Compilation

CM4 - Typage & Vérification de type

ISTIC, Université de Rennes 1
Sebastien.Ferre@irisa.fr

COMP, M1 info

Plan

Plan

Typer c'est quoi ?

Typer c'est :

- classer et normaliser les objets élémentaires
  - ex : entiers, booléens, flottants, caractères
- afin de savoir les composer
  - ex : tableaux, structures
- en se protégeant contre des erreurs courantes
  - ex : `tab.champ`, `entier[i]`, `entier + booléen`

Langage de types :

1. types élémentaires : ex, entier, booléen
2. types composés : ex, tableau, structure, fonction

→ *fait partie ou non du langage source*

Typer c'est quoi ?

Typer c'est :

- classer et normaliser les objets élémentaires
  - ex : entiers, booléens, flottants, caractères
- afin de savoir les composer
  - ex : tableaux, structures
- en se protégeant contre des erreurs courantes
  - ex : `tab.champ`, `entier[i]`, `entier + booléen`

Langage de types :

1. types élémentaires : ex, entier, booléen
2. types composés : ex, tableau, structure, fonction

→ *fait partie ou non du langage source*

Pourquoi typer ?

- rôle d'abstraction
  - les expressions `1`, `1+2`, `(3*x)+1` sont toutes de type entier *utilisables partout où un entier est attendu*
  - analogie : une ampoule à vis peut être utilisée dans n'importe quelle douille à vis
- prévenir les erreurs de type
  - ex : entier utilisé là où un booléen est attendu
  - analogie : une ampoule à vis dans une douille à baionnette
  - analogie : une prise male USB dans une prise femelle 220V !

Remarque

La machine ne voit que des emplacement mémoires et des octets !

Histoire des langages de programmation

- apparition pragmatique en C et Fortran [1960s]
  - conditionne allocation mémoire + détection erreurs
- formalisation très complète en ML [1970s]
  - type = propriété abstraite des expressions
    expressions entières
  - règles de propagation des types
    - la somme de deux entiers est un entier
    - si T est un tableau de chaines et i est un entier, alors $T[i]$ est une chaine
  - le tout forme une théorie logique avec axiomes et règles

Distinctions entre langages (1/3)

- vérification (C, Java)
  déclarations de type des variables
  - ex : `int x;` ⇒ `x + 1` est correct
- inférence (ML)
  déduction des types d'après l'utilisation des variables
  - ex : `x + 1` ⇒ `x : int`

Distinctions entre langages (2/3)

- typage statique (ML) vs dynamique (Lisp, Smalltalk) vs mixte (C, Java)
  - statique : tout est vérifié à la compilation
    avantage : les erreurs sont détectées plus tôt
  - dynamique : tout est vérifié à l'éxécution
    avantage : offre plus de flexibilité
- typage fort (ML, Lisp, Java) vs faible (C)
  - typage fort : aucune opération mal typée n'est permise
  - certaines vérifications sont possibles à la compilation
    ex. : `tab["toto"]`
  - d'autres doivent être faites à l'éxécution
    ex. : dépassement des bornes d'un tableau
- ATTENTION : les deux aspects sont orthogonaux

Distinctions entre langages (2/3)

- typage statique (ML) vs dynamique (Lisp, Smalltalk) vs mixte (C, Java)
  - statique : tout est vérifié à la compilation
    avantage : les erreurs sont détectées plus tôt
  - dynamique : tout est vérifié à l'éxécution
    avantage : offre plus de flexibilité
- typage fort (ML, Lisp, Java) vs faible (C)
  - typage fort : aucune opération mal typée n'est permise
  - certaines vérifications sont possibles à la compilation
    ex. : `tab["toto"]`
  - d'autres doivent être faites à l'éxécution
    ex. : dépassement des bornes d'un tableau
- ATTENTION : les deux aspects sont orthogonaux

Distinctions entre langages (3/3)

notation littérale des valeurs d'un type

- seulement pour les types élémentaires (C)
  - entier : `1`, `127`
  - chaine : `"Hello"`
  - arbre (type composé) :
    - déf. type : `struct tree { int val; struct tree* left, right; }`
    - création valeur : code nécessaire . . . . . .
- pour la plupart des types (ML, Prolog)
  - entier, chaine : idem C
  - arbre :
    - déf. type : `type tree = Leaf of int | Node of tree * int * tree`
    - notation valeur : `Node(Leaf(2), 1, Leaf(3))`

Typage

3 ingrédients

1. constructions du langages source (syntaxe du langage)
   grammaire du langage source
2. langage de types (élémentaires et complexes)
3. système de type (association entre les deux)
   grammaire attribuée dont les attributs contiennent des expressions de type

