, 11, 100, 101, 110, 111, \ldots\}$$

- les identifiants du langage de programmation Pascal (l'alphabet est composé des lettres latines minuscules, majuscules, des chiffres et du souligné)
- les expressions arithmétiques (l'alphabet est composé d'un symbole représentant les nombres, des symboles représentants les opérations et des parenthèses)
- l'ensemble de tous les programmes C (l'alphabet est composé d'un symbole représentant les nombres, d'un symbole représentant les identifiants, des symboles représentants les mots clés du langages, des accolades ouvrantes et fermantes, etc.)
- etc.

La hiérarchie de Chomsky

- Noam Chomsky a identifié quatre familles de langages ayant un pouvoir d'expression croissant mais avec des algorithmes de vérification de l'appartenance d'un mot en complexité croissante :

  type 3 Langages rationnels
  type 2 Langages algébriques
  type 1 Langages contextuels
  type 0 Langages récursivement énumérables

- Ces ensembles de langages sont inclus les uns dans les autres
  $type3 \subset type2 \subset type1 \subset type0$
- Ces langages sont définis par les outils permettant de les représenter et par les algorithmes/outils permettant de vérifier l'appartenance d'un mot

Les opérations sur les langages

Outre les opérations classiques sur les ensembles (intersection, union, complémentaire), il y a trois autres opérations sur les langages :

la concaténation soit deux langages L et M, LM représente tous les mots xy tels que x est un mot de L et y un mot de M

la puissance soit un langage L, $L^0 = \{\epsilon\}$ et $L^n = LL^{n-1}$ pour tout $n > 0$

la fermeture de Kleene soit un langage L, $L^* = \cup_{i \geq 0} L^i$

INSA ROUEN NORMANDIE

Clôture de ces langages

Opération	type 3	type 2	type 1	type 0
Union	clos	clos	clos	clos
Intersection	clos	pas de cloture	clos	clos
Complémentaire	clos	pas de cloture	clos	pas de cloture
Concaténation	clos	clos	clos	clos
Puissance	clos	clos	clos	clos
Fermeture de Kleene	clos	clos	clos	clos

INSA ROUEN NORMANDIE

Qu'est-ce qu'une grammaire (formelle) ?

- Un formalisme permettant de représenter un langage formel

Constituant d'une grammaire

- un ensemble de (symboles) terminaux appartenant à l'alphabet du langage à représenter (notés conventionnellement par des minuscules)
- un ensemble de (symboles) non terminaux (notés conventionnellement par des majuscules)
- un axiome qui est l'un des non terminaux, conventionnellement noté S
- des règles de production. Une règle associe deux suites de terminaux et non terminaux (que l'on nomme partie/membre gauche et droit) séparées par $\rightarrow$

Grammaire linéaire à gauche (ou à droite)

- Une grammaire est linéaire si :
  - les membres gauches des règles sont constitués d'un seul non terminal
  - les membres droits des règles possèdent des terminaux et au plus un non terminal
  - par exemple : $A \rightarrow aBc$
- Une grammaire est linéaire à gauche (respectivement à droite) si c'est une grammaire linéaire et le non terminal de la partie droite est le premier élément (respectivement le dernier élément) : $A \rightarrow Bc$
- Une grammaire linéaire à gauche (ou, exclusivement, à droite) permet représenter un langage rationnel (type 3)

Les nombres naturels pairs représentés en binaire

$S \rightarrow 0 \quad A \rightarrow 1 \quad A \rightarrow A1$

$S \rightarrow A0 \quad A \rightarrow A0$

que l'on peut noter :

$S \rightarrow 0|A0$

$A \rightarrow 1|A0|A1$

Grammaire non contextuelle (ou algébrique)

- Une grammaire est non contextuelle si les membres gauches des règles sont constitués d'un seul non terminal
- Par exemple : $A \rightarrow aBcC$
- Une grammaire non contextuelle permet de représenter les langages algébriques (type 2)

Les expressions arithmétiques

$S \rightarrow S + S | S - S | S * S | S / S | nb | (S)$

INSA ROUEN NORMANDIE

Grammaire contextuelle

- Une grammaire est contextuelle si les règles sont de la forme [a] :

$$\alpha X \beta \rightarrow \alpha \gamma \beta$$

- Une grammaire contextuelle permet de représenter les langages contextuels (type 1)

a. on représente conventionnellement à l'aide de lettres grecs minuscules une succession quelconque de terminaux et de non terminaux

$a^n b^n c^n$ avec $n > 0$ (exemple de wikipédia)

1. $S \rightarrow aSBC$
2. $S \rightarrow aBC$
3. $CB \rightarrow HB$
4. $HB \rightarrow HC$
5. $HC \rightarrow BC$
6. $aB \rightarrow ab$
7. $bB \rightarrow bb$
8. $bC \rightarrow bc$
9. $cC \rightarrow cc$

Dérivation 1 / 2

- On peut vérifier par dérivation qu'une suite de symboles s est un mot m d'un langage L représenté par une grammaire G si en partant de l'axiome de G et en appliquant à chaque fois une règle (remplacer la partie gauche par la partie droite correspondante) on arrive à obtenir s

10110 est bien un nombre pair !

- La grammaire :
  1. $S \rightarrow A0$
  2. $A \rightarrow A0$
  3. $A \rightarrow A1$
  4. $A \rightarrow \epsilon$
- Les dérivations :
  $S \Rightarrow_1 A0 \Rightarrow_3 A10 \Rightarrow_3 A110 \Rightarrow_2 A0110 \Rightarrow_3 A10110 \Rightarrow_4 10110$

INSA ROUEN NORMANDIE

Exercice

- Sachant que la grammaire suivante engendre le langage $a^n b^n c^n$ (avec $n > 0$) :

1. $S \rightarrow aSBC$
2. $S \rightarrow aBC$
3. $CB \rightarrow HB$
4. $HB \rightarrow HC$
5. $HC \rightarrow BC$
6. $aB \rightarrow ab$
7. $bB \rightarrow bb$
8. $bC \rightarrow bc$
9. $cC \rightarrow cc$

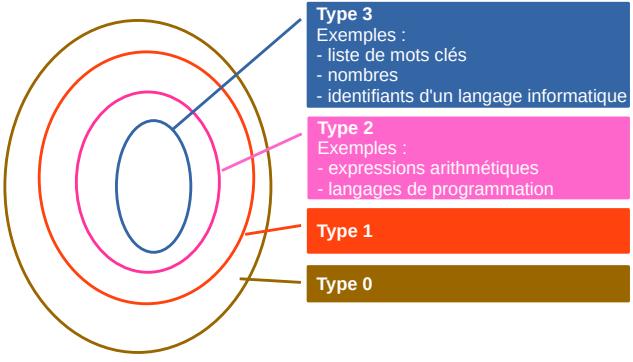
- Montrer que *aabbcc* est un mot de ce langage

INSA ROUEN NORMANDIE

Conclusion

Nous avons vu dans ce cours

- Langage formel
- Grammaire formelle
- Dérivation



INSA ROUEN NORMANDIE