

An SMT-Based Concolic Testing Tool for Logic Programs (System Description)

Sophie Fortz¹, Fred Mesnard², Etienne Payet², Gilles Perrouin¹, Wim Vanhoof¹, and German Vidal³

¹ Université de Namur, Belgique

² LIM - Université de la Réunion, France

³ MiST, DSIC, Universitat Politècnica de València

Abstract. Concolic testing mixes symbolic and concrete execution to generate test cases covering paths effectively. Its benefits have been demonstrated for more than 15 years to test imperative programs. Other programming paradigms, like logic programming, have received less attention. In this paper, we present a concolic-based test generation method for logic programs. Our approach exploits SMT-solving for constraint resolution. We then describe the implementation of a concolic testing tool for Prolog and validate it on some selected benchmarks.

Keywords: Concolic Testing · Logic Programming · SAT/SMT solving.

1 Introduction

Concolic testing is a well-established validation technique for imperative and object-oriented programs [5,6,12]. However, it has been less explored in the context of functional and logic programming languages. This is particularly unfortunate since it is becoming increasingly popular to *compile* programs in other programming languages to either a functional or a logic formalism. The main advantage of this approach is that so-called declarative programs have a simpler semantics and are thus much more appropriate for analysis, verification, validation, etc. In particular, Constraint Horn clauses (i.e., logic programs with constraints) are gaining popularity as an intermediate language for verification engines (see, e.g., [2,4]). In this context, it is unfortunate that there are no powerful concolic testing tools for logic programming languages. We thus aim at improving this situation with the design and implementation of an efficient SMT-based concolic testing tool for logic programs.

A notable exception in this context is the work on concolic testing developed by Mesnard, Payet and Vidal [14,9], later extended in [10,11]. Let us illustrate this approach with a simple example.

Example 1. Consider the following logic program:

$$\begin{array}{ll} (\ell_1) & p(a). \\ (\ell_2) & p(s(X)) \leftarrow q(X). \\ (\ell_3) & q(a). \end{array}$$

where ℓ_1, ℓ_2, ℓ_3 are (unique) clause labels.

The notion of *coverage* in [9] requires considering test cases so that the clauses defining each predicate are unified in all possible ways. For instance, in this example, we aim at producing a call to predicate $p/1$ that unifies with no clause, another one that unifies only with the head of ℓ_1 , another one that unifies only with the head of ℓ_2 , and another one that unifies with both the heads of ℓ_1 and ℓ_2 .

For this purpose, an (iterative) process starts with an arbitrary test case (an atomic goal), e.g., $p(a)$. We then evaluate in parallel both $p(a)$ (the *concrete* goal) and $p(X)$ (the *symbolic* goal), where X is a fresh variable, using a concolic execution extension of SLD resolution. Two key points of concolic execution are that not all nondeterministic executions of $p(X)$ are explored (but only those that corresponds to the executions of the concrete goal $p(a)$), and that we record the clauses unifying with both the concrete and the symbolic goal at each resolution step (in order to compute alternative test cases).

Now, in order to look for alternative test cases, [9] introduces the notion of *selective unification* problem: given an atom A , sets of clauses H^+ and H^- , and a set of variables $G \subseteq \text{Var}(A)$, find a substitution σ such that i) $A\sigma$ unifies with the heads of the clauses in H^+ , ii) it does not unify with the heads of the clauses in H^- , and iii) $G\sigma$ becomes ground. In this example, $p(a)$ only unifies with clause ℓ_1 , while $p(X)$ unifies with both ℓ_1 and ℓ_2 . Therefore, one now considers the following missing cases:

- an instance $p(X)\sigma$ of $p(X)$ for some substitution σ such that $p(X)\sigma$ unifies no clause, e.g., $p(b)$;
- another instance that unifies both ℓ_1 and ℓ_2 (unfeasible if we want the argument of p to be ground);
- and one more instance that unifies only ℓ_2 , e.g., $p(s(a))$.

Each case is formalized as a selective unification problem and then solved using a specific algorithm.⁴ Now, we add the two new test cases, $p(b)$ and $p(s(a))$, and then repeat the process until no new test cases are added.

We note that, in the example above, only *positive* constraints are considered (which are denoted by means of substitutions). However, *negative* constraints (such as, e.g., $X \neq a$) cannot be expressed in the framework of [9].

Actually, the approach in [9] suffers from some limitations. On the one hand, the algorithms for solving the unification problems above are computationally very expensive, which makes this approach impractical for large programs. On the other hand, the process is, in some cases, unnecessarily incomplete because of the lack of negative information, as witnessed by the following example:

Example 2. Consider now the following logic program:

$$\begin{array}{ll} (\ell_1) & p(a). \\ (\ell_2) & p(X) \leftarrow q(X). \\ (\ell_3) & q(b). \end{array}$$

⁴ More details on selective unification can be found in [10,11].

Given the initial call $p(a)$, the first alternative test case computed by the approach of [9] is $p(b)$ which only unifies the second clause. Now, $p(b)$ is unfolded to $q(b)$, which succeeds. The next computed test case is then $p(X)\sigma$ where σ binds X to any term different from b so that $q(X)\sigma$ fails (since test cases for failures are also required). However, it does not take into account the fact that X must be different to a in $p(a)$ in order to reach $q(a)$ since negative constraints cannot be represented within the framework of [9]. Hence, one could generate $p(a)$ so that concolic testing stops because $p(a)$ was already considered. The generated test cases are then $p(a)$ and $p(b)$. With the intended coverage definition, though, one would also expect a test case like $p(c)$ which is first unfolded using the second clause and then fails. Therefore, the concolic testing framework of [9] is unnecessarily incomplete here.

