

软件分析

过程间指针分析

熊英飞
北京大学

上下文敏感的指针分析



- 能否做精确的上下文敏感的指针分析?
- 域敏感的指针分析或者考虑二级指针的分析：不能
- Tom Reps等人2000年证明这是一个不可判定问题
  - 两个上下文无关语言的交集是否为空是不可判定的
    - 该问题的一个子问题P-PCP也是不可判定的
  - 上下文敏感的指针分析可以用CFL可达性求解
  - 域敏感的指针分析也可以用CFL可达性求解
  - 可以把P-PCP规约到上下文敏感和域敏感的指针分析

解决方法



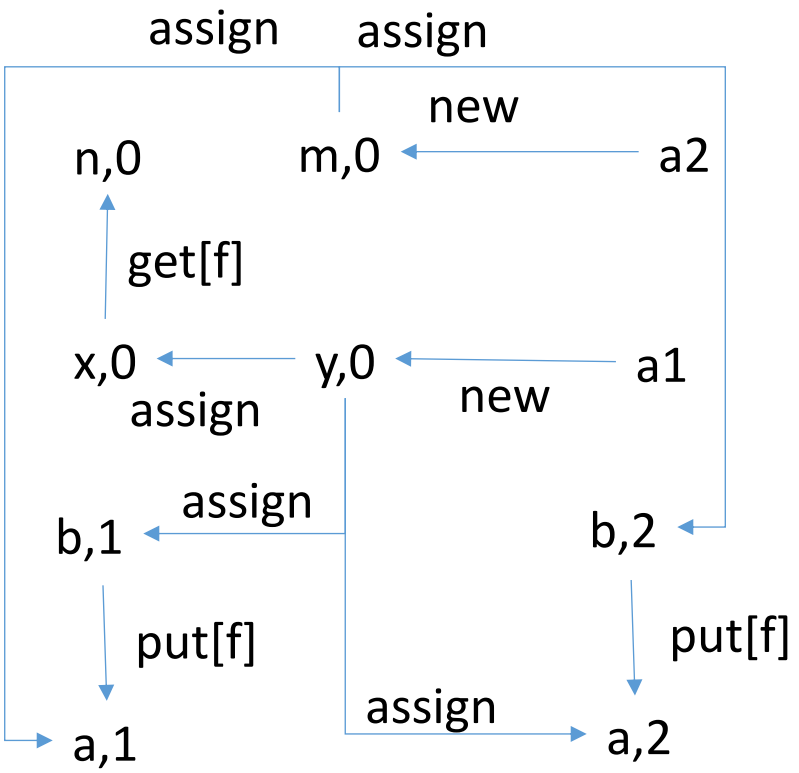
- 降低上下文敏感性：把被调方法根据上下文克隆n次
- 降低域敏感性：把域展开n次

降低上下文敏感性

FlowTo= new (assign | put[f] Alias get[f])*
PointsTo = (assign | get[f] Alias put[f])* new
Alias = PointsTo FlowTo

```
Main(): //0
y = new A();//a1
m=new A();//a2
SetF(m, y); //1
x=y;
SetF(y, m); //2
n=x.f;

SetF(a, b):
a.f=b;
```



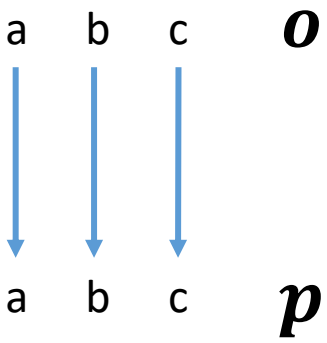
4 如果不克隆方法，会认为n指向a1

降低域敏感性

- 思路：转换成CFL可达性问题来进行精确的上下文敏感分析
- 遵循标准的多变量、流非敏感分析转换方法

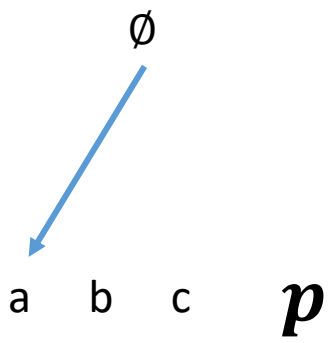
约束的转换

- 对于约束$\boldsymbol{p} \supseteq \boldsymbol{o}$
- 形成如下可达性图



常量约束的转换

- 对于约束$\boldsymbol{p} \supseteq \{a\}$
- 按标准方法应该所有节点都连一条到$\boldsymbol{p}.a$的边
- 该方法等价于从空集连一条到$\boldsymbol{p}.a$的边



降低域敏感性

- 但是，原分析中还有全称量词
  - $\forall x \in \boldsymbol{a},\ \boldsymbol{x.next} \supseteq \boldsymbol{b}$
- 这类约束无法直接转换成CFL可达性的图表示
  - 需要在图上动态加边
- 通过降低域敏感性来去掉全称量词

