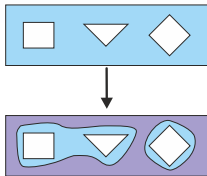


Анализ моделей в теории строк для преобразования и ускорения Рефал-программ



Непейвода А.Н.
Совместное совещание по языку Рефал
ИПС им. А.К. Айламазяна РАН
и МГТУ им. Н.Э. Баумана
17 июня 2023 г.



Бескванторный SMT-фрагмент

- Можно объявлять параметры:
(`declare-fun x () String`).
- Далее параметры связываются допущениями (аксиомами): (`assert (str.contains x "a")`).

Описывает конечное число пропозициональных формул над операторами и предикатами.

- Существование модели — поиск контрпримеров, атак, ошибок (например, в конколической интерпретации или фаззинге).
- Несуществование модели — возможные оптимизации и верификация безопасности.



Теория строк: базовые структуры

- Основная операция — конкатенация, базовый предикат — равенство.
 - $X_1 ++ X_2 ++ \dots ++ X_n$, X_i могут включать параметры.
 - Экзистенциальная теория — теория уравнений в словах.
- Дополнительный предикат — отношение подстроки $X_2 \preceq X_1$.
- $X_2 \preceq X_1$ — выразим через экзистенциальную теорию уравнений в словах.
- $\neg X_2 \preceq X_1$ — не выразим через равенство.
- Дополнительная операция — замена подстроки в строке $replace(X_1, X_2 \mapsto X_3)$.
- Естественное расширение — предикат вхождения слова в регулярный язык.
 - Не ломает разрешимость экзистенциальной теории уравнений в словах.
 - Позволяет выразить отрицания для отношений подстроки, но только при константном втором аргументе.



Противоречивые строковые модели

Эксперимент: порождение малых случайных противоречивых моделей в теории строк.

- 3–5 строковых параметра;
- 3–15 аксиом;
- все строки в операторе замены и в предикате проверки на подстроку константны и имеют длину не больше 5;
- отсеиваются тривиальные противоречия (такие, как несовпадения префиксов констант в уравнениях).

Противоречивость от 20% до 30% таких малых моделей не определяется солверами `svs5` и `Z3`!



Теории и сложность анализа

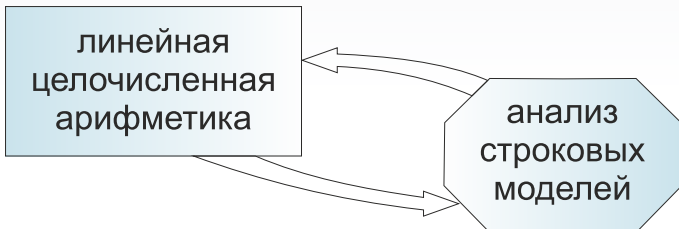
Здесь \mathcal{EST} — экзистенциальная теория строк (или теория уравнений в словах).

Теория	Replace All		Concat		счёт букв	счёт длин	REGEX	Сложность
	Const	\forall	Linear	\forall				
\mathcal{EST}	X	X	✓	✓	X	X	✓	PSPACE
$\mathcal{EST}+\text{len}$	X	X	✓	✓	X	✓	X	???
$\mathcal{EST}+\text{count}$	X	X	✓	✓	✓	X	X	Неразр.
$\mathcal{EST}+\text{repl}$	✓	X	✓	✓	X	X	X	Неразр.
$\text{repl}+\text{SL}$	✓	✓	✓	X	X	X	X	EXPSPACE
$\text{repl}+\text{len}$	✓	✓	✓	X	X	✓	X	Неразр.
$\text{repl}+\text{count}$	✓	✓	✓	X	✓	X	X	Неразр.

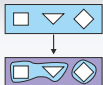


D2

Эксперименты с противоречивыми моделями



- Противоречивость условий на длины уравнений (12/50).
- Противоречивость условий на кратность букв (28/50).
- Рекурсивный анализ моделей как программ на языке Рефал (25/50).
- Совместно 2 и 3 (41/50).



Понятия образца и уравнения

Плоский образец — строка в смешанном алфавите переменных и констант.

Уравнение в словах — равенство между строками в смешанном алфавите.

