Vers un langage typé pour la programmation modulaire

UMONS

23 juin 2017

1 Introduction - OCaml

2 Programmation fonctionnelle et typage

3 DOT

4 Implémentation - RML

Introduction - OCaml

OCaml

Commençons par parler d'OCaml...

- langage fonctionnel : les fonctions sont des citoyens de première classe.
- statiquement typé : les types sont assignés à la compilation, non à l'exécution. $\neq$ dynamiquement typé (Python).
- dispose d'algorithmes d'inférence de type : pas obligé d'assigner manuellement un type à un terme (contrairement à C ou Java).
- divisé en deux « langages » :
  - langage de base : fonctions, enregistrements, tuples, variables, ...
  - langage de modules : modules, foncteurs.
- **Problème :** les deux langages possèdent des algorithmes d'inférence de type différents et sont gérés différement par le compilateur.

But du mémoire

- Etudier un calcul théorique où les deux langages (et en particulier les règles de typage et de sous-typage) sont unifiés. Calcul choisi : DOT (2016).
- Représenter les modules et les enregistrements de la même manière.
- (Apport principal de ce mémoire). Implémenter un langage de surface, proche d'OCaml, un algorithme de typage et un algorithme de sous-typage pour écrire des programmes DOT. Langage : RML

Programmation fonctionnelle et typage

λ-calcul non typé

Calcul = syntaxe (termes) + sémantique (règles d'évaluation, non discutée).
Soit V un ensemble infini dénombrable de variables. La syntaxe du λ-calcul est définie comme le plus petit ensemble Λ tel que

1. $V \subseteq \Lambda$
2. $\forall u, v \in \Lambda, u\, v \in \Lambda$
3. $\forall x \in V, \forall u \in \Lambda, \lambda x.\, u \in \Lambda$

$\Leftrightarrow$

$t ::=$		terme
	x	var
	$t\ t$	app
	$\lambda x.\, t$	abs

- $u\, v \rightarrow$ application d'une fonction u à un paramètre v.
- $\lambda x.\, u \rightarrow$ définition d'une fonction (anonyme) qui prend un paramètre x et retourne u.

Typage

- Classer les termes en fonction de leur nature $\rightarrow$ notion de type.
- Exemple avec le terme $u\,v$: u doit être une fonction qui prend un paramètre v d'un certain type T_1 et retourne un terme d'un type T_2. On note $T_1 \rightarrow T_2$ le type de u (type flèche).
- Formellement et de manière générale, on se donne τ, un ensemble de types, les éléments étant notés T, et on définit une **relation de typage** entre Λ et τ. On note $t : T$ pour dire que t est de type T. On dit aussi que t est bien typé.

λ-calcul simplement typé - Syntaxe des types

Soit $\mathcal{B}$ un ensemble de types de base, ses éléments étant notés B.

$T ::=$		types
	B	base
	$T \to T$	type des fonctions

Contexte et jugement de typage, différence de syntaxe

- $\lambda x.\, t$: type de x ? $\Rightarrow$ on annote la variable avec son type. $\lambda x.\, t$ devient $\lambda x : T.\, t$.
- un terme t peut contenir des variables $\Rightarrow$ pour typer t, besoin de connaître le type de celles-ci $\Rightarrow$ **contexte ou environnement de typage** Γ = suite (x_i, T_i) où x_i est une variable et T_i est le type de x_i.
- $t : T$ devient $\Gamma \vdash t : T$ (appelé **jugement de typage**).

λ-calcul simplement typé - Règles de typage

$$\frac{(x : T) \in \Gamma}{\Gamma \vdash x : T} \quad \text{(T-VAR)} \qquad \frac{\Gamma, x : T_1 \vdash t : T_2}{\Gamma \vdash \lambda x : T_1 . t : T_1 \to T_2} \quad \text{(T-ABS)}$$

$$\frac{\Gamma \vdash u : T_1 \to T_2 \qquad \Gamma \vdash v : T_1}{\Gamma \vdash u\, v : T_2} \quad \text{(T-APP)}$$

λ-calcul simplement typé - Arbre de dérivation

Une succession d'utilisations de règles de typage mène à **un arbre de dérivation**.

$$\dfrac{\dfrac{\dfrac{(\text{T-VAR})\ \dfrac{x : T_1 \to T_2 \in \Gamma}{\Gamma \vdash x : T_1 \to T_2} \qquad \dfrac{y : T_1 \in \Gamma}{\Gamma \vdash y : T_1}\ (\text{T-VAR})}{\Gamma, x : T_1 \to T_2, y : T_1 \vdash x\,y : T_2}\ (\text{T-APP})}{\Gamma, x : T_1 \to T_2 \vdash \lambda y : T_1 . x\,y : T_1 \to T_2}\ (\text{T-ABS})}{\Gamma \vdash \lambda x : T_1 \to T_2 . \lambda y : T_1 . x\,y : (T_1 \to T_2) \to T_1 \to T_2}\ (\text{T-ABS})$$

Pour le λ-calcul simplement typé : il existe au plus un arbre de dérivation.
Pas toujours le cas (exemple : DOT).

