

프로그램 분석 기초

최광훈

September 13, 2023

Chapter 1

프로그램 분석 개요

1.1 목적

프로그램을 개발하거나 프로그램의 보안 취약점을 탐지할 때 다양한 프로그램 분석 방법을 사용하여 개발 생산성을 높이거나 꼼꼼하게 보안 취약점을 탐지하고 있다. 특정 프로그램 분석을 지원하는 오픈소스 소프트웨어도 다양하게 개발되어 개발자나 보안 담당자가 이러한 프로그램을 사용고 있다. 하지만 프로그램 분석에 대한 기초 지식이 없다면 이 프로그램을 제대로 사용하는데도 한계가 있다. 프로그램 분석에 대한 기초 지식을 갖추고 있다면 이러한 프로그램을 사용하는 것뿐만 아니라 자신의 목적에 맞추어 수정하는 것도 가능할 것이다. 하지만 프로그램 분석을 직접 배울곳이 마땅치 않고 프로그램 분석 기법과 밀접하게 관련된 프로그래밍언어론과 컴파일러를 공부할 수 있는 곳도 제한되어 있다. 설사 그러한 과목들을 배울 수 있다하더라도 프로그램 분석 기법을 배우는데 한계가 있다.

프로그램 분석은 어려운 주제이다. 여러 이유를 생각해볼 수 있다. 우선, 모든 프로그램 분석 기법은 복잡한 수학에 기반을 두고 있기 때문이다. 그리고, 분석 대상이 되는 프로그램을 작성한 프로그래밍언어 자체가 복잡하기 때문이다. 마지막으로 우리가 분석하고 싶은 품질 속성이 다양하기 때문이다.

프로그램 분석 기법을 쉽게 익힐 수 있도록 이 강의 노트를 구성하고자 한다. 우선, 수학을 사용하여 프로그램 분석을 설명하는 것을 최소한으로 제한하고 실행 가능한 명세를 통해서 설명한다. 그리고 간단한 프로그래밍언어 WHILE을 정의

하여 이 언어로 작성된 프로그램을 대상으로 분석 기법을 설명한다. 마지막으로 어휘 분석, 구문 분석, 동적 의미, 타입 체킹, 정적 분석, 기호 실행으로 분석 기법을 한정하여 설명한다.

1.2 미리 알아두면 좋을 내용

프로그램 분석 기법에 대한 이 강의 노트를 공부하기 위해서 자료 구조와 기초 프로그래밍 지식이 필요하다. 리스트, 트리, 그래프와 같은 자료 구조를 이해하고 프로그래밍할 수 있어야 한다.

하스켈 프로그래밍언어로 프로그램 분석 기법에 대한 실행 가능한 명세를 작성 한다. 이 언어를 사용한 이유는 수학 표기법을 사용하여 명세를 작성한 것에 비해 접근하기 쉽고, 하스켈 프로그램이 간결하여 명세로 삼기에 적합하기 때문이다. 표 1.1에 각 분석 방법과 그 방법을 구현한 하스켈 프로그램의 라인 수를 볼 수 있다. 88 라인에서 250 라인으로 여섯 가지 분석 방법을 작성하였다.

분석 기법 명세	라인
어휘 분석	88
구문 분석	228
동적 의미	123
정적 의미(타입 검사)	145
정적 분석	250
기호 실행	208

Table 1.1: 분석 방법과 하스켈로 작성한 실행 가능한 명세의 라인 수

하스켈 프로그래밍언어에 대한 공식 사이트는 <https://haskell.org>이다. 다양한 레퍼런스를 이 웹 사이트에서 찾을 수 있다. 하스켈 기초 프로그래밍을 배우기 위해서 헬싱키 대학에서 만든 무크 사이트를 이용할 수 있다. 이 무크 사이트 내용으로 구성된 유튜브 동영상 강의도 하스켈 프로그래밍을 배우기에 도움이 될 것이다.

