

СОДЕРЖАНИЕ

Список сокращений и условных обозначений	8
Терминология	9
Введение	10
1 Анализ решений в области восстановления поведенческих моделей	14
1.1 Критерии сравнительного анализа	14
1.1.1 Метод извлечения трасс	14
1.1.2 Алгоритм восстановления модели	15
1.1.3 Применимость к реальным проектам и возможность автоматизации	16
1.1.4 Доступ к исходному коду	17
1.2 Обзор работ по извлечению спецификаций	17
1.3 Результаты анализа	21
2 Задача извлечения поведенческих моделей и анализ путей её решения . .	24
2.1 Поиск проектов	25
2.2 Подготовка проектов к анализу	26
2.3 Извлечение трасс	27
2.4 Восстановление поведенческой модели	28
2.5 Общий подход	28
3 Проектирование	30
3.1 Общая схема работы	30
3.2 Получение и подготовка проектов	31
3.2.1 GitHub	31
3.2.2 Maven Central	33

3.2.3 Сборка	33
3.3 Извлечение трасс	34
3.3.1 Статический подход	34
3.3.2 Динамических подход	38
3.4 Восстановление поведенческих моделей	39
3.5 Результат проектирования	40
4 Реализация	41
4.1 Общая архитектура	41
4.2 Пакет repository	41
4.3 Пакет analysis	42
4.4 Пакет storage	42
4.5 Пакет inference	42
5 Тестирование	43
Заключение	44
Список использованных источников	47

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБО- ЗНАЧЕНИЙ

- КА - конечный автомат
- ООП - объектно-ориентированное программирование
- JVM - Java Virtual Machine
- MBT - Model Based Testing
- PBT - Property Based Testing
- ПО - программное обеспечение
- СА - статический анализ
- ДА - динамический анализ
- CFG - Control Flow Graph
- ICFG - Inter-procedural Control Flow Graph
- ВМ - виртуальная машина

ТЕРМИНОЛОГИЯ

- символьное исполнение
- обратная совместимость
- динамический анализ
- статический анализ
- reflection
- мокирование
- concolic
- SMT-решатель
- фаззинг
- фреймворк
- identity hash code – уникальное числовое значение, которое определяется для каждого объекта на основе его памяти адреса

ВВЕДЕНИЕ

С развитием области анализа программ при решении большого количества задач в этом направлении разработчики все чаще стали сталкиваться с необходимостью использования формальных спецификаций. Формальная спецификация - описание поведения программы на специальном или её исходном языке, включающее в себя такие детали как пред и пост условия вызовов, состояния и переходы между ними, что и делает спецификации идеальным кандидатом для применения в области анализа ПО. Например, спецификации не заменимы в символьном исполнении. Там они используются для аппроксимации поведения внешних библиотек или сложных частей программы, тем самым ускоряя анализ или вовсе делая его возможным в определенных местах. Реализацию данного подхода можно увидеть в USVM¹ и UtBot[1]. Еще один пример использования формальных спецификаций - Taint анализ, когда в них отмечаются потенциальные места ввода и утечки информации. Наличие подобных данных позволяет анализаторам сфокусироваться на проверке размеченных мест, представляющих возможные ошибки, тем самым сильно повышая эффективность[2]. Кроме этого, спецификации могут использоваться в MBT и PBT в качестве тестового оракула, отвечая на вопросы о корректности состояния ПО и переходов между ними, а также о том, удовлетворяют ли входные и выходные данные для различных методов соответствующим предикатам. Сами формальные спецификации при этом могут быть представлены с помощью подмножества используемых языков программирования, как в UtBot, или на специальных языках спецификации, например LibSL[3].

¹<https://github.com/UnitTestBot/usvm>

При всем этом формальные спецификации в общем случае не постав-
ляются с программными библиотеками или другим ПО, что заставляет задум-
аться о способах их создания. Полностью ручное составление спецификаций
это довольно монотонная работа, при этом требующая от человека высокой
квалификации в области разработки и анализа программ. Однако составление
спецификаций можно частично автоматизировать.

Часть формальной спецификации можно найти в исходном коде библио-
тек или ПО. Если приводить в пример языки в парадигме ООП, то это классы
и интерфейсы программы, их поля и сигнатуры методов. Другую же часть
спецификации, описывающую поведение программы, а именно состояния и
переходы между ними, можно попытаться извлечь из реальных примеров
использования.

Именно автоматизации получения поведенческих моделей библиотек
посвящена данная работа. В рамках её выполнения будет разработан ком-
плексный подход для получения моделей и для его апробации реализована
утилита, позволяющая автоматически получать общедоступные Java проекты
и с помощью методов статического и динамического анализа извлекать из них
сценарии работы определенной библиотеки с целью последующего восстанов-
ления поведенческой модели в виде КА. На данный момент комплексных
автоматических решений для подобных задач нет, однако имеются работы в
области восстановления КА из трасс и описаны способы получения трасс.
Основные задачи, решаемые в этой работе, это автоматизация, интеграция и
применение на практике существующих наработок в области восстановления
поведенческой модели библиотек.

Важно сказать, что предполагается ручная обработка полученных с помощью реализованной утилиты автоматов. Необходимость этого является следствием фундаментальных ограничений алгоритмов восстановления КА. В дополнение к этому предполагается использование пользовательских репозиторий, которые не гарантируют корректного применения библиотек. Также свои ограничения накладывает статический и динамический анализ. Использование points-to анализа в статическом подходе не позволяет точно различать объекты, методы которых вызываются, что влияет на корректность получаемых трасс. Также при использовании СА, по крайней мере в рамках данной работы, не будут предприниматься попытки обработать многопоточную работу и получить данные о состоянии программы в моменты вызовов. Что касается ДА, то здесь на результат могут содержательность имеющихся тестов, точки входа в программу и необходимое для запуска окружение.

В первом разделе представлен обзор существующих решений, связанных с задачами поиска проектов, извлечением из них трасс вызовов и восстановлением модели. Также в данном разделе будет уделено внимание предшествующей работе по данной теме. На основе первого раздела сделан выбор в пользу определенных подходов и решений, составляющих общий подход и используемых в реализации инструмента. Во втором разделе формулируются требования к создаваемому инструменту и описываются пути решения каждой из задач, стоящих на пути реализации. Третий раздел посвящен разработке подхода. В нем будет подробно описана задуманная схема работы, способы извлечения трасс и их восстановления, детали работы с репозиториями. Четвертый раздел содержит описание реализации инструмента. Пятый раздел посвящен

тестированию полученной утилиты. Тестирование заключается в сравнении получаемых автоматов с несколькими заготовленными эталонами, а также демонстрацией получаемых КА для некоторого набора библиотек. Помимо этого, будет проанализировано количество успешно автоматически собранных проектов и полученных различными методами трасс. В заключении проведен анализ полученных результатов, отмечены преимущества и недостатки предложенного подхода, а также рассмотрены пути развития.

