ABSTRACT

Program transformation through the repeated application of simple rewrite rules is conducive to formal verification. In practice, program transformation oftentimes requires data to be moved throughout the program structure. This article explores the use of higher-order rewrite rules as the mechanism for accomplishing such data movement. The effectiveness of higher-order rewrite rules is demonstrated by showing how they can be used to perform field distribution within a Java class loader. An approach to formal verification of a higher-order strategic implementation of a class loader is also briefly discussed.

KEY WORDS

program transformation, higher-order rules, distributed data problem, Java class loader, Sandia Secure Processor

1 Transformations

The transformation of program structures through rewriting is an active area of research [2][3][6][7]. A primary goal is to enable rewriting to be used as a formal and automatable mechanism for high-assurance software development. Driving this research is the idea that the repeated application of a properly constructed set of simple rewrite rules can effect a major change in the structure of a program. The implications of this span the entire software life cycle, ranging from the development of implementations from formal specifications all the way to software maintenance.

Equational reasoning lies at the heart of rewriting-based development and offers promise that the transformation of programs through rewriting can be scaled to real-world systems. A set of equations, called an equational theory, can be used to capture both domain as well as problem specific knowledge. The languages in which such substitutions are taking place must be referentially transparent with respect to the substitutions being performed. The question now remains as to how this knowledge is to be applied.

The fully general application of equational theories immediately leads to undecidability. However, rewriting provides restrictions that can result in a decidable theory. In rewriting, the symmetric equality relation = is replaced by the anti-symmetric rewrite relation \( \rightarrow \). For example, the equation \( s = t \) states that the terms \( s \) and \( t \) may be freely substituted in place of one another. In contrast, the rewrite rule \( s \rightarrow t \) states that \( s \) may be replaced by \( t \), but does not specify that \( t \) may be replaced by \( s \).

In a pure rewriting framework, the order in which rewrite rules are applied is left unspecified. For a confluent and terminating system of rewrite rules, any application order is guaranteed to produce the same normal form. Unfortunately, the manipulation of software generally involves sets of rewrite rules that are neither terminating nor confluent. In this setting, it is necessary to make explicit the way in which rewrite rules should be applied. One approach is to introduce additional function symbols to control the application of rewrite rules. This approach yields sets of rewrite rules that are not very reusable, and does not scale well to real-world problems [2].

2 Strategic Programming

In strategic programming [7], the control problem is solved by making explicit a number of combinators capable of specifying both which rewrite rules should be applied as well as where they should be applied. These combinators can be used together with rewrite rules to form expressions called strategies. In this context, rewrite rules themselves are referred to as strategies. Strategies that are fully parameterized on the rewrite rules to be applied are called generic strategies. Generic strategies can be reused by any problem domain provided the control they provide is suitable for the problem at hand. In some systems [2], the user is provided with a fixed set of generic strategies. In other systems [7][10][11], users may define their own generic strategies.

A common class of generic strategy is one that defines the traversal of a structure. Examples of these generic traversals include TDL which will traverse a structure in a top-down left-to-right fashion (i.e., outside-in), and BUL which will traverse a structure in a bottom-up left-to-right fashion (i.e., inside-out). Generic traversals such as TDL and BUL are parameterized on other strategies. For example, if \( s \) denotes a strategy (e.g., a rewrite rule) developed...
for a particular problem domain, then the strategic expression \( TDL(s) \) denotes a strategy that will traverse a structure from top to bottom and attempt to apply \( s \) at every point along the way.

Strategic systems typically will also provide several primitive combinators capable of controlling rule application. The two most common such combinators are those that enable the sequential and conditional composition of rewrite rules to be expressed. The semi-colon is typically used to denote the sequential composition combinator and the symbol \( \triangleleft \) is typically used to denote the left-biased conditional combinator. Let \( s_1 \) and \( s_2 \) denote two strategies. The strategy \( s_1 \triangleleft s_2 \) denotes the left-biased conditional composition of \( s_1 \) and \( s_2 \). When applied to a term \( t \), the strategy \( s_1 \triangleleft s_2 \) will behave like \( s_1 \) if \( s_1 \) can be successfully applied to \( t \), otherwise it will behave like \( s_2 \).

