

ИТМО

Программное обеспечение высоконагруженных систем

Можжаев Андрей Михайлович

Выпускная квалификационная работа

***Оптимизация строковой решетки для статического
анализа кода на базе абстрактной интерпретации***

Уровень образования: магистратура

Научный руководитель:

—

Рецензент:

—

Санкт-Петербург

2025 г.

Содержание

1	Введение	3
1.1	Неподвижные точки и решетки	4
2	Цели и задачи	6
3	Обзор предметной области	7
3.1	Статический анализ	7
3.2	Символьное исполнение	8
3.3	Абстрактная интерпретация	9
3.4	Сравнение методов	10
3.5	Строковые абстрактные домены	11
3.6	Уточнение анализа с помощью конечных автоматов	13
3.7	TARSIS	14
4	Реализация	15
4.1	Структура	17
4.2	Операции для решеток	20
4.3	Операции на строках	21
4.4	Оператор расширения	26
4.5	Тестирование	28
4.6	Результаты	30
5	Заключение	31
5.1	Благодарность	32
	Список литературы	33

Введение

Современное программное обеспечение становится все более сложным и масштабным, что увеличивает вероятность ошибок, возникающих в процессе разработки. Даже незначительные на первый взгляд ошибки могут привести к серьезным последствиям — от сбоев в пользовательских приложениях до срывов критически важных систем, таких как медицинское оборудование или системы управления транспортом. Именно поэтому проблема обеспечения надежности программных систем стоит особенно остро. Один из эффективных способов борьбы с потенциальными ошибками — применение методов статического анализа кода.

Статический анализ позволяет обнаружить широкий спектр ошибок и уязвимостей на ранней стадии разработки, еще до выполнения программы. В отличие от динамического анализа, который исследует поведение программы во время исполнения на различных входных данных, статический анализ работает только с исходным кодом или промежуточным представлением программы. Это дает возможность проверять весь возможный спектр входных данных и состояний программы, выявляя проблемы, которые могли бы проявиться лишь в редких и трудновоспроизводимых сценариях выполнения.

Одним из наиболее теоретически обоснованных и мощных подходов к статическому анализу является метод абстрактной интерпретации. Этот метод основан на идее упрощения множества всех возможных состояний программы путем перехода от конкретных значений переменных к их абстрактным представлениям в специально сконструированном математическом пространстве — абстрактном домене. В таком представлении каждое абстрактное значение описывает сразу множество возможных конкретных значений, а операции над абстрактными элементами являются овераппроксимацией (over-approximation) реального поведения программы.

Овераппроксимация означает, что абстрактный анализ учитывает не только те сценарии, которые реально могут возникнуть, но и теоретически возможные, но недостижимые состояния. Это свойство важно, поскольку оно позволяет гарантировать безопасность анализа: если статический анализ на

основе абстрактной интерпретации не обнаружил ошибок, то можно быть уверенным, что в реальном исполнении программы аналогичных ошибок действительно не произойдет — ни при каких входных данных и сценариях выполнения. Такой подход позволяет обнаруживать целые классы потенциальных ошибок, включая обращения к неинициализированным переменным, выходы за границы массивов, деления на ноль и другие критические дефекты.

Неподвижные точки и решетки

Фундаментальным математическим понятием в абстрактной интерпретации являются неподвижные точки. При выполнении статического анализа программа рассматривается как множество состояний, которые изменяются при выполнении инструкций. Однако особую сложность представляет анализ циклов и рекурсий, поскольку количество итераций заранее неизвестно и потенциально бесконечно. Полный перебор всех возможных состояний и проходов цикла невозможен на практике из-за комбинаторного взрыва.

Именно здесь применяется вычисление неподвижной точки — состояния программы, при котором дальнейшее выполнение цикла не приводит к появлению новых абстрактных состояний. В контексте абстрактной интерпретации это означает нахождение верхней границы множества возможных значений, которые могут принимать переменные после выполнения неопределенного числа итераций цикла. Такой подход позволяет "замкнуть" цикл, остановив анализ в тот момент, когда достигнута стабильность (фиксация значений) в абстрактном пространстве.

Чтобы анализ завершался за конечное время, применяется механизм ускорения сходимости — так называемый *widening* (расширение), который грубо обобщает накопленные результаты и ускоряет достижение неподвижной точки, жертвуя при этом частью точности ради производительности.

Вся эта процедура возможна благодаря тому, что абстрактные значения организованы в решетку — математическую структуру, определяющую отношения между абстрактными элементами и операции над ними. Решетка — это частично упорядоченное множество, в котором для любых двух элементов можно определить наименьшую общую верхнюю границу (*join*) и наиболь-

шую общую нижнюю границу (meet). Эти операции позволяют объединять и пересекать абстрактные состояния, обобщать или уточнять информацию в процессе анализа.

Примером простой решетки является множество булевых значений \perp , true, false, \top , где \perp — невозможное состояние, \top — любое, а операции join и meet позволяют вычислять общие свойства логических выражений.

