UNIVERSITY OF CALGARY

Characterization of Logging Usage:
An Application of Discovering Infrequent Patterns via anti-unification

by

Narges Zirakchianzadeh

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

CALGARY, ALBERTA

August, 2016

UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "Characterization of Logging Usage: An Application of Discovering Infrequent Patterns via anti-unification" submitted by Narges Zirakchianzadeh in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE.

Dr. Robert J. Walker
Supervisor
Department of Computer Science

Dr. Jörg Denzinger
Examiner
Department of Computer Science

Dr. Christian J Jacob
Examiner
Department of Computer Science

Date

Abstract

Acknowledgements

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Abbreviation	Definition
AST	Abstract Syntax Tree
AU	Anti-unification
AUAST	Anti-unification Abstract Syntax Tree
HOAUMT	Higher-order Anti-unification Modulo Theories
LJC	Logged Java Class

Chapter 1

Introduction

Understanding the similarities and differences of a set of source code fragments is a potentially complex problem that has various actual or potential applications in program analysis. For example: detecting code clones [Bulychev and Minea, 2009], automating source code reuse [Cottrell et al., 2008], recommending replacements for APIs between various versions of a software library [Cossette et al., 2014], collating application programming interface (API) usage patterns, and automating the merge operation of various branches in a version control system. As a specific application, the focus of this study is on characterizing where logging is used in the source code by understanding the structural correspondences and differences of a set of source code fragments enclosing logging calls within a software system or from different software systems.

Logging is a conventional programming practice to record an application's state and/or actions during the program's execution [Gupta, 2005], and log system analysis assist developers in diagnosing the presence or absence of a particular event, understanding the state of an application, and following a program's execution flow. The importance of logging has been identified by its various applications in software development and maintenance tasks such as problem diagnosis [Lou et al., 2010], system behavioral understanding [Fu et al., 2013], quick debugging [Gupta, 2005], performance diagnosis [Nagaraj et al., 2012], easy software maintenance [Gupta, 2005], and troubleshooting [Fu et al., 2009].

Developers can perform logging in various ways, as they make different decisions about where and what to log. For example, they can apply logging to record the occurrence of every event of an application and use logging calls at the start and end of the body of every method in the source code [Clarke et al., 1999a,b]. However, three main problems are associated with excessive logging. First, it can generate a lot of redundant information that might be confusing and misleading

for developers to perform system log analysis, masking significant information. Second, excessive logging is costly. It requires extra time and effort to write, debug, and maintain the logging code. Third, it can cause system resource overhead and thus the application performance will be negatively affected. On the other hand, insufficient logging may result in losing some necessary run-time information for software analysis. Therefore, logging should be done in an appropriate manner to be effective.

Despite the importance of logging for software development and maintenance, few studies have been conducted on understanding logging usage in real-world applications since logging has been considered as a trivial task [Clarke et al., 1999a,b]. However, the availability of several complex frameworks (e.g., log4j,SLF4J) that assists developers to log suggests that effective logging is not a straightforward task to perform in practice. In addition, Yuan et al. [2012b] showed that developers spend a great effort to modify logging practices as after-thoughts, which indicates that it is not that simple for developers to perform logging practices efficiently in their first attempt.

In this research, I would like to understand where developers log in practice in a detailed way. The location of logging calls has a great impact on the quality of logging as it helps developers to trace the code execution path to identify the root causes of an error in log system analysis. However, no previous work has focused on characterizing where logging calls are used in real-world applications. In this study, I address this gap by developing an automated approach to detect the detailed structural correspondences and differences in the usage of logging within a system and between systems.

1.1 Programmatic support for logging

A typical logging call takes parameters including a log text message and verbosity level. A log text message consists of static text to describe the logged event and some optional variables related to the event. The verbosity level is intended to classify the severity of a logged event such as a debugging note, a minor issue, or a fatal error. Figure 1.1 provides examples of logging calls from

the log4j framework in descending order of severity. The fatal level designates a very severe error event that will likely lead the application to terminate. The error level indicates that a non-fatal but clearly erroneous situation has occurred. The warn level indicates that the application has encountered a potentially harmful situation. The info level designates important information that might be helpful in detecting root cases of an error or to understand the application behaviour. The debug level designates useful information to debug an application and is usually used by developers only during the development phase. In general, verbosity level is used for classification to avoid the overhead of creating large log files in high performance code.

```
log. fatal (”Fatal  Message %s”, variable);
log. error (”Error  Message %s”, variable);
log.warn(”Warn Message %s”, variable);
log. info (”Info  Message %s”, variable);
log.debug(”Debug Message %s”, variable);
```

Figure 1.1: Logging call examples from the log4j framework.

1.2 Broad thesis overview

In this study, I aim to provide a concise description of where logging is used in the source code by constructing generalizations that represents the detailed structural similarities and differences between entities that make use logging calls. My study investigates the location of logging calls from the point of view of logged Java methods (LJM) which are the Java methods enclosing logging calls in source code. To investigate how to construct generalizations using the syntax and semantics of the Java programming language, I looked to previous research conducted by Cottrell et al. [2008] that determines the detailed structural correspondences between two Java source code fragments through the application of approximated anti-unification such that one fragment can be integrated

to the other one for small scale code reuse. However, my problem context is different as I need to generalize a set of source code fragments with special attention to logging calls. Therefore, my approach must take into account the logging calls when performing the generalization task via the determination of structural correspondences.

[RW: Good.] My approach employs a hierarchical clustering algorithm to create a generalization hierarchy from a set of LJMs using a measure of similarity. It uses an approximated anti-unification algorithm to construct a structural generalization representing the similarities and differences between a pair of LJMs. My anti-unification approach proceeds in three steps. First, it uses the Jigsaw framework [Cottrell et al., 2008] to determine all potential correspondences between the two LJMs using a measure of similarity that relies on structural correspondences along with a simple knowledge of semantic equivalences in the Java language specifications. Second, it develops a greedy selection algorithm to approximate the best anti-unifier to my problem by determining the most similar correspondence for each substructure in my structures, applying some constraints in determining correspondences to prevent the anti-unification of logging calls with anything else. Third, it constructs an anti-unifier through the anti-unification of two structures and develops a measure of structural similarity between them.

[RW: I'm not sure if it is necessary to overview the empirical studies here. You mention them below. To be revisited later.]

1.3 Thesis Statement

The thesis of this work is to determine the detailed structural similarities and differences between entities of the source code that make use of logging calls to provide a concise description of where logging calls do occur in real-world software systems.

1.4 Thesis Organization

The remainder of the thesis is organized as follows. Chapter 2 motivates the problem of understanding where to use logging calls in the source code through a scenario in which a developer attempts to perform a logging task. This scenario outlines the potential problems she may encounter and illustrates that the current logging practice is insufficiently supported.

Chapter **??** provides background information on important concepts and past research that are key to my work. I described abstract syntax trees (ASTs), which are the basic structure I will use for describing software source code. I provide a definition of anti-unification and explain why its basic forms do not adequately address my problem. Then I define higher-order anti-unification modulo theories (HOAUMT), an extension to anti-unification that can be applied on an extended form of AST to solve my problem. Afterwards I discuss the Jigsaw framework, an existing tool that could assist us in the determination of potential structural correspondences by applying HOAUMT on extended structures **[RW: This may need some modification to introduce the fact that you do some studies to validate Jigsaw, etc.]**

Chapter 7 describes my proposed approach as a set of novel algorithms, and its implementation as a plug-in to the Eclipse integrated development environment (IDE).

Chapter 8 presents two empirical studies. The first study is conducted to evaluate the accuracy of my approach and its implemented tool by conducting an experiment on 10 sample logged Java methods. The second study is conducted to characterize the location of logging usage in three open-source software systems.

Chapter 9 discusses the results and findings of my work, threats to its validity, and the remaining issues. Chapter 10 describes related work to my research problem and how it does not adequately address the problem. Chapter 11 concludes the dissertation and presents the contributions of this study and future work. Additional materials appended to the end of this dissertation is provided in Appendix A.

Chapter 2

Motivational Scenario

Printing messages to the console or to a log file is an integral part of a software development that can be used to test, debug, and understand what is going on inside an application. In Java programming language, print statements are commonly used to print something on console. However, the availability of tools, frameworks and APIs for logging that offers more powerful and advanced Java logging features, flexibility, and improvement on logging quality suggests that using print statements is not sufficient for real-world applications.

The logging framework offers many more features, which is not possible using print statements. In most logging frameworks (e.g., log4j, SLF4j, java.util.logging), different verbosity levels of logging are available for use. That is, by logging at a particular log level, messages will get logged for that level and all levels above it and not for the levels below. As an example, debug log level messages can be used in a test environment, while error log level messages can be used in a production environment. This feature not only produces fewer log messages but also improves the performance of an application. In addition, most logging frameworks allow the production of formatted log messages, which makes it easier to monitor the behaviour of a system by a developer. Furthermore, in the case of working on a server side application, the only way to know what is going on inside the server is by monitoring log file. Although logging is a precious practice for software development and maintenance, it imposes extra time and energy on developers to write, test, and run the code, while affecting the application performance. Since latency and speed is major concerns for most software systems, it becomes necessary to understand and learn logging in great detail in order to perform logging in an efficient manner.

To illustrate the challenges that lie in effectively performing logging practices in software systems, consider a scenario in which a developer is asked to log an event-based mechanism of a

text editor tool written in the Java programming language. Consider the developer trying to log a Java class of this system shown in Figure 2.1 using the log4j logging framework. She knows that components of this application register with the EditBus class to receive messages reflecting changes in the application's state, and the EditBus class maintains a list of components that have requested to receive messages. That is, when a message is sent using this class, all registered components receive it in turn. Furthermore, any classes that subscribe to the EditBus and implements the EBComponent interface defines the method EBComponent.handleMessage(EBMessage) to handle a message sent on the EditBus. To perform this logging task several fundamental questions might appears in her mind which are mostly related to where and what to log.

The first solution she can come up with is to simply log at the start and end of every method. However, she believes that logging at the start and end of the addToBus(EBComponent), removeFromBus(EBComponent), and getComponents() methods are useless, producing redundant information. She assumes that the more she logs, the more she performs file I/O which slows down the application. Therefore, she decides to log only important information which is necessary to debug or troubleshoot potential problems if they happen. She proceeds to identify the information needed to be logged and then decides on where to use logging calls. She thinks that it is important to log the information related to a message sent to a registered component, including the message content and the transmission time, to troubleshoot the potential problems that might happen in sending messages. She simply wants to begin with using a logging call at the start of the send() method (line 2 of Figure 2.2) to log the information. However, she realizes that this logging call does not allow her to log the information she wants, as the time variable is not initialized in the beginning of this method; thus, she proceeds to examine the body of the send() method line-by-line and uses another logging call after the time variable is initialized inside an **if** statement that checks the value of the variable time is not invalid (shown in lines 9–11 of Figure 2.3).

She also believes that it is important to log an error if any problems happen in sending messages to the components. She decides to use a **try/catch** statement as it is a common way to handle

```
1  public class EditBus {
2      private static ArrayList components = new ArrayList();
3      private static EBComponent[] copyComponents;
4
5      private EditBus() {
6      }
7
8      public static void addToBus(EBComponent comp) {
9          synchronized(components) {
10             components.add(comp);
11             copyComponents = null;
12         }
13     }
14
15     public static void removeFromBus(EBComponent comp) {
16         synchronized(components) {
17             components.remove(comp);
18             copyComponents = null;
19         }
20     }
21
22     public static EBComponent[] getComponents() {
23         synchronized(components) {
24             if (copyComponents == null) {
25                 EBComponent[] arr = new EBComponent[components.size()];
26                 copyComponents =
27                     (EBComponent[])components.toArray(arr);
28             }
29         }
30         return copyComponents;
31     }
32
33     public static void send(EBMessage message) {
34         EBComponent[] comps = getComponents();
35         for(int i = 0; i < comps.length; i++) {
36             EBComponent comp = comps[i];
37             long start = System.currentTimeMillis();
38             comp.handleMessage(message);
39             long time = (System.currentTimeMillis() - start);
40         }
41     }
42 }
```

Figure 2.1: The EditBus class.

```
1 public static void send(EBMessage message){
2     //logging call
3     EBComponent[] comps = getComponents();
4     for (int i = 0; i < comps.length; i++) {
5         EBComponent comp = comps[i];
6         long start = System.currentTimeMillis();
7         comp.handleMessage(message);
8         long time = (System.currentTimeMillis() - start);
9     }
10 }
```

Figure 2.2: The developer's initial determination of the usage of logging calls for the send(EBMessage) method.

```
1 public static void send(EBMessage message) {
2     // logging call
3     EBComponent[] comps = getComponents();
4     for(int i = 0; i < comps.length; i++) {
5         EBComponent comp = comps[i];
6         long start = System.currentTimeMillis();
7         comp.handleMessage(message);
8         long time = (System.currentTimeMillis() - start);
9         if (time != 0){
10            // logging call
11        }
12    }
13 }
```

Figure 2.3: The developer's second determination of the usage of logging calls for the send(EBMessage) method.

exceptions in the Java programming language. She creates a **try/catch** block to capture the potential failure in sending messages and uses a logging call inside the **catch** block to log the exception (shown in lines 2–16 of Figure 2.4). However, she realizes that using this logging call would not allow her to reach the desired functionality as it does not reveal to which component the problem is related. Thus, she decides to relocate the **try/catch** block inside the **for** statement to log an error in case of a problem in sending messages to any components (shown in lines 5–15 of Figure 2.5).

Figure 2.6 shows the developer's final determination of the usage of logging calls to perform logging task of the EditBus class. By making proper decisions about where to use logging calls, the developer is in good position to proceed to write the logging messages by examining the remaining conceptually complex questions. Which information should I log? How to choose the format of log message? Which information goes to which level of logging? If she had reached this point more easily and quickly, she would have had more time and energy to make decisions about the remaining issues to complete the logging practice in a timely and appropriate manner.

2.1 Summary

This motivational scenario highlights the problems a developer may encounter in performing a logging task. The core problem she faces in this scenario is the difficulty in understanding where to use logging calls that enable her to log the desired information. However, having an understanding of how developers usually log in similar situations might assist her to make informed decisions about where to use logging calls more quickly, and so she could pay more attention to the remaining, conceptually complex issues to complete the logging task.

```
1  public static void send(EBMessage message){
2      try {
3          // logging call
4          EBComponent[] comps = getComponents();
5          for(int i = 0; i < comps.length; i++) {
6              EBComponent comp = comps[i];
7              long start = System.currentTimeMillis();
8              comp.handleMessage(message);
9              long time = (System.currentTimeMillis() - start);
10             if (time != 0){
11                 // logging call
12             }
13         }
14     } catch(Throwable t) {
15         // logging call
16     }
17 }
```

Figure 2.4: The developer's third determination of the usage of logging calls for the send(EBMessage) method.

```
1  public static void send(EBMessage message) {
2      // logging call
3      EBComponent[] comps = getComponents();
4      for (int i = 0; i < comps.length; i++) {
5          try {
6              EBComponent comp = comps[i];
7              long start = System.currentTimeMillis();
8              comp.handleMessage(message);
9              long time = (System.currentTimeMillis() - start);
10             if (time != 0) {
11                 // logging call
12             }
13         } catch(Throwable t) {
14             // logging call
15         }
16     }
17 }
```

Figure 2.5: The developer's fourth determination of the usage of logging calls for the send(EBMessage) method.

```
1  public class EditBus {
2      private static ArrayList components = new ArrayList();
3      private static EBComponent[] copyComponents;
4
5      private EditBus() {
6      }
7
8      public static void addToBus(EBComponent comp) {
9          synchronized(components) {
10             components.add(comp);
11             copyComponents = null;
12         }
13     }
14
15     public static void removeFromBus(EBComponent comp) {
16         synchronized(components) {
17             components.remove(comp);
18             copyComponents = null;
19         }
20     }
21
22     public static EBComponent[] getComponents() {
23         synchronized(components) {
24             if (copyComponents == null) {
25                 EBComponent[] arr = new EBComponent[components.size()];
26                 copyComponents = (EBComponent[])components.toArray(arr);
27             }
28         }
29         return copyComponents;
30     }
31
32     public static void send(EBMessage message) {
33         // logging call
34         EBComponent[] comps = getComponents();
35         for(int i = 0; i < comps.length; i++) {
36             try {
37                 EBComponent comp = comps[i];
38                 long start = System.currentTimeMillis();
39                 comp.handleMessage(message);
40                 long time = (System.currentTimeMillis() - start);
41                 if (time != 0) {
42                     // logging call
43                 }
44             } catch(Throwable t) {
45                 // logging call
46             }
47         }
48     }
49 }
```

Figure 2.6: The developer's final determination of the usage of logging calls for the EditBus class.

Chapter 3

Background: Abstract Syntax Trees and Anti-unification

A programming language is described by the combination of its syntax and semantics. The syntax concerns the legal structures of programs written in the programming language, while the semantics is about the meaning of every construct in that language. Furthermore, the abstract syntactic structure of source code written in a programming language can be represented as an *abstract syntax tree* (AST), in which nodes are occurrences of syntactic structures and edges represent nesting relationships. Since ASTs will be the form in which we represent and analyze source code, we need a means to generalize sets of ASTs in order to understand their commonalities while abstracting away their differences. The theoretical framework of anti-unification is presented as that means.