3 Data Distribution

Strategic programming systems [2][6][7] have been successful in a number of software development areas and are being applied to increasingly complex activities. The distributed data problem (DDP) is common [12] and occurs when data found in one section of a program structure must be passed to another section. Accumulation is a standard approach taken in first-order systems to solve instances of the DDP. This approach involves the creation of auxiliary structures such as lists to hold data as well as accompanying lookup functions to extract data from such lists. Parameterization is then commonly used as the mechanism to transport these lists to the appropriate portions of a program.

A novel approach to the DDP is based on higher-order strategies [12][13]. The idea is to dynamically produce strategies that contain accumulated data rather than collect the data in auxiliary structures. The advantage of this approach is that data can be distributed via the application of strategies to terms (which is a primitive operation in a strategic framework). Thus, higher-order strategies provide an elegant approach for solving instances of the distributed data problem.

The strategic language TL [12][13] was developed to explore the implications of higher-order strategies. HATS is an IDE for strategic programming that implements a restricted dialect of TL. All examples presented here have been implemented in HATS. HATS runs on Windows NT/2000/XP and Unix-based platforms and is freely available [4].

The motivating application for this work is the Sandia Secure Processor (SSP), a subset of the Java Virtual Machine (JVM) developed at Sandia National Laboratories for use in high-consequence embedded applications. An application program for the SSP is called a ROM image and consists of a collection of class file-like structures stored on a read-only memory. In the SSP, all the structures used during execution must be present in the ROM image prior to execution. This enables class loading activities of the SSP to be performed statically, prior to execution. Under these conditions, the functionality of the class loader is well-suited to a strategic implementation.

In the discussion that follows, we assume that an application consists of one or more Java class files having the structure defined in Lindholm and Yellin [8]. For the purposes of this discussion it is important to know that class files contain:

1. A class entry denoting the name of the class;
2. A constant pool whose entries contain a full description of the fields that are explicitly used within the class. This description forms a key that is unique within an application; and
3. A fields section containing all of the fields, both static and instance, declared within the class.

The remainder of the paper is as follows: Section 4 gives some basics of rewriting. In Section 5 the table-entry distribution problem is introduced as our running example. Section 6 picks up the table-entry distribution problem for the SSP. Section 7 touches on our research in verification. Section 8 concludes.

4 Rewriting

The examples presented in this paper are written in TL [12]. TL is a higher-order strategic programming language for transforming parse trees. Thus, in the context of TL, we use the words “structure”, “tree”, and “term” to denote a parse tree. For a given grammar, the notation \( B[\alpha'] \) denotes a parse tree corresponding to the derivation \( B \Rightarrow \alpha' \) where the nonterminals in \( \alpha \) have been subscripted yielding \( \alpha' \). Consider the BNF grammar production defining the structure of a simplified entry in a Java constant pool:

\[
\text{cp_entry} ::= \text{class field type}
\]

A term corresponding to a constant pool entry having this form would be written as follows:

\[
\text{cp_entry}[\text{class1 field1 type1}]
\]

Similarly, consider the BNF grammar production defining a greatly simplified Java class file structure consisting of a class name, a constant pool and a fields section, as with:

\[
\text{classfile} ::= \{\text{class constant_pool field_section}\}
\]

A term corresponding to such a class structure would be written as follows:

\[
\text{classfile}[\{\text{class1 constant_pool1 field_section1}\}]
\]

In TL, a first-order rewrite rule can have the form:

\[
r : \text{lhs} \rightarrow \text{rhs}
\]

where \( \text{lhs} \) and \( \text{rhs} \) denote terms and \( r \) is an optional label. The expression \( r(t) \) denotes the application of the strategy
\( \text{lhs} \to \text{rhs} \) to the term \( t \). We will only consider the application of strategies to ground terms (i.e., terms containing no subscripted nonterminal symbols in leaf positions). When expressing strategy application, application is a left-associative implicit operator and parenthesis can be used to override precedence or to enhance readability. For example, \( r t \) denotes the application of \( r \) to \( t \) and has the same meaning as \( r(t) \).

A second-order rewrite rule has the form: \( \text{lhs}_2 \to \text{rhs}_2 \), where the \( \to \) symbol is right-associative. The application of the second order strategy \( \text{lhs}_2 \to \text{rhs}_2 \) to a term \( t \) will yield a first-order strategy of the form \( \text{lhs}_1' \to \text{rhs}_1' \) where \( \text{lhs}_1' \) and \( \text{rhs}_1' \) are instances respectively of \( \text{lhs}_1 \) and \( \text{rhs}_1 \) as determined by variable bindings resulting from the (successful) match between \( \text{lhs}_2 \) and \( t \).