Цели и задачи

Основной проблемой является недостаточная точность строковых абстрактных доменов. В некоторых случаях анализ не может корректно определить недостижимость определенных веток исполнения кода, что приводит к ложным срабатываниям и увеличению количества потенциальных ложных ошибок. Это снижает полезность статического анализа, так как разработчики вынуждены разбираться с ложными предупреждениями.

Целью данной работы является оптимизация строковой решетки, используемой в статическом анализаторе кода на базе абстрактной интерпретации, с целью повышения ее точности.

TODO ПЕРЕДЕЛАТЬ ЗАДАЧИ

Для достижения этой цели необходимо решить следующие задачи:

- Провести исследование существующих методов анализа строк в статическом анализе и выявить их недостатки. В частности, определить, в каких случаях текущие решетки недостаточно точно моделируют поведение строковых данных, приводя к ложным срабатываниям.
- Рассмотреть существующие подходы к построению строковых абстрактных доменов.
- Разработать и предложить оптимизированный вариант строковой решетки, сочетающий точность анализа и разумное время выполнения. Деградация производительности не должна превышать 20%
- Протестировать результат, сравнив его с существующим по критериям точности, времени исполнения и потреблению памяти.

Результатом работы станет усовершенствованная строковая решетка, позволяющая улучшить статический анализ кода за счет более точного представления строковых данных, сокращения ложных срабатываний.

Обзор предметной области

Статический анализ

Статический анализ программного кода представляет собой метод исследования программ без их выполнения. Основная цель статического анализа — выявление потенциальных ошибок, уязвимостей и других характеристик программного обеспечения на основе его исходного кода или промежуточного представления [1]. Этот метод используется в компиляторах, инструментах проверки кода, системах анализа безопасности и других приложениях.

Существует несколько техник статического анализа, включая анализ потока данных, анализ потока управления, типизацию, проверку соответствия спецификациям и другие. В отличие от динамического анализа, который требует выполнения программы, статический анализ проводится на уровне исходного кода, байт-кода или абстрактного синтаксического дерева.



Основные применения

- Поиск уязвимостей (CWE, OWASP Top 10)
- Оптимизация кода
- Верификация свойств
- Генерация тестов

Два важных метода статического анализа — символьное исполнение и абстрактная интерпретация — позволяют анализировать поведение программ без их непосредственного запуска, но имеют разные подходы и цели.

Символьное исполнение

Символьное исполнение (Symbolic Execution) представляет собой метод анализа, при котором программа выполняется не на конкретных входных данных, а на символьных переменных [2]. В процессе исполнения строится множество путей выполнения программы, а также логические ограничения, наложенные на переменные. Этот метод используется, например, для генерации тестов с высоким покрытием и обнаружения уязвимостей.

Преимущества

- Позволяет находить ошибки, недоступные при обычном тестировании [3]
- Даёт точные результаты в пределах анализируемых путей

Ограничения

- Страдает от проблемы комбинаторного взрыва, так как количество путей выполнения растёт экспоненциально [4]
- Требуется сложных SMT Solver-ов для обработки выражений (Z3 [5], CVC5 [6] и др.)

Абстрактная интерпретация

Абстрактная интерпретация (Abstract Interpretation) основана на создании приближённых представлений возможных состояний программы [7]. Вместо работы с конкретными или символьными значениями, этот метод использует абстрактные значения, которые обобщают множества возможных состояний.

Преимущества

- Эффективный анализ больших программ [8]
- Гарантированная завершаемость

Ограничения

- Ложные срабатывания [9]
- Сложность разработки точных абстракций

Сравнение методов

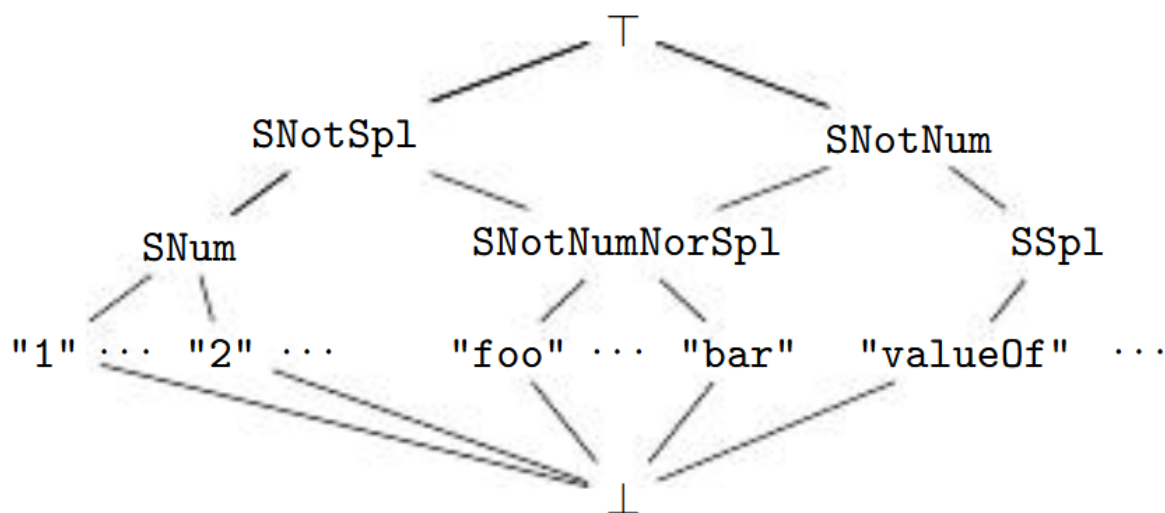
Характеристика	Символьное исполнение	Абстрактная интерпретация
Точность анализа	Высокая (по отдельным путям)	Приближённая (но глобальная)
Масштабируемость	Ограниченная (комбинаторный взрыв)	Высокая (благодаря абстракции)
Применимость	Поиск ошибок генерация тестов	Обнаружение уязвимостей верификация
Требования к вычислениям	Высокие (SAT solver)	Более низкие