In this paper, we design an improved concolic testing scheme that is based on [9] but adds support for negative constraints (so that the source of incompleteness shown in the example above is removed) and defines selective unification problems as constraints on Herbrand terms, so that an efficient SMT solver can be used (in contrast to [10,11]). We have implemented an SMT-based concolic testing tool for Prolog based on this design, where the SMT solver Z3 [3] is used to solve selective unification problems. A preliminary experimental evaluation has been conducted, which shows encouraging results in terms of execution time and scalability.

2 A Deterministic Operational Semantics

In this section, we recall a *deterministic* operational semantics for *definite* logic programs (i.e., logic programs without negation [8]). In particular, we consider the semantics in [9] which, in turn, follows the *local* operational semantics of [13], where backtracking is dealt with explicitly. Moreover, the semantics only considers the computation of the first answer for the initial goal. This is a design decision motivated by the fact that Prolog programs are often used in this way, so that one can measure the achieved coverage in a realistic way.

We refer the reader to [1] for the standard definitions and notations for logic programs. The semantics is defined by means of a transition system on *states* of the form $\langle \mathcal{B}_{\delta_1}^1 \mid \dots \mid \mathcal{B}_{\delta_n}^n \rangle$, where $\mathcal{B}_{\delta_1}^1 \mid \dots \mid \mathcal{B}_{\delta_n}^n$ is a sequence of goals labeled with substitutions (the answer computed so far, when restricted to the variables of the initial goal). We denote sequences with S, S', \dots , where ϵ denotes the empty sequence. In some cases, we label a goal \mathcal{B} both with a substitution and a program clause, e.g., $\mathcal{B}_{\delta}^{H \leftarrow \mathcal{B}}$, which is used to determine the next clause to be used for an SLD resolution step (see rules *choice* and *unfold* in Figure 1). Note that the clauses of the program are not included in the state but considered as global parameters since they are static. In the following, given an atom A and a logic program P , $\text{clauses}(A, P)$ returns the sequence of renamed apart program clauses c_1, \dots, c_n from P whose head unifies with A . A syntactic object s_1 is *more general* than a syntactic object s_2 , denoted $s_1 \leq s_2$, if there exists

$$\begin{array}{l}
\text{(success)} \quad \overline{\langle \text{true}_\delta | S \rangle} \rightarrow \langle \text{SUCCESS}_\delta \rangle \\
\text{(failure)} \quad \overline{\langle (\text{fail}, \mathcal{B})_\delta \rangle} \rightarrow \langle \text{FAIL}_\delta \rangle \qquad \text{(backtrack)} \quad \frac{S \neq \epsilon}{\langle (\text{fail}, \mathcal{B})_\delta | S \rangle \rightarrow \langle S \rangle} \\
\text{(choice)} \quad \frac{\text{clauses}(A, P) = (c_1, \dots, c_n) \wedge n > 0}{\langle (A, \mathcal{B})_\delta | S \rangle \rightarrow \langle (A, \mathcal{B})_\delta^{c_1} | \dots | (A, \mathcal{B})_\delta^{c_n} | S \rangle} \quad \text{(choice_fail)} \quad \frac{\text{clauses}(A, P) = \{\}}{\langle (A, \mathcal{B})_\delta | S \rangle \rightarrow \langle (\text{fail}, \mathcal{B})_\delta | S \rangle} \\
\text{(unfold)} \quad \frac{\text{mgu}(A, H_1) = \sigma}{\langle (A, \mathcal{B})_\delta^{H_1 \leftarrow \mathcal{B}_1} | S \rangle \rightarrow \langle (\mathcal{B}_1 \sigma, \mathcal{B} \sigma)_{\delta \sigma} | S \rangle}
\end{array}$$

Fig. 1. Concrete semantics

a substitution θ such that $s_1\theta = s_2$. $\mathcal{V}ar(o)$ denotes the set of variables of the syntactic object o . For a substitution θ , $\mathcal{V}ar(\theta)$ is defined as $\text{Dom}(\theta) \cup \text{Ran}(\theta)$, where Dom and Ran return the variables in the domain and range of a given substitution, respectively.

For simplicity, w.l.o.g., we only consider *atomic* initial goals. Therefore, given an atom A , an initial state has the form $\langle A_{id} \rangle$, where id denotes the identity substitution. The transition rules, shown in Figure 1, proceed as follows:

- In rules **success** and **failure**, we use constant SUCCESS_δ to denote that a successful derivation ended with computed answer substitution δ , while FAIL_δ denotes a finitely failing derivation; recording δ for failing computations might be useful for debugging purposes.
- Rule **backtrack** applies when the first goal in the sequence finitely fails, but there is at least one alternative choice.
- Rule **choice** represents the first stage of an SLD resolution step. If there is at least one clause whose head unifies with the leftmost atom, this rule introduces as many copies of a goal as clauses returned by function **clauses**. If there is at least one matching clause, unfolding is then performed by rule **unfold**. Otherwise, if there is no matching clause, rule **choice_fail** returns **fail** so that either rule **failure** or **backtrack** applies next.

Example 3. Consider the following logic program:

$$\begin{array}{lll}
p(a). & q(a). & r(a). \\
p(s(X)) \leftarrow q(X). & q(b). & r(c). \\
p(f(X)) \leftarrow r(X). & &
\end{array}$$

Given the initial goal $p(s(X))$, we have the following successful computation (for clarity, we label each step with the applied rule):⁵

$$\begin{aligned} \langle p(s(X))_{id} \rangle &\rightarrow_{\text{choice}} \langle p(s(X))_{id}^{p(s(X)) \leftarrow q(X)} \rangle \rightarrow_{\text{unfold}} \langle q(X)_{id} \rangle \\ &\rightarrow_{\text{choice}} \langle q(X)_{id}^{q(a)} \mid q(X)_{id}^{q(b)} \rangle \rightarrow_{\text{unfold}} \langle \text{true}_{\{X/a\}} \mid q(X)_{id}^{q(b)} \rangle \\ &\rightarrow_{\text{success}} \langle \text{SUCCESS}_{\{X/a\}} \rangle \end{aligned}$$

Therefore, we have a successful computation for $p(s(X))$ with computed answer $\{X/a\}$. Observe that only the first answer is considered.