Consider the abstract grammar shown in Table 1. Given this grammar, let us consider the second-order rewrite rule \( s^2 \) shown below:

\[
\begin{align*}
s^2: \text{g}[i_1, \text{data}_1] & \to \text{g}[i_2, i_1] \to \text{g}[i_2, \text{data}_1] \\
\end{align*}
\]

Informally speaking, the rule \( s^2 \) can be seen as a template for relating information between the terms \( \text{data}_1 \) and \( i_1 \) in a specific context. The application \( s^2(\text{g}[1, b], \text{g}[2, b]) \) yields the first-order strategy \( \text{g}[i_2, 1] \to \text{g}[i_2, b] \), and the application \( (\text{g}[i_2, 1] \to \text{g}[i_2, b])(\text{g}[2, 1]) \) yields \( \text{g}[2, b] \). In this instance, \( s^2 \) provides a vehicle for transferring data from \( \text{g}[1, b] \) to \( \text{g}[2, 1] \).

\[
\begin{align*}
g & ::= \mathrm{i \ data} \\
\mathrm{data} & ::= \mathrm{i \ | \ char} \\
i & ::= \mathrm{integer} \\
\end{align*}
\]

Table 1. A small abstract grammar

5 Table-Entry Distribution

The table-entry distribution problem involves the distribution of table entry data between a set of tables. Though presented in an idealized form here, this problem is interesting because field distribution within the class loading phase of the JVM can be seen as an instance of table-entry distribution.

Let \( T \) denote a a two-column table whose entries are of the form \((\text{key}, \text{data})\) where \( \text{data} \) is either a dash or a value of type integer. Given an entry \((k, d)\) if \( d \) is of type integer, then we say that the entry \((k, d)\) is resolved. Otherwise, \( d \) is a dash and the entry is unresolved.

**Definition 1** A table \( T \) is well-formed and only if its key space is unique.

**Definition 2** Given a set of well-formed tables \( S_T = \{T_1, T_2, ..., T_n\} \), a distribution step for \( S_T \) updates exactly one unresolved table entry in \( S_T \). The update causes the table entry to become resolved.

**Definition 3** A set of tables \( S_T \) is well-formed if and only if (1) every table in \( S_T \) is well-formed, and (2) every table entry in \( S_T \) can be resolved.

Given these definitions, the distribution step sequences are confluent and terminating for well-formed table sets. The term distribution-set refers to a set of ground rewrite rules capable of normalizing a given table set. Table 2 shows a well-formed table set \( S_T \) consisting of the three tables. The normal form of \( S_T \) is shown in Table 3. Table 4 shows a distribution-set for Table 2. The fixed point application of the distribution-set shown in Table 4 to the table set in Table 2 yields the table in Table 3.

From the perspective of correctness, we claim that distribution-sets, as shown in Figure 4, provide an attractive strategic solution to the table-entry distribution problem. This is primarily due to the direct relationship between the rules in the distribution-set and the definition of what it means to perform a distribution step. However, a drawback of the approach described is that it is highly

\[
\begin{align*}
r_1: (x_A, -) & \to (x_A, 1) \\
r_2: (y_A, -) & \to (y_A, 2) \\
r_3: (z_A, -) & \to (z_A, 3) \\
r_4: (x_B, -) & \to (x_B, 10) \\
r_5: (y_B, -) & \to (y_B, 20) \\
r_6: (z_B, -) & \to (z_B, 30) \\
r_7: (x_C, -) & \to (x_C, 100) \\
r_8: (y_C, -) & \to (y_C, 200) \\
r_9: (z_C, -) & \to (z_C, 300) \\
\end{align*}
\]

Table 4. A distribution-set
The field distribution problem, i.e., distributing field offset information among the constant pool entries in a Java application, is an instance of the table-entry distribution problem. A simplified version of the problem is presented here. We assume that the structure of Java class files has been simplified as defined by the grammar in Table 5; and all fields are associated strictly with offset addresses. We further assume that field entries in the constant pool table have been resolved (by an earlier transformation) into keys and that field entries in the field section table have been resolved into data. Field distribution can be realized by the higher-order strategies shown in Table 9.