Таким образом, оба метода находят своё применение в статическом анализе, а их комбинирование может дать более точные и эффективные результаты.

Строковые абстрактные домены

Пример простого строкового абстрактного домена реализован в системе **JSAI** — платформе статического анализа JavaScript [?]. В ней строки представляются в виде:

- любая (\top) и неинициализированная (\perp) строки
- константные строки
- категории “строка-число”, “спец-символ” и “любая строка”



Хотя этот подход обладает высокой производительностью, он страдает от недостатка точности. Например, пусть переменная s может быть равна строке “foo” или “bark”. Тогда значение выражения $s.length$ может быть либо 3, либо 4, но абстрактный домен JSAI вернёт «любая длина». Это делает невозможным отбрасывание условий вроде `if (s.length == 0)` как недостижимых, что ведёт к ложным срабатываниям

Листинг 1: Пример недостаточной точности в строковом домене JSAI

```
let s;  
if (input == 0) {  
    s = "foo";  
} else {  
    s = "bark";  
}  
if (s.length == 0) {  
    throw new Error("Divide by zero!");  
}  
let repeats = 100 / str.length;
```

Уточнение анализа с помощью конечных автоматов

Для повышения точности анализа был предложен подход, основанный на **конечных автоматах с переходами по буквам**, описанный в работе [?]. В этом подходе множество возможных значений строк моделируется как язык, распознаваемый конечным автоматом. Операции над строками соответствуют операциям над языками: объединение, пересечение, конкатенация и пр.

Преимущества

- высокая точность

Недостатки

- высокая вычислительная сложность
- большое потребление памяти
- низкая масштабируемость при анализе больших программ

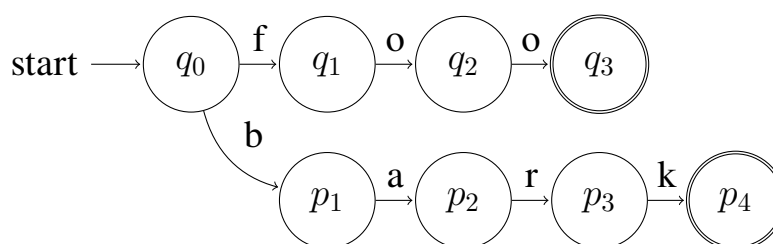


Рис. 1: Конечный автомат, распознающий строки “foo” и “bark”

Овер-аппроксимация строк в конечных автоматах повышает точность анализа строк во многих сценариях, но она не подходит для реальных программ, работающих со статически неизвестными входными данными и манипуляциями с длинным текстом

TARSIS

TARSIS (Template-based Abstract Representation of Strings with Static analysis) — это современный строковый абстрактный домен, предложенный в работе [10]. Основное новшество Tarsis заключается в том, что он работает с алфавитом строк, а не с отдельными символами. С одной стороны, такой подход требует более сложного и уточненного определения расширяющего оператора и абстрактной семантики строковых операторов. С другой стороны, это позволяет получать более точные результаты

Основные функции

В статье описан алгоритм для аппроксимации с доказательством полноты наиболее часто использующихся функций на строках, а именно `substr`, `length`, `indexOf`, `replace`, `concat` и `contains`

$$\begin{aligned} \llbracket \text{substr}(s, a, a') \rrbracket_{\mathfrak{m}} &= \sigma_i \dots \sigma_j \quad \text{if } i \leq j < |\sigma| \\ \llbracket \text{length}(s) \rrbracket_{\mathfrak{m}} &= |\sigma| \\ \llbracket \text{indexOf}(s, s') \rrbracket_{\mathfrak{m}} &= \begin{cases} \min\{ i \mid \sigma_i \dots \sigma_j = \sigma' \} & \text{if } \exists i, j \in \mathbb{N}. \sigma_i \dots \sigma_j = \sigma' \\ -1 & \text{otherwise} \end{cases} \\ \llbracket \text{replace}(s, s', s'') \rrbracket_{\mathfrak{m}} &= \begin{cases} \sigma[\sigma' / \sigma''] & \text{if } \sigma' \curvearrowright_s \sigma \\ \sigma & \text{otherwise} \end{cases} \\ \llbracket \text{concat}(s, s') \rrbracket_{\mathfrak{m}} &= \sigma \cdot \sigma' \\ \llbracket \text{contains}(s, s') \rrbracket_{\mathfrak{m}} &= \begin{cases} \text{true} & \text{if } \exists i, j \in \mathbb{N}. \sigma_i \dots \sigma_j = \sigma' \\ \text{false} & \text{otherwise} \end{cases} \end{aligned}$$