3 An SMT-Based Concolic Testing Procedure

In this section, we present our scheme to concolic testing of logic programs. Our concolic testing semantics performs both concolic execution and test case generation, which contrasts to [9] where they are kept as two independent stages. Moreover, it collects both positive and negative constraints on the input variables, so that it overcomes an important limitation of previous approaches (as explained in Section 1) and opens the door to making the overall process much more efficient by using a state-of-the-art SMT solver.

3.1 Auxiliary Functions and Notations

Let us first introduce some auxiliary definitions and notations which are required in the remainder of this section.

In the following, we let \bar{o}_n denote the sequence of syntactic objects o_1, \dots, o_n ; we also write \bar{o} when the number of elements is not relevant. Given an atom A , we let $\text{root}(A) = p/n$ if $A = p(\bar{t}_n)$. We also assume that every clause c has a corresponding unique label, which we denote by $\ell(c)$. By abuse of notation, we also denote by $\ell(\bar{c}_n)$ the set of labels $\{\ell(c_1), \dots, \ell(c_n)\}$. Moreover, we let $\text{hd}(H \leftarrow \mathcal{B}) = H$ and $\text{hd}(C) = \{\text{hd}(c) \mid c \in C\}$, where $H \leftarrow \mathcal{B}$ is a clause and C is a set of clauses.

We now introduce the auxiliary functions *neg_constr* and *alts*. First, function *neg_constr* is used to compute some negative constraints which will become useful to avoid the problem shown in Example 2.

Definition 1 (*neg_constr*). *Let A be an atom, $G \subseteq \text{Var}(A)$ a set of variables, and $\{\bar{H}_n\}$, $n > 0$, a set of atoms with $\text{Var}(\{\bar{H}_n\}) \cap \text{Var}(A) = \{\}$. Then,*

$$\text{neg_constr}(A, \{\bar{H}_n\}, G) = \forall \bar{X}_{k_1} A \neq H_1 \wedge \dots \wedge \forall \bar{X}_{k_n} A \neq H_n$$

where $\bar{X}_{k_i} = (\text{Var}(A) \setminus G) \cup \text{Var}(H_i)$, $i = 1, \dots, n$.

⁵ Note that a fact like “ $q(a)$.” is equivalent to a rule “ $q(a) \leftarrow \text{true}$.” and, thus, the unfolding of $q(X)$ returns **true** in the considered derivation.

For example, we have

$$\text{neg_constr}(p(X, Y), \{p(a, W)\}, \{X\}) = \forall Y, W \ p(X, Y) \neq p(a, W)$$

Function *alts* is used to encode a selective unification problem using both positive and negative constraints: Intuitively, given a call of the form $\text{alts}(A_0, \gamma, A', \mathcal{B}, \mathcal{B}', G)$, we are interested in new test cases that unify with each atom in each set of $\mathcal{P}(\mathcal{B}')$ ⁶ except for the set \mathcal{B} which is already *covered* by the current concrete goal (i.e., we do not want to produce an alternative test case that matches exactly the same clauses as the current test case that we are executing). Then, for each set $H^+ \in \mathcal{P}(\mathcal{B}')$ (the *positive* clauses) with $H^- = \mathcal{B}' \setminus H^+$ (the *negative* clauses), we look for a substitution σ such that $A'\sigma$ unifies with the atoms in H^+ but it does not unify with the atoms in H^- , while still grounding the variables in G . For each such substitution, we produce a new test case $A_0\sigma$. Formally,

Definition 2 (*alts*). *Let A_0, A' be atoms, $\mathcal{B}, \mathcal{B}'$ sets of atoms with $\text{Var}(\mathcal{B}, \mathcal{B}') \cap \text{Var}(A') = \{\}$, γ a (negative) constraint, and $G \subseteq \text{Var}(A_0)$ a set of variables. Then, we have*

$$\text{alts}(A_0, \gamma, A', \mathcal{B}, \mathcal{B}', G) = \left\{ A_0\sigma \left| \begin{array}{l} H^+ \in \mathcal{P}(\mathcal{B}'), \ H^+ \neq \mathcal{B}, \\ H^- = \mathcal{B}' \setminus H^+, \\ \alpha(A', \gamma, H^+, H^-, G) = \sigma \end{array} \right. \right\}$$

where function $\alpha(A', \gamma, H^+, H^-, G)$ returns a solution to the following constraint (represented as a substitution):⁷

$$\left(\begin{array}{l} \gamma \wedge \exists \overline{X}_{n_1}(A' = H_1) \wedge \dots \wedge \exists \overline{X}_{n_j}(A' = H_j) \\ \wedge \forall \overline{Y}_{n_1}(A' \neq H'_1) \wedge \dots \wedge \forall \overline{Y}_{n_k}(A' \neq H'_k) \end{array} \right)$$

with $H^+ = \{\overline{H}_j\}$, $H^- = \{\overline{H}'_k\}$, $\overline{X}_{n_i} = (\text{Var}(A') \setminus G) \cup \text{Var}(H_i)$, $i = 1, \dots, j$, and $\overline{Y}_{n_i} = (\text{Var}(A') \setminus G) \cup \text{Var}(H'_i)$, $i = 1, \dots, k$. Here, we look for a solution for the free variables of the above constraint: $\text{Var}(A') \cap G$.