1 АНАЛИЗ РЕШЕНИЙ В ОБЛАСТИ ВОССТАНОВЛЕНИЯ ПОВЕДЕНЧЕСКИХ МОДЕЛЕЙ

При изучении предметной области было выявлено, что на текущий момент нет исследований и инструментов, целью которых являются полностью автоматический процесс получения спецификаций, начиная от получения проектов, использующих заданную библиотеку, и заканчивая извлечением из нее поведенческой модели. Тем не менее, в данной области достаточное количество работ, сосредоточенных на методах извлечения трасс и последующего восстановления модели библиотек, подразумевающих применение подходов к подготовленным для анализа программам. В данном разделе рассмотрим и сравним существующие способы извлечения трасс из программ и алгоритмы восстановления поведенческих моделей в виде КА.

1.1 Критерии сравнительного анализа

Выделим определенные критерии, на которые будем обращать внимание при обзоре работ.

1.1.1 Метод извлечения трасс

Глобально методы можно поделить на статические и динамические. Первые подразумевают анализ исходного кода или байт-кода программы без его запуска. Динамические методы наоборот, предполагают запуск анализируемой программы. Статические методы уступают в точности, однако позволяют покрыть все возможные пути исполнения программы. Это позволяет находить ошибки в тех участках кода, которые не покрытых тестами и до которых исполнение не доходит при штатной работе программы. Динамические методы, в

свою очередь, за счет реального исполнения обеспечивают точность получаемых результатов, но с их помощью сложно получить все возможные состояния программы. Для восстановления поведенческой модели мы заинтересованы в получении как можно большего количества трасс, чему сопутствует использование статических методов, но в тоже время ошибки в трассах неизбежно приведут к ошибкам в модели. Таким образом, недостатки одного метода являются преимуществом другого и наоборот. В работах нас интересует, как авторы реализовали преимущества и нивелировали недостатки выбранных методов.

1.1.2 Алгоритм восстановления модели

Алгоритм восстановления непосредственно влияет на качество получаемых моделей. При этом он определяет, какие данные нам необходимо извлечь из программы. Например, базовый алгоритм *k-tail*[4] требует на вход исключительно последовательности вызовов и параметра *k*, определяющего длину соединяемых цепочек. Другой алгоритм, *gk-tail*[5], дополнительно требует на вход значения аргументов.

Также существуют различные алгоритмы, основанные на использовании инвариантов или состоянии программы в моменте вызова библиотеки. В рамках обзора важно обратить внимание, каких дополнительных усилий требует применение сложных алгоритмов восстановления, какие ограничения это накладывает и какой дает прирост в точности и полноте получаемых автоматов.

Перед тем, как перейти к обзору современных работ, следует уделить особое внимание алгоритмам *k-tail*[4] и *gk-tail*[5], на которых основано большинство современных методов восстановления автоматов из трасс. *K-tail* принимает на вход последовательность вызовов, полагая что трасса - это КА, где

переходами являются вызовы. Для каждого состояния рассматриваются хвосты длиной k (обычно равной один или два) и если эти хвосты эквивалентны, то они сливаются. Безусловным плюсом данного алгоритма является высокая точность при простоте применения - алгоритм не может породить модель, разрешающую несуществующие трассы, а также не требует дополнительных обработок входных данных. Но не смотря на то, что получаемая модель описывает корректные последовательности, получаемые состояния КА не отражают реальные и являются сильной аппроксимацией сверху реальной модели.

Gk-tail основан на k-tail, однако помимо самих вызовов, также учитывает информацию об аргументах вызовов и контекстных переменных. Сначала трассы, состоящие из одинаковых вызовов объединяются вместе с данными, накапливая множество возможных значений для переменных. Далее с помощью Daikon[6] (используется в оригинальной статье, возможно использование других подобных инструментов) выполняется вывод инвариантов для каждого перехода, основанный на накопленных данных. Затем применяется алгоритм k-tail, однако для потенциально сливаемых хвостов происходит проверка инвариантов на конфликты. Данный алгоритм более трудозатратный ввиду необходимости получения информации о значении аргументов и контекстных переменных, а также обязательного вывода инвариантов. Но взамен мы получаем более осмысленное деление на состояния, чем при использовании K-tail.

1.1.3 Применимость к реальным проектам и возможность автоматизации

Определенные подходы могут показывать отличные результаты и иметь минимальные недостатки, но при этом иногда они совершенно не применимы к реальным проектам, что обусловлено либо новизной, либо фундаментальными

ограничениями подхода. Также применение некоторых методов, в частности основанных на динамическом анализе, требует определенной ручной работы, например связанной с подготовкой анализируемой программы и её окружения. Это может сильно влиять на массовость применения подхода и его автоматизацию.

1.1.4 Доступ к исходному коду

Если авторы предоставляют доступ к инструментам, это позволяет убедиться в результатах проведенных экспериментов. И что не менее важно, появляется возможность применять и развивать разработанный в рамках исследований подход и инструмент.

1.2 Обзор работ по извлечению спецификаций

В работе «Static Specification Mining Using Automata-Based Abstractions»[7] авторы статически собирают трассы в виде последовательностей объектов одного типа, используя абстрактную интерпретацию[8]. Для получения последовательностей вызовов над конкретным экземпляром объекта в исследовании используется points-to анализ на основе алгоритма Андерсона (не чувствительный к потоку) и чувствительный к потоку access-paths анализ. При этом авторы не объединяют результаты применения анализов, а используют их в зависимости от потребности в максимально подробных и избыточных трассах (flow-insensitive) или точных и ограниченных (flow-sensitive). Для восстановления поведенческой модели авторы используют собственный подход, основанный на эвристических правилах слияния состояний. Предложенные алгоритмы выглядят интересно, однако сложно оценить их точность, так как в исследовании приводится сравнение результатов вари-

аций описанных алгоритмов между собой, хотя было бы уместно провести сравнение с классическим алгоритмом k-tail. В ограничениях подхода авторы описывают невозможность его применения для анализа проектов состоящих из десятков тысяч строк кода. Это ожидаемо, поскольку flow-sensitive подходы сталкиваются с проблемой взрыва состояний и применение access-paths анализа к реальной программе требует большого количества памяти даже при минимальной глубине анализа[9]. Авторы делают вывод, что получаемые поведенчески модели достоверно описывают поведение библиотек, однако являются сильной аппроксимацией сверху истинной модели и содержат множество лишних состояний и переходов. Однако предполагается, что выявление чистых функций в исходном коде библиотеки позволит избежать появления избыточных состояний, так как на этапе восстановления будет известно, что определенные вызовы не изменяют состояния программы, а значит конечную модель можно упростить. В статье явно упоминается разработанный инструмент для проведения анализа, однако ссылки на них не приводятся.