Скажем, что образец содержит (синтаксически) открытые переменные, если число различных переменных типа выражение, входящих в него, больше 1.

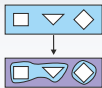
Самобытность Рефала — поддержка образцов с открытыми и кратными переменными типа выражение, а также условий, порождающих уравнения в словах.



Представление моделей в Рефале

- Логические функции
- Равенство
- Предикат подстроки
- Оператор замены
- Классическое сопоставление слева направо (С)
- Рефал-сопоставление с образцом (О)
- Предварительная специализация (КМР-подобный) (Р)

No	&	$X_1 = X_2$	$X_2 \preceq X_1$	$replace(X_1, X_2 \mapsto X_3)$	Успехи
1	С	С	С	С	20
2	С	С	О	О	25
3	О	О	О	О	11
4	О	С	Р	Р	9
5	О	С	С	С	11



Внутренний язык суперкомпилятора MSCP-A

$$C_i = \langle C_i, P_i \rangle$$

- C_i — параметризованные состояния стека.
- P_i — предикаты на параметры в КНФ, содержащие литералы двух типов.
 - Отрицательные условия в форме $(X_i \neq \Phi_i)$, где Φ_i может содержать как связанные значения (параметры), так и пробегающие все возможные значения (переменные).
 - Уравнения в словах $\Delta_1 = \Delta_2$, где Δ_i могут содержать только параметры, но не переменные.

Каждая конфигурация соответствует модели в теории строк \Rightarrow противоречивые модели влекут отсечение пути вычисления.

D1





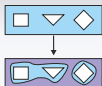
Нётеровость

- Множество путей, порождаемых обобщённым состоянием, включает в себя множества путей обобщаемых состояний. Т.е. определяется возрастание $\mathcal{C}_{k_i} \trianglelefteq \mathcal{C}_g$ по вложению.
- Нет гарантии, что возрастающая последовательность $\mathcal{C}_{g_1} \trianglelefteq \mathcal{C}_{g_2} \trianglelefteq \dots \trianglelefteq \mathcal{C}_{g_i} \trianglelefteq \dots$ когда-нибудь стабилизируется.

В алгебрах, содержащих ассоциативные операции, существуют бесконечные возрастающие последовательности, не обладающие нётеровостью, например:

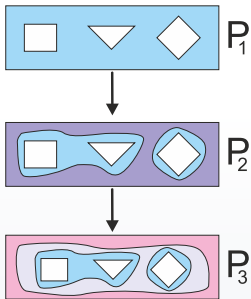
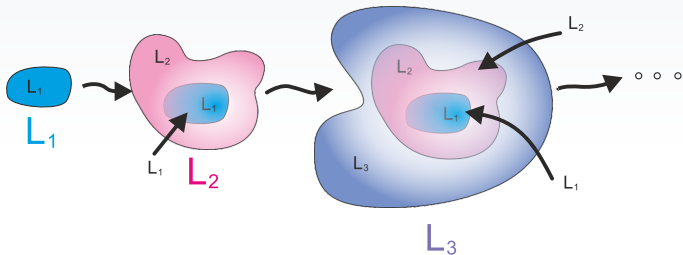
$x_0 x_0, x_1 x_0 x_0 x_1, \dots, x_{k+1} x_k \dots x_0 x_0 \dots x_k x_{k+1}, \dots$

Над данными в свободной алгебре (древесными термами) таких последовательностей не существует, если возрастание $T_i \trianglelefteq T_j$ определяется наличием подстановки σ такой, что $T_j \sigma = T_i$.

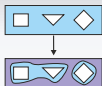


Нётеровость

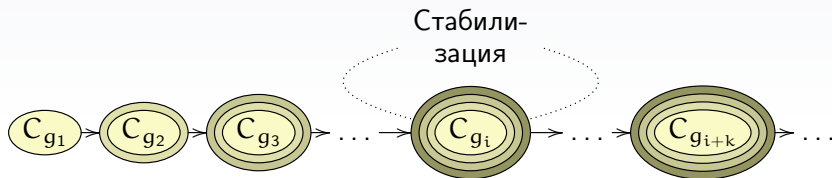
D3



Эквивалентно артиновости (фундированности, Well-Founded-Ordering) обратного отношения (\mathcal{C}_{k_i} к \mathcal{C}_g).



