Eötvös Loránd Tudományegyetem
Informatikai Kar
Programozási Nyelvek és
Fordítóprogramok Tanszék

Applying slicing algorithms on large code bases

Tibor Brunner
doktorandusz

Olivér Hechtl
programtervező informatikus MSc

Budapest, 2017

Contents

Chapter 1

Introduction

"Everyone knows that debugging is twice as hard as writing a program in the first place. So if you're as clever as you can be when you write it, how will you ever debug it?" - Brian Kernighan

Nowadays, there are a lot of tools available for debugging. A developer can examine call stacks, variable informations, and so on. There are a lot of times when a bug is when some variable does not behave like the writer of that code part would expect. When the programmer discovers it, he must follow back the path of assignments, and find out where did it get an unexpected value. Program slicing targets this kind of debugging. We can define a program slice as a part of the program which contains a selected statement, and all other statements which influences it, or gets influenced by it. These two types are respectively called backward and forward slicing. Basically this is what many programmers do intuitively, when facing with bugs. In this thesis, I'll introduce a few approaches for computing slices, and describe a working prototype tool application, which can analyze C++ programs, and compute slices of them. I've used the help of the Clang/LLVM compiler infrastructure, which builds the AST of the program and provides an interface to analyse it using C++.

Chapter 2

Slicing, methods and efficiency

2.1 About slicing

For better understanding programs, programmers organize code into structures. They write sub-problems into functions, and organize variables and data to structs. Also, with object oriented design, they put these into classes. These are all good for separating the data, and the procedures on data. But these are not helpful when we need to examine a flow of data in the program. Slicing gets useful in this scenario. This is a program analysis technique introduced by Mark Weiser[1]. In his paper, he wrote: "Program slicing is a decomposition based on data flow and control flow analysis". We define slicing as a subset of the program, which only includes the statements which have transitive control or data dependency regarding the selected statement.

2.2 Types of slicing

There is two different type of slicing known: static and dynamic. While dynamic slicing gets the statements which could affect the selected statement at a particular execution of the program with a fixed input, static slicing examines it statically, including all possible statements which could affect that selected statement. In this thesis, I'll focus on static slicing methods. There are two different subtypes of static slicing, backward and forward. They are indicating the relevant statements' direction from our selected statement.

2.3 Methods for slicing

We can construct slices via various methods on different representations of the program. All of these are using some kind of graph structures, which can be traversed through for searching the transitive data dependences.

2.3.1 Dataflow equations

The first method is created by there is the control flow graph, which contains basic blocks as nodes, which contains consecutive statements without any transfers of control, and edges which are the jumps connecting these blocks.

2.3.2 Information flow relations

2.3.3 PDG based graph reachability

Chapter 3

LLVM/Clang infrastructure

3.1 About Clang

3.2 The Clang AST

3.3 AST Matchers

Chapter 4

Implementation and algorithm

4.1 The approach

4.2 Building the PDG

4.2.1 Control dependences

4.2.2 Data dependences

4.3 Implementing slicing

Bibliography

[1] M. Weiser, Program slicing, IEEE Transactions on Software Engineering, 10(4):352-357, 1984.