Plan

Langage de types

- directement sous forme de syntaxe abstraite
  $\Rightarrow$ pour être indépendant de la syntaxe concrète de tel ou tel langage
- on retrouve peu ou prou les mêmes expressions de types d'un langage à l'autre
  - expressions élémentaires
  - expressions complexes

Expressions de types élémentaires

- booléen : valeurs vrai ou faux
- entier : signés ou non, courts ou longs, etc.
- flottants : à precision simple ou double
- chaine : ASCII ou UTF
- vide : information vide, 1 seule valeur
  - `unit` en ML, `void` en C
- exec : type des instructions, des procédures et effets de bord
  - cas particulier du type vide

Expressions de type complexes

- tableau(A,B) : indices de type A et valeurs de type B

a_1	a_2	a_3	...
b_1	b_2	b_3	...

  - en général, A = entier
  - sinon, tables de hachage
  - en C : B`[]`
  - en ML : B `array`, `(`A`,` B`)` `Hashtbl.t`

Remarque

tableau est un constructeur de type, et les paramètres A et B peuvent être des expressions de type quelconques.

Expressions de type complexes

- fonction(A,B) : fonctions à un argument de type A et un résultat de type B
  - en C : $B\,(A)$
  - en ML : $A \rightarrow B$

Expressions de type complexes

- struct($a_1 : A_1, \ldots, a_n : A_n$) : structure, enregistrement (*record*), type produit
  - les a_i sont des noms de champs et les A_i leurs types
  - les valeurs définissent tous les champs (type produit)
  - en C : `struct { A1 a1; ...; An an }`
  - en ML : `{ a1 : A1; ...; an : An }`

Expressions de type complexes

- union($a_1 : A_1, \ldots, a_n : A_n$) : union, type somme
  - les a_i sont des noms de champs ou constructeurs
  - les valeurs définissent un seul champ (type somme)
  - en C : `union { A1 a1; ...; An an }`
  - en ML : a_1 `of` A_1 `|` `...|` a_n `of` A_n

Expressions de type complexes

- pointeur(A) : pointeur ou référence sur un A
  - les valeurs sont des adresses d'emplacements mémoires de type A
  - en C : $A*$
  - en ML : A `ref`

Expressions de type

Ce sont les types les plus courants.
Il manque :

- les types des langages OO (ex., classes)
  - classe $\approx$ structure dont les champs sont des fonctions (méthodes)
  - objet $\approx$ structure(id, classe, attributs)
- les définitions de type
  - ex : type dictionnaire = tableau(chaine,chaine)

AST des expressions de types

Les expressions de type peuvent être représentées par des arbres de même nature que les AST :

- .....
- .....
- .....

Discussion sur les tableaux

- dans tableau(*A*,*B*)
  - *A* n'est pas un intervalle de positions
  - ne pas confondre le "10" avec *A* dans *B*[10] (positions 0..9)
- on pourrait avoir tableau(*A*,Inf,Sup,B)
  - avec Inf et Sup les bornes du tableau
  - mais on ne sait pas vérifier le respect de ces bornes de façon statique (à la compilation)
  - ces bornes serviront à produire le code d'adressage (et de vérification dynamique) lors de la traduction du code source en code intermédiaire/cible

Discussion sur les fonctions

- Les fonctions peuvent être combinées récursivement de façon arbitraire
  - fonction en argument ou en résultat
  - ex : fonction(fonction(A,B), fonction(fonction(B,C), fonction(A,C)))
    en ML : $(A \to B) \to (B \to C) \to (A \to C)$
  - c'est le type de la composition de fonction
- Cela ne gêne pas la vérification de type, ni ne la complexifie !
- Par contre, cela implique un schéma d'exécution particulier, pour manipuler les fonctions passées en paramètre
  - valeurs d'ordre supérieur

Discussion sur les fonctions

- Les fonctions peuvent être combinées récursivement de façon arbitraire
  - fonction en argument ou en résultat
  - ex : fonction(fonction(A,B), fonction(fonction(B,C), fonction(A,C)))
    en ML : $(A \to B) \to (B \to C) \to (A \to C)$
  - c'est le type de la composition de fonction
- Cela ne gêne pas la vérification de type, ni ne la complexifie !
- Par contre, cela implique un schéma d'exécution particulier, pour manipuler les fonctions passées en paramètre
  - valeurs d'ordre supérieur

Ordre d'un type

Definition (Ordre d'un type)

- L'ordre d'un type peut être défini de façon récursive comme suit :
  - $Ordre(fonction(A, B)) = Max(Ordre(A) + 1, Ordre(B))$
  - $Ordre(A) = 0$, pour tout type non fonctionnel
- Un type A est dit d'ordre supérieur si $Ordre(A) > 0$ (fonctions)
- Une fonction de type T est dite d'ordre supérieur si $Ordre(T) > 1$ (fonctions de fonctions)

Fonctions d'ordre supérieur ?