For example, we have

$$\text{alts}(p(X), p(X) \neq p(a), q(X), \{\}, \{q(b)\}, \{X\}) = \{p(c)\}$$

since

$$\alpha(q(X), p(X) \neq p(a), \{\}, \{q(b)\}, \{X\}) = \exists X(p(X) \neq p(a) \wedge q(X) \neq q(b))$$

and the selected solution is $X = c$ represented as a substitution $\{X/c\}$, where c is an arbitrary constant which is different from the previous ones (a and b).

⁶ Here, we denote by $\mathcal{P}(C)$ the powerset of a set C .

⁷ Function α returns an arbitrary solution when the considered constraint is satisfiable and fails otherwise.

3.2 Concolic Testing Semantics

Let us now consider our concolic testing semantics. In this work, we consider that the initial goal is terminating for a given *mode*, which is a reasonable assumption since non-terminating test cases are not very useful in practice. Essentially, a mode is a function that labels as “input” or “output” the arguments of a given predicate, so that input arguments are assumed to be ground at call time, while output arguments are usually unbound (see, e.g., [1]). Here, we assume a fixed *mode* for the predicate in the initial (atomic) goal. Different modes can also be considered by performing concolic testing once for each mode of interest. Therefore, our test cases should make the input arguments ground (to ensure terminating concrete executions). For clarity, we assume in the remainder of this paper that all input arguments (if any) are in the first consecutive positions of an atom. Moreover, given a predicate p/n with input arguments $\{1, \dots, m\}$, $m \leq n$, we let $\text{var}_{in}(p(\bar{t}_n)) = \text{Var}(t_1) \cup \dots \cup \text{Var}(t_m)$, i.e., the variables in the input positions, and $\text{out}_{vars}(p(\bar{t}_n)) = p(\bar{t}_m, X_{m+1}, \dots, X_n)$, i.e., the atom that results from replacing the output arguments by distinct fresh variables.

In the concolic testing semantics, *concolic states* have now the form $\langle S \parallel S' \rangle$, where S and S' are sequences of (possibly labeled) concrete and symbolic goals, respectively. In the context of logic programming, the notion of *symbolic* execution is very natural: the structure of both S and S' is the same, and the only difference (besides the labeling) is that some atoms might be less instantiated in S' than in S . To be precise, symbolic goals in a concolic state are now labeled as follows: $\mathcal{B}_{\sigma, \pi, A_0, \gamma, G}$, where

- σ is the substitution computed so far;
- π is the current trace⁸ (which is required to avoid considering the same computations once and again);
- A_0 is the initial atomic goal (which is required to compute alternative test cases in function *alts*);
- γ is a (negative) constraint that allows us to avoid unification with the heads of some clauses (to avoid the problem shown in Example 2);
- G is the set of variables that must be ground in test cases (i.e., the variables in the *input* arguments of the initial atomic goal).

Our concolic testing semantics considers two *global parameters* that are left implicit in the transition rules: *Traces* and *TestCases*. The first one, *Traces*, is used to store the already explored execution traces, so that we avoid computing the same test cases once and again. Here, a *trace* represents a particular execution by the sequence of clause labels used in the unfolding steps of this execution.

The second one, *TestCases*, stores the computed test cases, where the already processed test cases are distinguished by underlining them (i.e., \underline{A} means that the test case A has been already processed by the concolic testing semantics).

Concolic testing consists of an iterative process where *TestCases* is initialized with some arbitrary (atomic) goal, e.g., $\text{TestCases} = \{A\}$, *Traces* is initialized

⁸ Traces are sequences of clause labels, with ϵ the empty trace.

to the empty set, and it proceeds as follows:

```

repeat
  let  $p(\overline{t}_n, \overline{X}_m) \in TestCases$ 
   $TestCases \leftarrow (TestCases - \{p(\overline{t}_n, \overline{X}_m)\}) \cup \{p(\overline{t}_n, \overline{X}_m)\}$ 
  execute  $\langle p(\overline{t}_n, \overline{X}_m)_{id} \parallel p(\overline{Y}_{n+m})_{id, \epsilon, p(\overline{Y}_{n+m}), true, \{\overline{Y}_n\}} \rangle$ 
until all atoms in  $TestCases$  are underlined

```

where we assume that the first n arguments of p are its input arguments, and concolic states are executed using the semantics in Figure 2 (see below).

In general, though, this iterative process might run forever, even if all considered concolic executions are terminating.⁹ Essentially, our algorithm aims at full *path coverage*, so the required number of test cases is typically infinite. Therefore, in practice, one usually sets a bound in the number of iterations, a time limit, or a maximum term depth so that the domains of terms and atoms become finite. In the implemented tool, we consider the last approach (see Section 4).

Finally, let us briefly describe the rules of the concolic testing semantics shown in Figure 2:

Rules *success*, *failure*, *backtrack*, and *unfold* are straightforward extensions of the same rules in the concrete operational semantics of Figure 1.

As for rules *choice* and *choice_fail*, if we only look at the first component of concolic states, they are identical to their counterpart in Figure 1; indeed, the concolic testing semantics is a conservative extension of the standard operational semantics. Regarding the symbolic components, there are several notable differences:

- First, although the symbolic goal is only unfolded using the clauses matching with the concrete goal, we also determine the set of clauses matching the symbolic goal, \overline{d}_k . This information will be useful in order to compute alternative test goals (using the auxiliary function *alts*).
- Moreover, we update the current trace (from π to $\pi.\ell(c_i)$ in rule *choice*) and the negative constraint from γ to $\gamma \wedge \gamma'$, where γ' is used to ensure that the symbolic goal, A' , only matches the same clauses as the concrete goal A .
- Finally, observe how the global parameters $TestCases$ and $Traces$ are updated in these rules when $\pi \notin Traces$, i.e., when the considered execution path is considered for the first time (otherwise, we just continue the concolic execution without modifying $TestCases$ nor $Traces$). Note that we use `outvars` here to ensure that the output arguments are unbound.

