A Symbolic Model Checker for Database Programs

Angshuman Jana, Md. Imran Alam and Raju Halder
Indian Institute of Technology Patna, India
{ajana.pcs13, imran.pcs16, halder}@iitp.ac.in

Keywords: Model Checking, Database Program, Boolean Program, Verification, Refinement.

Abstract: Most of the existing model checking approaches refer mainstream languages without considering any database statements. As the result, they are not directly applicable to database applications for verifying their correctness. On the other hand, few works in the literature address the verification of database applications focusing atomicity constraints, transaction properties, etc. In this paper, as an alternative, we propose the design of a symbolic model checker for database programs to verify integrity properties defined over database attributes. The proposed model checker is designed based on the following key modules: (i) Abstraction, (ii) Verification, and (iii) Refinement.

1 INTRODUCTION

Model checking is an algorithmic method for proving that a system satisfies its specification (Clarke and Emerson, 1981; Wang et al., 2006). As stated in (Jhala and Majumdar, 2009), the goal of the model checking research is to expand the scope of automated techniques for program reasoning, both in the scale of programs handled and in the richness of properties that can be checked. Model checker receives application source codes and exhaustively explores their execution states, searching for possible violations of properties of interest. For examples, properties may include simple assertions, which state that a predicate on program variables holds whenever the computation reaches a particular location, or global invariants, that state certain predicates hold on every reachable state (e.g. each array access is within bounds), or termination properties.

Based on the state representations, model checkers are of two types: (i) enumerative (in which individual states are represented) and (ii) symbolic (in which sets of states are represented using constraints). State space explosion is the critical limitation of the enumerative techniques, which led researches to explore symbolic algorithmic approaches. The symbolic model checking approach manipulates the representation of sets of states, rather than individual state, and performs state exploration through the symbolic transformation of these representations (Queille and Sifakis, 1982). For example, the constraint $0 \leq a \leq 9 \land 6 \leq b \leq 10$ represents the set of all states over $a, b$ satisfying the constraint, which implicitly represents a list of 50 possible states. Therefore, the symbolic model checking algorithms are more efficient as compared to the enumerative approaches.

Over the past, several enumerative model checkers, e.g. Verisoft (Chandra et al., 2002), SPIN (Holzmann, 1997), Cmc (Yang et al., 2006), etc. are developed. On the other hand, many tools like SLAM (Ball and Rajamani, 2002), CBMC (Clarke et al., 2003), F-Soft (Ivancic et al., 2005), JavaPathFinder (Anand et al., 2007), etc. are developed based on symbolic algorithms. Notably, the above tools are built for the verification of various mainstream languages without considering any database statements. Even though intensive research is already done, researchers have not paid much attention towards database programs embedding queries and data-manipulation commands. Few works in the literature include symbolic model checking of stored procedure and SQL queries for automatic validation according to the specification (Diana et al., 2012), explicit-state model checker DPF for the verification of database atomicity constraints in web applications (Gligoric and Majumdar, 2013), etc.

In this paper, as an alternative, we extend the existing symbolic model checking algorithm for imperative language to the case of database programs. In particular, we aim at verifying the correctness of database programs respecting integrity properties defined on the underlying databases.

To summarize, our contributions in this paper are:
1. Abstraction of database programs into boolean programs.

2. Generation of verification conditions (VCs) form the Control Flow Graph (CFG) of boolean database programs and their verification using SMT solver.

3. Counter Example Guided Abstract Refinement (CEGAR) of boolean database programs.

Roadmap: In section 2, we discuss the current state-of-the-art in the literature. In section 3, we describe a motivating example. In section 4, we recall some preliminaries. We introduce our approach in section 5. Section 6 provides an overall tool architecture. Finally section 7 concludes the work.