Авторы статьи «Automatic mining of specifications from invocation traces and method invariants»[10] подробно рассмотрели и сравнили четыре алгоритма восстановления моделей из трасс. При этом был рассмотрен базовый алгоритм k-tail, предложены улучшения для алгоритма Contractor[11], основанного на получении КА из инвариантов, а также разработаны новые подходы: SEKT и TEMI, заключающиеся в извлечении поведенческой модели из трасс, усиленных инвариантами, и инвариантов, усиленных трассами, соответственно. В рамках исследования авторы получали трассы и инварианты с помощью инструмента динамического анализа Daikon[6]. Daikon очень мощный и

полезный инструмент, однако его применение сложно автоматизировать для сторонних проектов, так как даже чтобы получить полные трассы и полезные инварианты из собственного целевого проекта, нужно проделать определенную нетривиальную работу. В статье очень большое внимание уделено сравнению методов восстановления КА. Авторы ввели метрики *precision* и *recall*, где под *precision* понимается доля трасс, сгенерированных по восстановленной модели и подходящих под эталонную модель, а *recall* определяется как доля сгенерированных трасс по эталонной модели, не противоречащих восстановленной. В результате все методы, включая *k-tail*, показали *precision* близкий к 100 процентам для девяти эталонных моделей библиотек. Что касается *recall*, *k-tail* и SEKT показали результат от 20% до 60%, имея примерно одинаковые значения в рамках конкретной библиотеки. TEMI и Contractor++ показали лучшие результаты, достигая значений 100% для некоторых библиотек, однако также сохранялся большой разброс и худшие результаты были на уровне 40%. Стоит заметить, что не смотря на очевидное преимущество более сложных алгоритмов, *k-tail*, требующий минимальные входные данные, показывает конкурентноспособный результат. В данном исследовании авторы не делятся реализацией алгоритмов, хоть и детально описывают принцип их работы.

Интересный метод восстановления, а также его реализацию² в открытом доступе предлагают авторы статьи «*Inferring Extended Finite State Machine models from software executions*»[12]. Авторы развивают идею алгоритма *gk-tail* и предлагают использовать информацию о значении аргументов в анализируемых вызовах. Однако новизна заключается в том, что для поиска конфликтов

²<https://github.com/neilwalkinshaw/mintframework>

слияния применяются обучаемые классификаторы. Под конфликтами понимаются слияния таких трасс, где из одного и того же состояния при одних и тех же вызовах осуществляется переход в отличные друг от друга состояния. Это говорит о том, что на самом деле начальное состояние было не одно и то же. В `gk-tail` поиск конфликтов между инвариантами происходит локально для отдельных участков трасс длиной k , из за чего можно объединить состояния, где позже возникает конфликт. В предлагаемом подходе классификаторы используют трассы как источник данных для обучения, что позволяет осуществлять поиск конфликтов из всей совокупности данных. Также классификаторы избавляют от необходимости использовать тяжеловесные утилиты по типу `Daikon`, неявно выполняя задачу вывода инвариантов. Благодаря одновременному учету всех трасс обеспечивается высокий уровень обобщенности получаемой поведенческой модели. Однако при этом на пользователя ложится задача подбора алгоритма для классификации данных, поскольку разные алгоритмы могут показывать разный результат в зависимости от входных данных. Авторы в своем исследовании приводят сравнение алгоритмов, а также предоставляют в реализованном инструменте возможность удобно его менять. Что касается получения трасс, в исследовании используются трассы полученные из двух проектов с помощью `Daikon`. Отдельно стоит поблагодарить авторов за реализацию алгоритмов `k-tails` и `gk-tails` в предоставляемом инструменте.

Еще один выделяющийся подход реализован в инструментах `Tautoko`[13] для генерации тестов и `ADABU`[14], представленных в соответствующих исследованиях. `ADABU` решает задачу получения трасс и состояний программы на основе инструментации и реализует предложенный в исследовании метод

восстановления поведенческой модели. Общий подход заключается в отслеживании состояния программы до и после вызова библиотеки. После сбора трасс, собранные состояния классифицируются по определенным правилам, тем самым образуя состояния КА. Tautoko же является генератором тестов, позволяющим получить ранее не обнаруженные варианты поведения библиотеки. Авторы предлагают реализованный подход как решение проблемы ограниченного набора тестов при использовании инструментов по типу Daikon. К сожалению, в исследованиях не представлены сравнения с существующими алгоритмами восстановления и инструментами генерации тестов. Однако представленных результатов достаточно, чтобы убедиться в работоспособности предложенных подходов, а наличие инструментов в открытом доступе³ делает полученные результаты очень полезными. Но важно отметить, что инструменты реализованы в 2012 году и не получали никаких обновлений и они скорее всего не актуальны для анализа современных версий Java. Также данный подход имеет ограничение в виде необходимости работы над конкретными проектами для извлечения трасс, так как требуется плотное взаимодействие с исполняемыми файлами анализируемого ПО.

1.3 Результаты анализа

Все описанные работы предлагают работоспособные решения, подтвержденные авторами в рамках проведенных экспериментов. Однако нигде не уделяется внимания автоматизации решения – зачастую авторы извлекают трассы из одного и того же проекта (даже в рамках разных исследований разных авторов). Причиной этого является применение чисто динамических

³<https://www.st.cs.uni-saarland.de/models/>

подходов для получения информации о состоянии программы в момент вызовов, что принуждает к плотному взаимодействию с бинарными файлами программы и её необходимым окружением, а это довольно трудозатратно. Одна из описанных работ[7] использует подход на основе статического анализа и имеет потенциал для автоматизации, однако авторы применяют flow-sensitive алгоритм для анализа псевдонимов, что делает подход неприменимым для реальных проектов. Краткий итог сравнения представлен в Таблица 1.