3.3 Concolic Testing in Practice

Consider again the logic program from Example 2:

$$\begin{array}{ll}
 (\ell_1) & p(a). \\
 (\ell_2) & p(X) \leftarrow q(X). \\
 (\ell_3) & q(b).
 \end{array}$$

⁹ In principle, we assume that concrete goals are terminating and, thus, concolic executions are terminating too.

$$\begin{array}{l}
 \text{(success)} \frac{}{\langle \text{true}_\delta \mid S \parallel \text{true}_{\theta, \pi, A_0, \gamma, G} \mid S' \rangle \rightsquigarrow \langle \text{SUCCESS}_\delta \parallel \text{SUCCESS}_\theta \rangle} \\
 \text{(failure)} \frac{}{\langle (\text{fail}, \mathcal{B})_\delta \parallel (\text{fail}, \mathcal{B}')_{\theta, \pi, A_0, \gamma, G} \rangle \rightsquigarrow \langle \text{FAIL}_\delta \parallel \text{FAIL}_\theta \rangle} \\
 \text{(backtrack)} \frac{S \neq \epsilon}{\langle (\text{fail}, \mathcal{B})_\delta \mid S \parallel (\text{fail}, \mathcal{B}')_{\theta, \pi, A_0, \gamma, G} \mid S' \rangle \rightsquigarrow \langle S \parallel S' \rangle} \\
 \text{(choice)} \frac{\begin{array}{l} \text{clauses}(A, P) = \overline{c_n} \wedge n > 0 \wedge \text{clauses}(A', P) = \overline{d_k} \\ \wedge \gamma' = \text{neg_constr}(A', \text{hd}(\{\overline{d_k}\}) \setminus \text{hd}(\{\overline{c_n}\}), G) \end{array}}{\begin{array}{l} \langle (A, \mathcal{B})_\delta \mid S \parallel (A', \mathcal{B}')_{\theta, \pi, A_0, \gamma, G} \mid S' \rangle \\ \rightsquigarrow \langle (A, \mathcal{B})_\delta^{c_1} \mid \dots \mid (A, \mathcal{B})_\delta^{c_n} \mid S \\ \parallel (A', \mathcal{B}')_{\theta, \pi, \ell(c_1), A_0, \gamma \wedge \gamma', G} \mid \dots \mid (A', \mathcal{B}')_{\theta, \pi, \ell(c_n), A_0, \gamma \wedge \gamma', G} \mid S' \rangle \end{array}} \\
 \text{where} \\
 \text{TestCases} \leftarrow \text{TestCases} \cup \text{out_vars}(\text{alts}(A_0, \gamma, A', \text{hd}(\{\overline{c_n}\}), \text{hd}(\{\overline{d_k}\}), G)) \\
 \text{Traces} \leftarrow \text{Traces} \cup \{\pi\} \\
 \text{if } \pi \notin \text{Traces} \\
 \text{(choice_fail)} \frac{\text{clauses}(A, P) = \{\} \wedge \text{clauses}(A', P) = \overline{d_k} \wedge \gamma' = \text{neg_constr}(A', \text{hd}(\{\overline{d_k}\}), G)}{\langle (A, \mathcal{B})_\delta \mid S \parallel (A', \mathcal{B}')_{\theta, \pi, A_0, \gamma, G} \mid S' \rangle \rightsquigarrow \langle (\text{fail}, \mathcal{B})_\delta \mid S \parallel (\text{fail}, \mathcal{B}')_{\theta, \pi, A_0, \gamma \wedge \gamma', G} \mid S' \rangle} \\
 \text{where} \\
 \text{TestCases} \leftarrow \text{TestCases} \cup \text{out_vars}(\text{alts}(A_0, \gamma, A', \{\}, \text{hd}(\{\overline{d_k}\}), G)) \\
 \text{Traces} \leftarrow \text{Traces} \cup \{\pi\} \\
 \text{if } \pi \notin \text{Traces} \\
 \text{(unfold)} \frac{\text{mgu}(A, H_1) = \sigma \wedge \text{mgu}(A', H_1) = \rho}{\begin{array}{l} \langle (A, \mathcal{B})_\delta^{H_1 \leftarrow B_1} \mid S \parallel (A', \mathcal{B}')_{\theta, \pi, A_0, \gamma, G} \mid S' \rangle \\ \rightsquigarrow \langle (\mathcal{B}_1 \sigma, \mathcal{B} \sigma)_{\delta \sigma} \mid S \parallel (\mathcal{B}_1 \rho, \mathcal{B}' \rho)_{\theta \rho, \pi, A_0 \rho, \gamma \rho, \text{Var}(G \rho)} \mid S' \rangle \end{array}}
 \end{array}$$

Fig. 2. Concolic testing semantics

and the initial call $p(a)$, so that we initialize TestCases to $\{p(a)\}$ and, thus, start concolic testing with $\langle p(a)_{id} \parallel p(Y)_{id, \epsilon, p(Y), \text{true}, \{Y\}} \rangle$, where we assume that the argument of $p/1$ is an input argument and it is expected to be ground. Concolic testing then computes the following derivation:¹⁰

$$\begin{array}{l}
 \langle p(a)_{id} \parallel p(Y)_{id, \epsilon, p(Y), \text{true}, \{Y\}} \rangle \\
 \rightsquigarrow^{\text{choice}} \langle p(a)_{id}^{\ell_1} \mid p(a)_{id}^{\ell_2} \parallel p(Y)_{id, \ell_1, p(Y), \text{true}, \{Y\}}^{\ell_1} \mid p(Y)_{id, \ell_2, p(Y), \text{true}, \{Y\}}^{\ell_2} \rangle \\
 \rightsquigarrow^{\text{unfold}} \langle \text{true}_{id} \mid p(a)_{id}^{\ell_2} \parallel \text{true}_{id, \ell_1, p(Y), \text{true}, \{Y\}}^{\ell_1} \mid p(Y)_{id, \ell_2, p(Y), \text{true}, \{Y\}}^{\ell_2} \rangle \\
 \rightsquigarrow^{\text{success}} \langle \text{SUCCESS}_{id} \parallel \text{SUCCESS}_{id} \rangle
 \end{array}$$

Moreover, in the first step, we add ϵ to Traces and update TestCases as follows:

$$\begin{array}{l}
 \text{TestCases} \leftarrow \text{TestCases} \\
 \cup \text{out_vars}(\text{alts}(p(Y), \text{true}, p(Y), \{p(a), p(X)\}, \{p(a), p(X)\}, \{Y\}))
 \end{array}$$

¹⁰ In this example, we often use clause labels instead of the actual clauses for clarity.