In this chapter, ASTs are described briefly in Section 3.1, along with their more concrete counterparts, concrete syntax trees. Anti-unification is summarized in Section 3.2, starting with its most basic form, first-order anti-unification, and progressing to the form that we will make use of, higher-order anti-unification modulo equational theories.

3.1 Concrete syntax trees and abstract syntax trees

A concrete syntax tree is a tree (i.e., a kind of graph) $T = (V, E)$ whose vertices V (equivalently, nodes) represent the syntactic structures (equivalently, syntactic elements) of a specific program written in a specific programming language and whose directed edges E represent the nesting relationships amongst those syntactic structures. Non-leaf nodes in a concrete syntax tree (also called a parse tree) represent the grammar productions that were satisfied in parsing the program it represents; leaf nodes represent the concrete lexemes, such as literals and keywords.

We focus on the Java programming language and we make use of the grammar in the language specification [Gosling et al., 2012, Chapter 18] to determine the form of the concrete syntax trees.

```
1 public class HelloWorld {
2     public static void main(String[] args) {
3         System.out.println("Hello world!");
4     }
5 }
```

Figure 3.1: A simple example Java program.

Non-leaf node names are represented by names in "camel-case" written in italics. Consider the trivial program in Figure 3.1; its concrete syntax tree is represented in Figure 3.2.

Beyond the fact that the concrete syntax tree is rather verbose and thus occupies a lot of space even for a trivial example, we can see two key problems with it: (1) there are a multitude of redundant nodes such as *Expression1*, *Expression2*, and *Expression3* that are present solely for purposes of creating an unambiguous grammar; and (2) there are no nodes that express the concepts of "method declaration" and "method invocation" that should be obviously present in the example program.

To address these problems, concrete syntax trees are converted to abstract syntax trees (ASTs). An AST is similar in concept to a concrete syntax tree but it does not generally represent the parsing steps typed to differentiate different kinds of syntactic structure. The node types are chosen to represent syntactical concepts; we use the grammar presented for exposition by Gosling et al. [2012], which differs markedly from the grammar they propose in Chapter 18 for efficient parsing. Note that a given node type constrains the kinds and numbers of child nodes that it possesses. The AST derived from the concrete syntax tree of Figure 3.2 is shown in Figure 3.3. Note that, although we know that (for any normal program), System refers to the class java.lang.System and out is a static field on that class; however, non-normal programs can occur and a pure syntactic analysis cannot rule out that System is a package and that out is a class therein declaring a static method println(String).

This is still verbose, so in practice we elide details that are implied or otherwise trivial, to arrive at a more abstract AST as shown in Figure 3.4.

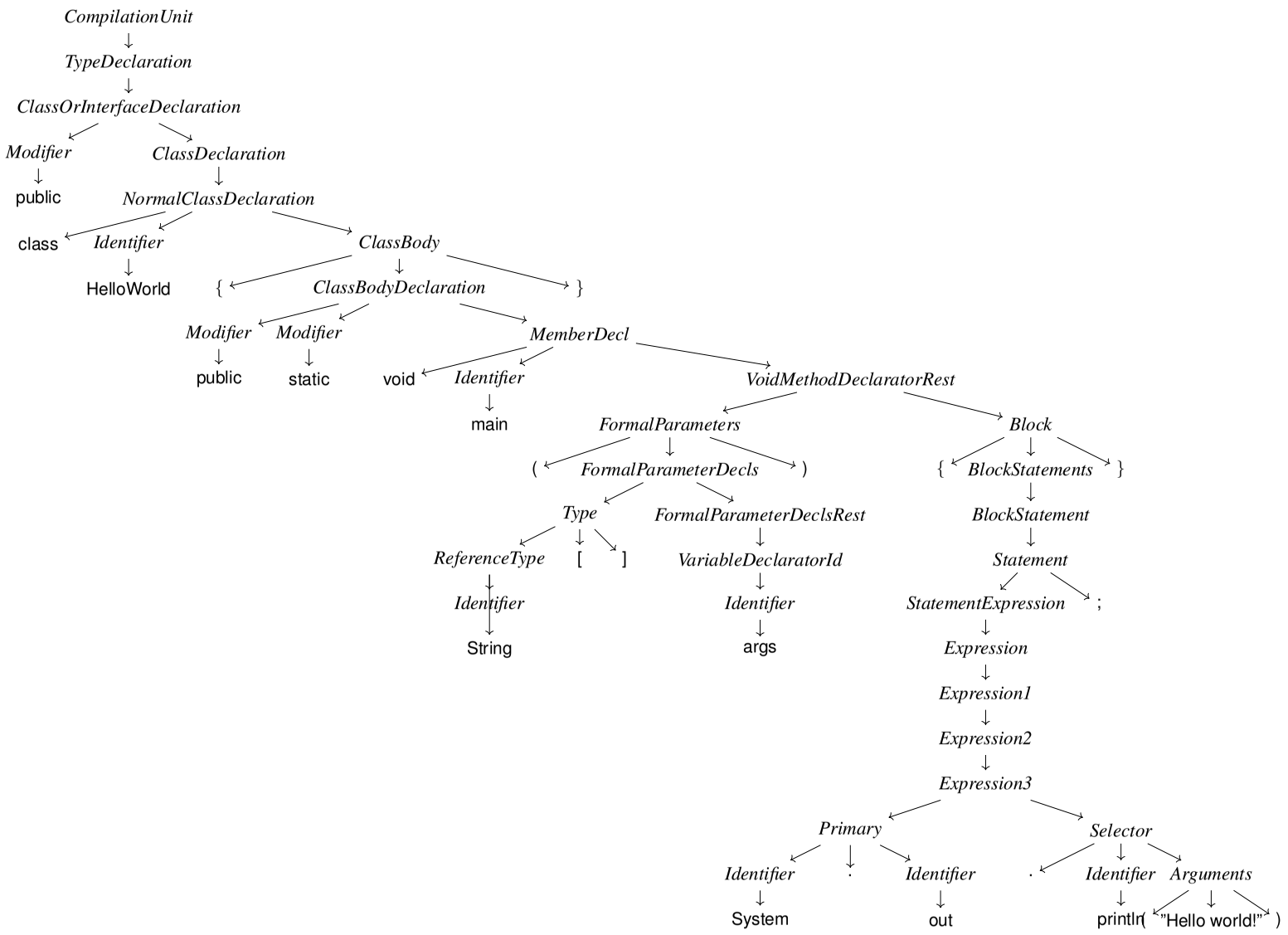


Figure 3.2: The concrete syntax tree for the program of Figure 3.1.

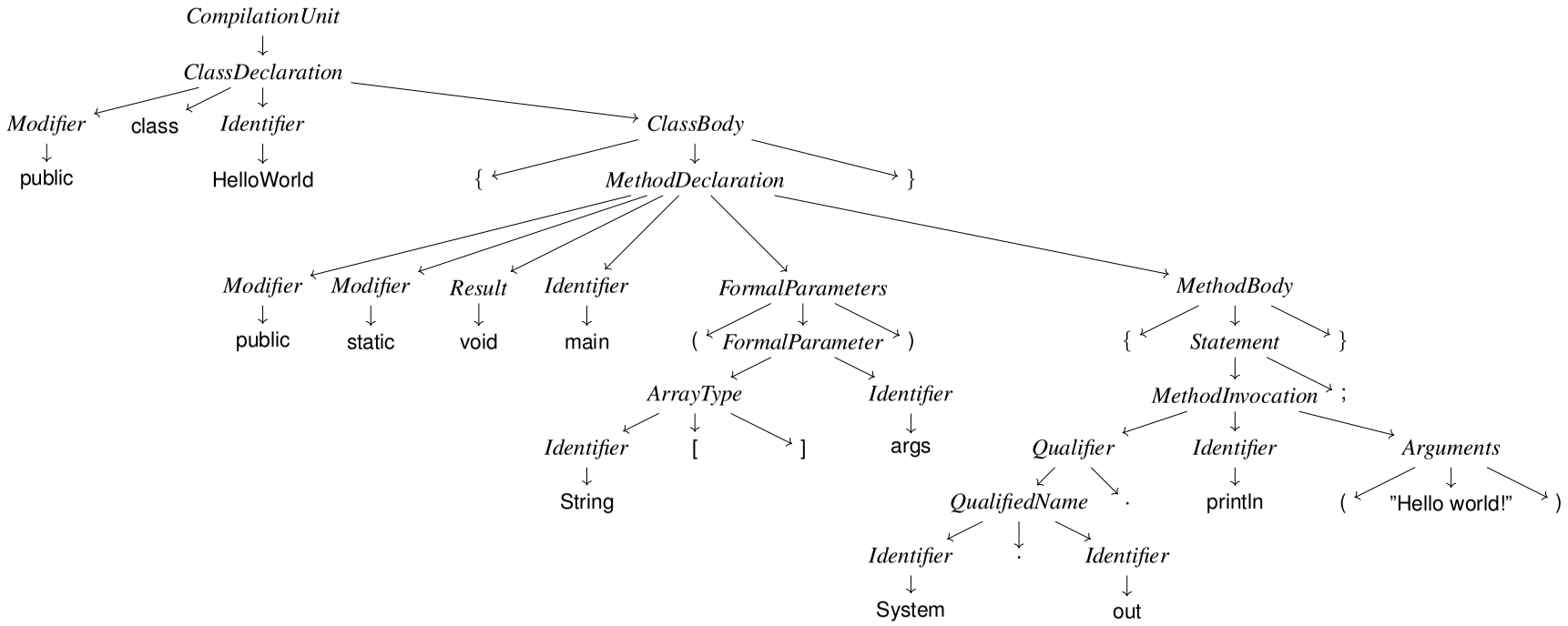


Figure 3.3: The abstract syntax tree derived from the concrete syntax tree of Figure 3.2.

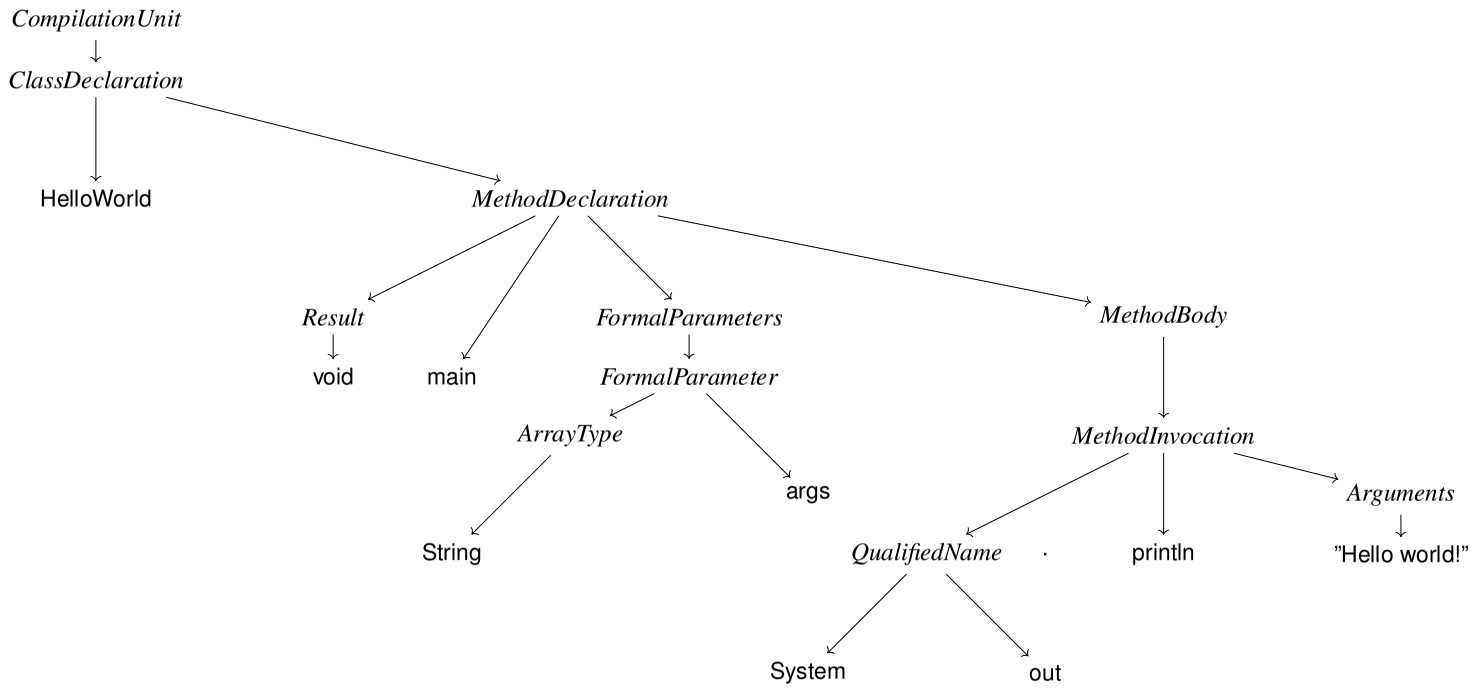


Figure 3.4: A more abstract, abstract syntax tree derived from the concrete syntax tree of Figure 3.2.

3.2 First-order anti-unification

This section defines terms, substitutions, applying a substitution to a term, and instances and anti-instances of a term, as the requirements needed to describe anti-unification theory (and its dual, unification theory).

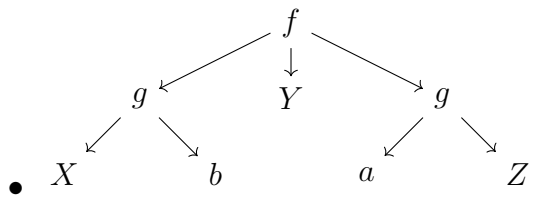
Definition 3.2.1 (Term)**.** A term is defined to be a variable, a constant, or a function symbol followed by a list of terms as the arguments of the function. [Note that function symbols without the subsequent list of terms do not constitute terms.]

Function symbols taking n arguments are called n-ary function symbols; 0-ary function symbols are called constants. The identifiers starting with a lowercase letter are used to represent function symbols (e.g., $f(a, b)$, $g(a, b)$) and constants (e.g., a, b), while variables are represented by identifiers starting with an uppercase letter (e.g., X, Y). The following are examples of a term:

- Y
- a
- $f(X, c)$
- $f(g(X, b), Y, g(a, Z))$

Note that for any term there is a unique, equivalent tree and vice versa: constants and (first-order) variables are leaf nodes, while function symbols are non-leaf nodes; a function with given arguments is represented by a non-leaf node (representing the function symbol) with directed edges pointing to leaf nodes representing each argument. For example:

- Y
- a
- $X \leftarrow f \rightarrow c$



Definition 3.2.2 (Substitution)**.** A substitution is a set of mappings, each from a variable to a term.

Definition 3.2.3 (Applying a substitution)**.** Applying a substitution to a term results in the replacement of all occurrences of each variable in the term, by its corresponding term as defined in the substitution.

As an example, applying the substitution $\Theta = (X \rightarrow a, Y \rightarrow b)$ to the term $f(X, Y)$ results in the replacement of all occurrences of the variable X by the term a and all occurrences of the variable Y by the term b, and thus $f(X, Y) \xrightarrow{\Theta} f(a, b)$.

Definition 3.2.4 (Instance & anti-instance)**.** a is an instance of a term X and X is an anti-instance of a, if there is a substitution Θ such that the application of Θ on X results in a ($X \xrightarrow{\Theta} a$).

Definition 3.2.5 (Unifier)**.** A unifier is a common instance of two given terms.

Unification usually aims to create the *most general unifier* (MGU); that is, U is the MGU of two terms such that for all unifiers U' there exists a substitution Θ such that $U \xrightarrow{\Theta} U'$. Unification aims to make a more concrete structure in essence, whereas what we need is a more generalized structure, which leads to the use of the dual of unification, called *anti-unification*.

Definition 3.2.6 (Anti-unifier)**.** X is an anti-unifier (or generalization) for a and b, if X is an anti-instance for a and an anti-instance for b under substitutions Θ_1 and Θ_2, respectively ($X \xrightarrow{\Theta_1} a$ and $X \xrightarrow{\Theta_2} b$).

An anti-unifier contains common pieces of the original terms, while the differences are abstracted away using variables. An anti-unifier for a pair of terms always exists since we can anti-unify any two terms by the anti-instance X, i.e., a single variable. However, anti-unification usually

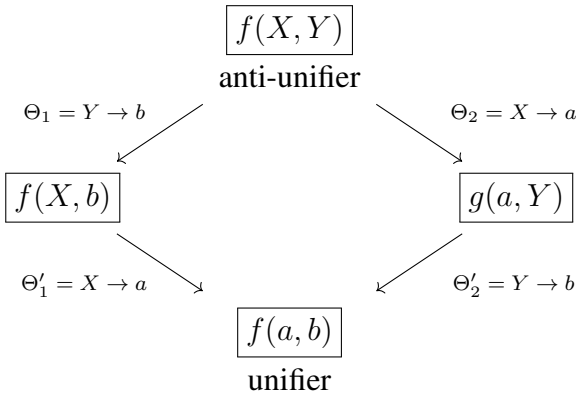


Figure 3.5: Unification and anti-unification of the terms $f(X,b)$ and $f(a,Y)$.

aims to find the *most specific anti-unifier* (MSA), that is, A is the MSA of two structures where there exists no anti-unifier A' such that $A \xrightarrow{\Theta} A'$.