2 RELATED WORKS

Chandra et al. (Chandra et al., 2002) proposed a tool Verisoft which pioneered the idea of execution-based stateless model checking of software. It uses a scheduler which is able to exhaustively explore all possible interleaving of the processes executions. In (Holzmann, 1997), author proposed SPIN, a tool which supports model-based verification of distributed software systems. It has been used to detect design errors in applications ranging from high-level descriptions of distributed algorithms to detailed code for controlling telephone exchanges. Another execution based model checker Cmc (Yang et al., 2006) is proposed for C programs. The Cmc tool explores different executions by controlling schedules at the level of the OS scheduler. Ball and Rajamani (Ball and Rajamani, 2002) introduced first Counter Example Guided Abstract Refinement (CEGAR)-based symbolic model checker SLAM for C programs. The SLAM tool works in the following three steps: (i) abstracting programs into the form of boolean programs, (ii) verifying properties using model checking algorithm on the boolean programs, and (iii) refinement of the boolean programs based on CEGAR. The model checker Magic (Chaki et al., 2004) was proposed to enable modular verification of concurrent, message passing C programs. Ivanic et al. (Ivanic et al., 2005) proposed a model checker F-Soft that used predicate abstraction along with other abstract domains that efficiently yield the kinds of invariants needed to check standard runtime errors in C programs. JavaPathFinder tool (Anand et al., 2007) is the model checker for Java programs that modifies the Java Virtual Machine to implement systematic search over different thread schedules. Several other model checkers CBMC (Clarke et al., 2003), Chess (Musuvathi and Qadeer, 2007), etc. are also developed in the literature.

In (Paleari et al., 2008), the authors proposed a dynamic approach to detect race condition vulnerabilities on web-based applications. The approach analyzes a log file of a single run and identifies dependencies among SQL queries based on the set of relations and attributes that are read/written. QED (Martin and Lam, 2008) is a first model checker for Java web applications that systematically explore sequences of requests to a web application and looks for taint-based vulnerabilities in web applications. Artzi et al. (Artzi et al., 2010) proposed an explicit-state model checker for PHP web applications. The model checker generates test inputs for web applications, monitors the applications for crashes, and validates that the output conforms to the HTML specification. Petrov et al. (Petrov et al., 2012) developed dynamic race detector tool for web applications. It is mainly formulation of a happens-before relation to capture the asynchronous behavior of most commonly-used web platform constructs. In (Gligoric and Majumdar, 2013), the authors proposed an explicit-state model checker DPF for verifying atomicity violations in web applications. The model checker interposes between the program and the database layer and precisely tracks the effects of queries made to the database. Another symbolic model checker is proposed in (Diana et al., 2012) for the verification of stored procedure and SQL queries w.r.t. their specification. The specification is expressed in CTL temporal logic. In (Scully and Chipala, 2017), the authors introduced Sqlcache, the automatic compiler optimization for database result caching. The main aim of Sqlcache is to analyze web applications for compatibility checking between queries and updates, by instrumenting updates with cache invalidations.

3 MOTIVATING EXAMPLE

Consider the database program depicted in Code Snippet 1. The code implements a module which performs withdrawal of cash from an ATM system.

The function withdraw() updates the balance of authorized customers after satisfying two conditions: (i) maximum withdrawal amount (amt) request of the customer should be less than or equal to 10,000 USD for each transaction, and (ii) minimum balance (minbal) will be more than or equal to 1000 USD after each successful transaction. Now consider the following integrity constraint on the database attribute balance:

The minimum account balance of all customers should be 1000 USD.

Our proposed framework will allow us to verify whether the database program satisfies or violates
such integrity constraints.

4 PRELIMINARIES

In this section, we recall the notions of predicate abstraction, weakest preconditions and boolean program.

Predicate Abstraction (Ball et al., 2001): The predicate abstraction algorithm generates a finite state abstraction from a large or infinite state system. It is an automated abstraction technique in which the abstract domain is constructed from a given set of predicates over program variables. It is used to reduce the state space of a program. The basic idea in predicate abstraction is to remove some variables from the program by just keeping information about a set of predicates about them. Let \( \mathcal{R} \) be a region. Consider the predicate abstraction domain which is parameterized by a fixed finite set \( \Pi \) of first order formulas. The predicate abstraction of \( \mathcal{R} \) with respect to \( \Pi \) is the smallest region \( \mathcal{A}(\mathcal{R}, \Pi) \) which contains \( \mathcal{R} \) and it is representable as a boolean combination of predicates from \( \Pi \):

\[
\mathcal{A}(\mathcal{R}, \Pi) = \bigwedge \{ \psi \mid \psi \text{ is a boolean formula over } \Pi \land \mathcal{R} \Rightarrow \psi \}
\]