Table 6 shows an application consisting of three simplified class files for which the constant pool entries are in resolved form and offsets have been computed for field entries in every field section table. The classes are named A, B, and C. The constant pool of A contains three field keys C, B, and A. The class A locally declares the fields x, y, and z. The computed hexadecimal offsets for these fields are :0004, :000C, and :0014 respectively.

Consider the field entry x : 0004 belonging to the field section of A and the key A x long belonging to constant pool of C. In this context, a distribution step is defined as A x long → :0004 long.

The higher-order solution to field distribution is realized using a few rewrite rules and two generic strategies: BUL and lcond_bul. Recall that the strategy BUL performs a bottom-up left-to-right traversal of a term structure. The expression BUL(s) t will result in the first-order rewrite s being applied to every sub-term encountered during the traversal of term t.

The strategy lcond_bul is similar to BUL except that lcond_bul will apply a higher-order strategy to the terms it encounters. The successful application of a higher-order strategy will not yield a term, but rather will produce another strategy whose order is one less than the strategy being applied. For example, let s^2 denote a second-order rewrite rule and let t denote a term. The successful application of s^2 resulting from the evaluation of lcond_bul(s^2)t produces a set of first-order strategies {s_1, s_2, ..., s_n}. However, the higher-order strategy framework of TL goes beyond this and enables the resulting (first-order) strategies to be composed using a binary strategic combinator. The generic strategy lcond_bul has been defined in such a way that it will use the conditional combinator, denoted by the symbol ◦ to compose the strategies produced. Thus, the resulting strategic expression is s_1 ◦ s_2 ◦ ... ◦ s_n.

Let us now consider the evaluation of the strategic expression: lcond_bul FIELD field_section_1 where

\[
\text{FIELD} = \{ f_{\text{entry}}[f_{\text{field_1}} \text{offset_1}] \\
\qquad \rightarrow cp_{\text{entry}}[A \text{ field_1 type_1}] \\
\qquad \rightarrow cp_{\text{entry}}[\text{offset_1 type_1}]
\]

and field_section_1 is the field section table belonging to the class file A. This strategic expression is taken from the Class and Field strategies defined in Table 9. Under the given conditions, the result of evaluation will yield the first-order strategy shown in Table 8.

Table 8. A field distribution-set strategy for the class file A

Given the appropriate instantiations of FIELD and field_section_1, the strategic expression lcond_bul FIELD field_section_1 can be used to generate field distribution-set strategies for the class files B and C. When taken collectively, these strategies form a distribution-set strategy for the Java application under consideration.

7 Assurance

Sandia requires that the development of the SSP produces strong evidence of the correctness of the system. The class loader is a weak link in the assurance chain of the SSP. Considerable effort is being devoted towards providing assurance in the translation performed by the SSP class loader.
Table 6. A simplified application consisting of the class files A, B, and C

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Constant Pool</th>
<th>Field Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>{ C x int, B y byte, A z long }</td>
<td>{ x:0004, y:0005, z:0014 }</td>
</tr>
<tr>
<td>B</td>
<td>{ A y long, B y byte, C z int }</td>
<td>{ x:0004, y:0005, z:0006 }</td>
</tr>
<tr>
<td>C</td>
<td>{ A x long, C y int, B z byte }</td>
<td>{ x:0004, y:0008, z:000C }</td>
</tr>
</tbody>
</table>

Table 7. Application after field distribution

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Constant Pool</th>
<th>Field Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>{ x:0004, y:0005, z:0014 }</td>
<td>{ x:0004, y:000C, z:0014 }</td>
</tr>
<tr>
<td>B</td>
<td>{ x:0004, y:0005, z:0006 }</td>
<td>{ x:0004, y:0005, z:0006 }</td>
</tr>
<tr>
<td>C</td>
<td>{ x:0004, y:0008, z:000C }</td>
<td>{ x:0004, y:0008, z:000C }</td>
</tr>
</tbody>
</table>

Formal reasoning techniques are being used to verify general properties of the class loader. The assurance provided by formal reasoning comes in the form of a mathematical proof and typically involves a model of the system under analysis rather than the actual system itself. General properties are stated in terms of theorems involving the model. The proof of theorems provides strong assurance that the model behaves as required. Assurance of the correctness of the system under analysis relies on (1) confidence that the proofs themselves are sound, and (2) confidence that the model faithfully describes the system (e.g., theorems that hold for the model hold for the system).