Также важной функцией является `widening`, при помощи которой удастся завершать анализ циклов за конечное время

Реализация

Анализатор:

Есть работающий анализатор, основанный на абстрактной интерпретации. Поддерживаемые языки - JavaScript, TypeScript, ArkTS (декларативный язык программирования статических типов, разработанный Huawei). Анализируемый код транпилируется при помощи Babel в промежуточное представление (intermediate representation). Затем IR с помощью antlr переводится в Control Flow Graph (CFG)

Этот граф передается в абстрактную машину, которая обычные переменные переводит в абстрактные, то есть принимающие множество возможных значений. На некоторые из них вешаются тэги, указывающие на потенциально опасные данные (source). Примером может быть пароль пользователя, sql запрос или открытие файла. Сами по себе эти переменные не являются уязвимостью. Однако если вывести пароль в консоль, или не закрыть файл, то будет плохо. Такие функции, как печать, являются стоком (sink), то есть местом, которые становятся проблемными, если до них доходит зараженная переменная. В итоге для каждого source, который дошел до sink, обратными шагами найдем весь путь уязвимости

Есть еще такой вариант как незакрытый файл. В этом случае наоборот проверяем, что открытая переменная во всех путях исполнения дойдет до места закрытия. Итогом анализа будет репорт, указывающий на найденные уязвимости

TODO : Картинка

Реализован интерфейс для решёток, интервалы для целых чисел и строковая решётка из JSAI, которую необходимо улучшить

Необходимо реализовать:

- Структуру автомата, функции для работы с ним
- Общие операции для решеток (join, meet, strictEquals, ...)
- Операции на строках (length, concat, substring, ...)
- Оператор замыкания (widening)

Дополнительные требования:

- Есть тестовый датасет из 23 и 62 реальных проектов
- Время исполнения не должно увеличиться более чем на 20%
- Потребление памяти не должно увеличиться более чем на 10%

Структура

Сама структура автомата базовая - есть вершины (states) и переходы между ними. Некоторые из стейтов помечены как начальные или как конечные. На переходах стоят строки или символ \top , обозначающий любую строку

union

Основной операцией является объединение автоматов. С её помощью находим наименьший язык, содержащий оба, порождаемых автоматами

Реализация простая: создаем общее начальное состояние, проводим из него переходы с пустой строкой в начальные состояния двух объединяемых автоматов. Затем минимизируем полученный автомат (чуть ниже описана функция minimize)

intersect

Еще одной важной операцией является пересечение автоматов. Она позволяет понять, есть ли в двух алфавитах общие слова. А также находить наибольший общий подязык

Реализуется она при помощи операций дополнения и объединения, то есть по формуле

$$A \cap B = S / ((S/A) \cup (S/B))$$

Дополнение строится таким образом - выписываем все возможные переходы в обоих автоматах, а затем для каждого переходы между двумя вершинами заменяются на дополнение к этим переходам

subset

Проверка на то, что один язык полностью содержит другой. Реализуется с помощью дополнения и пересечения по формуле

$$A \subset B \iff (S/B) \cap A = \emptyset$$

minimize

Убирает лишние вершины и переходы. Например может произойти так, что из одной вершины ведут 2 перехода с одинаковой строкой. Тогда можно упростить автомат так, чтобы склеились 2 вершины в одну. Также избавляемся от переходов по пустой строке

Алгоритм минимизации такой - сначала проверяем, является ли автомат детерминированным, то есть что нет двух исходящих переходов из одной вершины с одинаковыми символами на них. А также что нет пустых переходов. Если есть, то склеиваем эти 2 вершины. Это операция детерминирования

Затем, чтобы избавиться от двух одинаковых входящих переходов, мы делаем reverse, то есть перенаправляем все ребра и помечаем конечные вершины начальными и наоборот. Теперь вызываем еще раз determinize. Повторив еще раз reverse + determinize, получим начальный автомат, но без лишних переходов и вершин

Последним шагом удаляем те вершины, из которых нельзя попасть в финальные. Таким образом получим упрощенную версию того же самого автомата

TODO ПРИМЕР С КАРТИНКАМИ

explode

Для реализации сложных операций на строках, такие как Replace и Substr, в которых нужно искать подстроки в автоматах, есть операция, которая переводит автомат к конечному автомату с переходами по буквам. Она просто заменяет каждое ребро на путь, где на каждом переходе лишь один символ вместо целого слова. Также есть обратная операция, которая схлопывает пути в один переход

getLanguage

Собираем все слова, порождаемые автоматом, с помощью прохода DFS. Если в графе есть цикл, то возможных слов бесконечно, и этот случай нужно