The region \( \mathcal{A}(\mathcal{R}, \Pi) \) can be computed by recursively splitting as follows:

\[
\mathcal{A}(\mathcal{R}, \Pi) = \begin{cases} 
true, & \text{if } \Pi = \emptyset \text{ and } \mathcal{R} \text{ satisfiable} \\
false, & \text{if } \Pi = \emptyset \text{ and } \mathcal{R} \text{ unsatisfiable} \\
(p \land \mathcal{A}(\mathcal{R} \land p, \Pi')) \lor (\neg p \land \mathcal{A}(\mathcal{R} \land \neg p, \Pi')), & \text{if } \Pi = \{p\} \cup \Pi'
\end{cases}
\]

Weakest Preconditions (Ball et al., 2001): Given a program statement \( stmt \) and a predicate \( \psi \). The weakest precondition of \( stmt \) with respect to \( \psi \) is denoted as \( WP(stmt, \psi) \) which determines a predicate such that after successfully executing \( stmt \) on the initial state satisfying \( \psi \) results into the final state where \( \psi \) is true and \( stmt \) terminates. For example, given an assignment statement \( x = e \),

\[
WP(x = e, \psi) = \psi'
\]

where \( \psi' \) is obtained by replacing all occurrences of \( x \) in it with \( e \), i.e. \( \psi' \triangleq \psi[e/x] \).

Example 1. Consider the statement \( stmt \triangleq y = y + 1 \) and predicate \( \psi \triangleq y \leq 30 \). The weakest precondition of \( stmt \) with respect to \( \psi \) is computed as

\[
WP(y = y + 1, y \leq 30) = (y + 1) \leq 30 = (y \leq 29)
\]

Observe that, given a statement \( stmt \) and predicate \( \psi \in \Pi \), it may be the case that \( WP(stmt, \psi) \) is not in \( \Pi \). For instance, consider \( \Pi = \{(x < 5), (x = 2)\} \) and \( WP(x = x + 1, x < 5) = (x < 4) \) where the predicate \( (x < 4) \) is not in \( \Pi \). Therefore, the decision procedures (i.e. a theorem prover) can be used to strengthen or weaken the resulting precondition in order to express over the predicates in \( \Pi \) (Ball et al., 2001).

Boolean Programs (Ball and Rajamani, 2000): Boolean programs can be thought of as an abstract representation of a program that explicitly captures correlations between data and control, in which boolean variables can represent arbitrary predicates over the unbounded state of a program. The syntax is depicted in Figure 1. Boolean variables are either local or global with statically scoped as C program and variable declarations need not specify a type. A variable identifier is either a C-style identifier or an arbitrary string between the characters “{” and “}”. Two constants ‘0’ (false) and ‘1’ (true) are in the language. The expressions are built from these constants, variables and logical connectives. A parallel assignment statement allows the simultaneous assignment of a set of values to a set of variables. The statements ‘if’, ‘while’ and ‘assert’ can affect

Code Snippet 1: A Database Application PR

0. int withdraw() {
  1.   int acc_no, amt, bal, minbal=1000;
  2.   acc_no = read();
  3.   amt = read();
  4.   Statement con = DriverManager.getConnection("jdbc mysql :....... ", " scott ", " tiger "). createStatement();
  5.   bal = con.executeQuery("select balance form Account where Accno=acc_no");
  6.   if (amt <= 10000) { // Maximum withdraw amount
  7.     if (balance – amt ≥ minbal)
  8.       con.executeQuery("update Account balance = balance – amt where Accno = acc_no");
  9.       // ... do some work ...
 10.   } else
11.       ERROR_msg;
12.   bal = con.executeQuery("select balance form Account where Accno=acc_no");
13.   display (bal);
14.   // ... do some work ...
15.   return 0;
}
Given a predicate \( \psi \) point structs of \( P \) describe an abstraction of various programming con- 
cicates \( E \). Consider an assignment statement \( \text{Assignments (Ball et al., 2001):} \)