As an example, the anti-unifier of two given terms $f(X,b)$ and $f(a,Y)$ is the new term $f(X,Y)$, containing common pieces of the two original terms. The variable Y in the anti-unifier $f(X,Y)$ can be substituted by the term b to re-create $f(X,b)$ (with $\Theta_1 = Y \to b$) and the variable X in the anti-unifier can be substituted by the term a to re-create $f(a,Y)$ (with $\Theta_2 = X \xrightarrow{\Theta} a$), as depicted in Figure 3.5. In addition, the unifier $f(a,b)$ of the two terms can be instantiated by applying the substitutions $\Theta'_1 = X \xrightarrow{\Theta} a$ and $\Theta'_2 = Y \xrightarrow{\Theta} b$ on the terms $f(X,b)$ and $f(a,Y)$, respectively.

The MSA should preserve as much of common pieces of both original terms as possible; however, first-order anti-unification fails to capture complex commonalities as it restricts substitutions to only replace first-order variables by terms. That is, when two terms differ in function symbols, first-order anti-unification fails to capture common details of them. For example, the first-order anti-unifier of the terms $f(a,b)$ and $g(a,b)$ is X as depicted in Figure 3.6.

Higher-order anti-unification would allow us to create the MSA by extending the set of possible substitutions such that variables can be replaced not only by terms but also by function symbols in order to retain the detailed commonalities. For example, the higher-order anti-unifier of the terms $f(a,b)$ and $g(a,b)$ is $X(a,b)$ as depicted in Figure 3.7.

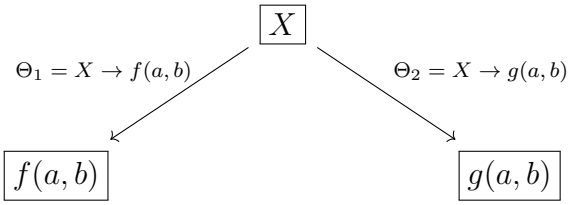


Figure 3.6: First-order anti-unification of the terms $f(a, b)$ and $g(a, b)$.

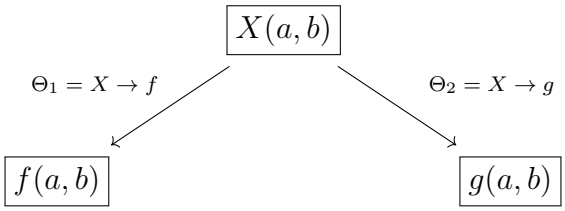


Figure 3.7: Higher-order anti-unification of the terms $f(a, b)$ and $g(a, b)$.

3.3 Higher-Order Anti-unification Modulo Equational Theories

In higher-order anti-unification modulo (equational) theories, a set of equational theories, which treat different structures as equivalent, is defined to incorporate background knowledge. Each equational theory $=_E$ determines which terms are considered equal and a set of these equations can be applied on higher-order extended structures to determine structural equivalences. For example, we have introduced an equivalence equation $=_E$, such that $f(X, Y) =_E f(Y, X)$ to indicate that the ordering of arguments does not matter in our context.

We have also introduced a theory, called NIL-theory, that adds the concept of a NIL structure, which permits a structure to be equated with nothing, and defines an equivalence equation $=_E$ for it. The NIL structure can be used to anti-unify two structures when a substructure exists in one but is missing from the other. However, some requirements should be taken to avoid the overuse of NIL structures such that the original structures must have common substructures but vary in the size for dissimilar substructures. For example, we can anti-unify the two structures b and $f(a, b)$ through the application of NIL-theory by creating the term NIL(NIL, b) which is $=_E$ to $f(b)$ and anti-unifying NIL(NIL, b) with $f(a, b)$ as depicted in Figure 3.8.

We have also defined a set of equivalence equations to incorporate semantic knowledge of

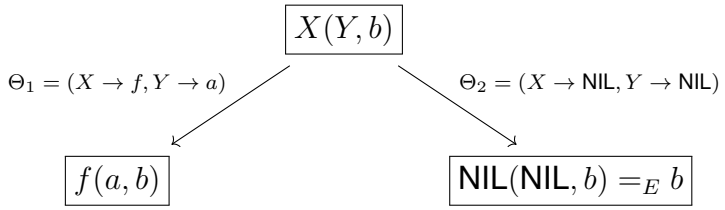


Figure 3.8: Anti-unification of the terms $f(a, b)$ and $\mathsf{NIL}(\mathsf{NIL}, b)$.

structural equivalences supported by the Java language specification as it provides various ways to define the same language specifications. These theories should be applied on higher-order extended structures to anti-unify AST structures that are not identical but are semantically equivalent. For example, consider **for**- and **while**-statements that are two types of looping structure in Java programming language: they have different syntax but semantically cover the same concept. Let us look at the code snippets **for**(i=0;i<10;i++) and **while**(i<10), whose ASTs can be represented as **for**(*Initializer*(i, 0); *LessThanExpression*(i, 10); *Updaters*(*PostIncrementExpression*(i))) and **while**(*LessThanExpression*(i, 10)), respectively. We could define an equivalence equation $=_E$ that allows the anti-unification of **for**- and **while**-statements. We also need to utilize the NIL-theory to handle the varying number of arguments as the **for**-loop has three arguments whereas the **while**-loop only has one. Using the NIL-theory we can create the structure **while**(NIL(NIL, NIL, NIL), *LessThanExpression*(i, 10), NIL(NIL, NIL)) that is $=_E$ to **while**(*LessThanExpression*(i, 10)) and construct the anti-unifier, $V_0(V_1(V_2, V_3)$, *LessThanExpression*(i, 10), $V_4(V_5(V_2)))$, as depicted in Figure 3.9.

However, defining complex substitutions in higher-order anti-unification modulo theories results in losing the uniqueness of MSA. For example, consider the terms $f(g(a, e))$ and $f(g(a, b), g(d, e))$. As described in Figure 3.10, two MSAs exist for these terms: we can anti-unify $g(a, e)$ and $g(a, b)$ to create the anti-unifier $g(a, X_0)$ and anti-unify $g(d, e)$ with the NIL structure to create the anti-unifier $Y(Z, X_1)$; or we can anti-unify $g(a, e)$ and $g(d, e)$ to create the anti-unifier $g(X_0, e)$ and anti-unify $g(a, b)$ with the NIL structure to create the anti-unifier $Y(Z, X_1)$.

Despite having multiple potential MSAs, we need to determine one single MSA that is the most

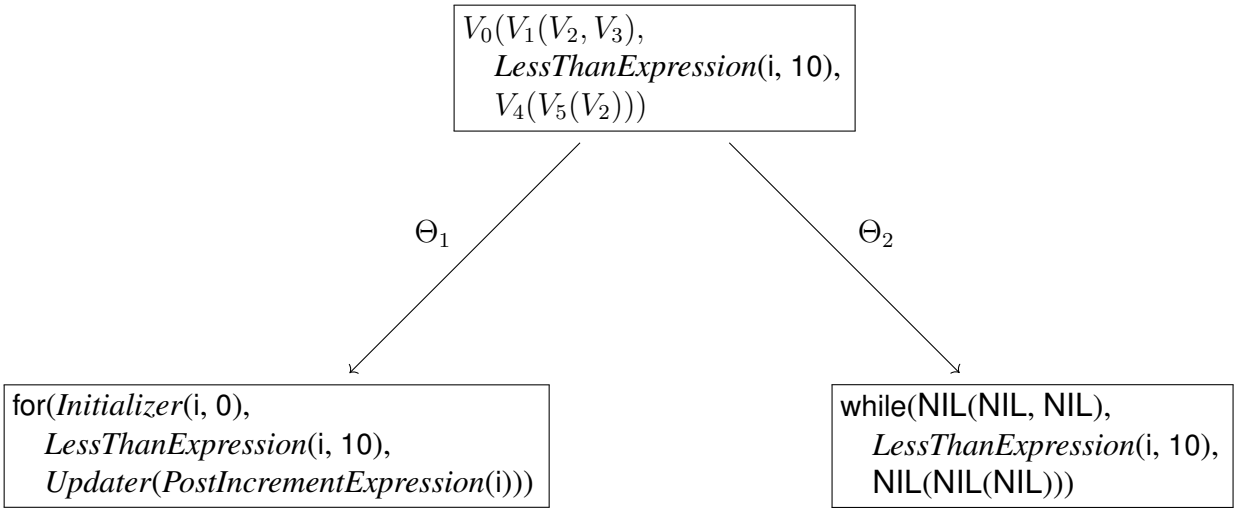


Figure 3.9: The anti-unification of the structures **for**(*Initializer*(i, 0), *LessThanExpression*(i, 10), *Updater*(*PostIncrementExpression*(i))) and **while**(NIL(NIL, NIL, NIL), *LessThanExpression*(i, 10), NIL(NIL, NIL)). The substitutions are defined as follows: $\Theta_1 = (V_0 \rightarrow$ for$, V_1 \rightarrow$ *Initializer*$, V_2 \rightarrow$ i$, V_3 \rightarrow 0, V_4 \rightarrow$ *Updater*$, V_5 \rightarrow$ *PostIncrementExpression*$)$; and $\Theta_2 = (V_0 \rightarrow$ while$, V_1 \rightarrow$ NIL$, V_2 \rightarrow$ NIL$, V_3 \rightarrow$ NIL$, V_4 \rightarrow$ NIL$, V_5 \rightarrow$ NIL$)$

appropriate in our context. However, the complexity of finding an optimal MSA is undecidable in general [Cottrell et al., 2008] since an infinite number of possible substitutions can be applied to every variable. Therefore, we need to use an approximation technique to construct one of the best MSAs that can sufficiently solve our problem.

3.4 Summary

We described abstract syntax trees (ASTs) as a standard syntactic representation of source code. Every AST can also be represented in a function format (and vice versa) which constitute the standard theoretical concept of terms. We demonstrated how the theoretical framework of anti-unification as a technique to construct a common generalization of two given terms, and hence of two ASTs. First-order anti-unification permits terms to be replaced with variables and vice versa, but it is limited in that low-level commonality can be discarded due to high-level differences. Higher-order anti-unification overcomes this by permit substitution relative to function symbols as well as terms. A further extension allows for insertion and deletion by declaring equivalence with

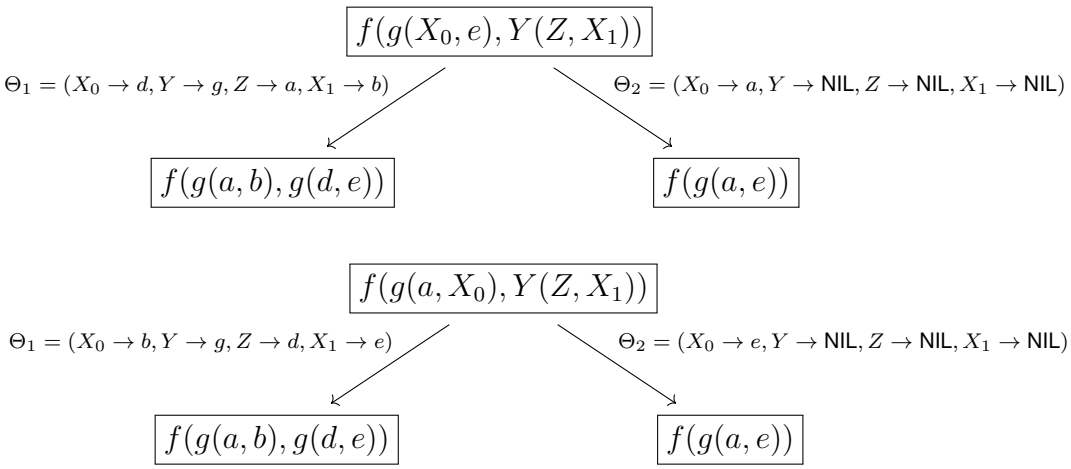


Figure 3.10: Higher-order anti-unification modulo theories of the terms $f(g(a, b), g(d, e))$ and $f(g(a, e))$, creating multiple MSAs.

the NIL structure, as well as other arbitrary equational theories to embed knowledge of semantic equivalence. Unfortunately, this higher-order anti-unification modulo theories approach leads to ambiguity and the potential for an infinite number of possible substitutions for every structural variable. To make use of that technique despite its weakness, we must apply an approximation technique to select amongst the best MSAs in order to reach a solution that is reasonable in practice.

Chapter 4

Background: Eclipse JDT and Jigsaw

In order to construct structural generalizations describing the commonalities and differences between logged Java methods (LJMs), we need a concrete framework for constructing and manipulating abstract syntax trees (ASTs). The Eclipse integrated development environment provides such a framework in its Java Development Tools (JDT) component. The details of our implementation will depend on certain details of Eclipse JDT, so we describe those in Section 4.1.

A framework exists for determining structural correspondences between ASTs, called Jigsaw [Cottrell et al., 2008]. We build atop that work in order to create our anti-unifiers. We describe Jigsaw in Section 4.2.

4.1 Eclipse JDT

The Eclipse Java Development Tools (JDT) framework provides APIs to access and manipulate Java source code via ASTs. An AST represents Java source code in a tree form, where the typed nodes represent instances of certain syntactic structures from the Java programming language. Each node type (in general) takes a set of child nodes, also typed and with certain constraints on their properties. Groups of children are named on the basis of the conceptual purpose of those groups; optional groups can be empty, which we can represent with the NIL element. Thus, any Java source code can be represented as a tree of AST nodes. For example, the simple AST structure of two sample logged Java classes in Figures 4.1 an 4.2 is shown in Figure 4.3

In the JDT framework, structural properties of each AST node can be used to obtain specific information of the Java element that it represents. These properties are stored in a map data structure that associates each property to its value; this data is divided into three types:

- *Simple structural properties:* These contain a simple value which has a primitive or

```
1 public abstract class EBPlugin extends EditPlugin implements EBComponent {
2     private Boolean seenWarning;
3
4     protected EBPlugin() {
5     }
6
7     public void handleMessage(EBMessage message) {
8         if (seenWarning) return;
9         seenWarning = true;
10        Log.log(Log.WARNING, this, getClassName() + " should extend EditPlugin not
              EBPlugin since it has an empty " + handleMessage());
11    }
12 }
```

Figure 4.1: A Java class that uses a logging call. This will be referred to as Example 1.

```
1 public static class Wrapper implements ActionListener {
2     private ActionContext context;
3     private String actionName;
4
5     public Wrapper(ActionContext context, String actionName) {
6         this.context = context;
7         this.actionName = actionName;
8     }
9
10    public void actionPerformed(ActionEvent evt) {
11        EditAction action = context.getAction(actionName);
12        if (action == null) {
13            Log.log(Log.ERROR, this, "Unknown action: " + actionName);
14        }
15        else
16            context.invokeAction(evt, action);
17    }
18 }
```

Figure 4.2: A Java class that uses a logging call. This will be referred to as Example 2.

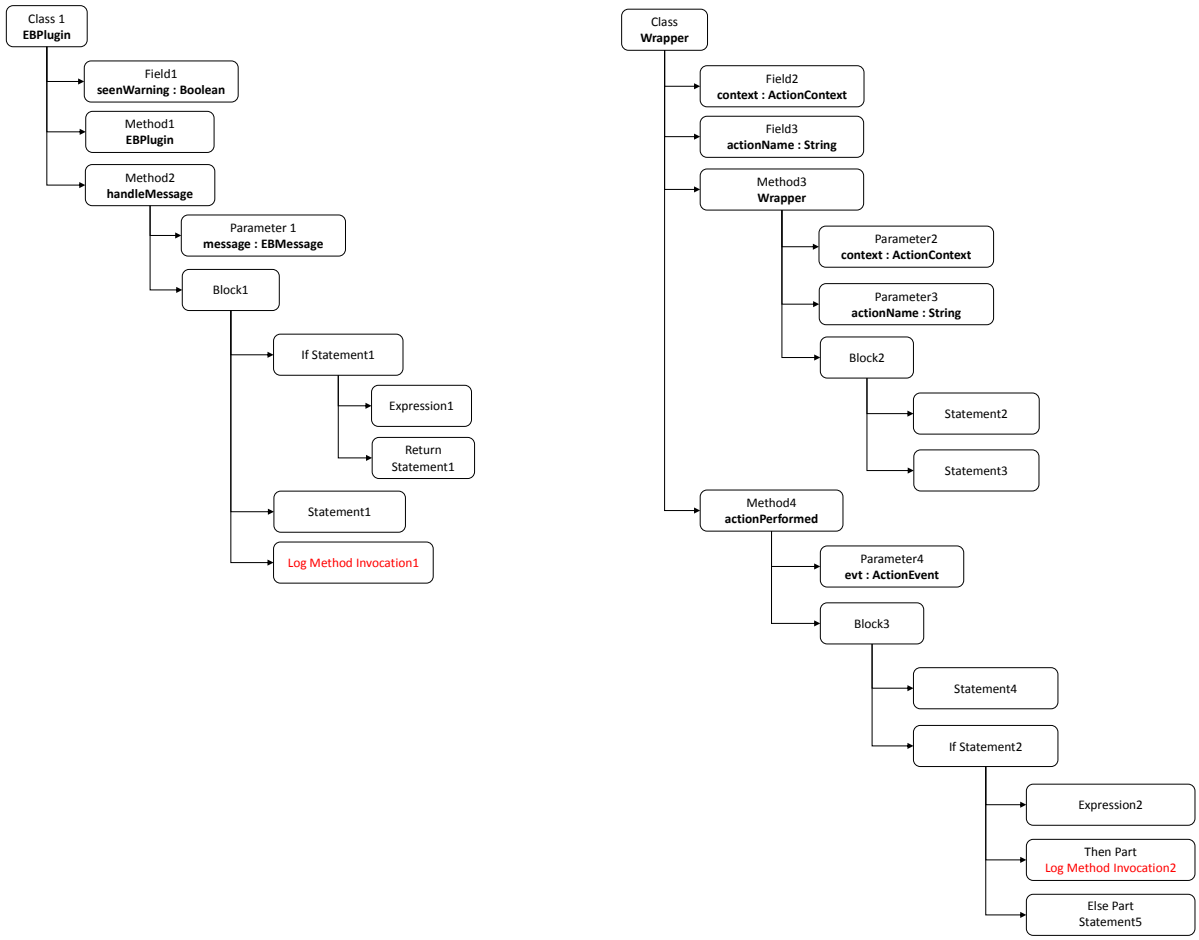


Figure 4.3: Simple AST structure of examples in Figures 4.1 and 4.2.

simple type or a basic AST constant (e.g., identifier property of a name node whose value is a String). For example, all the *Identifier* nodes in Figure 3.3 fall in this case; each references an instance of String representing the string that constitutes the identifier.