λ-calcul simplement typé - Sûreté

Théorèmes importants :

1. Préservation : si $t : T$ et $t \to t'$, alors $t' : T$.
2. Progression : si $t : T$, alors soit t est une valeur, soit $t \to t'$.

Preuves : lemmes intermédiaires + induction sur la dérivation $t : T$ ou $t \to t'$.

Sous-typage

- But : affiner notre relation de typage en définissant une relation binaire sur les types (dite de **sous-typage**).
- Notation : $S <: T$ (S est sous-type de T).
- Ajout d'une règle de typage qui crée un lien entre la relation de typage et la relation de sous-typage.

$$\frac{\Gamma \vdash t : S \qquad S <: T}{\Gamma \vdash t : T} \quad \text{(T-SUB)}$$

- La relation est souvent transitive et réflexive :

$$\frac{S <: T \qquad T <: U}{S <: U} \quad \text{(S-TRANS)} \qquad\qquad T <: T \quad \text{(S-REFL)}$$

Enregistrement

- enregistrement = ensemble de labels liés à des termes. Par exemple $\{l_1 = 5; l_2 = \lambda x.\, 42\}$.
- Les champs ne sont pas mutuellement dépendants !

$t ::=$	terme
x	var
$t\ t$	app
$\lambda x : T.\, t$	abs
$\{l_i = t_i\}^{1 \le i \le n}$	enreg
$t.l$	proj

$T ::=$	type
$T \to T$	fonction
$\{l_i : T_i\}^{1 \le i \le n}$	enreg

Sous-typage et enregistrement

$$\{l_i : T_i\}^{1 \leq i \leq n+k} <: \{l_i : T_i\}^{1 \leq i \leq n} \quad \text{(S-RCD-WIDTH)}$$

$$\frac{\{k_j : S_j\}^{1 \leq j \leq n} \text{ permutation de } \{l_i : T_i\}^{1 \leq i \leq n}}{\{k_j : S_j\}^{1 \leq j \leq n} <: \{l_i : T_i\}^{1 \leq i \leq n}} \quad \text{(S-RCD-PERM)}$$

$$\frac{\forall i \in \{1, ..., n\}, S_i <: T_i}{\{l_i : S_i\}^{1 \leq i \leq n} <: \{l_i : T_i\}^{1 \leq i \leq n}} \quad \text{(S-RCD-DEPTH)}$$

(T-SUB) et (S-RCD-WIDTH) $\rightarrow (\lambda p : \{x : \mathbb{R}\} . p.x) \{x = \pi; y = 42\}$ bien typé.

DOT

DOT - Description

- Ajoute les types à l'intérieur des enregistrements $\Rightarrow$ types dépendants $(x.A)$.
- Un enregistrement peut avoir des champs mutuellement dépendants grâce à une variable interne.
- Les fonctions deviennent dépendantes : le type de retour peut dépendre du paramètre.

DOT - Syntaxe des termes

$t ::=$	terme
x, y	var
$\lambda x : T . t$	abs
$x\, y$	app
$x.a$	champ proj
$let\ x\ =\ t\ in\ t$	let
$\nu(x : T^x)d$	rec

$d ::=$	decl
$\{a = t\}$	champ
$\{A = T\}$	type
$d \wedge d$	aggregation

DOT - Syntaxe des types

$S, T ::=$		type
	Top	top
	$Bottom$	bottom
	$\forall(x : S)T^x$	fonction
	$\{A : S..T\}$	type decl
	$\{a : T\}$	champ decl
	$S \wedge T$	inter
	$x.A$	type proj
	$\mu(x : T^x)$	rec

DOT - Quelques règles de typage et de sous-typage

...

$$\frac{\Gamma \vdash x : \{A : S..T\}}{\Gamma \vdash S <: x.A} \quad (<: \text{SEL}) \qquad \frac{\Gamma \vdash x : \{A : S..T\}}{\Gamma \vdash x.A <: T} \quad (\text{SEL} <:)$$

$$\frac{\Gamma \vdash x : T^x}{\Gamma \vdash x : \mu(z : T^z)} \quad (\text{VAR-PACK}) \qquad \frac{\Gamma \vdash x : \mu(z : T^z)}{\Gamma \vdash x : T^x} \quad (\text{VAR-UNPACK})$$

...