\[ \begin{align*}
\text{Commands:} & \\
B & := \psi^* P^* \\
\psi^* & := \text{id}\{id\} \text{ begin } \psi^* \text{ S end} \\
P^* & := \text{id}\{id\} \text{ begin } \psi^* \text{ S end} \\
S & := \text{stmt}^* \\
\text{stmt} & := \text{stmt}^* \text{ id} : \text{stmt} \\
\text{stmt} & := \text{stmt}^* \text{ skip} | \text{id} = \exp^* ; | \text{if}(\text{con}) \text{ then } S \text{ else } S | \text{while}(\text{con}) \text{ do } S \\
\text{con} & := \text{expr}^* | \text{iden} (\text{expr}) | \text{id}(\text{expr}) | \text{return} \\
\exp & := \exp^* \exp^* | \exp^* \exp | \exp^* \exp | \exp | \text{id} | \text{const} \\
\text{id} & := [a-zA-Z] [a-zA-Z+0-9]* | \{asring\} \\
\text{const} & := 0 \text{ | } 1
\end{align*} \]

A list of global variable declarations followed by a list of procedure definitions

Declaration of variables

Procedure definition

Sequence of statements

Non-deterministic choice

Logical connectives

False/True

the control flow of the language. Observe that the predicate of control statements is a decider which 
can be used to model non-determinism. A decider evaluates to ‘0’ or ‘1’ deterministically or ‘?’ which 
evaluates to ‘0’ or ‘1’ non-deterministically.

5 PROPOSED APPROACH

In this section, we illustrate our proposed model checker which consists of three key modules: (i) 
abstraction of database programs into their equivalent boolean programs, (ii) automatic VC generation 
from the Control Flow Graph (CFG) of the boolean programs and its verification using SMT, and (iii) 
abstraction refinement introducing additional predicates, by identifying spurious execution path in the 
original database program. Let us explain each of the modules in the following sub-sections:

5.1 PROGRAM ABSTRACTION

This section describes the abstraction of database programs into boolean programs using predicates.

Given a database program \( P \) and a set of pred- 
cicates \( \Psi = \{ \psi_1, \psi_2, \psi_3, \ldots, \psi_n \} \), let \( B \) \( P, \Psi \) generates a boolean program which consists of boolean 
variables and it has similar control flow as in \( P \). In particular, \( B \) \( P, \Psi \) contains \( n \) boolean variables 
\( B = \{ b_1, b_2, b_3, \ldots, b_n \} \) where each boolean variable \( b_i \) represents the predicate \( \psi_i \) (1 \( \leq i \leq n \)). Note that, 
\( B \) \( P, \Psi \) is guaranteed to be an abstraction of \( P \) in the 
sense that the set of execution traces of \( B \) \( P, \Psi \) is a 
superset of the set of execution traces of \( P \). We now 
describe an abstraction of various programming 
structures of \( P \) with respect to \( B \) below:

Assignments (Ball et al., 2001):

Consider an assignment statement \( x = e \) at program 
point \( l \) in \( P \) where \( e \) denotes an arithmetic expression. 
Given a predicate \( \psi \in \Psi \), a boolean variable \( b_i \) in \( B \) \( P, \Psi \) can have the value \( \text{true} \) after \( l \) if \( \text{WP}(x = e, \psi) = \text{true} \) before \( l \). Similarly, \( b_i \) can have the value \( \text{false} \) 
after \( l \) if \( \text{WP}(x = e, \neg \psi) = \text{false} \) before \( l \). Observe 
that, if neither of these predicates holds before \( l \) then 
\( b_i \) in \( B \) \( P, \Psi \) contains parallel assignment at \( l \) as:

\[ b_1 \ldots b_n =
\begin{align*}
& \text{T}_{\text{ch}}(\text{WP}(x = e, \psi_1), \text{WP}(x = e, \neg \psi_1)), \\
& \ldots, \\
& \text{T}_{\text{ch}}(\text{WP}(x = e, \psi_n), \text{WP}(x = e, \neg \psi_n))
\end{align*}
\]

where the function \( T_{\text{ch}} \) is defined as bellow:

\[ \text{bool } T_{\text{ch}}(\text{bool } pos, \text{bool } neg) \{ \\
\text{if (pos) } \text{return true;} \} \\
\text{if (neg) } \text{return false; } \\
\text{return unknown(); } \}
\]

The unknown function is defined as follows:

\[ \text{bool unknown}() \{ \\
\text{if (+) return true; } \\
\text{else return false; } \\
\}
\]

Observe that the unknown function uses the control 
expression “+” which non-deterministically deter- 
mines the result either \( \text{true} \) or \( \text{false} \).