여담이지만 하스켈과 같은 정통 함수형 프로그래밍언어를 한번 배우는 것은 설사 객체지향 프로그래밍언어를 주로 사용하는 사람들에게도 도움이 될 것이다. 최근 파이썬, C++, 자바, 자바스크립트, 코틀린, 스위프트와 같은 객체지향 프로그래밍언어에 람다, 다형 타입 (제네렉, 템플릿), 변경 불가능한 값(immutable value), 리스트 제시법(list comprehension)과 같은 함수형 프로그래밍언어의 핵

심 특징들을 도입하고 있는 추세이다. 객체지향 프로그래밍 언어의 구문에 함수형 프로그래밍언어 특징을 구겨넣다보니 구문도 복잡하고 그 의미도 이해하기 어려운 경우도 있다. 정통 함수형 프로그래밍언어에서 그 특징들을 경험한다면 파이썬 프로그래밍을 할때도 프로그래밍에 대한 관점이 달라질 것이다.

이 외에 집합(set)이나 릴레이션(relation)과 같은 기초적인 수학 개념을 알 필요가 있다.

1.3 이 강의 노트로 공부하는 방법

이 강의 노트를 공부하는 방법은, 하스켈 프로그램으로 미리 준비해놓은 실행 가능한 명세들을 독자가 사용하는 프로그래밍언어로 다시 작성함으로써 각 분석 방법을 주체적으로 이해하는 방법을 권한다. 이 강의 노트에서는 파이썬을 선택하여 설명 한다.

강의 노트에서 사용하는 하스켈 프로그램의 내용은 다음과 같다.

- WHILE 프로그래밍 언어 예제 프로그램
- 어휘 분석(lexical analyzier)
- 구문 구조 분석(Parser)
- 동적 의미(semantics)
- 정적 의미-타입체킹
- 정적 분석-자료 흐름 분석
- 기호 실행

이 프로그램을 깃허브 저장소에서 내려 받고, 빌드한다.

기본적으로 하스켈 빌드 시스템 도구 스택(stack)을 설치해야 한다. 추가로 기호 실행 분석에서 사용하는 산술 복합 논리식을 해결하는 풀이 엔진 (SMT - Satisfiability Module Theories - solver) Z3 라이브러리가 필요하다.

- 리눅스

```
$ sudo apt-get install libz3-dev
$ git clone https://github.com/kwanghoon/Lecture_SAV
$ cd Lecture_SAV/whilelang
$ stack build
```

- 윈도우즈

- z3-4.8.12 버전을 내려받기
- (D:\z3-4.8.12-x64-win 디렉토리 아래 bin에 라이브러리, include에 헤더 파일이 있다고 가정)

```
D:\> git clone https://github.com/kwanghoon/Lecture_SAV
D:\> cd Lecture_SAV/whilelang
D:\> stack build
--extra-include-dirs=D:\z3-4.8.12-x64-win\include
--extra-lib-dirs=D:\z3-4.8.12-x64-win\bin
```

하스켈 프로그램을 빌드한 다음 어휘 분석부터 기호 실행까지 실행하는 방법은 다음과 같다. 실행 파일을 통해 각 분석 방법을 다음과 같이 실행할 수 있다.

```
$ stack exec -- whilelang-exe --lex ./example/while2.while
$ stack exec -- whilelang-exe --parse ./example/while2.while
$ stack exec -- whilelang-exe --typecheck ./example/while2.while
$ stack exec -- whilelang-exe --dataflow ./example/while2.while
$ stack exec -- whilelang-exe --symexec ./example/while2.while
```

하스켈 프로그램을 read-eval-print 방식으로 실행할 수도 있다.

```
$ stack ghci --
ghci> let srcFile = "./example/while2.while"
ghci> doLexing srcFile
ghci> doParsing srcFile
ghci> doRun srcFile
```

```
ghci> doTypecheck srcFile  
ghci> doAnalysis srcFile  
ghci> doSymbolic srcFile
```