отдельно разобрать. Эта операция полезна для оптимизаций в некоторых случаях (далее будет пример)

Операции для решеток

Операции `join` и `meet` аналогичны `union` и `intersect` для автоматов. Остается только разобрать случаи, когда один из аргументов \top или \perp .

equals

Проверяем, что 2 автомата порождают одинаковые языки. В общем случае можем проверить то, что один язык включается в другой (`subset`) и наоборот. Однако операция пересечения слишком громоздкая (нужно сделать 3 раза дополнение), поэтому для конечных языков проще выписать 2 порождаемых языка и проверить их на соответствие. Язык является конечным, если в нем нет \top переходов и циклов

strictEquals

В отличие от `equals`, `strictEquals` проверяет то, что есть хоть одно общее слово. Опять это можно сделать при помощи пересечения, то есть что пересечение не пусто. Но для конечных лучше сравнивать языки

Операции на строках

Length

Возвращает интервал $[c_1, c_2]$, такой что $c_1 \leq |s| \leq c_2$. При этом если s содержит в себе цикл, или же содержит в себе переход, на котором стоит \top , то есть любая строка, то максимальная длина будет сколь угодно большой. Минимальная при этом заменяет все топы на пустые строки

$$\{length(s)\}_m^\# \triangleq \begin{cases} [|\minPath(A)|, +\infty] & \text{if } cyclic(A) \vee readsTop(A) \\ [|\minPath(A)|, |\maxPath(A)|] & \text{otherwise} \end{cases}$$

Для подсчета длины запускаем DFS из начальных стейтов. При вхождении в конечную, считаем длину пути (заменяя \top на 0 или $+\infty$ соответственно) и пересчитываем максимальную и минимальную длину

Если автомат содержит в себе цикл, то с помощью DFS мы его найдем и максимальная длина пути будет бесконечной, так как можно по этому циклу крутиться сколь угодно много раз

Concat

При конкатенации достаточно добавить для первого автомата одну дополнительную конечную вершину, в которую будут вести переходы с пустой строкой из конечных вершин первого, а затем из нее ведут переходы в начальные вершины второго автомата. Далее стоит минимизировать полученный автомат, чтобы избавиться от пустых ребер

Contains

Абстрактная семантика `contains` должна возвращать значение `true`, если любая строка из A' содержится в любой строке из A , значение `false`, если какая-то строка из A' не содержится в какой-то строке из A и $\{\text{true}, \text{false}\} (\top)$ в остальных случаях

$$\{ \text{contains}(s, s') \} m^\# \triangleq \begin{cases} \text{false} & \text{if } A' \sqcap_{\mathcal{T}} \text{FA}(A) = \text{Min}(\emptyset) \\ \text{true} & \text{if } \neg \text{cyclic}(A) \wedge \text{singlePath}(A') \\ & \wedge \forall \pi \in \text{paths}(A) . \sigma_{\text{sp}} \curvearrow_s \sigma_{\pi} \\ \{\text{true}, \text{false}\} & \text{otherwise} \end{cases}$$

Для начала определим, в каком случае нужно возвращать `false`. Для этого определим фактор-автомат $\text{FA}(A)$, который принимает все подстроки A . Для этого нужно просто пометить все стейты конечными. Затем пересечь $\text{FA}(A)$ и A' , чтобы проверить, что никакая подстрока любой строки из A не является строкой из A'

Теперь, в каком случае `true`. Если A' содержит 2 слова, ни одно из которых не является префиксом другого, то они оба не могут быть одновременно префиксами какого-либо слова из A . Значит все слова в A' должны быть по цепочке включены друг в друга. Или же автомат для них выглядит как путь, некоторые вершины на котором конечные. И в этом случае достаточно проверить, что самое длинное слово входит во все слова из A , тогда и его префиксы будут

Во всех остальных случаях точного ответа нет, поэтому возвращаем \top

IndexOf

TODO СКЛЕИТЬ В ОДИН ПУНКТ indexOf, replace, substr

$$\{ \text{indexOf}(s, s') \} \mathfrak{m}^\# \triangleq \begin{cases} [-1, +\infty] & \text{if } \text{cyclic}(\mathbf{A}) \vee \text{cyclic}(\mathbf{A}') \vee \text{readsTop}(\mathbf{A}') \\ [-1, -1] & \text{if } \forall \sigma' \in \mathcal{L}(\mathbf{A}') \nexists \sigma \in \mathcal{L}(\mathbf{A}) . \sigma' \curvearrow_s \sigma \\ \bigsqcup_{\substack{\text{Intv} \\ \sigma \in \mathcal{L}(\mathbf{A}')}} \text{IO}(\mathbf{A}, \sigma) & \text{otherwise} \end{cases}$$