In this case, $\mathcal{P}(\{p(a), p(X)\}) = \{\{\}, \{p(a)\}, \{p(X)\}, \{p(a), p(X)\}\}$, and we exclude the last element since this case is already considered. Therefore, following the definition of function *alts*, we consider the following three candidates for H^+ in order to compute alternative test cases:

- $H^+ = \{\}$: here, we have $H^- = \{p(a), p(X)\}$ and we should find a solution to $\alpha(p(Y), true, \{\}, \{p(a), p(X)\}, \{Y\})$, i.e., find a ground instance of $p(Y)$ that unifies with no clause. In particular, one must solve the following constraint:

$$\exists Y (true \wedge p(Y) \neq p(a) \wedge \forall X p(Y) \neq p(X))$$

which is trivially unfeasible.

- $H^+ = \{p(a)\}$: here, we have $H^- = \{p(X)\}$ and we should find a solution to $\alpha(p(Y), true, \{p(a)\}, \{p(X)\}, \{Y\})$, i.e., find a ground instance of $p(Y)$ that only unifies with the first clause. In this case, one must solve the following constraint:

$$\exists Y (true \wedge p(Y) = p(a) \wedge \forall X p(Y) \neq p(X))$$

which is also unfeasible.

- $H^+ = \{p(X)\}$: here, we have $H^- = \{p(a)\}$ and we should find a solution to $\alpha(p(Y), true, \{p(X)\}, \{p(a)\}, \{Y\})$, i.e., find a ground instance of $p(Y)$ that only unifies with the second clause. In this case, one must solve the following constraint:

$$\exists Y (true \wedge \exists X p(Y) = p(X) \wedge p(Y) \neq p(a))$$

which is satisfiable with model, e.g., $X = b, Y = b$ (represented by the substitution $\{X/b, Y/b\}$). Therefore, we add a new test case $p(b)$ to *TestCases*.

In the second iteration of the concolic testing algorithm, we consider the initial goal $p(b)$ and thus, we start concolic testing with $\langle p(b)_{id} \parallel p(Y)_{id, \epsilon, p(Y), true, \{Y\}} \rangle$. Concolic testing then computes then the following derivation:

$$\begin{aligned} & \langle p(b)_{id} \parallel p(Y)_{id, \epsilon, p(Y), true, \{Y\}} \rangle \\ & \quad \rightsquigarrow_{\text{choice}} \langle p(b)_{id}^{\ell_2} \parallel p(Y)_{id, \ell_2, p(Y), p(Y) \neq p(a), \{Y\}}^{\ell_2} \rangle \\ & \quad \rightsquigarrow_{\text{unfold}} \langle q(b)_{id} \parallel q(X)_{\{Y/X\}, \ell_2, p(X), p(X) \neq p(a), \{X\}} \rangle \\ & \quad \rightsquigarrow_{\text{choice}} \langle q(b)_{id}^{\ell_3} \parallel q(X)_{\{Y/X\}, \ell_2, \ell_3, p(X), p(X) \neq p(a), \{X\}}^{\ell_3} \rangle \\ & \quad \rightsquigarrow_{\text{unfold}} \langle true_{id} \parallel true_{\{Y/b\}, \ell_2, \ell_3, p(b), p(b) \neq p(a), \{Y\}} \rangle \\ & \quad \rightsquigarrow_{\text{success}} \langle SUCCESS_{id} \parallel SUCCESS_{\{Y/b\}} \rangle \end{aligned}$$

Moreover, in the second choice step, we add ℓ_2 to *Traces* and update *TestCases* as follows:

$$\begin{aligned} TestCases & \leftarrow TestCases \\ & \cup \text{out}_{vars}(\text{alts}(p(X), p(X) \neq p(a), q(X), \{q(b)\}, \{q(b)\}, \{X\})) \end{aligned}$$

Now, we have $\mathcal{P}(\{q(b)\}) = \{\{\}, \{q(b)\}\}$, and we exclude the last element since this case is already considered. Therefore, there is only one candidate for computing alternative test cases (i.e., $H^+ = \{\}$) and we should find a solution to

$$\alpha(q(X), p(X) \neq p(a), \{\}, \{q(b)\}, \{X\})$$

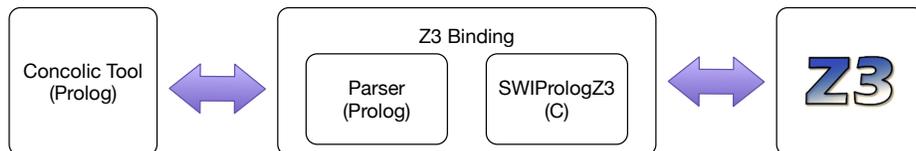


Fig. 3. Implementation Workflow

i.e., finding a ground instance of $q(X)$ that does not unify with $q(b)$ and, moreover, $p(X) \neq p(a)$ holds. In particular, one must solve the following constraint:

$$\exists X(p(X) \neq p(a) \wedge q(X) \neq q(b))$$

which is trivially feasible.¹¹ Note that the solution $X = a$ is now ruled out thanks to the negative constraint $p(X) \neq p(a)$, thus avoiding the problem shown in Example 2. A possible solution is, e.g., $X = c$, so that the next test case that we add to *TestCases* is $p(c)$.