- *Child structural properties:* These involve situations where the value is a single AST node (e.g., name property of a method declaration node). For example, the *ClassDeclaration* node in Figure 3.3 has a single child that represents its name as an *Identifier* node; this would be a child structural property.

- *Child list structural properties*: These involve situations where the value is a list of child nodes. For example, the *ClassDeclaration* node in Figure 3.3 can possess multiple *Modifier*s; these are recorded in the *ClassDeclaration* as a child list structural property.

As an example, the ASTs of the logging calls at line 10 of Figure 4.1 and line 13 of Figure 4.2 can be represented respectively as:

- *MethodCall*(
  QualifiedName(Log, *Identifier*(log)),
  Arguments(
    QualifiedName(Log, *Identifier*(WARNING)),
    ThisExpression(),
    AdditionExpression(
      MethodInvocation(*Identifier*(getClassName), *Arguments*()),
      StringLiteral(" should extend EditPlugin not EBPlugin since it has an empty "),
      MethodInvocation(*Identifier*(handleMessage), *Arguments*())))))

- *MethodInvocation*(
  QualifiedName(Log, *Identifier*(log)),
  Arguments(

QualifiedName(Log, *Identifier*(ERROR)),

ThisExpression(),

AdditionExpression(

StringLiteral("Unknown action: "),

Identifier(actionName))))

4.2 The Jigsaw framework

The Jigsaw tool was developed by Cottrell et al. [2008] to determine the structural correspondences between two Java source code fragments through the application of higher-order anti-unification modulo equational theories such that one fragment can be integrated to the other one for small-scale code reuse. Jigsaw could help determine potential candidate structural correspondences between AST nodes of logged Java classes by producing an augmented form of AST, called a *correspondence AST* (CAST), where each node holds a list of candidate correspondence connections between the two structures, each implicitly representing an anti-unifier. Jigsaw also provides a measure of structural similarity to indicate how similar the nodes involved in each correspondence connection are. The Jigsaw similarity function relies on structural correspondence along with simple knowledge of semantic equivalences supported by the Java language specification. It returns a value in $[0, 1]$ where zero indicates complete lack of similarity and one indicates perfect similarity. In addition, several semantical heuristics are used to improve the accuracy of similarity measurement by allowing the comparison of AST nodes that are not syntactically identical but are semantically related to each other.

For example, the similarity between names of AST nodes is measured using a normalized computation based on the length of the longest common substring. The comparison of **int** and **long** types is another example, where an arbitrary value of 0.5 is defined as the similarity value as they are not syntactically identical but are semantically related. In addition, the Jigsaw framework

also detects the structural correspondence between **for**-, enhanced-**for**-, **while**-, and **do**-loop statements; and **if** and **switch** conditional statements. As an example, Figure 6.2 shows the structural correspondence connections created by Jigsaw between the AST nodes of Examples 1 and 2 along with the similarity value for each correspondence connection.

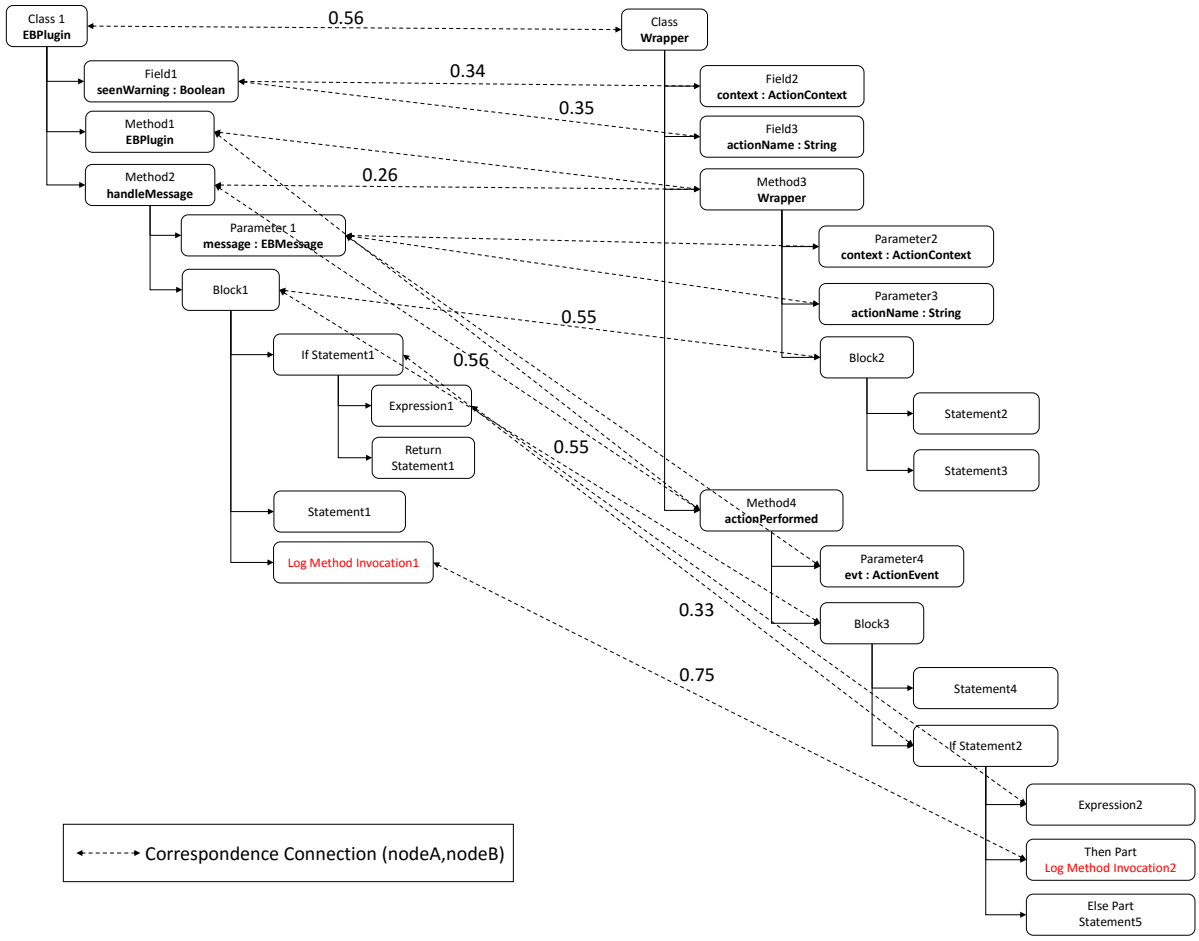


Figure 4.4: Simple CAST structure of examples in Figures 4.1 and 4.2. The links between AST nodes indicate structural correspondence connections created by the Jigsaw framework along with the similarity value.

However, the Jigsaw tool does not suffice to construct an anti-unifier that is the best fit to our application. In addition, the Jigsaw similarity function does not measure the similarity of two logged Java classes with a focus on logging calls, which is needed in our context. To address these

issues, we should develop a greedy selection algorithm to approximate the best anti-unifier by determining the best correspondence for each node. In the following chapter, we will discuss our approach to construct structural generalizations and our implementation by means of the higher-order anti-unification modulo theories and the Jigsaw framework.

4.3 Summary

[RW: Redo]

Chapter 5

An Assessment of Jigsaw

[RW: Describe here the procedure you used to select the examples, etc., how you tested Jigsaw, and what your findings were. At some point, you complained that Cottrell had not done something right ... do you have any evidence to demonstrate it? How does this affect your work? Such points can go in a discussion section towards the end of this chapter if they don't fit otherwise. Full details of examples can go in an appendix; here, just describe enough so people can get the point.]

Chapter 6

Anti-unification ASTs

[RW: New intro blurb needed. Basically, you need to point out how Jigsaw does not suffice, which hopefully you will have demonstrated in the previous chapter. It's not clear to me why this is separate from the next chapter. Perhaps combine them?]

Section 6.1 describes the development of an anti-unification algorithm for our application. **[RW: Describe testing/evaluation.]**

6.1 Constructing the AUAST

[RW: This is a concrete implementation, not a generic idea, at least not the way it is described. I strongly suggest that you give a generic description of your assumptions about ASTs then relate AUASTs to those, then talk about implementation details.] An anti-unifier AST (AUAST) is an extended form of AST that allows the insertion of variables in place of any node in the tree structure, including both subtrees and leaves, to indicate variations between original structures. The AUAST addresses the limitations of AST to construct an anti-unifier by adding the following structural properties:

- *Simple Variable Property*: an extension of simple property referring to two simple values to allow the insertion of variables in place of leaves.

- *Child Variable Property*: an extension of child property referring to two child AST nodes to allow the insertion of variables in place of subtrees.

The anti-unification of AUASTs of logging calls in Figures 4.1 and 4.2 is depicted in Figure 6.1. The structural variables X and Y are used to abstract away the structural variations. The substitutions are defined in Equations 6.1 and 6.2.

$$\Theta_1 = (X \rightarrow \text{WARNING}, Y \rightarrow \textit{additionExpression}(\textit{methodCall}(\textit{simpleName}(\text{getClassName}), \textit{arguments}()), \textit{stringLiteral}(\text{"should extend ..."}), \textit{methodCall}(\textit{simpleName}(\text{handleMessage}), \textit{arguments}())))) \quad (6.1)$$

$$\Theta_2 = (X \rightarrow \text{ERROR}, Y \rightarrow \textit{additionExpression}(\textit{stringLiteral}(\text{"Unknown action: "}), \textit{simpleName}(\text{actionName})))) \quad (6.2)$$

Applying higher-order anti-unification on AUAST structures could help to construct a structural generalization by maintaining the common pieces and abstracting the differences away using variables. However, it is not comprehensive enough to solve our problem as it does not consider background knowledge about AST structures, such as syntactically different but semantically relevant structures, missing structures, and different ordering of arguments. In the following section, we will look at an extension of anti-unification, higher-order anti-unification modulo theories, and how it can sufficiently address the limitations of anti-unification in our context.

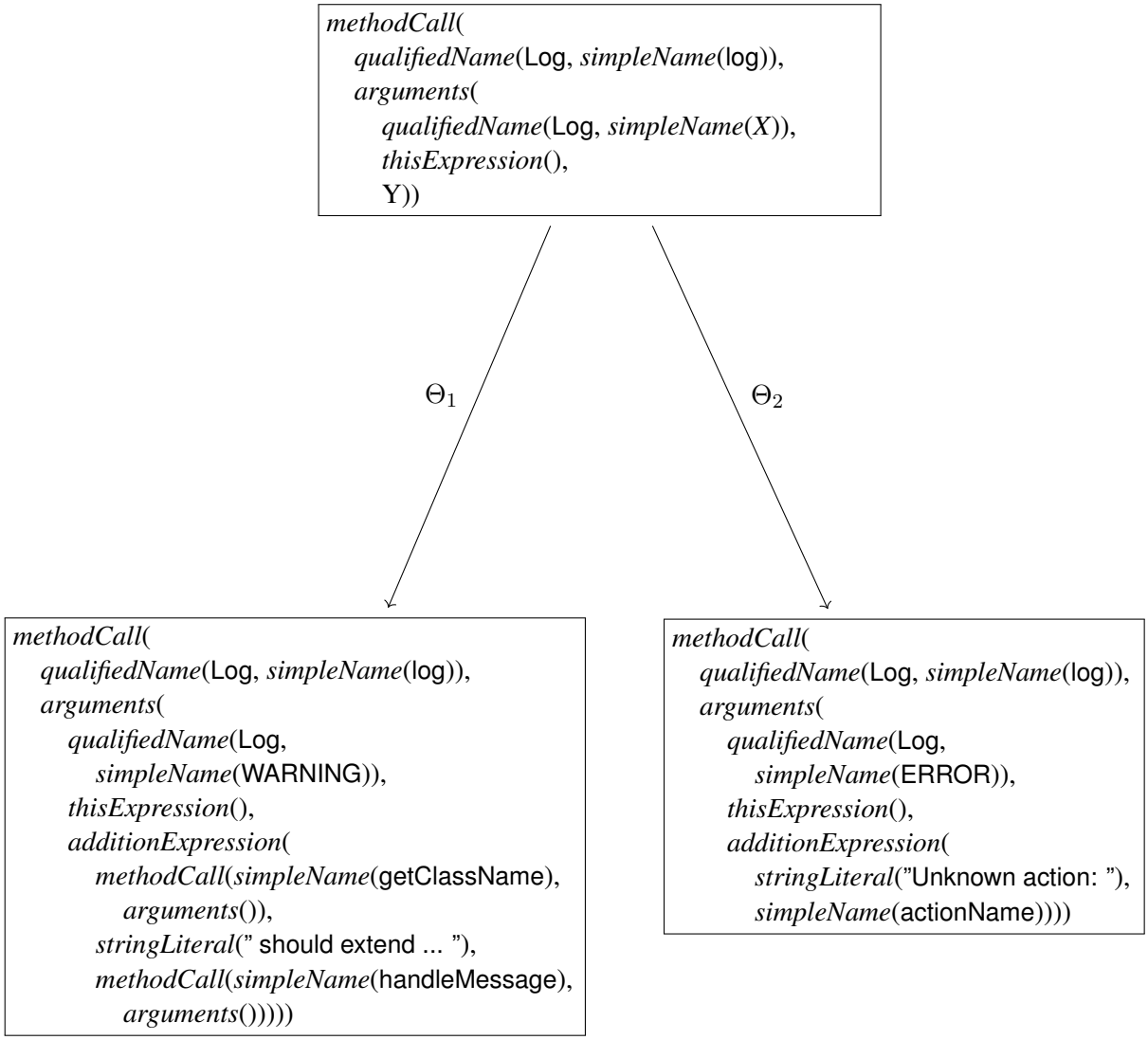


Figure 6.1: Anti-unification of the AUASTs of the logging calls in Examples 1 and 2.