WHILE 프로그램을 하스켈 어휘 분석과 구문 구조 분석 결과로 얻은 추상 구문 트리(AST, Abstract Syntax Tree)를 다른 프로그래밍언어에서 사용하기 위해서 이 트리를 JSON 형식으로 출력하는 방법을 제공한다.

```
$ stack exec -- whilelang-exe --json ./example/while2.while  
  
$ stack ghci --  
ghci> let srcFile = "./example/while2.while"  
ghci> doJson srcFile
```

파이썬 프로그램에서 이 JSON 형식의 분석 내용을 읽어서 원하는 분석을 진행할 수 있다.

Chapter 2

WHILE 프로그래밍언어

프로그램 분석 대상으로 C나 Java와 같은 프로그래밍언어를 사용하면 그 언어의 복잡성으로 프로그램 분석을 집중해서 설명하기 어려울 수 있다. 그러한 이유로 기초적인 특징만을 고려한 프로그래밍언어를 선택하고 그 언어로 작성한 프로그램들을 대상으로 한다. 이 강의노트에서 WHILE 프로그래밍 언어를 선택했다.

2.1 WHILE 프로그래밍 언어 구문 개요

WHILE 언어는 가장 기초적인 특징들로만으로 구성되어 있다. 문장으로 할당문, 조건문, WHILE 반복문, 복합문 (여러 문장들을 세미콜론으로 분류하여 나열), 생략문(SKIP), 입출력문, assert문이 있다. 문장내에 사용할 수 있는 식의 종류는 상수, 변수, 단항 연산과 이항 연산이 있다. 정수와 부울, 두 가지 종류의 값을 사용하며, 사칙연산과 나머지 연산, 비교 연산 두 종류(<, ==)와 논리곱, 논리합, 논리부정 연산이 있다.

WHILE 프로그램은 변수 타입들을 먼저 선언하고 뒤이어 문장들이 나오도록 구성되어 있다. 함수나 클래스는 제공하지 않는다.

정수를 입력받아 변수 x에 놓고 팩토리얼을 구해 출력하는 WHILE 예제 프로그램을 살펴보자. 변수 x와 z를 정수 타입으로 선언한다. 정수를 입력 받아 변수 x에 놓고, 변수 z는 1로 초기화한다. WHILE 반복문으로, 변수 x를 1씩 감소시키며 변수 z에 누적해서 변수 x를 곱하는 것을 변수 x의 값이 1보다 큰 동안 반복한다.

마지막으로 변수 z의 값을 출력한다.

```

1 int x;
2 int z;
3
4 read(x);
5
6 z = 1;
7 while (x > 1)
8 {
9     z = z * x;
10    x = x - 1
11 }
12
13 write(z)

```

더 많은 WHILE 프로그램 예제는 강의 웹 사이트에서 확인할 수 있다.

2.2 WHILE 프로그램을 자료구조로 표현: 추상구문

트리

이제 WHILE 프로그램을 자료구조로 표현하는 방법을 살펴보자. 일반적으로 프로그램에 분석 방법을 적용하려면 이 프로그램 텍스트를 먼저 추상구문트리(Abstract Syntax Tree, AST)로 표현하는 과정이 선행되어야 한다. 추상구문트리는 프로그램의 구조를 모두 담고 있는 트리 자료 구조이다.

WHILE 프로그램의 추상구문트리를 다음과 같은 하스켈 프로그램의 타입 선언으로 그 명세를 작성하려고 한다.

2번째 줄 data Prog에서 data는 새로운 타입을 정의하는 하스켈 키워드이고 뒤이어 나오는 Prog는 이때 정의하는 타입 이름이다.

WHILE 프로그램은 이 Prog 타입의 값으로 표현하려 한다. 뒤이어 Prog Decl Comms에서 동일한 이름이어서 혼동되지만 이때 Prog는 이 값에 붙이는 태그로 이해할 수 있다. 이 태그가 붙은 값은 Decl 타입의 값과 Comms 타입의 값이 뒤따라야만 비로서 Prog 타입의 값이 완성된다.