Таблица 1 — Анализ работ

Название	Извлечение трасс	Восстановление модели	Исх. код	Ограничения
Static Spec. Mining[7]	Абстрактная интерпретация	Из трасс	Нет	Невозможен анализ больших реальных проектов
SEKT/TEMI[10]	Daikon/аналоги	Из трасс и состояний	Нет	Зависимость от тестов
MINT[12]	Daikon/аналоги	Из трасс и состояний	Да	Зависимость от тестов
Tautoko[13]/ADABU[14]	Собственная инструментация и генерация тестов	Из трасс и состояний	Да	Частный подход к каждому проекту
Gk-tail	Daikon	Из трасс и состояний	Да, в MINT	Зависимость от тестов
K-tail	-	Из трасс	Да, в MINT	Получаемая модель далека от реальной

Все это наводит на мысль о необходимости создания комплексного автоматизированного решения для извлечения трасс и поведенческих моделей библиотек. Безусловно, для начала автоматизация потребует некоторых уступ-

ков в требованиях к качеству получаемых автоматах и решения специфичных проблем. Однако это положит начало развитию подобных автоматизированных методов и, возможно, позволит использовать извлечение автоматов в реальной жизни с меньшими усилиями.

2 ЗАДАЧА ИЗВЛЕЧЕНИЯ ПОВЕДЕНЧЕСКИХ МОДЕЛЕЙ И АНАЛИЗ ПУТЕЙ ЕЁ РЕШЕНИЯ

Задача извлечения поведенческих моделей библиотек состоит из нескольких составляющих:

- Поиск проектов, использующих заданную библиотеку
- Подготовка проектов к анализу
- Извлечение трасс и состояний
- Восстановление автоматов из трасс

Каждая из этих задач требует отдельного внимания, а также определенного уровня согласованности с остальными. Из анализа существующих решений в предыдущем разделе видно, что методы извлечения трасс могут зависеть от требований к масштабированию подхода и от самих используемых для анализа программ (исходный код, исполняемые файлы, наличие тестов, необходимое окружение). Метод извлечения трасс в свою очередь определяет применимые для восстановления моделей методы. Поэтому важно комплексно подходить к выбору путей решения каждой из указанных составляющих.

Выбор путей решения в рамках данной работы будет основан на гипотезе о том, что в текущем состоянии предметной области получение части спецификаций, связанных с поведенческими моделями библиотек, является очень трудозатратным и требует автоматизации. Даже при использовании инструментов, предоставленных авторами статей, пользователю необходимо самостоятельно найти подходящий проект. Затем, в случае применения динамических методов, убедиться в наличии тестов или заняться их генерацией и оценить содержательность получаемых трасс. Если говорить про статиче-

ские подходы, то готовых решений обнаружено не было. Пользователь будет вынужден самостоятельно разрабатывать анализы на основе фреймворков СА или изучать доступные символьные машины с целью применения их для сбора трасс. Только после успешного решения подобных задач и связанных с ними проблем, можно перейти к самому восстановлению трасс. К счастью, забегаю вперед, MINT[12] предлагает действительно удобный и рабочий модульный инструмент для применения собственного алгоритма восстановления, а также k-tail и gk-tail. Тем не менее, для получения КА на данный момент требуется преодолеть ряд сложных и не очевидных задач, требующих определенного погружения в область анализа ПО.

2.1 Поиск проектов

Для апробации подхода остановимся на Java проектах по нескольким причинам:

- Большинство исследований в области восстановления поведенческих моделей библиотек используют проекты на Java в качестве целевых, что позволяет перенимать практический опыт данных работ, а также их реализацию
- Для языка Java разработано большое количество библиотек, в том числе для статического и динамического анализа
- Не смотря на определенные недостатки, Java по прежнему является популярным языком, на котором реализовано множество проектов, число которых продолжает расти

В качестве источников проекта предлагается использовать GitHub и Maven Central Repository. Рассмотрим данные варианты подробнее.

GitHub является крупнейшим сервисом хранения программных репозиторий и на текущий момент содержит более 420 миллионов репозиторий[15]. При этом GitHub предоставляет возможности поиска по коду и фильтрации, реализованные в рамках API. При этом репозитории могут содержать артефакты программ, а именно JAR файлы. Их наличие позволит избежать этапа сборки, тем самым ускоряя процесс анализа проекта и потенциально повышая число успешно обработанных проектов, так как для статического анализа необходимы бинарные файлы.

Не смотря на большое количество проектов на GitHub, среди них могут быть как серьезные рабочие проекты, так и множество неработоспособных программ. Поэтому в качестве альтернативы есть возможность использовать Maven Central Repository, где среднее качество проектов должно быть значительно выше. Репозиторий Maven предлагает поиск библиотек по API. Для каждого пакета можно получить список зависимых от него, что и требуется для поиска примеров использования.

2.2 Подготовка проектов к анализу

Для применения статических подходов необходимы скомпилированные программы. Для динамических же необходим еще и исходный код, поскольку тесты редко упаковываются в JAR файл и находятся непосредственно с самим кодом. Из этого возникает необходимость выполнять сборку проектов перед извлечением трасс. Опираясь на то, что JVM решает вопрос зависимости от платформы, а Java гарантирует обратную совместимость, можно предположить, что при существовании зависимостей, указанных в проекте, и совпа-

дении версий сборщиков Maven или Gradle, будет возможно успешно собрать определенную долю проектов.

2.3 Извлечение трасс

Для извлечения трасс будут применяться как статический, так и динамический анализ. В рамках данной работы предполагается, что с помощью динамического подхода в автоматическом режиме будет получено сильно меньше трасс, чем с помощью статического. Более того, уже имеющиеся тесты не гарантируют целостность получаемых трасс ввиду возможного мокирования и искусственного создания объектов с помощью reflection, что будет необходимо учесть при проектировании инструмента.

Предлагаемый статический подход заключается в анализе межпроцедурного графа потока управления (в дальнейшем ICFG), поиску в нем вызовов искомой библиотеки и последующему применению Points-to анализа для определения объектов вызовов и формирования трасс. По умолчанию предполагается использования анализа на основе алгоритма Андерсона – для возможности применения такого подхода к реальным большим проектам мы вынуждены использовать не чувствительные к контексту или потоку алгоритмы.

Что касается динамического подхода, в рамках данной работы мы остановимся на сборе исключительно последовательностей вызовов, игнорируя значения аргументов и состояние программы. Сделано это с целью ограничить сложность работы, так как первоочередной целью является апробация идеи автоматизированного подхода к извлечению поведенческих моделей. Сбор информации должен осуществляться через инструментацию и запуск тестов. При отсутствии тестов в проекте или при недостаточном количестве полученных трасс предлагается воспользоваться методом фаззинга входных точек в

программу с помощью фаззинга. Он позволит обеспечить лучшую целостность получаемых трасс в случае успешного запуска и корректного выбора точек входа. Важно заметить, что выполнять запуск случайных сторонних программ не безопасно, поэтому предполагается, что использование динамических подходов должно производиться на специально выделенной для этого машине или на VM.