Нётеровость



Алгоритм обобщения корректен, если порождаемые им состояния являются нётеровыми относительно отношения вложения путей. Свойство нётеровости гарантирует, что вместо обобщения рано или поздно выполнится вложение.



Пример прошлого года

При обобщении $(\varepsilon)(x)$ и $(\mathbf{A})(\mathbf{A} x')$ получается заготовка $(x_1)(x_1 x_2)$. После чего параметры x_1 и x_2 сливаются как подряд идущие, даже если на x и x' есть рестрикция, запрещающая вхождения буквы \mathbf{A} , переносимая на x_2 .

Частичное решение: записать уравнение, устанавливающее связь между переменными после слияния — оказывается слабым, потому что негативные рестрикции всё равно теряются.



n -замкнутость

Пусть \mathcal{P} — образец (выражение); P_1, \dots, P_n — свободные фрагменты элементов его плоского разбиения (далее кратко СФР).

- Если x_j — единственная переменная, входящая в некоторый P_k , тогда x_j — 0-замкнутая.
- Если x_{j_1}, \dots, x_{j_k} — j_i -замкнутые переменные, входящие в P_k вместе с некоторой $x_{j_{k+1}}$, степень замкнутости которой неизвестна либо больше $\max(j_i) + 1$, тогда $x_{j_{k+1}}$ — $\max(j_i) + 1$ -замкнутая.

Дан образец P . n -ку (t_{n-1}, \dots, t_0) такую, что t_i — количество различных переменных замкнутости i , входящих в P , назовём мерой открытости $\mu(P)$.



Решения проблемы однозначности

Решение прошлого года гарантировало нётеровость только очень ограниченных классов образцов.

Лемма

Отношение возрастания языков ($\mathcal{L}(P_i) \subset \mathcal{L}(P_j)$) нётерово для образцов замкнутости $\leq k$.

Доказательство: индукцией по степени замкнутости.



Решения проблемы однозначности

Но как выяснилось, выполняется более сильное утверждение.

Теорема

Отношение возрастания языков ($\mathcal{L}(P_i) \subset \mathcal{L}(P_j)$) нётерово для образцов, матрица кратности которых имеет ненулевой определитель.

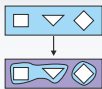
- Ограниченность множества различных переменных \Rightarrow кратности переменных — элементы конечного кортежа;
- Лемма Диксона \Rightarrow найдутся два кортежа, каждый элемент первого из которых не больше каждого из элементов второго. Построим различающую подстановку в «меньший» образец.
- Пустые подстановки \Rightarrow линейная зависимость между уравнениями на длины переменных.
- Нет пустых подстановок \Rightarrow возможна лишь переименовка.



Строковые модели и сопоставление в Рефале

В конфигурациях суперкомпилятора могут возникать почти произвольные условия (большое число повторных вхождений параметров, из-за особенностей прогонки).

Однако в самих образцах редко появляется больше трёх повторных переменных.



Строковые модели и сопоставление в Рефале

Что встречается:

- пассивные условия (уравнения) вида $x_1 = z_1 x_1 z_2 \dots x_n z_{n+1}$, где z_i — свежие переменные. Чаще $i = 1$. Таких условий может быть несколько.
- образцы с повторными переменными в разном контексте, «плавающими» внутри открытых. $z_1 \Phi_1 x_1 \Phi_2 z_2 \Phi_3 x_1 \Phi_4 z_3$ — чаще всего повторение однократное.
- образцы, состоящие из нескольких фрагментов, погружённых в открытые переменные. Пример: $(x_1)(x_3(x_2)x_4)z_1x_1x_2z_2$



Резюме

- Анализ программ на языке Рефал \Rightarrow анализ строковых моделей; суперкомпилятор \Rightarrow внутренний SMT-солвер в модели SLIA \Rightarrow снова суперкомпилятор.
- Эффективные Рефал-образцы и Рефал-условия \Rightarrow стремительная разработка достаточно эффективных программ на Рефале.
- Использование мощного сопоставления с образцом \Rightarrow анализ простых строковых моделей.