In the third (and last) iteration, no more alternatives are obtained (we omit the concolic testing derivation for brevity), so our algorithm produces three test cases for this example: $p(a)$, $p(b)$ and $p(c)$, achieving a full coverage.

4 Implementation and Experimental Evaluation

In this section, we describe the implementation of an SMT-based concolic testing tool for Prolog programs that follows the ideas presented so far.

We have implemented our tool in SWI-Prolog [15] and have used the C interface of SWI-Prolog in order to call the functions of the Z3 solver [3]. The scheme of the workflow is shown in Figure 3. Here, “Concolic Tool” is the main module (written in Prolog) which performs concolic execution. Once a constraint is built, it is transformed into an SMT well-formed string and sent to the module “SWIPrologZ3”, a module written in C that uses Z3’s library to interact with the Z3 solver and solve this constraint. The results are sent back to the main module “Concolic Tool”.

In contrast to the concolic testing semantics shown in Figure 2, we have implemented a *nondeterministic* version of concolic execution which is based on Prolog’s backtracking mechanism. Here, the information that must survive a backtracking step is inserted to the internal database using dynamic predicates and asserted for consistency.

Regarding the termination of concolic testing, we impose a maximum term depth for the generated test cases. Since the domain is finite and we do not generate duplicated test cases, termination is trivially ensured. Consider, for

¹¹ Here, we assume that the considered domain includes at least one more constant, e.g., c .

instance, the usual specification of natural numbers built from 0 and $s(_)$:

$$\begin{aligned} (\ell_1) \text{ nat}(0). \\ (\ell_2) \text{ nat}(s(X)) \leftarrow \text{nat}(X). \end{aligned}$$

where we assume that the argument of $\text{nat}/1$ is an *input* argument (and must be ground). Given an initial goal like $\text{nat}(0)$, in the first iteration of the algorithm we add, e.g., the following new test cases: $\text{nat}(1)$ and $\text{nat}(s(0))$, where the constant 1 is used to avoid matching any clause. When considering the second test case, we will generate the alternative test cases $\text{nat}(s(1))$ and $\text{nat}(s(s(0)))$. And the process goes on forever.

In this context, by setting a maximum term depth, e.g., 2, one can limit the generated test cases to only

$$\{\text{nat}(0), \text{nat}(1), \text{nat}(s(0)), \text{nat}(s(1)), \text{nat}(s(s(0))), \text{nat}(s(s(1)))\}$$

The maximum term depth is an input parameter of our concolic testing tool since it depends on the particular program and the desired code coverage.

Finally, we show some selected results from a preliminary experimental evaluation of our concolic testing tool. We aimed at addressing the following questions:

1. Q1: What is the performance of our technique on typical benchmarks? Here, the goal was to assess the viability of the proposed method by measuring its execution time on some selected benchmarks.
2. Q2: How does it compare to existing tools for concolic testing? In particular, we wanted to consider the tool `contest` [9], which is publicly available through a web interface.¹²

Benchmarks. We selected six subject programs from previous benchmarks [9] and from GitHub.¹³ We ran concolic testing between 3 and 100 executions on a MacBook Pro hexacore 2,6 Ghz with 16 GB RAM in order to get reliable results. Reported times, in seconds, are the average of these executions. Our results are reported in Table 1. Here, `concolic` refers to the tool presented in this paper, while `contest` refers to the tool introduced in [9]. The size of a subject program is the number of its source lines of code. The column `Ground Args` displays the number of arguments of the initial symbolic goal to ground, starting at the first position. `#TCs` refers to the number of generated test cases. A timeout for `contest` is set to 1000 seconds (the crash is an overflow).

Q1: Performance. The three first lines of Table 1 clearly show the influence of the maximum term depth in a typically recursive program. Our procedure handles recursion better than `contest` because we end up generating complex constraints that are more efficiently solved using an SMT solver than using the specific algorithms in [9], as expected. For simpler cases, though, interacting

¹² Moreover, a copy of the Prolog sources of `contest` were provided by its authors.

¹³ <https://github.com/Anniepoo/prolog-examples>

Table 1. Summary of experimental results

Subject program	size	Initial goal	Ground Args	Max Depth	time concolic	time contest	#TCs concolic	#TCs contest
Nat	2	nat(0)	1	1	0.050	0.0273	3	4
Nat	2	nat(0)	1	5	0.0897	0.1554	7	12
Nat	2	nat(0)	1	50	1.6752	19.5678	52	102
Generator	7	generate(empty,-A,-B)	1	1	1.4517	0.7096	9	9
Generator	7	generate(empty,T,-B)	2	1	1.3255	4.4820	9	9
Generator	7	generate(empty,T,H)	3	1	1.3211	crash	9	N/A
Activities	38	what_to_do_today(sunday,sunny,wash_your_car)	3	2	6.3257	timeout	122	N/A
Cannibals	78	start(config(3,3,0,0))	1	2	0.0535	timeout	2	N/A
Family	48	parent(dicky,X)	1	1	20.0305	64.1838	9	19
Monsters and mazes	113	base_score(will,grace)	2	2	0.2001	0.4701	6	7

with the solver is likely more expensive than performing the computations in [9]. However, as the complexity increases, our SMT-based technique is faster and scales better than *contest*. These preliminary experiments support our choice of using a powerful SMT solver for test case generation.