2.4 Восстановление поведенческой модели

Так как в рамках данной работы мы оперируем трассами, состоящими исключительно из последовательностей вызовов, то это сильно ограничивает выбор алгоритма восстановления. По этой причине предлагается использовать алгоритм k-tail. Тем не менее, алгоритм обладает рядом преимуществ:

- Он не уступает в точности более сложным алгоритмам, то есть сгенерированные по получаемому КА трассы будут корректны
- Простота алгоритма обеспечивает сравнительно высокую скорость работы в обмен на полноту, так как нет необходимости выводить инварианты и искать конфликты между сливаемыми состояниями
- Благодаря прозрачному принципу восстановления интерпретация получаемых КА будет проще

Для апробации общего подхода будем полагать, что возможностей данного алгоритма достаточно.

2.5 Общий подход

Из предложенных способов решения задач складывается общий подход. Мы должны получать проекты с GitHub и репозитория Maven, использующие заданную библиотеку. Далее необходимо подготовить проекты к извлечению

трасс, выполнив компиляцию исходного кода и тестов. При отсутствии тестов или малом количестве полученных трасс следует предпринять попытку фаззинга точек входа в программу. Затем необходимо извлечь трассы с применением описанных выше подходов и подать полученные результаты на вход алгоритму восстановления поведенческой модели в виде КА. Каждый из шагов имеет множество сопутствующих проблем и будет подробно рассмотрен в следующем разделе, посвященном проектированию инструмента.

3 ПРОЕКТИРОВАНИЕ

Перед реализацией предложенного подхода по получению поведенческих моделей библиотек необходимо сформулировать схему работы основных его составляющих и их взаимодействия друг с другом. В данном разделе рассмотрены общая схема работы инструмента, получение и подготовка проектов, извлечение трасс и восстановление из них поведенческих моделей.

3.1 Общая схема работы

На Рис. 1 в общих чертах представлен полный цикл работы разрабатываемого инструмента. Предполагается, что у пользователя будет возможность выбора режима работы, что позволит пропускать определенные этапы и, например, восстанавливать КА из заранее накопленных данных или извлекать трассы из уже подготовленных проектов, что не отражено на схеме. Цифрами обозначены основные части инструмента, которые будут рассмотрены в данной главе:

- 1) Получение и подготовка проектов
- 2) Извлечение трасс
- 3) Восстановление поведенческих моделей

Общая последовательность действий заключается в том, что изначально пользователь конфигурирует инструмент – задает целевую библиотеку, определяет параметры используемых алгоритмов, указывает пути к требуемому окружению и выбирает режим работы. Далее итеративно выполняется поиск проектов, их подготовка и извлечение трасс с последующим сохранением. После анализа заданного количества проектов или каких-либо других установленных условий остановки выполняется обработка полученных трасс и

восстановление КА. На более детальном уровне каждый шаг разобран в соответствующих подразделах.



Рисунок 1 — Общая схема работы

3.2 Получение и подготовка проектов

В качестве источника проектов предлагается использовать GitHub и Maven Central Repository. Каждый из них имеет свои механизмы поиска и получения проектов, поэтому рассмотрим их отдельно. Дополнительно должна быть предусмотрена возможность использования заранее подготовленных проектов. Также для анализа необходимы JAR файлы проектов и их зависимости, поэтому необходимо реализовать их автоматическую сборку.

3.2.1 GitHub

Для GitHub мы вынуждены использовать поиск по коду, так как других вариантов найти проекты, использующие заданную библиотеку, отсутствуют.

Пользователь может указать в качестве исходного кода импорт основного пакета библиотеки или момент её подключения в `build.gradle` или `pom` файле. При этом репозитории помимо исходного кода могут содержать другую полезную информацию:

- Тесты. Они необходимы для получения трасс с помощью динамического анализа.
- JAR файлы. Наличие бинарного файла увеличивает вероятность того, что проект рабочий, а также избавляет от необходимости выполнять его компиляцию для статического извлечения трасс. В случае применения динамических подходов необходимо компилировать программу для получения её зависимостей.
- Конкретные версии программы. При выпуске версии фиксируется соответствующий коммит, фиксирующий исходный код. Это также повышает вероятность того, что проект рабочий.

При реализации получения проектов должна быть предусмотрена возможность сбора всей полезной информации и фильтрация по ней, чтобы обеспечить возможность работы только с репозиториями, содержащими тесты или скомпилированные программы. Также необходимо отслеживать обработанные репозитории, так как в результатах поиска один проект может содержаться несколько раз.

Для поиска и получения проектов GitHub предоставляет бесплатное API. В случае отсутствия артефактов, получение проектов может быть реализовано через `git`.

3.2.2 Maven Central

Maven Central Repository содержит информацию об общедоступных библиотеках, а также об их зависимостях и зависящих от них проектах. Таким образом, есть возможность узнать где используется заданная библиотека – это нас и интересует. Для работы с Maven Repository пользователь должен указать библиотеку и её версию в определенном формате.

Получение проектов предлагается выполнять с помощью gradle, так как это позволит автоматически собрать её и разрешить зависимости. Получение исходного кода, хранящегося в Maven Central не имеет смысла, так как он не содержит тестов. На странице библиотек в Maven Central иногда расположены ссылки на GitHub репозитории, однако их нельзя получить с помощью API, а автоматическое получение и парсинг страниц может привести к блокировке на ресурсе.

Для получения информации о зависящих библиотек от заданной также не предусмотрен общий API. Однако после определенного исследования было обнаружено, что существует некий внутренний API, который реализует данную возможность. Чтобы убедиться в возможности использования данного API был сделан запрос на почту по контактам Maven, где сообщили, что проблемы могут возникнуть только в случае попытки массовой выгрузки всего Maven Central, что не предполагается в рамках проекта.

3.2.3 Сборка

В рамках работы будут рассматриваться только проекты, использующие Gradle и Maven в качестве системы сборки, так как они используются чаще всего и относительно легко поддаются автоматизации.

Для работы с Gradle существует Gradle Tooling API, позволяющий автоматически получать необходимую версию сборщика и программно вызывать необходимые задачи. При этом остается возможность использовать конкретную версию для всех проектов или использовать предустановленный Gradle. Maven не имеет подобного API, в связи с чем необходима предустановленная система сборки.