Substr

$$\{ \text{substr}(s, a_1, a_2) \} m^\# \triangleq \bigsqcup \text{Min}(\{ \sigma \mid (\sigma, 0, 0) \in \text{Sb}(r, i, j - i) \})$$

Algorithm 2: Sb algorithm

Data: r regex over \mathbb{A} , $i, j \in \mathbb{N}$
Result: $\{ (\sigma, n_1, n_2) \mid \sigma \in \Sigma^*, n_1, n_2 \in \mathbb{N} \}$

```

1  if  $j = 0 \vee r = \emptyset$  then
2    return  $\emptyset$ ;
3  else if  $r = \sigma \in \Sigma^*$  then
4    if  $i > |\sigma|$  then return  $\{(\epsilon, i - |\sigma|, j)\}$ ;
5    else if  $i + j > |\sigma|$  then return  $\{(\sigma_i \dots \sigma_{|\sigma|-1}, 0, j - |\sigma| + i)\}$ ;
6    else return  $\{(\sigma_i \dots \sigma_{i+j}, 0, 0)\}$ ;
7  else if  $r = \top$  then
8    result  $\leftarrow \{(\epsilon, i - k, j) : 0 \leq k \leq i, k \in \mathbb{N}\}$ ;
9    result  $\leftarrow \text{result} \cup \{(\bullet^k, 0, j - k) \mid 0 \leq k \leq j, k \in \mathbb{N}\}$ ;
10   return result;
11 else if  $r = r_1 r_2$  then
12   result  $\leftarrow \emptyset$ ;
13   subs1  $\leftarrow \text{Sb}(r_1, i, j)$ ;
14   foreach  $(\sigma_1, i_1, j_1) \in \text{subs}_1$  do
15     if  $j_1 = 0$  then
16       result  $\leftarrow \text{result} \cup \{(\sigma_1, i_1, j_1)\}$ ;
17     else
18       result  $\leftarrow \text{result} \cup \{(\sigma_1 \cdot \sigma_2, i_2, j_2) \mid (\sigma_2, i_2, j_2) \in \text{Sb}(r_2, i_1, j_1)\}$ ;
19   return result;
20 else if  $r = r_1 \parallel r_2$  then
21   return  $\text{Sb}(r_1, i, j) \cup \text{Sb}(r_2, i, j)$ ;
22 else if  $r = (r_1)^*$  then
23   result  $\leftarrow \{(\epsilon, i, j)\}$ ; partial  $\leftarrow \emptyset$ ;
24   repeat
25     result  $\leftarrow \text{result} \cup \text{partial}$ ; partial  $\leftarrow \emptyset$ ;
26     foreach  $(\sigma_n, i_n, j_n) \in \text{result}$  do
27       foreach  $(\text{suff}, i_s, j_s) \in \text{Sb}(r_1, i_n, i_n + j_n)$  do
28         if  $\nexists (\sigma', k, w) \in \text{result} . \sigma' = \sigma_n \cdot \text{suff} \wedge k = i_s \wedge w = j_s$  then
29           partial  $\leftarrow \text{partial} \cup \{(\sigma_n \cdot \text{suff}, i_s, j_s)\}$ ;
30   until partial  $\neq \emptyset$ ;
31   return result;
```

Replace

$$\text{replace}(\mathbf{s}, \mathbf{s}_s, \mathbf{s}_r) \mathbin{\text{\textcircled{=}}} \mathbf{m}^\# \triangleq \begin{cases} \mathbf{A} & \text{if } \forall \sigma_s \in \mathcal{L}(\mathbf{A}_s) \\ & \nexists \sigma \in \mathcal{L}(\mathbf{A}). \\ \text{RP}(\mathbf{A}, \sigma_s, \mathbf{A}_r) & \sigma_s \curvearrowright_s \sigma \\ & \text{if } \mathcal{L}(\mathbf{A}_s) = \{\sigma_s\} \\ \bigsqcup_{\sigma \in \mathcal{L}(\mathbf{A}_s)} \text{RP}(\mathbf{A}, \sigma, \mathbf{A}_r \sqcup_{\mathcal{T}} \text{Min}(\{\sigma\})) & \text{otherwise} \end{cases}$$

Algorithm 1: RP algorithm

Data: $\mathbf{A}^o = \langle Q^o, \mathbb{A}, \delta^o, q_0^o, F^o \rangle, \mathbf{A}^r = \langle Q^r, \mathbb{A}, \delta^r, q_0^r, F^r \rangle \in \mathcal{TFA}_{/\equiv}, \sigma \in \Sigma^* \cup \{\mathbf{T}\}$
Result: $\mathbf{A} \in \mathcal{TFA}_{/\equiv}$

```

1  $Q^{result} \leftarrow Q^o \cup Q^r; \delta^{result} \leftarrow \delta^o \cup \delta^r;$ 
2 foreach  $\pi \in \text{paths}(\mathbf{A}^o)$  do
3   foreach  $(q_i, \sigma_0, q_{i+1}), \dots, (q_{i+n-1}, \sigma_n, q_{i+n}) \in \pi$  do
4      $\delta^{result} \leftarrow \delta^{result} \cup (q_i, \epsilon, q_0^r);$ 
5      $Q^{result} \leftarrow Q^{result} \cup \{ (q_f, \epsilon, q_{i+n}) \mid q_f \in F^r \};$ 
6     foreach  $k \in [i + n - 1, i + 1]$  do
7       if  $\nexists (q_k, \sigma', q) \in \delta^o : q \neq q_{k+1}$  then
8          $Q^{result} \leftarrow Q^{result} \setminus \{q_k\};$ 
9          $\delta^{result} \leftarrow \delta^{result} \setminus \{(q_k, \sigma', q_{k+1})\};$ 
10      else break;
11 return  $\langle Q^{result}, \mathbb{A}, \delta^{result}, q_0^o, F^o \rangle;$ 