Q2: Comparison to other concolic testing tools. Besides scalability, which is already considered above, we noticed that our tool typically produces less test cases than *contest*. In principle, this can be explained by the fact that the algorithm in [9] allows one to also bind the *output* arguments of the initial goal. Consider, e.g., the following simple program:

$$\begin{aligned} & p(a, b). \\ & p(X, c). \end{aligned}$$

where the first argument is an input argument and the second one is an output argument. Here, *contest* might return four test cases:

- $p(a, a)$, which matches no clause;
- $p(a, b)$, which only matches the first clause;
- $p(a, c)$, which only matches the second clause;
- $p(a, Y)$, which matches both clauses.

In contrast, our tool would only return the test cases $p(a, Y)$, which matches both clauses, and $p(b, Y)$, which only matches the second clause, since the second argument is an output argument and so, it cannot be bound.

This is essentially a design decision, but we think it is more sensible to keep output arguments unbound in test cases.

On the other hand, we also noticed that, in some cases, our tool produced some test cases that were not generated by *contest*. This is explained by the problem illustrated in Example 2.

Threats to Validity. These experiments are preliminary and are therefore subject to validity threats. We mitigated internal validity by repeating our experiments several times and ensuring the validity of the produced test cases manually. We alleviated external validity by selecting programs of varying size publicly available on GitHub, though we cannot guarantee they are representative.

5 Conclusion

In this paper, we have designed an improved concolic testing scheme that is based on [9] but adds support for negative constraints and defines *selective unification problems* as constraints on Herbrand terms. Our approach overcomes some of the problems in previous approaches [10,11], mainly regarding the scalability of the technique as well as a potential source of incompleteness due to the lack of negative information. We have implemented an SMT-based concolic testing tool in SWI-Prolog that uses the Z3 library for C through the foreign language interface of SWI-Prolog. Our preliminary experimental evaluation has shown promising results regarding the scalability of the method in comparison to previous approaches.

As for future work, we plan to extend and improve our concolic testing tool. In particular, we will consider its extension to CLP programs [7]. Given the way selective unification problems are represented in this paper (as constraint satisfiability problems), dealing with constraints over domains other than Herbrand terms seems very natural. Finally, we will formally prove the correctness and completeness of the improved algorithms.

References

1. Apt, K.: From Logic Programming to Prolog. Prentice Hall (1997)
2. Bjørner, N., McMillan, K.L., Rybalchenko, A.: Program verification as satisfiability modulo theories. In: Fontaine, P., Goel, A. (eds.) Proc. of the 10th International Workshop on Satisfiability Modulo Theories (SMT 2012). EPiC Series in Computing, vol. 20, pp. 3–11. EasyChair (2013)
3. De Moura, L., Bjørner, N.: Z3: An efficient SMT solver. In: Proc. of the Theory and Practice of Software, 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems. pp. 337–340. TACAS’08/ETAPS’08, Springer-Verlag, Berlin, Heidelberg (2008)
4. Gange, G., Navas, J.A., Schachte, P., Søndergaard, H., Stuckey, P.J.: Horn clauses as an intermediate representation for program analysis and transformation. TPLP **15**(4-5), 526–542 (2015). <https://doi.org/10.1017/S1471068415000204>
5. Godefroid, P., Klarlund, N., Sen, K.: DART: directed automated random testing. In: Proc. of PLDI’05. pp. 213–223. ACM (2005)
6. Gómez-Zamalloa, M., Albert, E., Puebla, G.: Test case generation for object-oriented imperative languages in CLP. TPLP **10**(4-6), 659–674 (2010)
7. Jaffar, J., Maher, M.J.: Constraint logic programming: A survey. J. Log. Program. **19/20**, 503–581 (1994). [https://doi.org/10.1016/0743-1066\(94\)90033-7](https://doi.org/10.1016/0743-1066(94)90033-7)
8. Lloyd, J.: Foundations of Logic Programming. Springer-Verlag, Berlin (1987), 2nd Ed.

9. Mesnard, F., Payet, É., Vidal, G.: Concolic testing in logic programming. *TPLP* **15**(4-5), 711–725 (2015). <https://doi.org/10.1017/S1471068415000332>
10. Mesnard, F., Payet, É., Vidal, G.: On the completeness of selective unification in concolic testing of logic programs. In: Hermenegildo, M.V., López-García, P. (eds.) *Proc. of the 26th International Symposium on Logic-Based Program Synthesis and Transformation (LOPSTR 2016)*. Revised Selected Papers. *Lecture Notes in Computer Science*, vol. 10184, pp. 205–221. Springer (2017)
11. Mesnard, F., Payet, É., Vidal, G.: Selective unification in constraint logic programming. In: Vanhoof, W., Pientka, B. (eds.) *Proc. of the 19th International Symposium on Principles and Practice of Declarative Programming (PPDP 2017)*. pp. 115–126. ACM (2017)
12. Sen, K., Marinov, D., Agha, G.: CUTE: a concolic unit testing engine for C. In: *Proc. of ESEC/SIGSOFT FSE 2005*. pp. 263–272. ACM (2005)
13. Ströder, T., Emmes, F., Schneider-Kamp, P., Giesl, J., Fuhs, C.: A Linear Operational Semantics for Termination and Complexity Analysis of ISO Prolog. In: *LOPSTR'11*. pp. 237–252. Springer LNCS 7225 (2011)
14. Vidal, G.: Concolic execution and test case generation in Prolog. In: Proietti, M., Seki, H. (eds.) *Proc. of the 24th International Symposium on Logic-Based Program Synthesis and Transformation (LOPSTR 2014)*. Revised Selected Papers. *Lecture Notes in Computer Science*, vol. 8981, pp. 167–181. Springer (2015)
15. Wielemaker, J., Schrijvers, T., Triska, M., Lager, T.: SWI-Prolog. *TPLP* **12**(1-2), 67–96 (2012). <https://doi.org/10.1017/S1471068411000494>