Java гарантирует обратную совместимость со старыми версиями, а JVM решает большинство проблем зависимости от платформы. Благодаря этому можно рассчитывать на успешную сборку определенной доли программ. Тем не менее, есть несколько причин, по которой не все проекты могут быть собраны:

- Проект может использовать специфические зависимости, получение и сборка которых не может быть автоматизирована
- Проект может использовать зависимости, которые уже не доступны из заданных репозиториях библиотек

С этими факторами ничего нельзя поделать и мы вынуждены принять тот факт, что не все программы будут собраны.

3.3 Извлечение трасс

Для извлечение трасс предлагается использовать статический и динамический подходы. Каждый из них имеет сложную схему работы и определенные детали, что требует тщательной разработки методов извлечения.

3.3.1 Статический подход

Примерный алгоритм извлечения трасс с применением статического анализа представлен на Рис. 2. Общий подход заключается в том, что строится

межпроцедурный граф потока управления для программы. Также необходимо получить набор точек входа в программу. Из каждой такой точки выполняется обход ICFG, в рамках которого выполняется обнаружение и сбор вызовов заданной библиотеки. Для каждого вызова определяется объект, над которым он выполняется, и с помощью points-to анализа формируются трассы для каждого объекта. Полученные трассы сохраняются для последующего восстановления поведенческой модели.

Что касается получения точек входа, предлагается предоставить пользователю возможность создавать определенные фильтры для методов, содержащие следующие параметры:

- Теги метода
- Теги класса
- Регулярное выражение для названия метода
- Регулярное выражение для названия класса
- Модификатор доступа метода (для исследования всех публичных методов)
- Типы аргументов метода

Подобный набор должен обеспечить возможность поиска специфических точек входа. Как пример, это могут быть методы, отвечающие за обработку запросов по URL путям в различных фреймворках для создания веб-приложений.

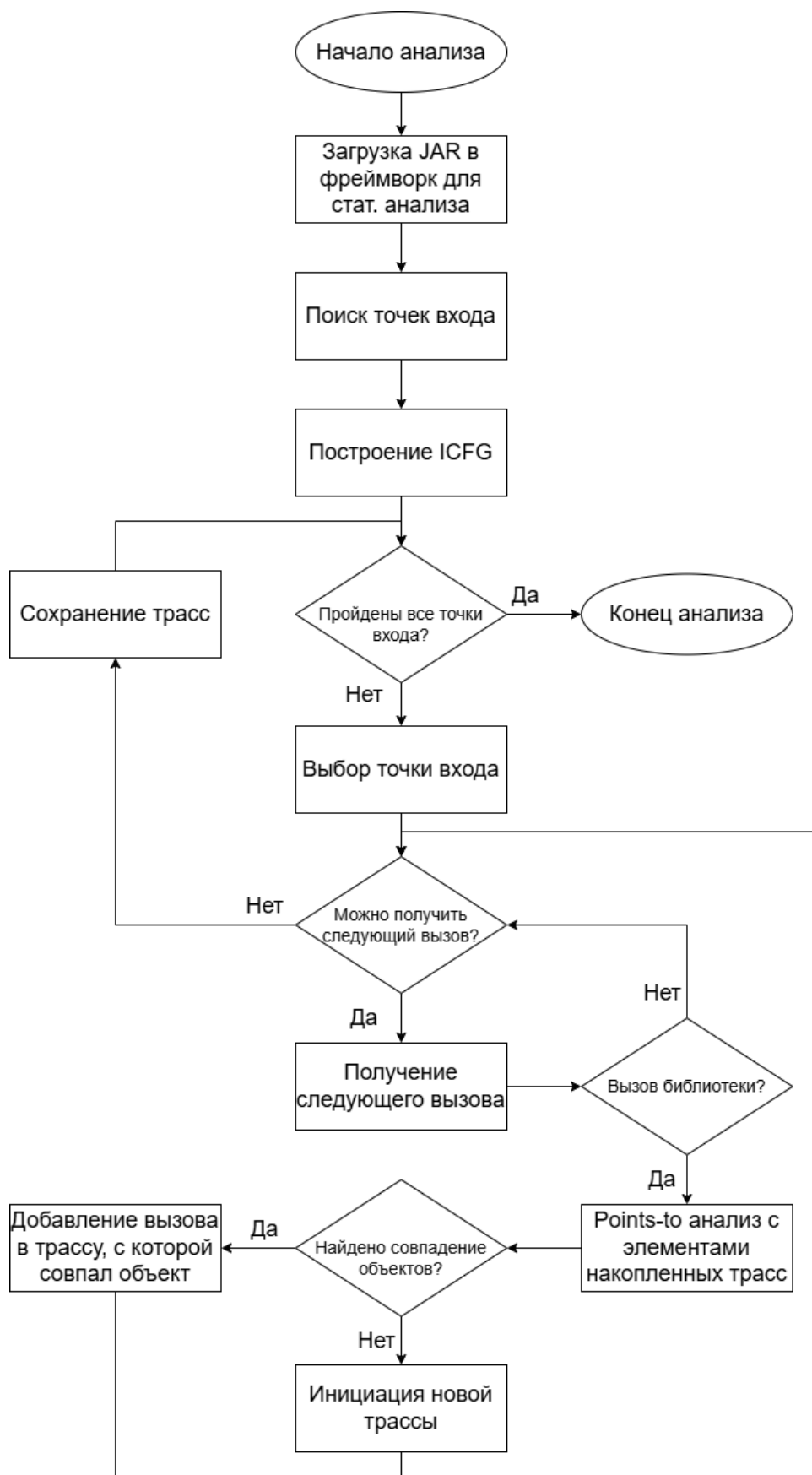


Рисунок 2 — Схема извлечения трасс с помощью статического анализа

Необходимость выполнения points-to анализа обусловлена тем, что об-
щая последовательность нерепрезентативна и не содержит в себе реальных
переходов и состояний библиотеки. Рассмотрим простой пример на Листинг 1.

```
1 FileOutputStream dest = new FileOutputStream(file);
2 ZipOutputStream zip1 = new ZipOutputStream(new BufferedOutputStream(dest));
3 ZipOutputStream zip2 = new ZipOutputStream(new BufferedOutputStream(dest));
4 zip1.write(1);
5 zip1.close();
6 zip2.write(1);
7 zip2.close();
```

Листинг 1 — Пример последовательностей вызовов

Извлекаемая трасса при игнорировании объектов будет выглядеть при-
мерно как на Листинг 2.

```
1 1. init
2 2. init
3 3. write
4 4. close
5 5. write
6 6. close
```