4번째 줄 type Decl에서 type은 새로운 타입 이름을 붙이는 하스켈 키워드이고 뒤이어 나오는 Decl는 이때 새로 붙이려는 타입 이름이다. 하스켈에서 리스트 타입은 [-] 대괄호를 사용한다. 그 안에 리스트 원소의 타입을 넣는다. 즉, [Decl]은 Decl 타입의 값을 원소로 하는 리스트의 타입을 뜻한다. 이 [Decl] 타입을

반복해서 나중에 언급해도 되나 짧게 Decl 타입이라 부르겠다는 것이 4번째 줄의 새로운 타입 이름 붙이기 선언이 의도하는 바이다.

이 Decl 타입으로 WHILE 프로그램의 선언들을 표현하려 한다. Decl 타입은 Decl 타입을 원소로 하는 리스트 타입인데, Decl 타입은 7번째 줄에 Type 타입과 VarName 타입의 쌍 타입으로 정의되어 있다.

Type 타입은 WHILE에서 사용하는 정수 타입과 부울 타입을 표현한다. 10번째 줄에 data Type을 선언하였고, 이때 두 가지 타입을 구분해서 표현하도록 각각 TyInt와 TyBool 두 종류의 태그를 도입했다. 두 태그 사이의 바(|)는 Type 타입의 값이 TyInt이거나 또는 TyBool 둘 중 하나를 선택 가능한 것임을 나타낸다.

예를 들어, “int x” 타입 선언을 표현한 하스켈 값은 Decl 타입의

```
[ (TyInt, "x") ]
```

이다.

VarName 타입은 변수 이름을 표현하는 타입으로 14번째 줄에 하스켈 문자열 타입인 String으로 정의되어 있다.

2번째 줄 Prog Decls Comms에서 Comms 타입으로 WHILE 프로그램 변수 선언에 뒤따라 나오는 문장들을 표현하려 한다.

5번째 줄 type Comms 선언도 4번째 줄과 같이 이해할 수 있다. Comm 타입은 17번째 줄에 선언한 WHILE 언어에서 허용하는 문장들을 모아놓은 타입이다. 8 가지 문장 종류를 구분하기 위해 CSkip, CSeq, CAAssign, CRead, CWrite, CIf, CWhile, CAssert 8가지 태그를 도입하였다.

CSkip 태그는 뒤이어 채워야 할 값이 없지만, CAAssign 태그는 VarName 타입의 값과 Expr 타입의 값이 함께 와야 비로서 Comm 타입의 완전한 문장이 된다.

예를 들어, 할당문 “x = 0”을 표현한 하스켈 값은 Comm 타입의

```
CAAssign "x" (ECst (CInt 0))
```

이다. 하스켈의 String 타입의 문자열 “x”는 VarName 타입의 값이기도 하다. 28-29번째 줄을 보면 data Expr 선언으로 WHILE 언어의 식을 선언한 타입이 있는데, 이 중 상수식을 표현하는 태그가 ECst이다. ECst 태그를 식으로 완성하기 위해서 Const 타입의 값이 필요한데, 34-36번째 줄에 data Const 선언으로 이 타입이 선

언되어 있다. 35번째 줄에 CInt 태그와 하스켈 정수 타입 Int로 WHILE 언어의 정수 값을 표현한다.

또 다른 할당문 예를 살펴보자. “ $x = x + 1$ ”은 변수 x 의 값을 1 증가시키는 할당문이다. 이를 표현한 하스켈 값은 Comm 타입의

```
CAssign "x" (EBinOp OpAdd (EVar "x") (ECst (CInt 1)))
```

이다. WHILE 언어의 식을 표현하는 Expr 타입(28-32번째줄)을 보면 이 할당문의 오른쪽 식에서 사용한 이진 연산 덧셈은 EBinOp OpAdd … …의 구조로 표현 한다. 이때 OpAdd는 Op 타입의 가능한 값 중 하나로 38-48번째 줄에 data Op 선언에서 39번째 줄에 있다.