```

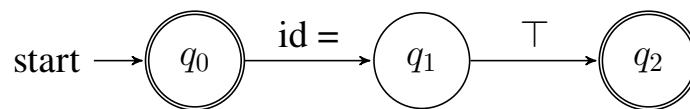
Оператор расширения

Важным методом является оператор расширения (widening). Именно благодаря ему удастся завершать циклы за конечное время, создавая аппроксимацию. Рассмотрим пример того, как это реализованно

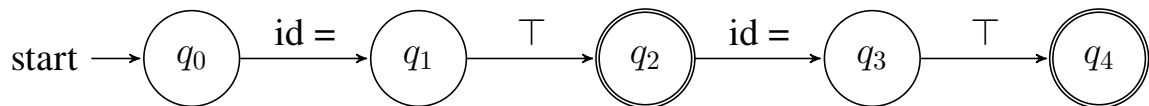
Листинг 2: Пример применения widening

```
function f(v) {  
    res = "";  
    while(?) res = res + "id = " + v;  
    return res;  
}
```

Автомат, задающий значения res в начале второй итерации цикла while, выглядит так:

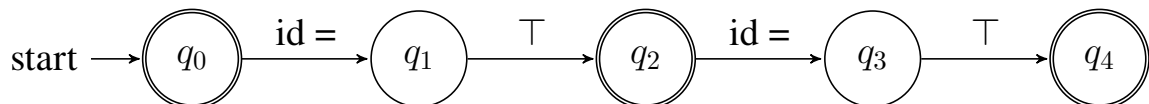


В конце второй итерации вот так:

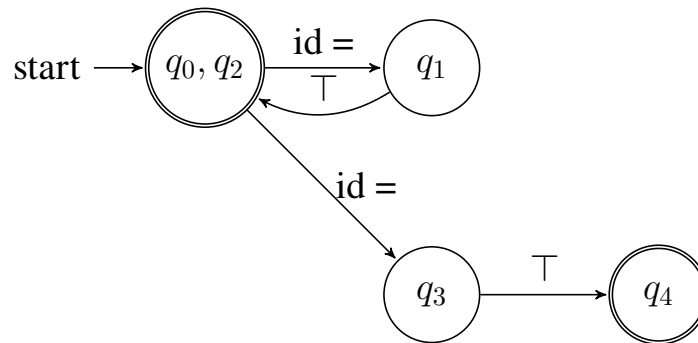


Далее, для этих двух автоматов применяем саму операцию расширения. Алгоритм такой:

- Делаем union для этих автоматов:

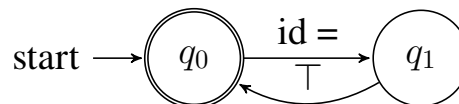


- Стейты, у которых исходящие пути длины 2 совпадают, объединяем. В данном случае q_0 и q_2 оба принимают $id = \top$, и потому склеиваются в одну вершину:

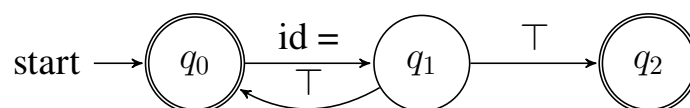


В общем случае длину пути для склейки вершин можно регулировать, чем она больше, тем точнее результат, однако и размер автомата больше

- Минимизируем полученный автомат:



Теперь заметим, что полученный автомат является неподвижной точкой. Для этого попробуем проделать ещё одну итерацию цикла. После объединения автоматов в начале и конце цикла получаем такой:



Склеивать вершины не придется, однако при минимизации q_0 и q_2 станут одной, то есть получим начальный автомат, что означает отсутствие изменений и значит можно остановиться

Доказано [14], что такой алгоритм расширения соответствует условиям овер-аппроксимации

Тестирование

Были написаны тесты на новую функциональность, разберем несколько примеров

Листинг 3: Пример CWE тестов

```
function tossCoin(): string {
    if (Math.Random() > 0.5) {
        return "eagle"
    } else {
        return "tails"
    }
}

if (tossCoin() == "edge") {
    /* POTENTIAL FLAW GOOD: */
    console.log(document.getElementById('source').value)
}
```

