СОДЕРЖАНИЕ

# СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

– 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

# ТЕРМИНОЛОГИЯ

– SMT-решатель

– условия корректности

# 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 ее 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 USVM[ССЫЛКА] и UtBot[ССЫЛКА]. 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[ЕСТЬ ЛИ ССЫЛКА НА ПРОЕКТ?]. 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.

При всем этом формальные спецификации в общем случае не поставляются с программными библиотеками или другим ПО, что заставляет задуматься 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 библиотек 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 извлечения поведенческих моделей библиотек посвящена данная работа. В рамках ее выполнения будет 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 восстановления поведенческой модели в виде КА. На данный момент комплексных 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 использования пользовательских репозиториев, 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 программы. Что касается ДА, то здесь на результат могут влиять точки входа в программу. Так как трассы будут собираться из запусков имеющихся или сгенерированных тестов, невозможно гарантировать корректность начального состояния библиотеки.

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, 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 извлечения 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-tails[1] требует на вход исключительно последовательности вызовов и параметра k, определяющего длину соединяемых цепочек. Другой алгоритм, gk-tails[2], 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.

### 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 ее окружения. Это может сильно влиять на массовость применения подхода и его автоматизацию.

## 1.2 Обзор работ по извлечению спецификаций

В работе «Static Specification Mining Using Automata-Based Abstractions»[3] авторы статически собирают трассы в виде последовательностей объектов одного типа, используя абстрактную интерпретацию[4]. 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 вариаций описанных алгоритмов между собой, хотя было бы уместно провести сравнение с классическим алгоритмом k-tails[1]. В ограничениях подхода авторы описывают невозможность его применения для анализа проектов состоящих из десяток 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[5]. Авторы делают вывод, что получаемые поведенчески модели достоверно описывают поведение библиотек, однако являются сильной аппроксимацией сверху истиной 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.

Тест[6]

ТЕст2[3]

## 1.3 Результаты анализа

# 2 ПУТИ РЕШЕНИЯ

## 2.1 подсекция

# 3 ПРОЕКТИРОВАНИЕ

# 4 РЕАЛИЗАЦИЯ

# 5 ТЕСТИРОВАНИЕ

# ЗАКЛЮЧЕНИЕ

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ


1. Biermann A., Feldman J. On the Synthesis of Finite-State Machines from Samples of Their Behavior // IEEE Transactions on Computers. 1972. № 6. cc. 592–597.
2. Lorenzoli D., Mariani L., Pezzè M. Automatic generation of software behavioral models // Proceedings - International Conference on Software Engineering. 2008. cc. 501–510.
3. Shoham S., Yahav E., J. S.F. Static Specification Mining Using Automata-Based Abstractions // IEEE Transactions on Software Engineering. 2008. т. 34, № 5. cc. 651–666.
4. Cousot P., Cousot R. Abstract Interpretation: A Unified Lattice Model for Static Analysis of Programs by Construction or Approximation of Fixpoints. 1977. cc. 238–252.
5. Lerch J. и др. Access-Path Abstraction: Scaling Field-Sensitive Data-Flow Analysis with Unbounded Access Paths (T). 2015. cc. 619–629.
6. Walkinshaw N., Taylor R., Derrick J. Inferring Extended Finite State Machine models from software executions // 20th Working Conference on Reverse Engineering (WCRE). 2013. cc. 301–310.