复习：域非敏感分析

```
Struct Node {
  int value;
  Node* next;
  Node* prev;
};
a = malloc();
a->next = b;
a->prev = c;
```

- 把所有struct中的所有fields当成一个对象
- 原程序变为
  - a=malloc();
  - a’=b;
  - a’=c;
  - 其中 a’ 表示 a->next 和 a->prev
- 约束中不会出现全称量词

域展开一次

```
Struct Node {
  int value;
  Node* next;
};
a = malloc();
a->next = b;
```

- 对于每个Node*的变量a，创建两个指针变量
  - **a**
  - **a->next**
- a=b产生
  - **a ⊇ b**
  - **a->next⊇b->next**
  - **b->next⊇a->next**
- a->next=b产生
  - **a->next ⊇ b**
  - **a->next⊇b->next**
  - **b->next⊇a->next**
- a=b->next产生
  - **a ⊇ b->next**
  - **a->next⊇b->next**
  - **b->next⊇a->next**

为什么两个方向都有？

约束中不含全程量词，可以用IFDS转成图并加上括号。

域展开两次

```
Struct Node {
  int value;
  Node* next;
};
a = malloc();
a->next = b;
```

- 对于每个Node*的变量a，创建3个指针变量
  - **a**
  - **a->next**
  - **a->next->next**
- a->next=b产生
  - **a->next ⊇ b**
  - **a->next->next⊇b->next**
  - **a->next->next⊆b->next**
  - **a->next->next⊇b->next->next**
  - **a->next->next⊆b->next->next**
- a=b->next产生
  - **a ⊇ b->next**
  - **a->next⊇b->next->next**
  - **a->next⊆b->next->next**
  - **a->next->next⊇b->next->next**
  - **a->next->next⊆b->next->next**

域敏感性 vs 上下文敏感性



- 降低上下文敏感性和降低域敏感性的研究历史上长期各自独立发展，学界对两者的统一和权衡研究还较少
  - Späth、Ali、Bodden: Context-, Flow-, and Field-Sensitive Data-Flow Analysis using Synchronized Pushdown Systems. POPL 2019.
- 降低上下文敏感性的工作
  - 通常基于CFL可达性的指向分析，通过引入新的表示和算法来尽量精确的进行上下文敏感分析
  - 如：用正则文法来模拟上下文敏感性所需的上下文无关文法
  - 参考： Manu Sridharan, Rastislav Bodík: Refinement-based context-sensitive points-to analysis for Java. PLDI 2006: 387-400
- 降低域敏感性的工作
  - 通常基于Andersen指向分析，通过引入新的表示和算法来尽量精确的进行域敏感分析
  - 如：引入*表示任意长的访问序列
  - 参考： Johannes Lerch, Johannes Späth, Eric Bodden, Mira Mezini:
  - Access-Path Abstraction: Scaling Field-Sensitive Data-Flow Analysis with Unbounded Access Paths (T). ASE 2015: 619-629
- 主流框架中更多采用精确域敏感性，用克隆实现上下文敏感性的方法

过程间分析-函数指针

```
interface I {
  void m();
}
class A implements I {
  void m() { x = 1; }
}
class B implements I {
  void m() { x = 2; }
}
static void main() {
   I i = new A();
   i.m();
}
```

如何设计分析算法得出程序执行结束后的x所有可能的值？

控制流分析



- 确定函数调用目标的分析叫做控制流分析
- 控制流分析是may analysis
  - 为什么不是must analysis?
- 控制流分析 vs 数据流分析
  - 控制流分析确定程序控制的流向
  - 数据流分析确定程序中数据的流向
  - 数据流分析在控制流图上完成，因此控制流分析是数据流分析的基础

Class Hierarchy Analysis

```
interface I {
  void m(); }
class A implements I {
  void m() { x = 1; } }
class B implements I {
  void m() { x = 2; } }
static void main() {
   I i = new A();
   i.m(); }
class C { void m() {} }
```

- 根据i的类型确定m可能的目标
- 在这个例子中，i.m可能的目标为
  - A.m()
  - B.m()
- 不可能的目标为
  - C.m()
- 分析结果为x={1,2}
- 优点：简单快速
- 缺点：非常不精确，特别是有Object.equals()这类调用的时候

Rapid Type Analysis

```
interface I {
  void m(); }
class A implements I {
  void m() { x = 1; } }
class B implements I {
  void m() { x = 2; } }
static void main() {
   I i = new A();
   i.m(); }
class C { void m() {
  new B().m();
} }
```

- 只考虑那些在程序中创建了的对象
- 可以有效过滤library中的大量没有使用的类

Rapid Type Analysis




```
interface I {
  void m(); }
class A implements I {
  void m() { x = 1; } }
class B implements I {
  void m() { x = 2; } }
static void main() {
   I i = new A();
   i.m(); }
class C { void m() {
  new B().m();
} }
```