The modeling and verification framework provided by ACL2 [5] is used to prove theorems about the class loader. ACL2 is a programming language based on the applicative subset of Common Lisp. In this language, users can build executable models of software systems. ACL2 is also a tool that assists users in proving theorems about their ACL2 programs. It has been used to prove the correctness of hardware implementations of microprocessors and floating-point algorithms [1] and more recently to verify parts of implementations of the JVM [1][9].

The class loader is modelled as a system described in terms of states and state transitions. In the context of the class loader, we model the class files of the Java application as our state and the transformations on this application as our state transitions. The JVM and the SSP provide the basis for formally understanding equivalence between states.

A model of TL is constructed by defining an abstract machine that controls the application of transformation rules. Each transformation rule modifies the state of the system. The abstract machine operates according to the following sequence: fetch the next transformation rule and node from the current state, apply the transformation, and return a new machine state.

Though we are exploring the verification of a number of properties of the class loader, our ultimate goal is to verify that the transformation rules preserve the meaning of the term to which they are applied (i.e., the class loader is correct). In the context of the SSP, the initial term is a set of class files, C₀, generated by a Java compiler. The semantics of this term is defined by the JVM specification [8]. Let \( \text{Eval}_{JVM} : \text{classfiles} \times \text{inputs} \rightarrow \text{outputs} \) denote the mapping from classfiles and inputs to outputs. \( \text{Eval}_{JVM} \) defines the behavior of the program encoded in the classfiles. The final term is a ROM image, which we denote \( C_{ROM} \). The semantics of this term is defined by the SSP hardware, \( \text{Eval}_{SSP} \). HATS accomplishes the conversion of \( C_0 \) to \( C_{ROM} \), as indicated by the notation \( C_{ROM} = T^*(C_0) \), where \( T^* \) denotes a sequence of transformations. In this notational framework, what must be shown for inputs I is:

\[
\forall(C_0, I)\text{Eval}_{JVM}(C_0, I) = \text{Eval}_{SSP}(T^*(C_0), I)
\]

The problem above can be decomposed by defining a sequence of normal forms, \( C_0, C_1, C_2, ... \) in the transformation of \( C_0 \) to \( C_{ROM} \). These normal forms are formally specified and are theorems within our verification framework. Constant pool normalization and field distribution are two examples of normal forms. Informally stated, in constant pool normalization all indirection is removed from the constant pool entries of the classfiles in \( C_0 \). Let \( T^1 \) denote the transformation that accomplishes this task. Similarly, let \( T^2 \) denote the normal form resulting from field distribution. At present, a sequence of five intermediate normal forms have been defined. For each normal form, there is an evaluation function, \( \text{Eval}_n \). Thus, the original correctness conjecture can be restated as a sequence of conjectures:

\[
\forall(C_0, I)\text{Eval}_{JVM}(C_0, I) = \text{Eval}_{1}(T^1(C_0), I) = \text{Eval}_{2}(T^2(T^1(C_0), I)) = \text{Eval}_{SSP}(T^*(C_0), I)
\]

where \( T^* \) is the composition of the individual transformations. This allows the proof to be constructed incrementally, and therefore, reduces the complexity of the proof.

To date, a class loader for the SSP has been completely implemented using HATS. Because Sandia Na-
tional Laboratories places a high priority on the provision of strong evidence that the system behaves correctly, attention is being turned to the verification of the transformations used. ACL2 proofs of a simplified version of HATS have been completed for the first two canonical forms. Using this as a basis, a complete model of HATS will be implemented and the remaining canonical forms verified.

8 Conclusion

In practice, program transformation oftentimes requires data to be distributed throughout a program structure. TL is a strategic programming language in which higher-order rules and strategies provide a mechanism for distributing data. Conceptually, the functionality of a higher-order strategy can be understood in terms of the first-order rules they ultimately produce. In the case of the table-entry distribution problem, a higher-order strategy can be used to construct distribution-sets which, in turn, can be used to resolve table entries. A similar approach can be used to solve the field distribution problem within a Java class loader.

The ability of higher-order strategies to automatically generate first-order strategies leads to elegant solutions to instances of the distributed data problem. In turn, this elegance can benefit efforts to formally verify the correctness of the strategies. Formal verification is a technique that can be used to provide strong evidence of the correctness of a software implementation such as a class loader. Such a level of assurance is important when considering high-consequence systems such as the SSP.

References