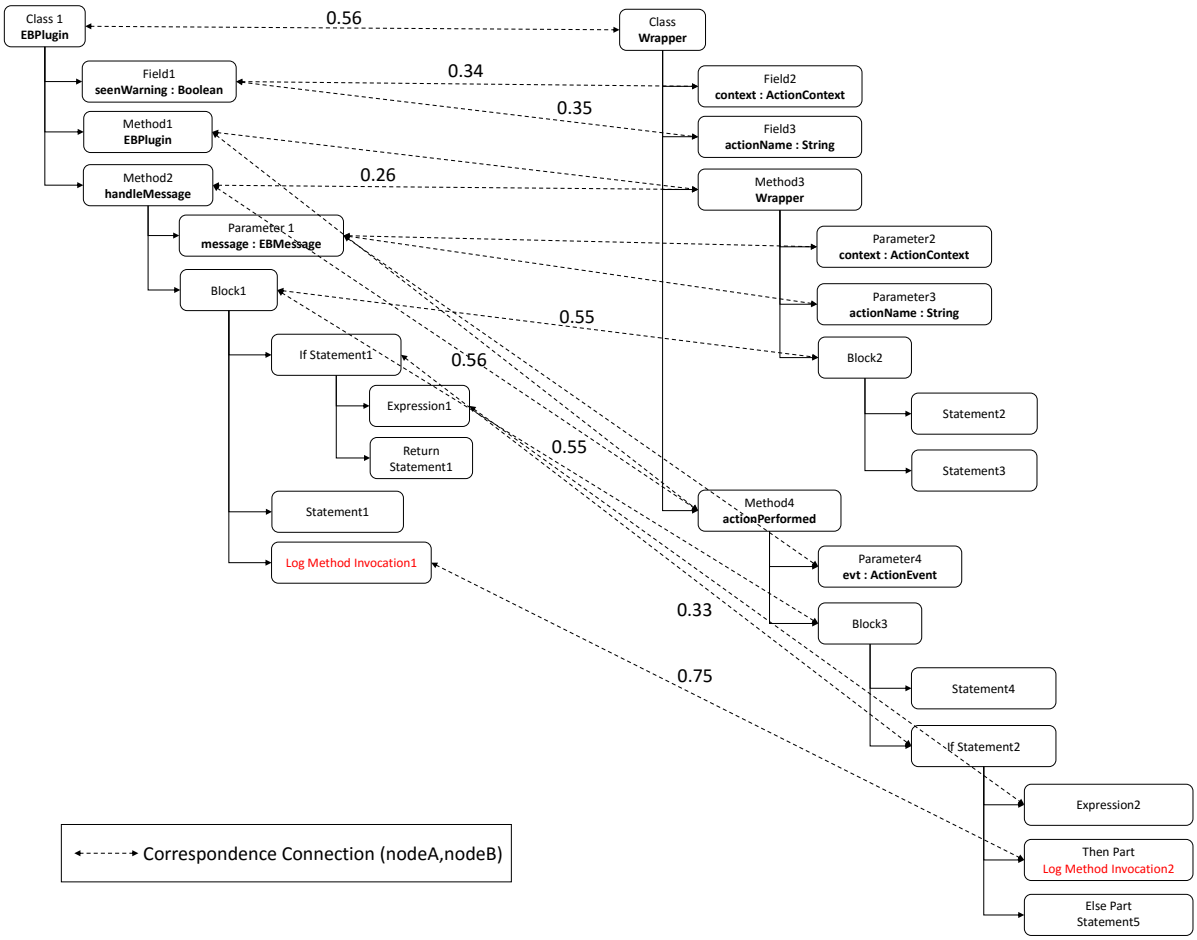


Figure 6.2: Simple CAST structure of examples in Figures 4.1 and 4.2. The links between AST nodes indicate structural correspondence connections created by the Jigsaw framework along with the similarity value.

However, the Jigsaw tool does not suffice to construct an anti-unifier that is the best fit to our application. In addition, the Jigsaw similarity function does not measure the similarity of two logged Java classes with a focus on logging calls, which is needed in our context. To address these issues, we should develop a greedy selection algorithm to approximate the best anti-unifier by determining the best correspondence for each node. In the following chapter, we will discuss our approach to construct structural generalizations and our implementation by means of the higher-order anti-unification modulo theories and the Jigsaw framework.

6.2 Evaluation

[RW: TBD]

6.3 Summary

In this chapter, we described anti-unification as a technique to construct a common generalization of two given terms. We have also introduced an extended form of anti-unification, which is called higher-order anti-unification modulo theories, where a set of equivalence equations can be applied on higher-order extended structures to incorporate background knowledge. In addition, we provided a brief description of AST that maps Java source code in a tree structure form, and why an extended form of it, named AUAST, is required to create higher-order structures specific to our problem context. Finally, we discuss the Jigsaw framework and how it could assist us in determining the potential structural correspondences.

Chapter 7

Anti-unification of Logged Java Classes

[RW: Lots of this stuff is now moved to earlier chapters. This chapter needs to be reworked as a result. It's not clear to me that "chapter3-2.tex" needs to be separate from this.] In Chapter **??** we provided background information on higher-order anti-unification modulo theories—a theoretical framework for constructing a generalization from two given structures—and we described how the Jigsaw tool applies this framework on AST structures of two given Java classes to determine potential structural correspondences between them. We now consider how these frameworks could help us (1) to generalize ASTs of two Java classes containing logging calls and (2) to develop a similarity measure with a focus on logging calls that can provide us with useful information for clustering LJCs in a later phase.

Recall the general point of this study: we aim to provide a concise description of where logging calls happen in the source code through constructing structural generalizations that represent the detailed structural similarities and differences of LJCs. To this end, we should develop an algorithm that:

- classifies LJCs into groups using a measure of similarity such that entities in each group has maximum similarity with each other and minimum similarity to other ones
- abstracts structural correspondences of LJCs of each group into a structural generalization representing the similarities and differences

To construct structural generalizations from a set of LJCs, we developed a prototype tool that applies the Jigsaw framework to find candidate correspondences between two ASTs, the HOAUMT to generalize the structures, and a modified version of the agglomerative hierarchical clustering algorithm to classify a set of LJCs using a measure of similarity. As explained in Section 6.1,

the AST structure should be extended to AUAST structure that allows the insertion of variables in place of any node, which is required for HOAUMT.

Our hierarchical clustering algorithm is a bottom-up approach that starts with singleton clusters, where each contains one AUAST. In every iteration, it merges the closest clusters which are the clusters with maximum similarity between their AUASTs. Therefore, we need to develop a measure of similarity between each pair of AUAST and then construct an anti-unifier when it is needed to merge two clusters.

The structural similarity between two given AUASTs is defined as the number of identical simple property values over total number of simple property values of the anti-unifier (see Section 7.5). To do so, we determine the best correspondences for each node and compute the structural matches between them. Our tool performs a sequence of 3 actions to determine the best correspondences between two AUASTs, outlined by the algorithm DETERMINE-BEST-CORRESPONDENCES: (1) it generates all possible candidate correspondence connections between ASTs of two AUASTs using the Jigsaw framework (line 1) (see Section 7.2); (2) it applies some constraints to prevent the anti-unification of logging calls with anything else (line 2) (see Section 7.3); (3) it determines the best correspondence for each node of AUASTs with the highest similarity and then removes the other correspondence connections involving those nodes (line 3) (see Section 7.4);

To construct an approximation of the best anti-unifier to our problem with a special attention to logging calls, a further step should be taken, which is anti-unification of each AUAST node with its best correspondence determined in the previous step through anti-unifying their structural properties (see Section 7.6). Figure 7.1 shows an overview of the general process of our anti-unification technique, as will be described in the following sections.

Algorithm 7.1 DETERMINE-BEST-CORRESPONDENCES($auastA$,$auastB$) determines best correspondences between the two AUAST nodes $auastA$ and $auastB$

DETERMINE-BEST-CORRESPONDENCE($auastA$**,** $auastB$**)**
1: JIGSAW-CORRESPONDENCE($auastA$, $auastB$)
2: APPLY-CONSTRAINS($auastA$, $auastB$)
3: DETERMINE-CORRESPONDENCES($auastA$)

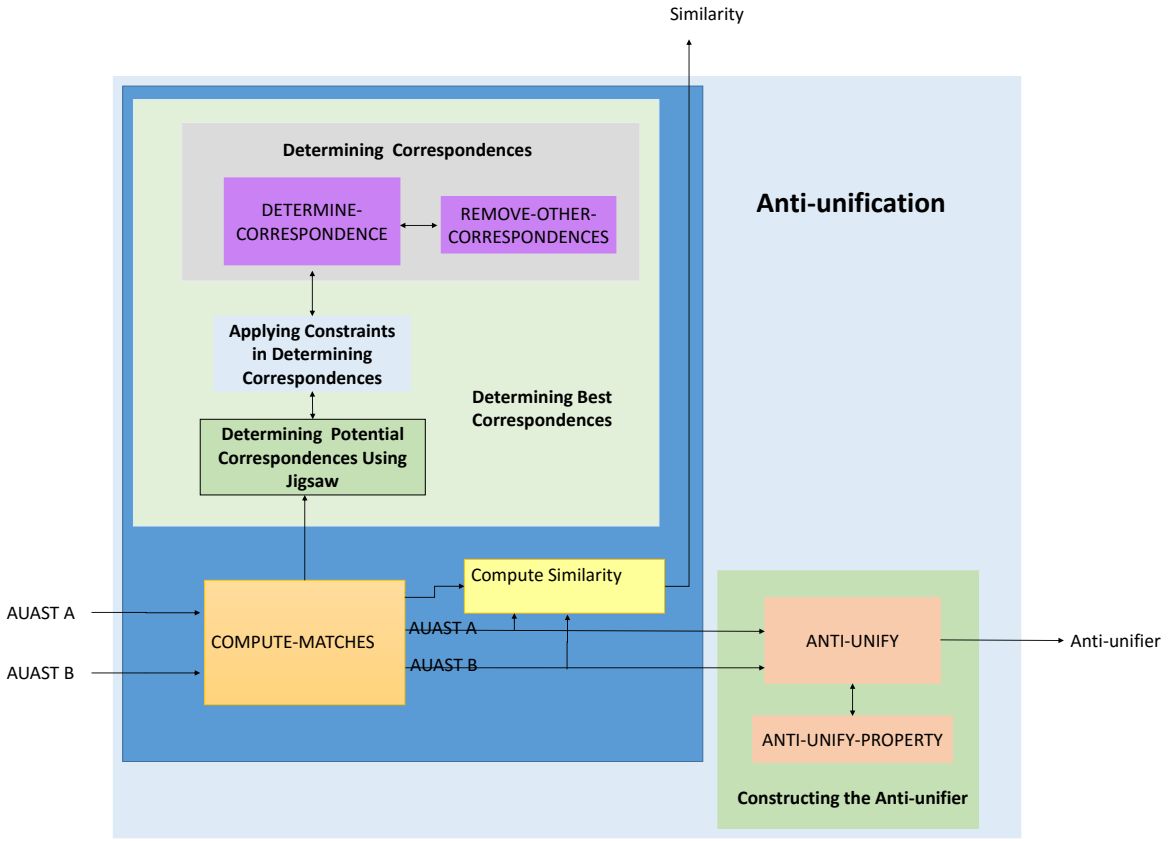


Figure 7.1: Overview of the system.

7.1 Constructing the AUAST

The goal of this phase is to construct an extension of the AST structure that would allow the creation of an antiunified structure. As described in Section **??**, an antiunified structure utilizes variables that must be substituted with a proper substructure to gain back to each original structure; however, the AST structure does not contain any variables thus an extended form of it is required, named AUAST, to address these limitation by allowing the insertion of variables in place of any node in the tree structure, including both subtrees and leaves, to indicate variations between original structures. the AUAST structure addresses the limitations of AST to construct an anti-unifier

by adding the following structural properties:

- **Simple Variable Property**: an extension of simple property referring to two simple property values to allow the insertion of variables in place of leaves.

- **Child Variable Property**: an extension of child property referring to two child nodes to allow the insertion of variables in place of subtrees.

We provide an example to demonstrate the AUAST structure, which is limited to log method invocation subtrees of the sample Java classes shown in Figure 7.2. The log method invocation nodes both contains `EXPRESSION`, `ARGUEMENTS`, and `NAME` structural properties which are made up of **`Log`**, **`Log`**, **`WARNING`** simple values for the AUAST1 and **`Log`**, **`Log`**, **`ERROR`** simple values for the AUAST2, respectively. The structural representation of the AUASTs as defined in Section **??** is `EXPRESSION[EXPRESSION[IDENTIFIER[`**`Log`**`]], ARGUMENTS[QUALIFIER[IDENTIFIER[`**`Log`**`]], NAME[IDENTIFIER[`**`WARNING`**`]]` for the AUAST1 and `EXPRESSION[EXPRESSION[IDENTIFIER[`**`Log`**`]], ARGUMENTS[QUALIFIER[IDENTIFIER[`**`Log`**`]], NAME[IDENTIFIER[`**`ERROR`**`]]` for the AUAST2, where the words capitalized represents subtrees and the words shown in bold represents leaves of the tree structure.

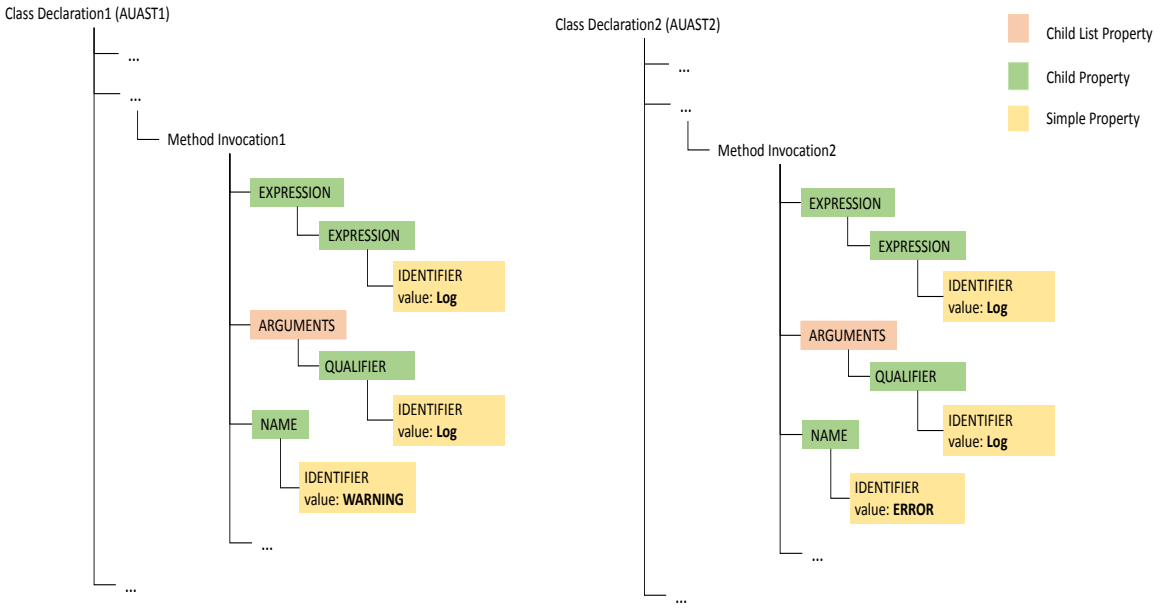


Figure 7.2: The AUASTs of log Method Invocation nodes from the Java classes in Figure 4.1 and Figure 4.2.

7.2 Determining Correspondences Using Jigsaw

Algorithm 7.2 JIGSAW-CORRESPONDENCE($auastA$, $auastB$) determines all the potential correspondences between nodes of two given AUASTs

JIGSAW-CORRESPONDENCE($auastA$, $auastB$)
1: **for do** $astA \in auastA$
2: **for do** $astB \in auastB$
3: $castA, castB \leftarrow$ JIGSAW-ANTIUNIFY$(astA, astB)$
4: **end for**
5: **end for**

7.3 Constraints in Determining Correspondences

To construct an anti-unifier of two AUASTs with a focus on logging calls, some constraints should be applied prior to determining the best correspondences. The first constraint (as described below)

should be applied to prevent the anti-unification of log method invocation nodes with any other type of node.

Constraint 1. A logging call should either be antiunified with another logging call or should be antiunified with "nothing".

This constraint creates a further constraint, which is:

Constraint 2. A structure containing a logging call should be antiunified with a corresponding structure containing another logging call or should be antiunified with "nothing".

To provide an example to illustrate it consider ASTs of two Java classes in Figure 6.2. Jigsaw creates a correspondence connection between the two log method invocation nodes and the two **if** statements. As is clear, the second **if** statement contains a logging call, while there is no corresponding logging call in the first one. According to the first constraint, two log method invocation nodes should be antiunified together. On the other hand, a correspondence connection is created between the two **if** statements; however, anti-unification of these statements includes anti-unifying their children nodes as well. Thus, statements inside the body of **if** statements must be antiunified with each other, indicating that log method invocation inside the body of **if** statement in the second example should be antiunified with "nothing", which is contrary to our first assumption. In order to comply with the first constraint, the correspondence connection between two **if** statements should be deleted, leading us to apply the second constraint.

Our approach applies these constraints by taking the following steps prior to determining correspondences:

1. Augment a property to AUAST node to mark log method invocation nodes and structures enclosing them as "logged".

2. Remove correspondence connections where one node is marked as "logged" and the corresponding node is not.

7.4 Determining Correspondences

As explained in Section 7.2, each node of the AUAST structure holds a list of candidate correspondence connections where each represents an anti-unifier. Despite having multiple potential anti-unifiers, we need to determine one single anti-unifier that is helpful to solve our problem. In general, higher order anti-unification modulo theories is undecidable [Cottrell et al., 2008]. That is, the complexity of determining the most optimal MSA is undecidable, but our desire is to create one of the best MSAs to approximate the optimal one that can sufficiently solve our problem, thus the anti-unification process should construct an anti-unifier that is the best approximate fit for our application. To this end, a greedy selection algorithm has been used, which is an approximation technique to determine the best correspondence for each node in the AUAST so constructing the anti-unifier that is approximately the best fit to our problem. As a result, each node can either be antiunified with its best correspondence in the other AUAST or with "nothing".

DETERMINE-CORRESPONDENCE algorithm greedily selects the most similar correspondence as the best fit for each node in AUAST. It takes one of the AUASTs, visiting the AUAST nodes therein to store all candidate correspondence connections between the two AUAST nodes in a list, which is sorted in a descending order based on the Jigsaw similarity measure (lines 1–8). The correspondence connection with the highest similarity value is determined as the best fit for the two nodes involved (lines 9–11); all other correspondence connections involving these two nodes are removed using REMOVE-OTHER-CORRESPONDENCES algorithm (line 10). This process terminates when no more correspondence connections is left in the list.