Conditional (Ball et al., 2001):

Consider a conditional statement if \( (\psi) \{ \text{stmt} \} \) else 
\( \{ \text{stmt} \} \) in \( P \). If the predicate \( \psi \) is evaluated to \( \text{true} \) 
in \( P \) then predicate \( \psi \) in the corresponding \( B \) \( P, \Psi \) 
should also be evaluated to \( \text{true} \). Similarly, \( \psi \) evaluated 
to \( \text{false} \) in \( P \) then predicate \( \psi \) in the corresponding 
\( B \) \( P, \Psi \) should also be evaluated to \( \text{false} \). In \( B \) \( P, \Psi \), the condition is encoded as:
if (*) {
    assume(ψ);
    ...
} else {
    assume(¬ψ);
    ...
}

Observe that, as "∗" in 'if' condition nondeterministically evaluates to either true or false, both execution paths will be explored. The functions assume(ψ) under 'if' and assume(¬ψ) under 'else' preserve the semantics of the conditional statement in P.

Databases Statements:

We now define the abstraction of database statements (SELECT, UPDATE, DELETE and INSERT) into their boolean form. To do so, let us first recall from (Halder and Cortesi, 2012) the formal syntax of database programs as below:

\[
Q := (A, ϕ)
\]

\[
A := \text{select}(v_a, f(\vec{c}), r(h(\vec{x})), \phi', g(\vec{c})) \mid \text{update}(\vec{v}_d, \vec{c}) \mid \text{insert}(\vec{v}_d, \vec{c}) \mid \text{delete}(\vec{v}_d)
\]

\[
τ := n \mid v_a \mid v_d \mid f_n(τ_1, τ_2, ..., τ_n),
\]

where \( f_n \) is an n-ary function.

\[
a_f := R_n(τ_1, τ_2, ..., τ_n) \mid 1 = τ_2,
\]

where \( R_n(τ_1, τ_2, ..., τ_n) \in \{\text{true, false}\} \)

\[
φ := a_f \mid ¬φ_1 \mid φ_1 \lor φ_2 \mid φ_1 \land φ_2 \mid ∀v_i φ \mid ∃v_i φ
\]

\[
c := Q[v = e] \text{ if } \text{cond then } c_1 \text{ else } c_2
\]

Where \( φ \) denotes an well-formed first order formula defined over constants (\( n \)), application variables (\( v_a \)) and database attributes (\( v_d \)). The SQL clauses GROUP BY, ORDER BY, DISTINCT/ALL and the aggregate functions (SUM, COUNT, MAX, MIN, AVG) are represented in the form of functions \( g() \), \( f() \), \( r() \), \( h() \) respectively parameterized with either none or one arithmetic expression \( e \) or an ordered sequence of arithmetic expressions \( \vec{c} \). The abstract syntax of a database statement is denoted by \( ⟨A, ϕ⟩ \) where \( A \) represents Action-part and \( ϕ \) represents Condition-part. The Action-part include SELECT, UPDATE, DELETE and INSERT. For example, consider the query \( \text{Q=update1 set sal=sal+100 where age\geq40} \). According to abstract syntax, \( Q \) is denoted by \( ⟨A, ϕ⟩ = (\text{update}(\vec{v}_d, \vec{c}), \phi) \), where \( \vec{v}_d = (\text{sal}) \) and \( \vec{c} = (\text{sal+100}) \) and \( φ = \text{age\geq40} \).