뒤이어 덧셈의 두 피연산자 “ x ”와 “1”에 해당하는 식이 나온다. 변수 x 는 Expr 타입의 EVar 태그를 사용하여 EVar “ x ”로 표현하고, 상수 1은 Expr 타입의 ECst 태그를 사용하여 ECst (CInt 1)로 작성한다. 이때 Const 타입의 정수를 나타내는 CInt 태그와 하스켈 정수 값 1로 표현한다.

지금까지 설명을 종합하면, 변수 x 를 선언하고, 0으로 초기화하고, 1 증가시키는 WHILE 프로그램을 하스켈로 표현하면 다음과 같다.

<pre>int x; x = 0; x = x + 1</pre>	<pre>Prog [(TyInt, "x")] ⇒ [CAassign "x" (ECst (CInt 0)), CAassign "x" (EBinOp OpAdd (EVar "x") (ECst (CInt 1)))]</pre>
------------------------------------	---

```

1   — Program
2   data Prog = Prog Decls Comms
3
4   type Decls = [ Decl ]
5   type Comms = [ Comm ]
6
7   type Decl = (Type, VarName)
8
9   — Type
10  data Type =
11      TyInt
12      | TyBool
13
14  type VarName = String
15
16  — Statement
17  data Comm =
18      CSkip
19      | CSeq Comm Comm
20      | CAssign VarName Expr
21      | CRead VarName
22      | CWrite Expr
23      | CIf Expr Comm Comm
24      | CWhile Expr Comm
25      | CAssert Expr
26
27  — Expression
28  data Expr =
29      ECst Const
30      | EVar VarName
31      | EBinOp Op Expr Expr
32      | EUUnaryOp Op Expr
33
34  data Const =
35      CInt Int
36      | CBool Bool
37
38  data Op =
39      OpAdd
40      | OpSub
41      | OpMul
42      | OpDiv
43      | OpMod
44      | OpLessThan — x < y
45      | OpEqual — x == y
46      | OpAnd
47      | OpOr
48      | OpNot

```

2.3 파이썬 구현

WHILE 프로그램의 추상구문트리를 표현하는 명세로 하스켈 타입을 선언하였다.

이 명세에서 의도하는 바를 파이썬으로 구현해본다.

type T = ... 또는 data T = ... 형태로 선언한 하스켈 타입은 파이썬 클래스로 옮길 수 있다.

예를 들어, data Prog 선언은 Prog라는 이름을 갖는 클래스를 도입하고 그 안에 Prog Decls Comms에서 필요한 Decls를 저장하는 멤버 변수와 Comms를 저장하는

멤버 변수를 두는 형태로 파일 클래스를 구현할 수 있다.

type Decls 선언은 Declsl라는 이름의 파일 클래스를 선언하고 멤버 변수로 Decl 값의 리스트에 해당하는 파일 리스트 값을 저장할 수 있도록 구성하면 된다.

data Comm 선언은 8가지 서로 다른 태그가 있다. 이러한 형태의 명세는 Comm이라는 이름의 기반 클래스를 먼저 선언한다. 멤버 변수는 특별히 필요하지 않다. 각 태그에 대해 클래스를 도입하되 이 클래스는 기반 클래스를 상속 받도록 서브 클래스로 구성한다. 각 태그에 수반해야 할 값들 (예를 들어 CSeq의 경우 두 개의 Comm 타입의 값들이 필요하다)을 서브 클래스의 멤버 변수로 두어 하스켈 명세를 충실히 구현할 수 있다.

Chapter 3

구문 분석

Chapter 4

의미 분석

Chapter 5

정적 분석

Chapter 6

기호 실행

Chapter 7

맺음말