REMOVE-OTHER-CORRESPONDENCES algorithm removes correspondence connections that are not selected as the best fit from three lists: the list of all correspondence connections (Line 5 and Line 12); the list of candidate correspondence connections of the first node involved in these connections(Line 6 and Line 13); the list of candidate correspondence connections of the second node involved in these connections(Line 7 and Line 14).

As an example, Figure 7.3 shows the correspondences between AUAST nodes after applying

Algorithm 7.3 DETERMINE-CORRESPONDENCE($auastA$) takes in an AUAST node and create a list of correspondence connections containing the best correspondence to each node in the AUAST.

DETERMINE-CORRESPONDENCE($auastA$**)**

```
1: list ← ()
2: nodes ← VISITOR(auastA)
3: for do node ∈ nodes
4:     for do ce ∈ correspondences[node]
5:         APPEND(ce, list)
6:     end for
7: end for
8: SORT(list)
9: for ce ∈ list do
10:    REMOVE-OTHER-CORRESPONDENCES(ce, list)
11: end for
12: return list
```

Algorithm 7.4 REMOVE-OTHER-CORRESPONDENCES(ce, $list$) Remove all other correspondences involving nodes of a particular correspondence connection or element (ce) from lists of correspondence connections.

REMOVE-OTHER-CORRESPONDENCES(ce**,** $list$**)**

```
1: list1 ← correspondences[nodeA[ce]]
2: list2 ← correspondences[nodeB[ce]]
3: for ce1 ∈ list1 do
4:     if ce1 ≠ ce then
5:         REMOVE(ce1, list)
6:         REMOVE(ce1, correspondences[nodeA[ce1]])
7:         REMOVE(ce1, correspondences[nodeB[ce1]])
8:     end if
9: end for
10: for ce2 ∈ list2 do
11:    if ce2 ≠ ce then
12:        REMOVE(ce2, list)
13:        REMOVE(ce2, correspondences[nodeA[ce2]])
14:        REMOVE(ce2, correspondences[nodeB[ce2]])
15:    end if
16: end for
```

the constraints and DETERMINE-BEST-CORRESPONDENCE algorithm on the list of correspondence connections created by the Jigsaw framework in Figure 6.2.

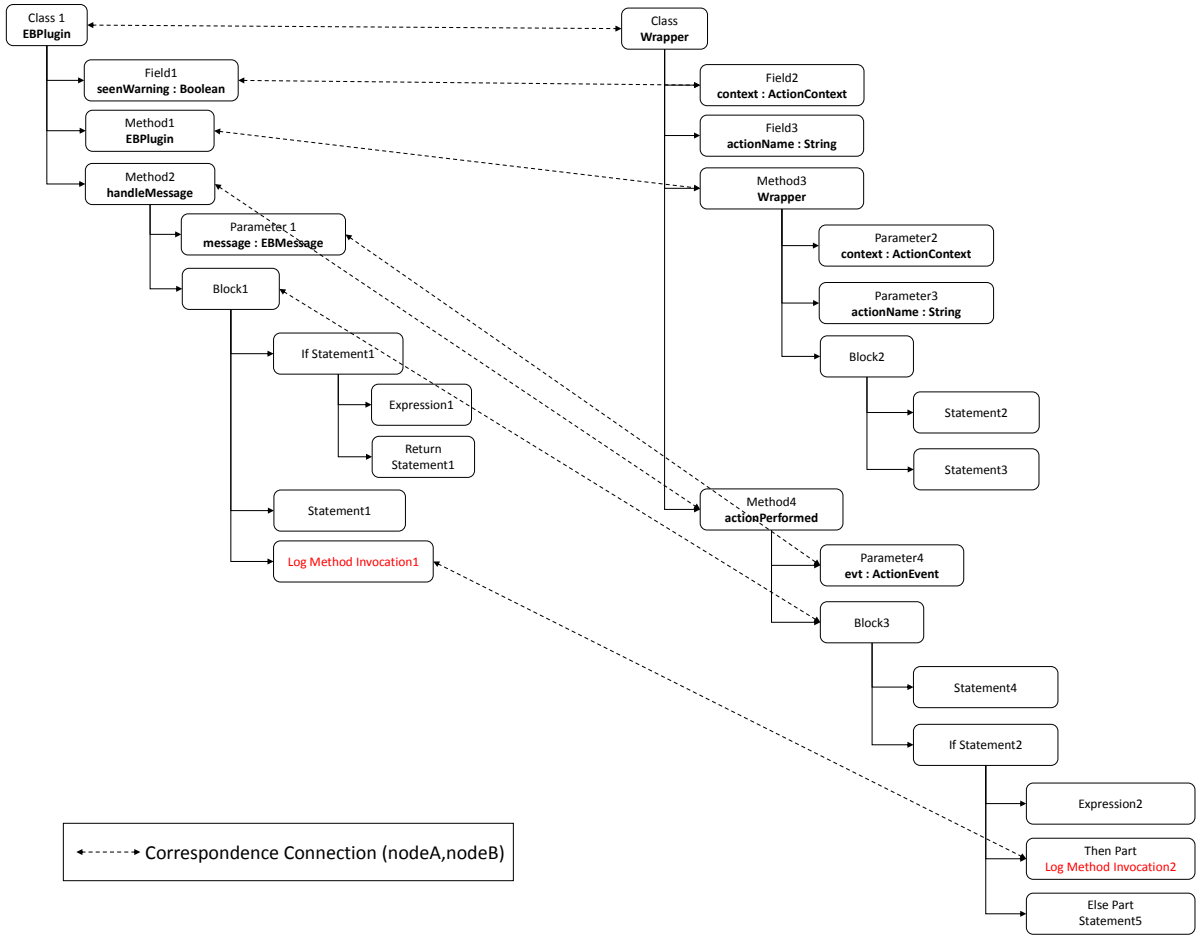


Figure 7.3: Simple AUAST structures constructed from the ASTs in Figure 6.2. Links between AUAST nodes indicate structural correspondences selected as the best fit

7.5 Computing Similarity

Similarity computation is particularly important for the clustering phase that relies on accurate estimation of distance between logged Java classes. The notion of similarity can differ depending on the given context. That is, similarity between certain features could be highly important for a particular application, while it is not for the other one. The utility of a similarity function can

be determined based on how good it enables us to produce accurate results for a particular task. In this study, a similarity measure is needed to classify Java classes that use logging calls based on structural similarity between them. The structural similarity of two AUASTs can be defined as the number of identical simple structural property values over total number of simple structural property values of the anti-unifier.

Algorithm 7.5 COMPUTE-MATCHES($auastA$, $auastB$), determines the matches between two AUASTs via a recursive traversal of structural properties

```
    COMPUTE-MATCHES(auastA, auastB)
 1: if auastB ≠ NULL then
 2:     DETERMINE-BEST-CORRESPONDENCE(auastA, auastB)
 3: end if
 4: matches ← 0
 5: for property ∈ properties[ant − unifier] do
 6:     valueA ← value[property]
 7:     valueB ← value[GETCORRESPONDENCE(valueA)]
 8:     if property instanceof SimpleProperty or property instanceof SimpleVariableProperty
    then
 9:         matches ← matches + JIGSAW-MATCHES(valueA, valueB)
10:     else if property instanceof ChildProperty or property instanceof ChildVariableProperty
    then
11:         matches ← matches + COMPUTE-MATCHES(valueA, valueB)
12:     else if property instanceof ChildListProperty then
13:         for nodeA ∈ valueA do
14:             nodeB ← GETCORRESPONDENCE(nodeA)
15:             matches ← matches + COMPUTE-MATCHES(nodeA, nodeB)
16:         end for
17:     end if
18: end for
19: return matches
```

The number of matches between $auastA$ and $auastB$ is computed via the COMPUTE-MATCHES algorithm through a recursive traversal of structural properties of the nodes. First, the best correspondences are selected using the DETERMINE-BEST-CORRESPONDENCE algorithm. For simple and simple variable structural properties, the number of matches is computed re-using the Jigsaw similarity function that computes the number of matches between the property values(Lines 8-9). For child and child variable structural properties, the number of matches is computed recursively

for the child node and is propagated to the parent(Lines 10-11). For child list structural properties, the number of matches is computed for each child node recursively and is propagated to the parent node(Lines 12-17). All matches are summed up to compute total number of matches between the two AUASTs. Then the following equation is used to compute the structural similarity between $auastA$ and $auastB$:

$$similarity = \frac{2 * matches}{|auastA| + |auastB|} \tag{7.1}$$

Where total number of simple values for $auastA$ and $auastB$ is computed via COMPUTE-MATCHES($auastA$) and COMPUTE-MATCHES($auastB$), respectively. The similarity function returns a value between 0 and 1 where indicate zero and total class matching, respectively.

7.6 Constructing the anti-unifier

Once the best correspondences has been determined between AUAST nodes, we construct a new antiunified AUAST by traversing AUAST structures recursively and anti-unifying the structural properties. The new antiunified structure is a generalization of two original structure, called anti-unifier, where common structural properties are represented by copy, and differences in structural properties are represented by structural variables. The variables may be inserted in place of any node in AUAST including both subtrees and leaves and can be substituted with proper original substructures to gain back to original structures.

Anti-unification of two AUAST nodes is performed through anti-unification of their structural properties, via the ANTIUNIFY algorithm. For each structural property of $auastA$ and $auastB$, where there is no corresponding property in the other AUAST, a structural variable property is created through anti-unifying the structural property with the NIL structure via the ANTIUNIFY-PROPERTY algorithm and added to properties of the anti-unifier (Lines 3-6 and Lines 13-17); if both nodes has the same property but with different property values, a structural variable property is created via the ANTIUNIFY-PROPERTY algorithm and appended to the anti-unifier structural

properties (Lines 7-8); otherwise, if the two nodes has the same exact structural property, a copy of one of them is added to the anti-unifier structural properties (Lines 10-11).

Algorithm 7.6 Input into ANTIUNIFY($auastA$, $auastB$) are two AUAST nodes. This algorithm construct an antiunified AUAST node through anti-unification of input node's structural properties.

```
    ANTIUNIFY(auastA, auastB)
 1: anti − unifier ← Null
 2: for propA ∈ properties[auastA] do
 3:     valueA ← value[property]
 4:     if CONTAINS(auastB, propA) = NULL then
 5:         ADDPROPERTY(anti − unifier, ANTIUNIFY-PROPERTY(propA, NIL))
 6:     else if valueA ≠ value[CONTAINS(auastB, propA)] then
 7:         ADDPROPERTY(anti − unifier, ANTIUNIFY-PROPERTY(propA, CONTAINS(auastB, propA))
 8:     else
 9:         ADDPROPERTY(anti − unifier, propA)
10:     end if
11: end for
12: for propB ∈ properties[auastB] do
13:     if CONTAINS(auastA, propB) = NULL then
14:         ADDPROPERTY(anti − unifier, ANTIUNIFY-PROPERTY(propB, NIL))
15:     end if
16: end for
17: return anti − unifier
```

Anti-unification of structural properties $propA$ and $propB$ is performed via the ANTIUNIFY-PROPERTY algorithm. If $propA$ is a simple property, a simple variable property is constructed referring to two simple values (Lines 2-3); If structural property is a child property, a child variable structure is constructed (Line 5); if structural property is a child list property, for each child of $propA$ and $propB$, where there is no correspondence in the other AUAST, an antiunified node is created through anti-unifying the child node with the NIL structure via ANTIUNIFY algorithm and added to the value of the antiunified child list property; otherwise, the child node is antiunified with its best correspondence (Lines 6-18).

For example, we supply ANTIUNIFY algorithm with the log method invocation nodes from the AUASTs in Figure 7.2. `EXPRESSION` and `ARGUMENTS` are similar in both AUASTs thus a copy of them will be added to structural properties of the antiunified AUAST (Line 10); however, the simple value of Name property is different in both structures thus a call to ANTIUNIFY-

Algorithm 7.7 ANTIUNIFY-PROPERTY($propA$, $propB$) takes two structural properties and creates an antiunified structural property.

```
    ANTIUNIFY-PROPERTY(propA, propB)
 1: property ← Null
 2: if propA instanceof SimpleProperty then
 3:     property ← CREATE-SIMPLE-VARIABLE-PROPERTY(propA, propB)
 4: else if propA instanceof ChildProperty then
 5:     property ← CREATE-CHILD-VARIABLE-PROPERTY(propA, propB)
 6: else if propA instanceof ChildListProperty then
 7:     for child ∈ value[propA] do
 8:         if correspondence[child] ≠ NULL then
 9:             APPEND(children, ANTIUNIFY(child, correspondence[child]))
10:         else
11:             APPEND(children, ANTIUNIFY(child, NIL))
12:         end if
13:     end for
14:     for child ∈ value[propB] do
15:         if correspondence[child] = NULL then
16:             APPEND(children, ANTIUNIFY(child, NIL))
17:         end if
18:     end for
19:     value[property] ← children
20: end if
21: return property
```

PROPERTY on Line 8 will return a simple structural variable. Figure 7.4 shows the antiunified AUAST, where the annotation `WARNING-or-ERROR` is used to represent the simple structural variable that must be substituted with either `WARNING` or `ERROR` simple value to gain back to each original AUAST structure. The structural representation of the antiunified AUAST is `EXPRESSION[EXPRESSION[IDENTIFIER[`**`Log`**`]], ARGUMENTS[QUALIFIER[IDENTIFIER[`**`Log`**`]], NAME[IDENTIFIER[`**`WARNING-or-ERROR`**`]].`

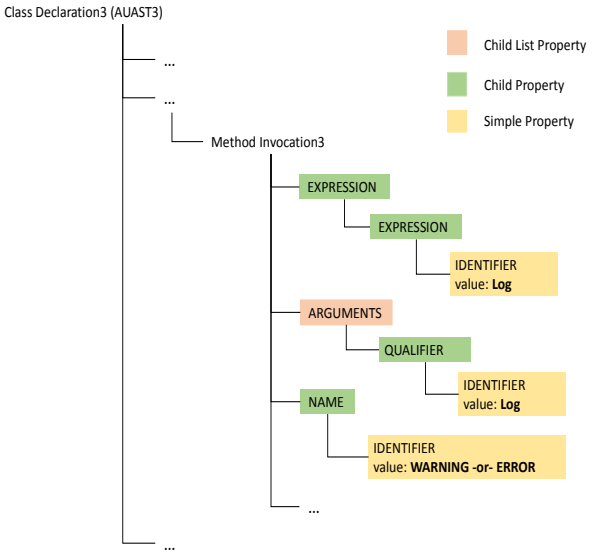


Figure 7.4: The anti-unifier (AUAST3) constructed from log Method Invocation AUAST nodes in Figure 7.2

Figure 7.5 shows a simple view of the antiunified AUAST constructed from the two AUASTs in Figure 7.3, where "$a \langle\rangle b$" represents that the two subtrees a and b are antiunified with each other in the anti-unifier and "a-or-b" represents a simple structural variable that must be substituted with either a or b simple value to recover each original structure.