- 三个集合
  - 程序中可能被调用的方法集合Methods，初始包括main
  - 程序中所有的方法调用和对应目标Calls→Methods
  - 程序中所有可能被使用的类Classes
- Methods中每增加一个方法
  - 将该方法中所有创建的类型加到Classes
  - 将该方法中所有的调用加入Call，目标初始为根据当前Classes集合类型匹配的方法
- Classes中每增加一个类
  - 针对每一次调用，如果类型匹配，把该类中对应的方法加入到Calls→Methods
  - 把方法加入到Methods当中

分析结果i.m()调用A.m()

Rapid Type Analysis

```
interface I {
  void m(); }
class A implements I {
  void m() { x = 1; } }
class B implements I {
  void m() { x = 2; } }
static void main() {
   I i = new A();
   I j = new B();
   i.m(); }
class C { void m() {
  new B().m();
} }
```

- 分析速度非常快
- 精度仍然有限
- 在左边的例子中，得出i.m的目标包括A.m和B.m
- 如何进一步分析出精确的结果?

精确的控制流分析CFA



- 该算法没有名字，通常直接称为CFA (control flow analysis)
- CFA和指针分析需要一起完成
  - 指针分析确定调用对象
  - 调用对象确定新的指向关系
- 原始算法定义在λ演算上
- 这里介绍算法的面向对象版本

CFA-算法

```
interface I {
  I m(); }
class A implements I {
  I m() { return new B(); } }
class B implements I {
  I m() { return new A(); } }
static void main() {
   I i = new A();
   if (…) i = i.m();
   I x = i.m();
}
```

- 首先每个方法的参数和返回值都变成图上的点
  - 注意this指针是默认参数
- 对于方法调用

```
f() {…
  x = y.g(a, b)
…}
```

- 生成约束
  - $\forall y \in \mathbf{f\#y}.\ \forall m \in \text{targets}(y, g),$
    $\mathbf{f\#x} \supseteq \mathbf{m\#ret}$
    $\mathbf{m\#this} \supseteq \text{filter}\ (\mathbf{f\#y}, \text{declared}(m))$
    $\mathbf{m\#a} \supseteq \mathbf{f\#a}$
    $\mathbf{m\#b} \supseteq \mathbf{f\#b}$
- 约束求解方法和Andersen指针分析算法类似

根据调用对象和方法名确定被调用方法

方法的声明类

保留符合特定类型的对象

CFA-计算示例

```
interface I {
  I f(); }
class A implements I {
  I f() { return new B[1](); } }
class B implements I {
  I f() { return new A[2](); } }
static void main() {
   I i = new A[3]();
   if (…) i = i.f();
   I x = i.f();
}
```

- **main#i** ⊇{3}
- $\forall i \in$ **main#i**, $\forall m \in$ targets(i, f),
  - **main#i**⊇**m#ret**
  - $\mathbf{m\#this} \supseteq \mathrm{filter}(\mathbf{main\#i}, \mathrm{declared(m)})$
- $\forall i \in$ **main#i**, $\forall m \in$ targets(i, f),
  - **main#x**⊇**m#ret**
  - $\mathbf{mthis} \supseteq \mathrm{filter}\left(\mathbf{main\#i}, \mathrm{declared(m)}\right)$
- **A.f#ret** ⊇{1}
- **B.f#ret** ⊇{2}
-
- 求解结果
  - **main#i**={1,2,3}
  - **main#x**={1, 2}

CFA



- 以上CFA算法是否是上下文敏感的?
- 不是，因为每个方法只记录了一份信息，没有区分上下文
- 用克隆的方法处理上下文敏感性
- 基于克隆方法的CFA也被称为k-CFA
  - 上下文不敏感的CFA称为0-CFA

作业

给定下面的程序，进行流非敏感的指针分析，域展开一次，请画出可达性图。

```
Struct Node {
  int value;
  Node* next;
};
a = malloc(); //1
b = malloc(); //2
a->next = b;
b->next = malloc(); //3
```

参考文献



- 基于CFL的指向分析：
  - Thomas W. Reps: Program Analysis via Graph Reachability. ILPS 1997: 5-19
- Steensgaard分析：
  - Lecture Notes: Pointer Analysis
    - Jonathan Aldrich
    - https://www.cs.cmu.edu/~aldrich/courses/15-819O-13sp/resources/pointer.pdf
- 控制流分析：
  - Lecture Notes: Object-Oriented Call Graph Construction
    - Jonathan Aldrich
    - https://www.cs.cmu.edu/~aldrich/courses/15-819O-13sp/resources/callgraph.pdf
- 基于展开域的指针分析：
  - Neil D. Jones, Steven S. Muchnick: Flow Analysis and Optimization of Lisp-Like Structures. POPL 1979: 244-256