- fonction(chaine, fonction(entier, fonction(entier, chaine))) ?

  ......
- fonction(fonction(chaine,exec), fonction(tableau(entier,chaine), exec)) ?

  ......
- fonction(entier, fonction(fonction(entier,entier), fonction(entier,entier))) ?

  ......

Remarque

Les fonctions d'ordre 2 sont communes dans les langages fonctionnels (ex., map, fold). Les fonctions d'ordre 3 ou plus sont rarrissimes en pratique.

Fonctions d'ordre supérieur ?

- fonction(chaine, fonction(entier, fonction(entier, chaine))) ?

  ......
- fonction(fonction(chaine,exec), fonction(tableau(entier,chaine), exec)) ?

  ......
- fonction(entier, fonction(fonction(entier,entier), fonction(entier,entier))) ?

  ......

Remarque

Les fonctions d'ordre 2 sont communes dans les langages fonctionnels (ex., map, fold). Les fonctions d'ordre 3 ou plus sont rarrissimes en pratique.

Curryfication des fonctions

Definition (Curryfication)

La curryfication est le codage d'une fonction n-aire en une imbrication de fonctions unaires.

- type C : $R(A_1,...,A_n)$
  $\rightarrow$ fonction(A_1, ..., fonction(A_n,R))
- appel C : $f(x_1,...,x_n)$
  $\rightarrow f(x_1)(x_2)...(x_n)$

Centralité du type fonction

Rôle central des fonctions :

- tableau(A,B) peut être assimilé à une fonction (de A vers B) définie en extension (au cas par cas)
- dans struct($a_1 : A_1, \ldots, a_n : A_n$), les champs a_i peuvent être assimilés à des fonctions
  - a_i : fonction(struct($a_1 : A_1, \ldots, a_n : A_n$), A_i)
- dans union($a_1 : A_1, \ldots, a_n : A_n$), les constructeurs a_i peuvent être assimilés à des fonctions
  - a_i : fonction(A_i, union($a_1 : A_1, \ldots, a_n : A_n$))

Remarque

Le λ-calcul est un formalisme Turing-complet qui ne connaît que les fonctions. Même les entiers et les booléens y sont codés par des fonctions !

Centralité du type fonction

Rôle central des fonctions :

- tableau(A,B) peut être assimilé à une fonction (de A vers B) définie en extension (au cas par cas)
- dans struct($a_1 : A_1, \ldots, a_n : A_n$), les champs a_i peuvent être assimilés à des fonctions
  - a_i : fonction(struct($a_1 : A_1, \ldots, a_n : A_n$), A_i)
- dans union($a_1 : A_1, \ldots, a_n : A_n$), les constructeurs a_i peuvent être assimilés à des fonctions
  - a_i : fonction(A_i, union($a_1 : A_1, \ldots, a_n : A_n$))

Remarque

Le λ-calcul est un formalisme Turing-complet qui ne connaît que les fonctions. Même les entiers et les booléens y sont codés par des fonctions !

Exemples d'expressions de types en C et ML

. . . . . .

Plan

Vérification de types

Exemple courant de vérification contextuelle non exprimable dans la grammaire :

⇒ analyse sémantique

⇒ grammaire attribuée

attribut	description	type
BT	“bien typé”	booléen
TS	table des symboles	tableau(ident,type)
type	expression de type	type

- ident est le type des identificateurs des programmes sources
- type est le type des expressions de type
  - les valeurs sont des types (AST) !

Le type des types

Comment définir le type des types ?

```
define type =
  union(booleen : vide,
        entier : vide,
        ...,
        pointeur : type,
        tableau : struct(indice : type, valeur : type),
        fonction : struct(param : type, result : type),
        struct : tableau(chaine, type),
        union : tableau(chaine, type))
```

On retouve la distinction «en texte» / «en actes» pour les types, parallèle à celle des programmes :

- «en texte» : le type est une valeur du type `type` ci-dessus et apparaît dans un programme «en texte» (source)
- «en actes» : le type est une propriété d'une valeur manipulée par un programme «en actes»

Le type des types

Comment définir le type des types ?

```
define type =
  union(booleen : vide,
        entier : vide,
        ...,
        pointeur : type,
        tableau : struct(indice : type, valeur : type),
        fonction : struct(param : type, result : type),
        struct : tableau(chaine, type),
        union : tableau(chaine, type))
```

On retouve la distinction «en texte» / «en actes» pour les types, parallèle à celle des programmes :

- «en texte» : le type est une valeur du type `type` ci-dessus et apparaît dans un programme «en texte» (source)
- «en actes» : le type est une propriété d'une valeur manipulée par un programme «en actes»

Exemple : déclarations et expressions

Grammaire attribuée pour la vérification de type d'un petit langage de déclarations et d'expressions.