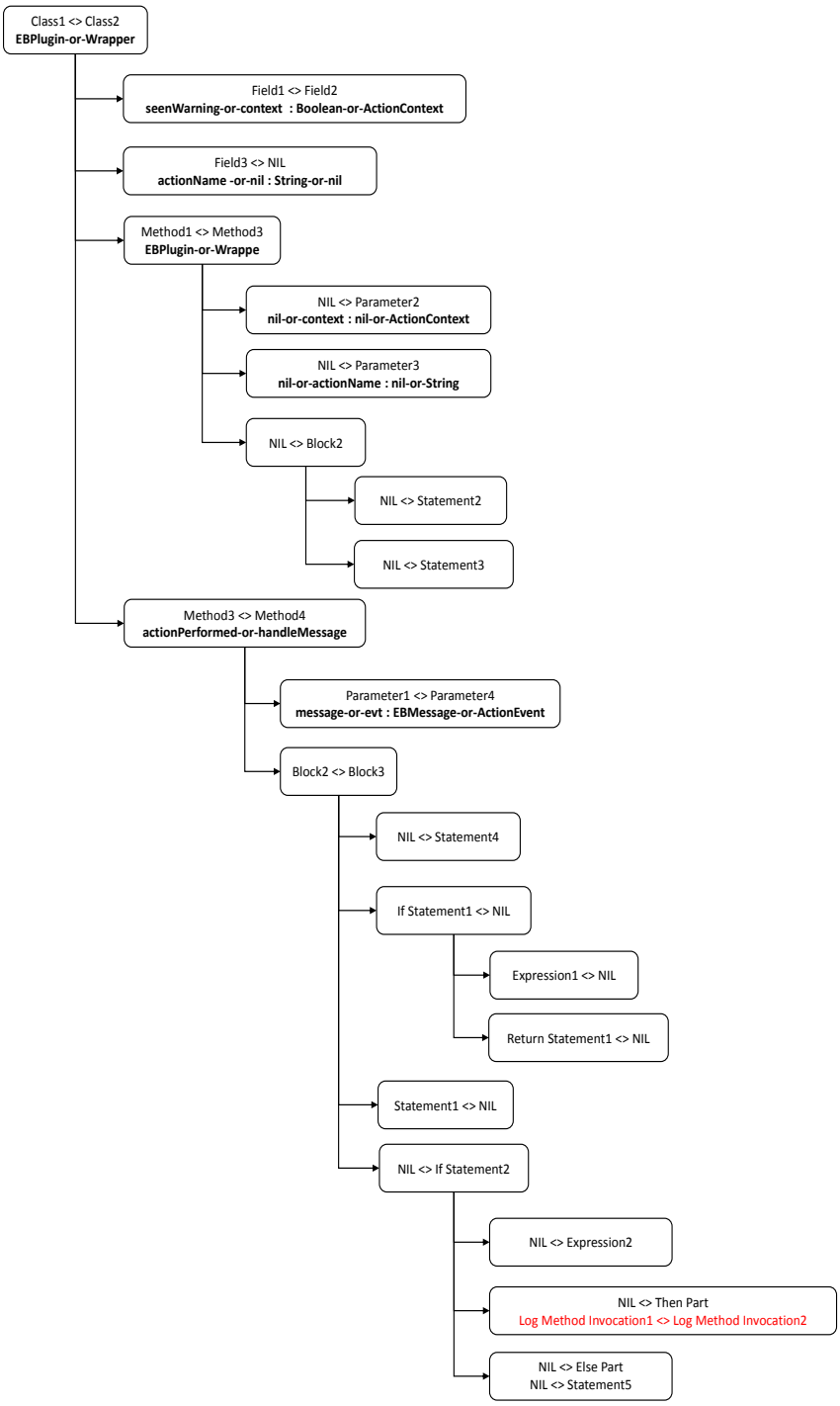


Figure 7.5: Simple antiunified AUAST structure of the two AUASTs in Figure 7.3

7.7 Multiple logging calls

Problem: There might be some cases that our approach is not able to anti-unify logging calls in two input seeds, when there is more than one logging call in a logged Java class. For example, consider the logged Java classes in Figures 7.6 and 7.7. Figure 7.8 shows the simple AUASTs for these examples and all potential correspondence connections between the AUAST nodes. Figure 7.9 shows the correspondence connections selected as the best match using our greedy algorithm. To anti-unify method1 with method3, we should anti-unify their structural properties; thus, log1 should be antiunified with log3 and log4 should be antiunified with "nothing" since there is no corresponding logging call in the body of method1, while there is a corresponding logging call for log4 in the body of method2 (log2).

```
1 public class test1{
2         public void method1(){
3                 ...
4                 Log.log();
5                 ...
6         }
7         public void method2(){
8                 ...
9                 Log.log();
10                ...
11        }
12 }
```

Figure 7.6: A Java class that utilizes multiple logging calls. This will be referred to as Example 1.

```
1 public class test2{
2         public void method3(){
3                 ...
4                 Log.log();
5                 ...
6                 Log.log();
7         }
8 }
```

Figure 7.7: A Java class that utilizes multiple logging calls. This will be referred to as Example 2.

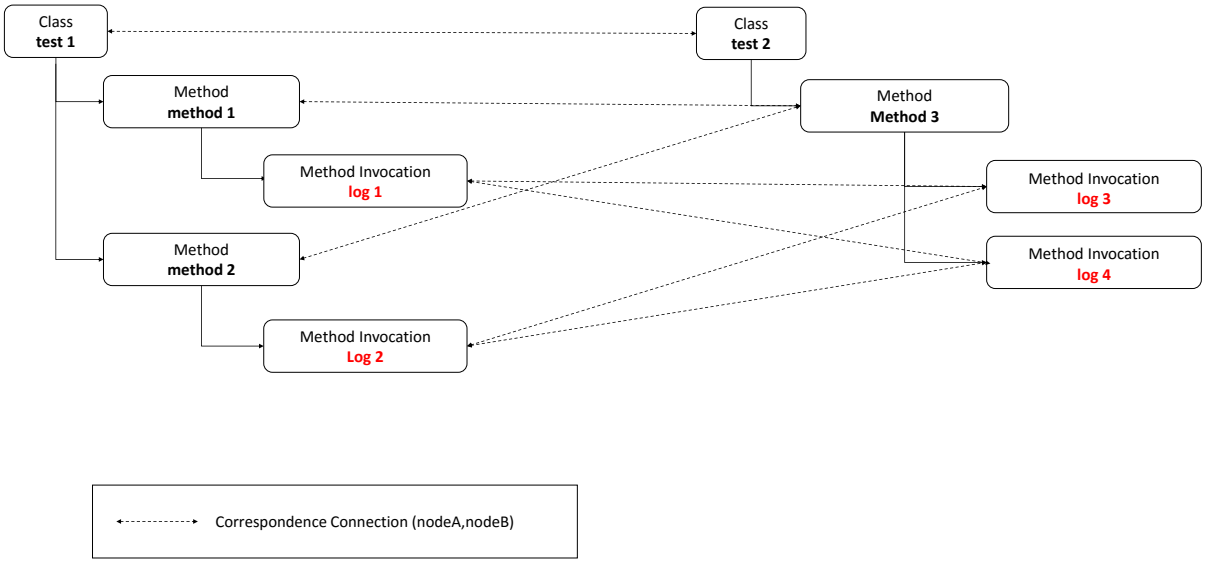


Figure 7.8: Simple AUAST structure of examples in Figures 7.6 and 7.7. Links between AUAST nodes indicate potential candidate structural correspondences detected by the Jigsaw framework.

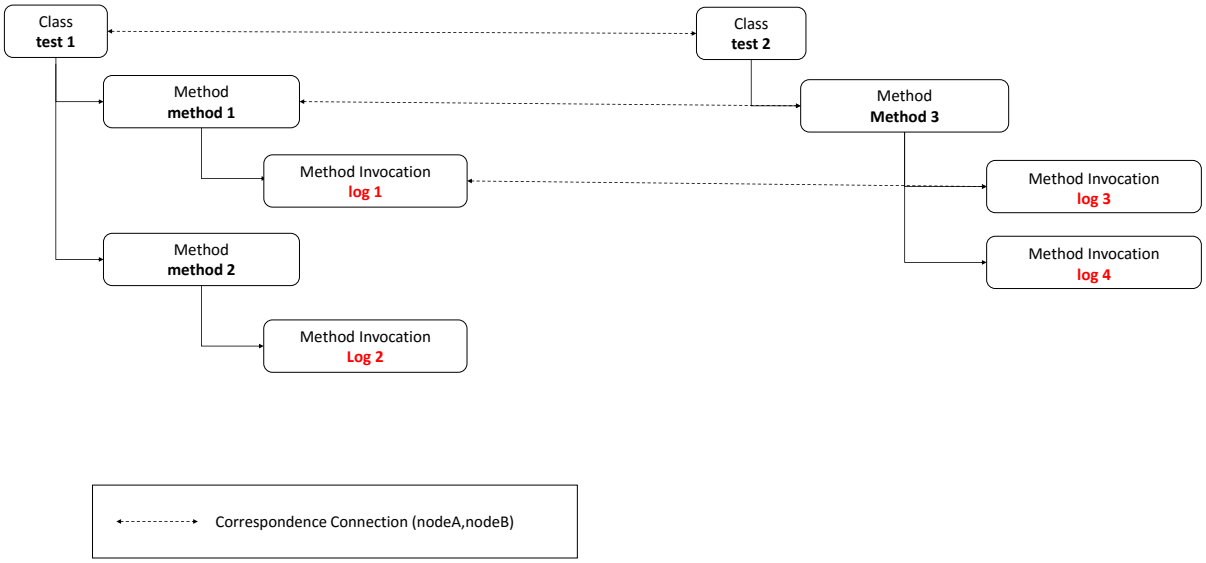


Figure 7.9: Simple AUAST structure of examples in Figures 7.6 and 7.7. Links between AUAST nodes indicate structural correspondences selected as the best match using our greedy algorithm.

Suggested Solution: We can split these cases into more than one case, where each logged Java class contains only one logging call. To do so, we need to create a copy of logged Java class for each logging call by maintaining the logging call and removing the other ones. For example, we need to create two copies for each logged Java class of Examples 1 and 2 as depicted in Figures 7.10 and 7.11, respectively.

```
1  public class test1{
2          public void method1(){
3                  ...
4                  Log.log();
5                  ...
6          }
7          public void method2(){
8                  ...
9                  //removed
10                 ...
11         }
12 }
13
14 public class test1{
15         public void method1(){
16                 ...
17                 //removed
18                 ...
19         }
20         public void method2(){
21                 ...
22                 Log.log();
23                 ...
24         }
25 }
```

Figure 7.10: Create multiple copies of Example 1 for each logging call.

```
1 public class test2{
2         public void method3(){
3                 ...
4                 Log.log();
5                 ...
6                 //removed
7         }
8 }
9
10 public class test2{
11        public void method3(){
12                ...
13                //removed
14                ...
15                Log.log();
16        }
17 }
```

Figure 7.11: Create multiple copies of Example 2 for each logging call.

7.8 Antiunifying a set of AUASTs

- PROBLEM: anti-unifying a set of AUASTs of LJCs
- SOLUTION: Developing a modified version of a hierarchical agglomerative clustering algorithm (illustrated in Figure 7.12) as described below:

  1. Start with singleton clusters, where each cluster contains one AUAST
  2. Compute the similarity between clusters in a pairwise manner
  3. Find the closest clusters (a pair of clusters with maximum similarity)
  4. Merge the closest cluster pair and replace them with a new cluster containing anti-unifier of AUASTs of the two clusters
  5. Compute the similarity between the new cluster and all remaining clusters

  – Repeat Steps 3,4, and 5 until the similarity between closest clusters becomes below a predetermined threshold value

- The similarity between a pair of clusters is defined as the similarity between their AUASTs
- Determine the similarity threshold value through informal experimentation

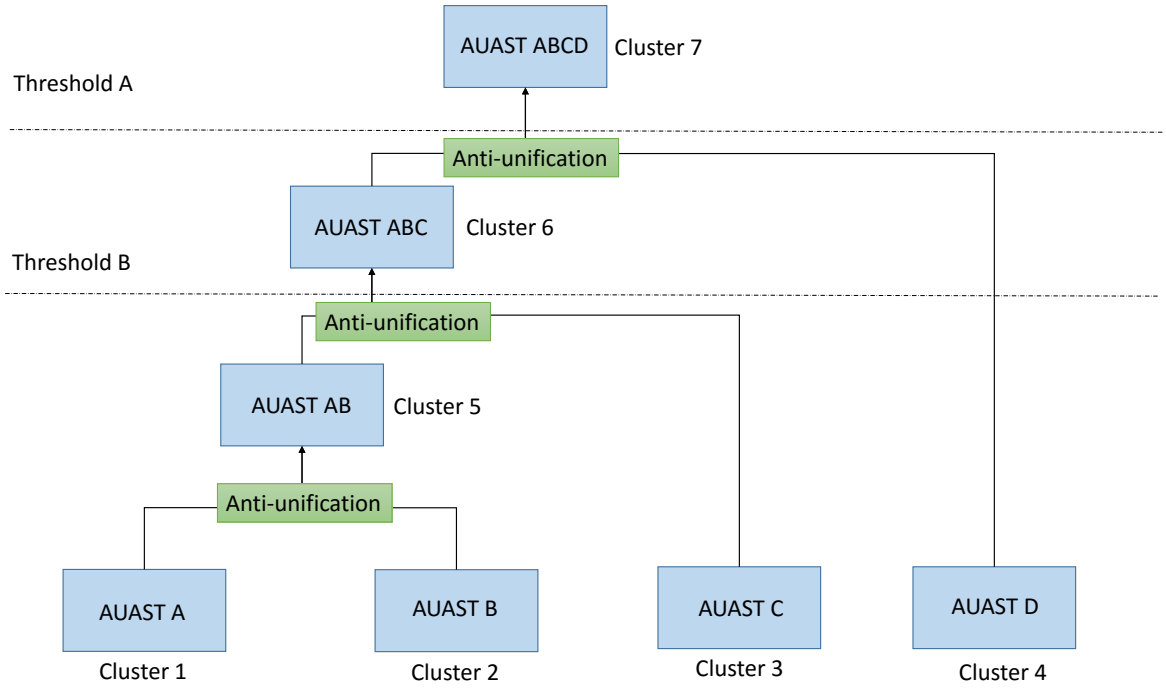


Figure 7.12: Anti-unification of 4 AUAST nodes using an agglomerative hierarchical clustering algorithm. The threshold value indicates the number of clusters we will come up with.

Chapter 8

Evaluation

- Two empirical studies were conducted

8.1 Experiment 1

- First experiment is conducted to evaluate the accuracy of our approach and tool
- It addresses the following research questions:
  - RQ1: can our tool determine the structural similarities and differences between logged Java classes correctly?
  - RQ2: can our tool compute the similarity between logged Java classes correctly?
- To do so, 10 logged Java classes were selected randomly from jEdit v4.2 pre 15 (2004), as our test set
- We apply our tool on the test set to create generalizations and to compute the similarity value between logged Java classes in a pairwise manner
- To address the first research question we compute the following measurements for each test case:
  - the number of correspondences that our tool detects correctly
  - the total number of correspondences
- To determine the correct correspondences we performed a manual investigation
- To address the second research question we compute:

  - the number of similarity values between logged Java classes that are computed correctly by our tool
  - the total number of comparisons

- The correct similarity value for each comparison is calculated manually by
- The results taken form our tool are compared with the results taken by manual investigation using the JUnit testing framework

8.2 Experiment 2

- Second experiment is conducted to address the following research questions:
  - RQ3: what structural similarities and differences do logged Java classes have?
  - RQ4: Is it possible to find common patterns in where logging calls do occur?
- To do so, we applied our tool on the source code of three open-source full systems that make use of logging to determine the patterns on a per-system class-granularity basis analysis
- These systems are different from the system that the test set is selected from

8.3 Results

- I will describe the results taken form the second experiment

8.4 Lessons learned

- I will describe our findings

Chapter 9

Discussion

9.1 Threats to validity

- Our goal is to recognize the limitations and pitfalls of our approach and its developed tool support
- The first potential thread to validity of our characterization study is the degree to which our sample set of software systems is a good representation of all real-world logging practices. To address this issue we selected software systems that:
  - are different in terms of application
  - are among the most popular applications in their own product category
  - has long history in software development
- Secondly, our manual investigation to find the correct correspondences might be biased due to human errors. To limit the bias
  - other people can be involved to double check the accuracy of manual work in the future work
- However, these results are still promising

9.2 Our tool output

- We investigated the cases where our tool fails and we found that the failures are due to:
  - the assumptions taken in developing the algorithms
  - the fundamental limitations and complexities in determining the detailed structural similarities and differences

- The are some issues that our tool is not able to handle perfectly during generalization:
  - maintaining the correct ordering of statements inside the method bodies
  - resolving all the conflicts that happen in determining the best correspondences
  - producing executable generalizations

9.3 Theoretical foundation

- anti-unification and its extensions has several theoretical and practical applications:
  - analogy making [Schmidt, 2010]
  - determining lemma generation in equational inductive proofs [Burghardt, 2005]
  - detecting the construction laws for a sequence of structures [Burghardt, 2005]
- Using higher-order anti-unification modulo theories in our application, which is undecidable in general, leads us to take approximations suitable to our context
- The set of equational theories should be developed particularly for the structure used in each problem context

Chapter 10

Related Work

In this chapter, we review related work to the topics of our study including: the application of logging in real-world software systems (Section 10.1), determining correspondences in the source code (Section 10.2), data mining approaches to extract API usage patterns (Section 10.3), anti-unification and its application to detect strcutural correspondences and construct generalizations (Section 10.4), and clustering (Section 10.5).

10.1 Usage of logging

Logging is a conventional programming practice to record a software system's runtime information that can be used in post-modern analysis to trace the root causes of systems' activities. Log analysis is most often performed for failure diagnosis, system behavioral understanding, system security monitoring and performance diagnostics purposes as described below:

- **Log analysis for failure diagnosis:** Xu et al. [2009] use statistical techniques to learn a decision tree based signature from the console logs and then utilize the signature to diagnose anomalies. SherLog [Yuan et al., 2010] uses failure log messages to infer the source code paths that might have been executed during a failure.

- **Log analysis for system behavior understanding:** Fu et al. [2013] present an approach for understanding system behavior through contextual analysis of logs. They first extracted execution patterns reflected by a sequence of system logs and then utilized the patterns to find contextual factors from logs that causes a specific system behavior. The Linux Trace Toolkit [Yaghmour and Dagenais, 2000] was created to record and analyze system behavior by providing an efficient kernel-level event logging infrastructure. A

more flexible approach is taken by DTrace [Cantrill et al., 2004] which allows dynamic modification of kernel code.

- **Log analysis for system security monitoring:** Bishop [1989] proposes a formal model of system's security monitoring using logging and auditing. Peisert et al. [2007] have developed a model that demonstrates a mechanism for extracting logging information to detect how an intrusion occurs in software systems.

- **Log analysis for performance diagnosis:** Nagaraj et al. [2012] developed an automated tool to assist developers in diagnosis and correction of performance issues in distributed systems by analyzing system behaviors extracted from the log data.