Given a set of predicate \( E \) consisting of the integrity constraints under consideration, the abstraction of database statements into their equivalent boolean form involves the following two major tasks:

1. Computation of weakest precondition of database statements w.r.t. \( ψ \in E \): We define below the computation of weakest precondition for various database statements for a given post condition.

\[
\text{WP}(\text{select}(v_a, f(\vec{c}), r(h(\vec{x})), \phi', g(\vec{c})), \phi), ψ) = ψ
\]
\[
\text{WP}(\text{update}(\vec{v}_d, \vec{c}), ψ) = ((ψ ∧ ¬ϕ) ∨ (ψ[\vec{c}/\vec{v}_d] ∧ ϕ))
\]
\[
\text{WP}(\text{insert}(\vec{v}_d, \vec{c}), ψ) = ψ[\vec{c}/\vec{v}_d] ∨ ψ
\]
\[
\text{WP}(\text{delete}(\vec{v}_d), ψ) = ψ
\]

2. Conversion of database statements into boolean form: Given a database statement \( Q \) and a postcondition \( ψ \), suppose \( \text{WP}(Q, ψ) = ψ'. \) Let us now define \( \text{BP} \) for database statements to convert into their equivalent boolean form w.r.t. \( ψ \) below:

\[
\text{BP}(\text{select}(v_a, f(\vec{c}), r(h(\vec{x})), \phi', g(\vec{c})), \phi), ψ') = \text{skip}
\]
\[
\text{BP}(\text{update}(\vec{v}_d, \vec{c}), ψ) = \{ b = φ \ ? : \text{true, if } ψ' \in E 
\]
\[
\text{BP}(\text{insert}(\vec{v}_d, \vec{c}), \phi), ψ') = \{ b = *, \text{if } ψ' \in E 
\]
\[
\text{BP}(\text{delete}(\vec{v}_d), \phi), ψ') = \text{skip}
\]

Since SELECT and DELETE operations do not violate any integrity constraints defined over database states at row level, the corresponding statements are replaced by ‘skip’. On the other hand, in case of UPDATE and INSERT, the following two cases may arise: (i) if \( ψ' \) already exists in \( E \) and \( ϕ \) is evaluated to true, this means that the update operation is taking place which may lead the updated values possibly to violate the integrity properties. Therefore, the corresponding boolean variable \( b \) in the boolean program is assigned to ‘∗’. Otherwise, the evaluation of \( ϕ \) to false indicates no update operation (hence, no property violation) and therefore \( b \) is assigned to true, and (ii) if \( ψ' \) does not exist in \( E \) (with or without applying strengthen or weaken operation) then we replace the given database statement by “skip”.

Let us illustrate this using an example as follows.

**Example 2.** Let \( E=\{\text{sal} \leq 8000\} \) and its corresponding boolean variable be \( b \). We assume that the initial database is in consistent state, and therefore, the boolean variable \( b \) is set to true.

Consider the database table \( \text{Emp} \) which stores employees information and the following update statement:

\[
Q=\text{update } \text{Emp set sal=sal+100 where age>30}
\]

The abstract syntax of \( Q \) is represented as

\[
Q = (\text{update}(\text{sal}), (\text{sal+100}), \text{age} > 30)
\]
5.2 VERIFICATION

In this section, we describe the generation of verification conditions and their satisfiability. After converting the program into a boolean form, we will verify whether all possible execution paths respect the defined properties. For this objective, we perform the below steps:

- Control Flow Graph (CFG) construction of boolean programs.
- VC generation from CFGs of boolean programs.
- Satisfiability checking of the VCs using SMT.

The Control Flow Graph $G_{BL}$ of boolean program $BL$ is constructed by following the similar approach proposed in (Ball and Rajamani, 2000). After constructing $G_{BL}$, we list all possible execution paths. We convert each path $p_i$ into a verification formula $f_i$ by anding all boolean statements encountered along that path. The generated $f_i$ is supplied to SMT solver for the satisfiability checking. If the formula $f_i$ is satisfied, then the corresponding path $p_i$ satisfies the property and the algorithm terminates. Otherwise, we check whether the corresponding path $p_i$ is a feasible path in the original program. If yes, we conclude a violation of the property and an error trace is generated. Otherwise, a refinement of the boolean program is performed which we describe in the next section.