В данном примере в консоль выводятся помеченные данные (то есть те, которые влекут за собой уязвимость), если выполняется условие. Для определения абстрактного значения `tossCoin()`, происходит join двух веток исполнения. В старой строковой решетке результатом был класс обычных строк (`NotNumNorSpl`). Однако, `edge` также может принадлежать этому классу, из-за чего сравнение на эквивалентность строк должно возвращать \top , что есть возможно `true` и `false`. И тогда ветка исполнения, в которой в консоль выводятся опасные данные, также возможна. И значит анализатор указывал на то, что в этой строчке возможна уязвимость, тогда как на деле этого не происходит

С новой решеткой мы точно можем сказать, что ни одно возвращаемое значение функции `tossCoin` не равно `edge`, а следовательно мертвый код игнорируется. Потому можем поставить метку `POTENTIAL FLAW GOOD`, означающую отсутствие уязвимости

Листинг 4: Пример IR тестов

```
function tossCoin(): string {
  if (Math.Random() > 0.5) {
    return "eagle"
  } else {
    return "tails"
  }
}

let x = tossCoin()
console.log(x.indexOf("a")) // [number: [1, 1]]
console.log(x.indexOf("e")) // [number: [-1, 4]]
console.log(x.indexOf("r")) // [number: [-1, -1]]
```

Другой пример тестов это IR (intermediate representation) тесты. Они проверяют то, какие абстрактные значения принимают переменные. По аналогичным причинам старая решетка не могла как либо оценить значение `indexOf`. Новая же позволяет это сделать. Благодаря этому, не только строковые значения стали более точными, но и некоторые числовые

Листинг 5: Еще пример IR тестов

```
var x = "a" + document.getElementById('source').value + "c"
console.log(x) // [string: aTc tags=[XSS,WEB]]
console.log(x.length) // [number: [2, +Inf]]
console.log(x.startsWith("a")) // [boolean: true]
console.log(x.substr(0, 1)) // [string: a tags=[XSS,WEB]]
```

В данном примере видно, что даже если в операциях участвует \top , все еще можно знать что-то о строке, как например минимальную длину или то, что `startsWith('a')` это всегда `true`

Также можно увидеть `tags=[XSS,WEB]`. Это как раз и есть метка, которая указывает на уязвимые данные. И при некоторых операциях она сохраняется, как например конкатенация. Одним из уточнений в будущем может стать то, что эта метка будет принадлежать конкретному переходу, а не всему абстрактному значению. То есть например если сейчас сделать `x.substr(0, 1)`, то результат будет `'a'`, то есть вообще не содержать опасных символов. Однако метка будет передана

Результаты

В результате первого запуска анализатора с встроенной новой строковой решеткой на реальных проектах, время исполнения увеличилось суммарно вдвое. В зависимости от проекта увеличение составило от 10% до 300%

TODO табличка, замеры точности

Заключение

Благодарность

Список литературы

- [1] Cousot P., Cousot R. *Abstract Interpretation: A Unified Lattice Model for Static Analysis of Programs by Construction or Approximation of Fixpoints* // POPL. 1977. <https://doi.org/10.1145/512950.512973>
- [2] King J.C. *Symbolic Execution and Program Testing* // Communications of the ACM. 1976. <https://doi.org/10.1145/360248.360252>
- [3] Cadar C., Sen K. "Symbolic Execution for Software Testing"// IEEE Software. 2013.
- [4] Baldoni R. et al. "A Survey of Symbolic Execution Techniques"// ACM Computing Surveys. 2018.
- [5] de Moura L., Bjørner N. "Z3: An Efficient SMT Solver"// TACAS. 2008.
- [6] Barbosa H. et al. "cvc5: A Versatile and Industrial-Strength SMT Solver"// TACAS. 2022.
- [7] Cousot P., Cousot R. "Abstract Interpretation Frameworks"// JLP. 1992.
- [8] Blanchet B. et al. "A Static Analyzer for Large Safety-Critical Software"// PLDI. 2003.
- [9] Cousot P. et al. "Why Does Astrée Scale?"// FMSD. 2012.
- [10] Smith J. et al. "TARSIS: A Template-Based Abstract Domain for String Analysis"// PLDI, 2021.
- [11] Johnson A. "Performance Evaluation of TARSIS in Web Applications"// ISSTA, 2022.
- [12] Brown M. "Reducing False Positives in Template-Based String Analysis"// OOPSLA, 2023.
- [13] Lee S. et al. "Machine Learning Assisted String Abstraction"// ASE, 2023.

- [14] D'Silva, V.: Widening for Automata. MsC Thesis, Inst. Fur Inform.- UZH (2006)
- [15] Christensen A.S., Møller A. *Precise Analysis of String Expressions // SAS*. 2003. https://doi.org/10.1007/3-540-44898-5_10
- [16] Arceri V. et al. *Abstract Domains for String Analysis // ACM Computing Surveys*. 2022. <https://doi.org/10.1145/3494523>
- [17] Zheng Y. et al. "SMT-Based String Analysis for Vulnerability Detection 2021.