Листинг 2 — Пример некорректной трассы

При восстановлении поведенческой модели окажется, что мы допускаем
запись после того, как закрыли ZipOutputStream. Очевидно, такая модель не
соответствует действительности. Применив points-to анализ мы рассчитываем
получить две идентичные трассы как на Листинг 3.

```
1 1. init
2 2. write
3 3. close
```

Листинг 3 — Пример правильной трассы

Предполагается применять points-to анализ на основе алгоритма Андер-
сена, так как это один самых точных алгоритмов, которые могут быть примени-
мы в для всей программы межпроцедурно. Конечно, существуют гораздо более
точные алгоритмы, однако ввиду экспоненциальной сложности их использо-

вание в таких условиях невозможно. В качестве инструмента для статического анализа выбран Soot⁴, так как он предоставляет возможности для построения ICFG, конкретные реализации points-to анализа, а также предлагает средства для инструментирования программ, что будет необходимо для динамического подхода по извлечению трасс.

3.3.2 Динамических подход

Извлечение трасс с помощью динамического анализа заключается в предварительной инструментации на уровне байт-кода и последующем запуске программы. Инструментация может быть выполнена с помощью Soot и заключается во вставке специального кода после каждого вызова библиотеки, который позволяет получить информацию о вызове и объекте. Для того, чтобы точно различать экземпляры объектов между собой, может быть использован identity hash code.

Чтобы получить трассы после инструментации, программу необходимо запустить. Для этого предлагается использовать тесты при их наличии и фаззинг входных точек в программу. Входные точки предлагается получать тем же образом, что и при статическом анализе.

В качестве инструмента для фаззинга предлагается использовать Jazzer⁵. Он поставляется в виде собранного JAR файла и имеет режим автоматического фаззинга для указанных точек входа без предварительной подготовки целевого проекта. Для генерации данных Jazzer создает случайные массивы байт, интерпретируемые как необходимый объект и в последствии их мутирует для увеличения покрытия кода.

⁴<https://github.com/soot-oss/soot>

⁵<https://github.com/CodeIntelligenceTesting/jazzer>

Общая схема динамического извлечения трасс представлена на Рис. 3.

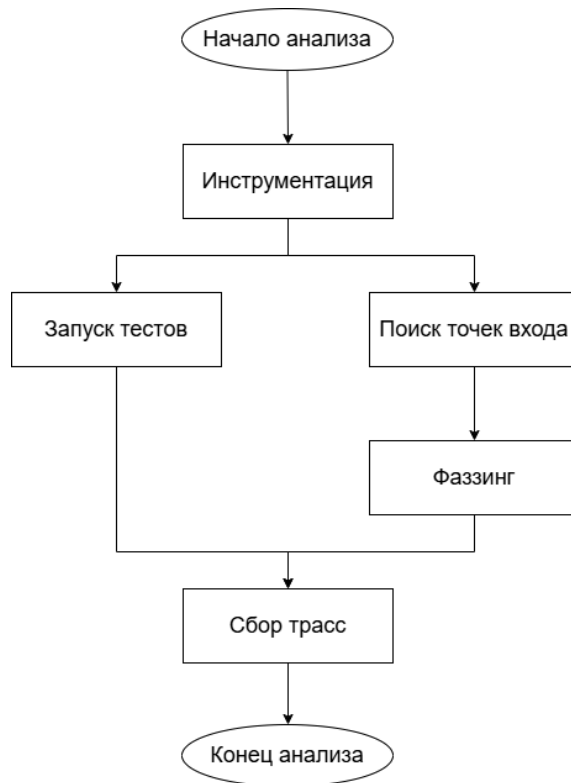


Рисунок 3 — Схема извлечения трасс с помощью динамического анализа

3.4 Восстановление поведенческих моделей

Из накопленных трасс необходимо восстановить поведенческую модель в виде КА. Для этого необходимо обработать полученные трассы и сформировать входные данные для алгоритма k-tail. Что касается обработки, то ввиду использования двух подходов для извлечения трасс, мы можем реализовать несколько режимов работы:

- Для восстановления КА полученные разными способами трассы объединяются в единое множество
- Для каждого способа получения трасс строятся соответствующие КА
- Динамические трассы являются верификатором для статических, подтверждая их в случае совпадения динамической трассы с подпоследовательностью статической

Можно попытаться уменьшить количество некорректных трасс. Трассы могут повторяться между собой, и если считать количество таких повторений для каждой трассы, мы получим определенное частотное распределение. Чтобы исключить погрешности, можно воспользоваться фильтрацией по абсолютному или относительному порогу.

Получаемый в результате работы алгоритма КА может содержать ложные конечные состояния из-за слияний. Для правильного определения состояния необходимо их обойти и обозначить как конечные лишь те, из которых отсутствуют какие либо переходы. После этого конечные состояния следует объединить, что приведет к сокращению состояний КА без потери информации о переходах.

Вместо собственной реализации алгоритма k-tail предлагается использовать фреймворк MINT, рассмотренный в рамках обзора существующих работ в области восстановления поведенческих моделей. Помимо этого, MINT содержит реализации алгоритмов, использующих информацию о переменных, что будет полезно при развитии инструмента.

Получаемые с помощью MINT КА выводятся в dot формате. Для более универсального представления, предлагается транслировать их в JSON.

3.5 Результат проектирования

В результате рассмотрены основные задачи и способы их решения, включая используемые инструменты и различные детали, которые необходимо учесть при реализации. Для общего понимания представлена схема предлагаемого подхода, а также схемы наиболее сложных его частей.

4 РЕАЛИЗАЦИЯ

Создание инструмента для применения предложенного метода извлечения поведенческих моделей библиотек заключается в реализации каждой из его частей, разобранных ранее, и их связи между собой и пользователем. В данном разделе приведена общая архитектура инструмента и детали реализации его модулей, включая различные инженерные проблемы и пути их решения.

4.1 Общая архитектура

Реализованный инструмент может быть использован как консольное приложение или как библиотека, с помощью которой пользователь сможет реализовывать свои сценарии анализа или пользоваться отдельными её частями.