Illustration on motivational example: Consider the boolean program in the code snippet 2. Its CFG is depicted in Figure 1. Consider the CFG path $p$ $\triangleq 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 8 \rightarrow 12 \rightarrow 13 \rightarrow 16$ (denoted by blue color). The generated VC along this path is $b = \phi ? * : true // Accno=acc_no$. We pass the negation of this $f$ (i.e., $\neg f$) to SMT solver which reports “satisfied”. This shows that there exist some solution for which the predicate $b$ evaluates to false. Therefore, the model checker indicates that the program $PR$ may violate the integrity constraint. However, observe that this path is not a feasible path in $PR$, because $b$ which corresponds to $bal \geq minbal$ is false. This indicates that the abstraction is not precise enough and there is a need of refinement.

5.3 COUNTER EXAMPLE GUIDED ABSTRACT REFINEMENT

The process of refining the abstraction is done using a method called counterexample-guided abstraction refinement (CEGAR) (Wang et al., 2006). If the boolean program contains an error path $p_i$ and this path is not feasible in the original program, then the refinement of the Boolean program will be initiated to eliminate this false error paths. More specifically, we determine suitable predicates, by analyzing the infeasibility of paths in the original database program which illustrates feasible paths in the corresponding boolean program. We then refine the abstraction by...
adding these new predicates to $E$. Let us illustrate below the application of CEGAR on the motivating example:

**Illustration on motivational example:** Consider the boolean program in code snippet 2. As we already observed in section 5.2 that this abstraction is not precise enough, in order to refine this abstraction we discover the predicate ($amt \leq 10000$) form $PR$ and we add it to $E$. This results $E = \{bal \geq minbal, amt \leq 10000\}$ and $B = \{b, b_1\}$ where $b$ and $b_1$ corresponds to $bal \geq minbal$ and $amt \leq 10000$ respectively. Given the refined boolean program w.r.t. these new $E$ and $B$, observe that after verifying, by following the same procedure as before, we get the error path $p_1 \triangleq 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 8 \rightarrow 12 \rightarrow 13 \rightarrow 16$ which is also not a feasible path in $PR$ because the predicate ($amt \leq 10000$) is always true along the path. By initiating the refinement process once again, we discover a new predicate ($balance - amt \geq minbal$). Now considering the new $E = \{bal \geq minbal, amt \leq 10000, balance - amt \geq minbal\}$ and $B = \{b, b_1, b_2\}$, we get a refined boolean program depicted in code snippet 3. We now generate the VC from the CFG of this code snippet 3, which is encoded as “$f_2 \triangleq ((b_1 == true) \land (b_1 == true \land b == true) \land (b_2 == true \land b_1 == true \land b == true) \land ((b \land \phi) \lor (\neg b \land \phi) \lor (b \land \neg \phi) \Rightarrow b == true) \Rightarrow b == true$”. The SMT solver reports unsatisfied when $\neg f_2$ is passed as input. Therefore, the model checker indicates that $PR$ is safe w.r.t. the integrity constraint $bal \geq minbal$.

6 TOWARDS IMPLEMENTATION

In general, the proposed framework accepts as inputs a database program and properties of interest, and this verifies whether the input program respects the properties by generating safe/unsafe message as output. The overall architecture of our proposed framework is shown in Figure 2. We have identified the following key modules which play important roles in implementing the proposed framework:

1. **Proformat:** This modules preprocess input database programs and annotates them by adding line numbers to all statements.

2. **Abstraction:** The module “Abstraction” abstracts the database program into a boolean program using a set of predicates.
3. **Verification**: This module at first constructs CFG of the boolean program. After that, it generates VC from CFG. Finally, it tests the satisfiability of each VC using SMT solver.

4. **Refinement**: The primary task of this module is to discover additional predicates which refine the abstraction avoiding the existence of the spurious paths in the corresponding boolean database programs.

7 **CONCLUSIONS AND FUTURE WORKS**

In this paper, we proposed a symbolic model checking algorithm for verifying the correctness of database applications with respect to integrity properties. We design our model checker based on the following key modules: (i) Abstraction, (ii) Verification and (iii) Refinement. We are currently implementing our proposed model checker, as per the description provided in the tool architecture, in a modular way to support scalability.

Acknowledgement

This work is partially supported by the research grant (SB/FTP/ETA-315/2013) from the Science and Engineering Research Board (SERB), Department of Science and Technology, Government of India.

REFERENCES