DOT - Remarques

- Pas de règles de sous-typage pour les types récursifs $\Rightarrow$ (VAR-PACK, VAR-UNPACK) + règles de sous-typage.
- Règles de sous-typage dépendent du typage (e.g. SEL $<:$ et $<:$ SEL). **Pas usuel**.
- Problème des mauvaises bornes. Si $x : \{A : Top..Bottom\} \in \Gamma$, alors $\Gamma \vdash S <: T$ pour tous S, T par (SEL $<:$), ($<:$ SEL) et (S-TRANS).

Implémentation de DOT en OCaml - RML

Implémentation - Remarques

1. Aucun article ne décrit l'implémentation d'algorithmes de typage et de sous-typage pour DOT.
2. Besoin de définir un langage de surface (parseur).
3. RML implémente des sucres syntaxiques pour écrire plus facilement des programmes (fonction à plusieurs arguments, applications avec plusieurs arguments), gérés lors du parsing.
4. RML implémente un algorithme d'inférence de type pour les modules (terme $\nu(x)d$ en plus de $\nu(x : T)d$). Dans DOT, la définition d'un enregistrement récursif est obligatoirement liée à un type.

RML - Module = Enregistrement

```
let module Point2D = struct(point)
  type t = { x : Int.t ; y : Int.t }
  let add = fun(p1 : point.t, p2 : point.t) ->
    { x = Int.plus p1.x p2.x ; y = Int.plus p1.y p2.y }
end;;
```

- $\nu(x)d \rightarrow$ struct d end
- $\mu('self : T) \rightarrow$ enregistrement
- $\{a = t\} \rightarrow$ let a = t
- $\{A = T\} \rightarrow$ type a = t
- $\lambda x : T_1 . t \rightarrow$ fun(x : T_1) $\rightarrow$ t

RML - Module = Enregistrement - Type

```
sig(point)
  type t = sig('self) (* enregistrement *)
    val x : Int.t
    val y : Int.t
  end .. sig('self)
    val x : Int.t
    val y : Int.t
  end
  val add : forall(p1 : point.t) (* fonction *)
            forall(p2 : point.t)
    sig('self)
      val x : Int.t
      val y : Int.t
    end
end
```

RML - Foncteur = Fonction

```
let module MakePoint2D =
  fun(typ : sig(self)
    type t
    val plus : self.t -> self.t -> self.t
  end) ->
  struct(point)
    type t = { x : typ.t ; y : typ.t }
    let add = fun(p1 : point.t, p2 : point.t) ->
      let x' = typ.plus p1.x p2.x in
      let y' = typ.plus p1.y p2.y in
      { x = x' ; y = y' }
end;;
```

RML - Foncteur = Fonction - Type

```
forall(typ : sig(self)
  type t = Nothing .. Any
  val plus : self.t -> self.t -> self.t
end) sig(point)
  type t = sig('self) (* Enregistrement *)
    val x : typ.t (* Borne inférieure *)
    val y : typ.t
  end .. sig('self)
    val x : typ.t (* Borne supérieure *)
    val y : typ.t
  end
  (* Fonction add... *)
end
```

Implémentation - Pas si facile...

Passer de la théorie à l'implémentation n'est pas facile !

1. Gestion des variables libres et liées (utilisation d'AlphaLib).
2. Un même jugement de typage peut être dérivé de plusieurs manières.
3. Indécidabilité du sous-typage $\Rightarrow$ il n'existe pas d'algorithme qui termine sur chaque entrée et qui soit correct et complet.
4. Une même question $S <: T$ peut être posée plusieurs fois (ex : liste avec tail).

Implémentation - Problèmes

1. DOT n'est pas stable par insertion de binding locaux variable-variable.

```
let module M = struct(self)
  type t = Int.t
  let a : self.t = 5
end;;

(* Type M_local.t.
   Or M_local n'existe pas en dehors de
   la liaison locale.
*)
let M_local = M in M_local.a;;
```

Implémentation - Problèmes

1. DOT n'est pas stable par insertion de binding locaux variable-variable.
2. (SEL $<$:) et ($<$: SEL) nécessite un algorithme intermédiaire appelé best_bound[1].

$$\frac{\Gamma \vdash x : \{A : L..U\}}{\Gamma \vdash x.A <: U} \quad \text{(SEL-}<\text{ :)}$$

devient

$$\frac{\Gamma \vdash x : T \qquad \Gamma \vdash T <: \{A : L..U\} \qquad \Gamma \vdash U <: U'}{\Gamma \vdash x.A <: U'} \quad \text{(UN-SEL-}<\text{ :)}$$

1. Ce dernier n'est pas évident !

Travail futur

Travail futur

- Résoudre le problème des questions déjà posées (en cours, plusieurs idées).
- Améliorer les algorithmes et montrer qu'ils sont complets et corrects.
- Evaluateur.
- Interpréteur interactif.

Merci