Основой инструмента являются несколько пакетов:

- 1) `analysis` – пакет, содержащий реализацию извлечения трасс с помощью статического и динамического анализа, а именно построение ICFG, `points-to` анализ, фаззинг и сбор точек входа в программу
- 2) `inference` – пакет, обеспечивающий обработку трасс и восстановление из них поведенческих моделей
- 3) `repository` – пакет, реализующий получение репозитория и работу с ними
- 4) `storage` – пакет, отвечающий за работу с БД
- 5) `workflow` – пакет, связывающий все реализованные возможности между собой с помощью определенных сценариев работы

4.2 Пакет `repository`

Данный пакет решает задачи поиска, получение и сборки проектов из GitHub и Maven Central Repository. Также в рамках этого пакета реализована

возможность запуска тестов, имеющихся в репозитории. Рассмотрим каждую из задач подробнее.

4.3 Пакет analysis

4.4 Пакет storage

4.5 Пакет inference

5 ТЕСТИРОВАНИЕ

ЗАКЛЮЧЕНИЕ

В результате работы предложен метод автоматического извлечения поведенческих моделей библиотек из открытых программных репозиториях. Метод заключается в применении статических и динамических подходов к полученным с GitHub или Maven Central Repository проектам для извлечения трасс и последующем применении алгоритма восстановления КА. Для апробации метода реализован инструментарий⁶, позволяющий запускать автоматические сценарии для получения поведенческих моделей в виде КА и предоставляющий элементы управления настройками поиска, анализа, восстановления и используемого окружения.

Результат исследования показал применимость предложенного подхода. С помощью поиска по GitHub и Maven Central Repository была решена задача получения общедоступных программ, использующих заданную библиотеку. Для извлечения трасс применены статический и динамический анализ. Статический заключается в обходе ICFG и применении Points-to анализа на основе алгоритма Андерсена, а динамический – в инструментации и запуске тестов или фаззинга входных точек в программу на основе Jazzer. Восстановление поведенческих моделей выполнено с применением алгоритма k-tail, реализованного в рамках инструмента MINT.

Представленный метод и инструмент требуют дальнейшего развития в нескольких направлениях.

Одна из проблем в том, что получаемые КА являются сильной аппроксимацией сверху и на сложных библиотеках имеют мало общего с истинными поведенческими моделями. Это является следствием используемого алгоритма

⁶<https://github.com/kechinvv/LibMiner>

восстановления и игнорирования состояния программы, включая значения аргументов. Для решения этого необходимо провести работу над инструментацией, реализовав сбор значений аргументов и доступных в контексте переменных. Это позволит применять более сложные алгоритмы восстановления, обеспечивающие лучшую обобщенность модели. При этом важно сохранить простоту применения инструментации для работы в автоматическом режиме. Другой способ получения информации о состоянии программы, причем сразу в форме предикатов - использование символьного исполнения, но это также объект для отдельного исследования.

Еще одна проблема, требующая решения – отсутствие тестов в проектах, их малое количество или плохая репрезентативность в контексте получаемых трасс. В данной работе для решения этой проблемы применяется фаззинг, но, вероятно, лучший результат покажет генерация тестов, направленная на получения трасс, содержащих поведение заданной библиотеки (схожее с Tautoko). Однако, это отдельная сложная задача, требующая соответствующего исследования.

Сильным ограничением для автоматизации динамических методов является зависимость программ от определенного внешнего окружения - базы данных, ПО, переменные окружения. Идеальным решением этого было бы мокирование подобного кода, интегрированное в задачу генерацию тестов. Это бы обеспечило высокую долю успешно исполняемых тестов. На текущий момент нет инструментов, позволяющих это осуществить – в связи со сложностью решения и узким направлением его применения актуален вопрос релевантности исследований по данной теме.

Не смотря на указанные недостатки, требующие дальнейшего развития и дополнительных исследований, предложенный метод является комплексным решением задачи получения поведенческих моделей. При этом для его применения требуется сравнительно небольшое участие человека. Что касается необходимости вручную обрабатывать полученные КА – это ограничение всех существующих подходов к восстановлению, на данный момент не имеющее другого решения, кроме как использовать различные алгоритмы и экспериментировать с настройкой их параметров. Хочется надеяться, что инструменты, позволяющие автоматизировать получение поведенческих моделей, в будущем смогут улучшить опыт разработки и применения анализов, основанных на использовании формальных спецификаций.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Ivanov D. и др. UTBot Java at the SBST2022 tool competition. 2023. сс. 39–40.
2. Ицыксон В.М. и др. Моделирование поведения функций стандартной библиотеки в задачах анализа программ // Информационно-управляющие системы. 2024. № 4. сс. 24–39.
3. Itsykson V. LibSL: Language for Specification of Software Libraries // Programmnaya Ingeneria. 2018. т. 9, № 5. сс. 209–220.
4. Biermann A., Feldman J. On the Synthesis of Finite-State Machines from Samples of Their Behavior // IEEE Transactions on Computers. 1972. № 6. сс. 592–597.
5. Lorenzoli D., Mariani L., Pezzè M. Automatic generation of software behavioral models // Proceedings - International Conference on Software Engineering. 2008. сс. 501–510.
6. Ernst M.D., Perkins J.H., Guo P.J. The Daikon system for dynamic detection of likely invariants // Sci. Comput. Program. 2007. т. 69. сс. 35–45.
7. Shoham S., Yahav E., J. S.F. Static Specification Mining Using Automata-Based Abstractions // IEEE Transactions on Software Engineering. 2008. т. 34, № 5. сс. 651–666.
8. Cousot P., Cousot R. Abstract Interpretation: A Unified Lattice Model for Static Analysis of Programs by Construction or Approximation of Fixpoints. 1977. сс. 238–252.
9. Lerch J. и др. Access-Path Abstraction: Scaling Field-Sensitive Data-Flow Analysis with Unbounded Access Paths (T). 2015. сс. 619–629.
10. Krka I., Brun Y., Medvidovic N. Automatic mining of specifications from invocation traces and method invariants // Proceedings of the 22nd ACM

SIGSOFT International Symposium on Foundations of Software Engineering.
2014. т. 9, № 5. сс. 178–189.

11. Caso G. и др. Automated Abstractions for Contract Validation // IEEE Trans. Software Eng. 2012. т. 38. сс. 141–162.
12. Walkinshaw N., Taylor R., Derrick J. Inferring Extended Finite State Machine models from software executions // 20th Working Conference on Reverse Engineering (WCRE). 2013. сс. 301–310.
13. Dallmeier V., Knopp N., Mallon C. Automatically Generating Test Cases for Specification Mining // IEEE Transactions on Software Engineering. 2012. т. 38, № 2. сс. 243–257.
14. Dallmeier V. и др. Mining object behavior with ADABU // WODA. 2006. с. .
15. About [электронный ресурс]. URL: <https://github.com/about>.