Jiang et al. [2009] study the effectiveness of logging in problem diagnosis. Their study shows that customer problems in software systems with logging resolve faster than those without logging by investigating the correlations between failure root causes and diagnosis time. Despite the importance of logging for software development and maintenance, few studies have been conducted in pursuit of understanding logging usage in real-world software. Yuan et al. [2012b] provides a quantitative characteristic study to investigate log message modifications on four open-source software systems by mining their revision history. Their study shows that developers spend a great effort to modify logging calls as after-thoughts, which indicates that they are not satisfied with the log quality in their first attempt. They also characterize where developers spend most of their time in modifying the log messages.

Yuan et al. [2012a] studies the problem of lack of log messages for error diagnosis and suggests to log when generic error conditions happens. LogEnhancer [Yuan et al., 2012c] automatically enhances existing log message by detecting important variable values and inserting them into the log messages. However, these studies only consider code snippets containing bugs that are needed to be logged and do not consider other code snippets containing no bugs but still need to be logged. Moreover, these studies mainly research log message modifications and potential enhancements of

them, however, the focus of this study is on understanding where logging calls are used in the source code.

10.2 Correspondence

Several studies have been conducted to find similarities and differences between the source code fragments. Baxter et al. [1998] develop an algorithm to detect code clones in source code that uses hash functions to partition subtrees of ASTs of a program source code and then find common subtrees in the same partition through a tree comparison algorithm. Apiwattanapong et al. [2004] present a top-down approach to detect differences and correspondences between two versions of a Java program, through comparison of the control flow graphs created from the source code. Holmes et al. [2005] recommends relevant code snippet examples from a source code repository for the sake of helping developers to find examples of how to use an API by heuristically matching the structure of the code under development with the source code in the repository. Coogle [Sager et al., 2006] is developed to detect similar Java classes through converting ASTs to a normalized format and then comparing them through tree similarity algorithms. However, none of these approaches determines the detailed structural correspondences needed in our context.

Umami [Cossette et al., 2014] presents a new approach, called Matching via Structural generalization (MSG), to recommend replacements for API migration. He used the Jigsaw tool to find structural correspondences, however, their proposed algorithm does not suffice to our context since it does not construct a generalization to represent structural similarities and differences. It also does not take the required constraints in determining correspondences needed to solve our problem.

10.3 API usages patterns

Various data mining approaches has been used to extract API usages patterns out of the source code such as unordered pattern mining and sequential pattern mining [Robillard et al., 2013].

Unordered pattern mining, such as association rule mining and itemset mining, extracts a set of API usage rules without considering their order [Agrawal et al., 1994]. CodeWeb [Michail, 2000] uses data mining association rules to identify reuse patterns between a source code under development and a specific library. PR-Miner [Li and Zhou, 2005] uses frequent itemset mining to extract implicit programming rules from source code and detect violations. The sequential pattern mining technique is different from the unordered one in the way that it considers the order of API usage. As an example, MAPO [Xie and Pei, 2006] combines frequent subsequence mining with clustering to extract API usage patterns from the source code. The other technique for extracting API usage patterns is through statistical source code analysis. For example, PopCon [Holmes and Walker, 2008] is a tool developed to help developers understanding how to use APIs in their source code through calculating popularity statistics for each API of a library. Acharya et al. [2007] present a framework to extract API usage scenarios as partial orders. Specifications were extracted from frequent partial orders. They adapted a compile time model checker to generate control-flow-sensitive static traces of APIs, from which API usage scenarios were extracted. However, none of these approaches suffice to determine the detailed structural correspondences.

10.4 Anti-unification

Anti-unification is the problem of finding the most specific generalization of two terms. First-order syntactical anti-unification was introduced by Plotkin [1970] and Reynolds [1970] independently. Burghardt and Heinz [1996] extend the notion of anti-unification to E-anti-unification to incorporate background knowledge to syntactical anti-unification, which is required for some applications. anti-unification has been applied in various studies for program analysis. Bulychev and Minea [2009] suggest an anti-unification algorithm to detect clones in ASTs. Their approach consists of three stages: first, identifying similar statements through anti-unification and classifying them into clusters; second, determining similar sequences of statements with the same Cluster identifier; third, refining candidate statement sequences using an anti-unification based similarity

measurement to generate final clones. However, their approach does not construct a generalization by determining the structural correspondences. Cottrell et al. [2007] propose Breakaway to automatically determine structural correspondences between a pair of abstract syntax trees (ASTs) to create a generalized correspondence view. However, their approach does not allow us to detect the best structural correspondence for each node suited to our problem. Cottrell et al. [2008] develop Jigsaw to help developers integrate small-scale reused source code into their own code by determining structural correspondences through the application of higher-order anti-unification modulo theories. However, considering the limitations of our study in determining correspondences, their approach does not suffice to construct a structural generalization needed in our context.

10.5 Clustering

Clustering is an unsupervised machine mining technique that aims to organize a collection of data into clusters, such that intra-cluster similarity is maximized and the inter-cluster similarity is minimized [Karypis et al., 1999, Grira et al., 2004]. We divided existing clustering approaches into two major categories: partitional clustering and hierarchical clustering. Partitional clustering try to classify a data set into k clusters such that the partition optimizes a pre-determined criterion [Karypis et al., 1999]. The most popular partitional clustering algorithm is k-means, which repeatedly assigns each data point to a cluster with the nearest centroid and computes the new cluster centroids accordingly until a pre-determined number of clusters is obtained [Bouguettaya et al., 2015]. However, k-means clustering algorithm is not a good fit to our problem since it requires to predefine the number of clusters we want to come up with, which is not reasonable in our context.

Hierarchical clustering algorithms produce a nested grouping of clusters, with single point clusters at the bottom and an all-inclusive cluster at the top [Karypis et al., 1999]. Agglomerative hierarchical clustering is one of the main stream clustering methods [Day and Edelsbrunner, 1984] and has applications in document retrieval [Voorhees, 1986] and information retrieval from a search engine query log [Beeferman and Berger, 2000]. It starts with singleton clusters, where

each contains one data point. Then it repeatedly merges the two most similar clusters to form a bigger one until a pre-determined number of clusters is obtained or the similarity between the closest clusters is below a pre-determined threshold value. Hierarchical clustering algorithms work implicitly or explicitly with the $n \times n$ similarity matrix such that an element in row i and column j represents the similarity between the i^{th} and the j^{th} clusters [Karypis et al., 1999].

There are various versions of agglomerative hierarchical algorithms that mainly differ in how they update the similarity between clusters. There are various methods to measure the similarity between clusters, such as single linkage, complete linkage, average linkage, and centroids [Rasmussen, 1992]. In the single linkage method, the similarity is measured by the similarity of the closest pair of data points of the two clusters. In the complete linkage method, the similarity is computed by the similarity of the farthest pair of data points of the two clusters. In the average linkage method, the similarity is measured by the average similarity of all pairwise similarities of data points of the two clusters. In the centroids methods, each cluster is represented by a centroid of all data points in the cluster, and the similarity between two clusters is measured by the similarity of the clusters' centroids. However, in our application, each cluster is composed of one AUAST, and the similarity between two clusters is measured by the similarity between the clusters' AUASTs, which is computed via anti-unification.

10.6 Summary

Despite the great importance of logging and its various applications in software development and maintenance, few studies have focused on understanding logging usage in the source code. Some work has been done on characterizing log messages modifications made by developers and to help them enhance the content of log messages. However, to the best of our knowledge, no study has been conducted on characterizing where logging is used in the source code through determining structural correspondences. Several data mining and statistical source code analysis techniques have been used to extract API usage patterns, however, none of them enable us to determine the

detailed structural correspondences between source code fragments. On the other hand, using higher-order anti-unification modulo theories and an agglomerative hierarchical clustering algorithm allow us to construct structural generalizations that describe the similarities and differences between logged Java classes and classifying logged Java classes into groups based on the structural correspondences, respectively.

Chapter 11

Conclusion

- Determining the detailed structural similarities and differences between source code fragments is a complex task
- It can be applied to solve several source code analysis problems, for example, characterizing logging practices
- logging is a pervasive practice and has various applications in software development and maintenance
- However, it is a challenging task for developers to understand how to use logging calls in the source code
- We have presented an approach to characterize where logging calls happen in the source code by means of structural generalization
- We have developed a prototype tool that:
  - detects potential structural correspondences using anti-unification
  - uses several constraint to remove the correspondences that are not suited to our application
  - determines the best correspondences with the highest similarity
  - constructs the structural generalizations using anti-unification
  - classifies the entities using a measure of similarity
- An experiment is conducted to evaluate our approach and tool
- Our experiment found that ...

- An experiment is conducted to characterize logging usage in three software systems
- In summary, our study makes the following contributions:
  -
  -

11.1 Future Work

- Future extensions could be applied to resolve the pitfalls of this study:
  - Data flow analysis techniques: to resolve the problem of inaccurate statement ordering
  - Further analysis: to detect and resolve all the conflicts happen in deciding the best correspondences
- To further validate our findings from the source code analysis:
  - a survey can be conducted to ask developers on the factors they consider when they want to decide on where to log
- Characterizing logging usage could be a huge step towards
  - improving logging practices by providing some guidelines that might help developers in making decisions about where to log.
  - developing recommendation support tools:
    * to save developers'time and effort
    * to improve the quality of logging practices

Bibliography

Mithun Acharya, Tao Xie, Jian Pei, and Jun Xu. Mining api patterns as partial orders from source code: from usage scenarios to specifications. In *Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, pages 25–34. ACM, 2007.

Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

Taweesup Apiwattanapong, Alessandro Orso, and Mary Jean Harrold. A differencing algorithm for object-oriented programs. In *Proceedings of the 19th IEEE international conference on Automated software engineering*, pages 2–13. IEEE Computer Society, 2004.

Ira D Baxter, Andrew Yahin, Leonardo Moura, Marcelo Sant'Anna, and Lorraine Bier. Clone detection using abstract syntax trees. In *Software Maintenance, 1998. Proceedings., International Conference on*, pages 368–377. IEEE, 1998.

Doug Beeferman and Adam Berger. Agglomerative clustering of a search engine query log. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 407–416. ACM, 2000.

Matt Bishop. A model of security monitoring. In *Computer Security Applications Conference, 1989., Fifth Annual*, pages 46–52. IEEE, 1989.

Athman Bouguettaya, Qi Yu, Xumin Liu, Xiangmin Zhou, and Andy Song. Efficient agglomerative hierarchical clustering. *Expert Systems with Applications*, 42(5):2785–2797, 2015.

Peter Bulychev and Marius Minea. An evaluation of duplicate code detection using anti-unification. In *Proc. 3rd International Workshop on Software Clones*. Citeseer, 2009.

Jochen Burghardt and Birgit Heinz. Implementing anti-unification modulo equational theory. arbeitspapier 1006, 1996.

Bryan Cantrill, Michael W Shapiro, Adam H Leventhal, et al. Dynamic instrumentation of production systems. In *USENIX Annual Technical Conference, General Track*, pages 15–28, 2004.

Siobhán Clarke, William Harrison, Harold Ossher, and Peri Tarr. The dimension of separating requirements concerns for the duration of the development lifecycle. In *First Workshop on Multi-Dimensional Separation of Concerns in Object-oriented Systems (at OOPSLA)*, 1999a.

Siobhán Clarke, William Harrison, Harold Ossher, and Peri Tarr. Subject-oriented design: towards improved alignment of requirements, design, and code. *ACM SIGPLAN Notices*, 34(10):325–339, 1999b.

Bradley Cossette, Robert Walker, and Rylan Cottrell. Using structural generalization to discover replacement functionality for API evolution. Technical Report 2014-745-10, Department of Computer Science, University of Calgary, Calgary, Canada, May 2014.

Rylan Cottrell, Joseph J. C. Chang, Robert J. Walker, and Jörg Denzinger. Determining detailed structural correspondence for generalization tasks. In *Proceedings of the European Software Engineering Conference/ACM SIGSOFT International Symposium on Foundations of Software Engineering*, pages 165–174, 2007. doi: 10.1145/1287624.1287649.

Rylan Cottrell, Robert J. Walker, and Jörg Denzinger. Semi-automating small-scale source code reuse via structural correspondence. In *Proceedings of the ACM SIGSOFT International Symposium on Foundations of Software Engineering*, pages 214–225, 2008. doi: 10.1145/1453101.1453130.

William HE Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24, 1984.

Qiang Fu, Jian-Guang Lou, Yi Wang, and Jiang Li. Execution anomaly detection in distributed systems through unstructured log analysis. In *ICDM*, volume 9, pages 149–158, 2009.

Qiang Fu, Jian-Guang Lou, Qingwei Lin, Rui Ding, Dongmei Zhang, and Tao Xie. Contextual analysis of program logs for understanding system behaviors. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, pages 397–400. IEEE Press, 2013.

James Gosling, Bill Joy, Guy Steele, Gilad Bracha, and Alex Buckley. *The Java Language Specification*. Addison-Wesley, Java SE 7 edition, 2012. URL `http://docs.oracle.com/javase/specs/jls/se7/html/index.html`.

Nizar Grira, Michel Crucianu, and Nozha Boujemaa. Unsupervised and semi-supervised clustering: a brief survey. *A review of machine learning techniques for processing multimedia content*, 1:9–16, 2004.

Samudra Gupta. Pro apache log4j: Java application logging using the open source apache log4j api. *Apress®, USA*, 2005.

Reid Holmes and Robert J Walker. A newbie's guide to eclipse apis. In *Proceedings of the 2008 international working conference on Mining software repositories*, pages 149–152. ACM, 2008.

Reid Holmes, Robert J Walker, and Gail C Murphy. Strathcona example recommendation tool. In *ACM SIGSOFT Software Engineering Notes*, volume 30, pages 237–240. ACM, 2005.

Weihang Jiang, Chongfeng Hu, Shankar Pasupathy, Arkady Kanevsky, Zhenmin Li, and Yuanyuan Zhou. Understanding customer problem troubleshooting from storage system logs. In *FAST*, volume 9, pages 43–56, 2009.

George Karypis, Eui-Hong Han, and Vipin Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999.

Zhenmin Li and Yuanyuan Zhou. Pr-miner: automatically extracting implicit programming rules and detecting violations in large software code. In *ACM SIGSOFT Software Engineering Notes*, volume 30, pages 306–315. ACM, 2005.

Jian-Guang Lou, Qiang Fu, Shengqi Yang, Ye Xu, and Jiang Li. Mining invariants from console logs for system problem detection. In *USENIX Annual Technical Conference*, 2010.

Amir Michail. Data mining library reuse patterns using generalized association rules. In *Proceedings of the 22nd international conference on Software engineering*, pages 167–176. ACM, 2000.

Karthik Nagaraj, Charles Killian, and Jennifer Neville. Structured comparative analysis of systems logs to diagnose performance problems. In *Presented as part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, pages 353–366, 2012.

Sean Peisert, Matt Bishop, Sidney Karin, and Keith Marzullo. Toward models for forensic analysis. In *Systematic Approaches to Digital Forensic Engineering, 2007. SADFE 2007. Second International Workshop on*, pages 3–15. IEEE, 2007.

Gordon D Plotkin. A note on inductive generalization. *Machine intelligence*, 5(1):153–163, 1970.

John C Reynolds. Transformational systems and the algebraic structure of atomic formulas. *Machine intelligence*, 5(1):135–151, 1970.

Martin P Robillard, Eric Bodden, David Kawrykow, Mira Mezini, and Tristan Ratchford. Automated api property inference techniques. *Software Engineering, IEEE Transactions on*, 39(5): 613–637, 2013.

Tobias Sager, Abraham Bernstein, Martin Pinzger, and Christoph Kiefer. Detecting similar java classes using tree algorithms. In *Proceedings of the 2006 international workshop on Mining software repositories*, pages 65–71. ACM, 2006.

Ellen M Voorhees. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing & Management*, 22(6):465–476, 1986.

Tao Xie and Jian Pei. Mapo: Mining api usages from open source repositories. In *Proceedings of the 2006 international workshop on Mining software repositories*, pages 54–57. ACM, 2006.

Wei Xu, Ling Huang, Armando Fox, David Patterson, and Michael I Jordan. Detecting large-scale system problems by mining console logs. In *Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles*, pages 117–132. ACM, 2009.

Karim Yaghmour and Michel R Dagenais. Measuringandcharacter izing systembeh av ior usingkernel-leveleventlogging. 2000.

Ding Yuan, Haohui Mai, Weiwei Xiong, Lin Tan, Yuanyuan Zhou, and Shankar Pasupathy. Sherlog: error diagnosis by connecting clues from run-time logs. In *ACM SIGARCH computer architecture news*, volume 38, pages 143–154. ACM, 2010.

Ding Yuan, Soyeon Park, Peng Huang, Yang Liu, Michael M Lee, Xiaoming Tang, Yuanyuan Zhou, and Stefan Savage. Be conservative: enhancing failure diagnosis with proactive logging. In *Presented as part of the 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI 12)*, pages 293–306, 2012a.

Ding Yuan, Soyeon Park, and Yuanyuan Zhou. Characterizing logging practices in open-source software. In *Proceedings of the 34th International Conference on Software Engineering*, pages 102–112. IEEE Press, 2012b.

Ding Yuan, Jing Zheng, Soyeon Park, Yuanyuan Zhou, and Stefan Savage. Improving software diagnosability via log enhancement. *ACM Transactions on Computer Systems (TOCS)*, 30(1):4, 2012c.