. . . . . .

Extension de l'exemple

Pour chaque nouvelle construction du langage

- accès tableau `tab[i]`
- accès champ structure `point.x`

il suffit d'ajouter une règle syntaxique pour E et de définir les attributs : BT, TS et type.

Notation formelle et concise

Remarque

On peut adopter une présentation plus formelle et plus concise pour la vérification de type.

- La notation $E : \tau$ équivaut aux équations
  - `E.BT := vrai`
  - `E.type :=` τ
  - la table des symboles est implicite
- La notation $\frac{H1 \; H2}{C}$ signifie que si $H1$ et $H2$ sont vérifiés, alors C l'est aussi

Notations formelles pour l'exemple étendu

. . . . . .

Système formel

On vient de construire un système formel permettant de prouver si une expression est bien typée et de calculer son type

- $\frac{}{C}$ est un axiome
- $\frac{H1 \quad H2}{C}$ est une règle d'inférence

Du système formel à la grammaire attribuée

Un tel système formel se traduit directement en grammaire attribuée

- compilable par un compilateur de compilateurs (ANTLR)
- axiome : force la valeur de l'attribut "type"
  - ex : $\frac{}{\text{int} : \text{entier}}$ `E → int { E.type := entier }`
- règle : le type dépend des types à droite
  - + contraintes de "bon typage"
    1. les différentes occurences d'une variable doivent être égales
    2. les types à droite doivent "matcher" les constructeurs de type
    3. les conditions sur la TS doivent être vérifiées
  - ex :

$$\frac{t \ : \ \text{tableau}(\tau_1 = \text{entier}, \tau_2) \quad \text{"x"} \ : \ \tau_1 = \text{chaine}}{t[\text{"x"}] \ : \ \tau_2 = \text{erreur}}$$

    - contrainte non vérifiée : 2 valeurs différents pour τ_1
    - échec vérification : BT = faux, type = erreur

Extension de la vérification aux instructions

La vérification de type peut être étendue aux instructions

- instruction = expression de type 'exec'
- affectation

$$\frac{\mathit{ident} \ : \ \tau \quad E \ : \ \tau}{(\mathit{ident} := E) \ : \ \text{exec}}$$

- conditionnelle

$$\frac{E_1 \ : \ \text{booleen} \quad E_2 \ : \ \text{exec} \quad E_3 \ : \ \text{exec}}{(\text{si } E_1 \text{ alors } E_2 \text{ sinon } E_3) \ : \ \text{exec}}$$

- ...

Exemple complet de vérification de type

. . . . . .

Vérification de type statique/dynamique

La vérification de type peut être

- statique : lors de la compilation
  au seul vu du programme source
- dynamique : lors de l'exécution
  les valeurs des variables sont connues
- les deux

La vérification dynamique s'impose pour les propriétés non-décidables statiquement

- ex : respect des bornes d'un tableau

Typage statique fort

Certains langages tels que ML ont un système de typage statique et fort.

Definition (typage statique fort)

Dans un langage de programmation à typage statique fort, un programme bien typé (statiquement) ne peut pas échouer à la vérification dynamique.

- la vérification dynamique devient donc inutile
- il n'est plus nécessaire de représenter les types à l'exécution ni de faire des tests

$\rightarrow$ gain en sûreté ET en efficacité

Plan

Inférence de type

Principe

Laisser le compilateur inférer les types des variables plutôt que de demander au programmeur de les déclarer. Cela permet une plus grande concision des programmes.

- le type d'une variable est inféré à partir de ses contextes d'utilisation
  - ex : dans x + 1, x doit avoir le type entier
- la table des symbole est synthétisée par les expressions
- il faut vérifier qu'une même variable a toujours le même type !
  - ex : dans x + strlen(x), les 2 occurences de x ont des types incompatibles (entier vs chaine)
- on s'appuie sur le même système formel (axiomes et règles)

Exemple d'inférence

Typage de x + 1 = plus(x)(1) où :

- plus : fonction(entier, fonction(entier, entier))

. . . . . .

Exemple d'inférence avec erreur

Typage de x + strlen(x) où :

- strlen : fonction(chaine, entier)

. . . . . .

Autre exemple d'inférence

Typage de print((*objet).m1(x))(10) où :

- print : fonction(chaine, fonction(entier, exec))

. . . . . .