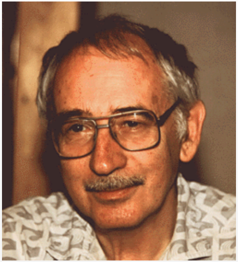
Краткое вступление



В.Ф. Турчин
(1931–2010)

- Функциональный язык Рефал
- Суперкомпиляция
- . . .

- Семинары **МЕТА**: 2008–2016 в Переславле-Залесском
- Приглашённые докладчики: Neil D. Jones, Simon Peyton-Jones

Теорема Турчина
в анализе формальных языков

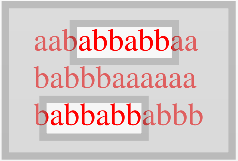
А. Непейвода, `a_nevod@mail.ru`

Методы анализа формальных языков

Коммутативные:

- Образы Париха;
- Вычисление функции мощности множества слов по длине.

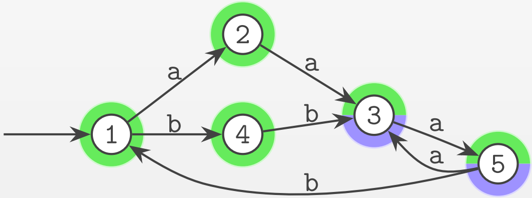
Некоммутативные:



- леммы о накачке;
- леммы о перескоке;
- утверждения о неизбежных подсловах.

Конечные системы переходов

- Конфигурация вычисления описывается только состоянием.
- Число состояний заранее ограничено конечным N.
- Метки на переходах — символы, читаемые из входа.



- Если язык бесконечен, то существуют компоненты сильной связности, длины которых ограничены тоже N.

Конечные системы переходов

- Если язык бесконечен, то существуют компоненты сильной связности, длины которых ограничены тоже N.
- Читая слово длины $N+1$, мы точно попадём в одно и то же состояние q_i дважды.
- Цикл из q_i в q_i можно проходить сколько угодно раз (в том числе и нисколько) $\Rightarrow$ существует «накачка».

$$\underbrace{a_1 a_2 \dots a_{k-1}}_{\text{путь из } q_0 \text{ в } q_i} \overbrace{a_k \dots a_{k+m}}^{\text{цикл из } q_i \text{ в } q_i} \underbrace{a_{k+m+1} \dots a_{k+m+n}}_{\text{путь из } q_i \text{ в } q_F}$$

- Выбираем самое первое попадание в цикл $\Rightarrow$ сумма длин $a_1 \dots a_{k-1}$ и $a_k \dots a_{k+m}$ ограничена N.
- Можно начинать отсчёт с любой позиции в слове, после которой есть подслово как минимум N букв, и внутри этого подслова тоже будет «накачка».

Перегруженность формальной нотации

Классическая лемма о накачке

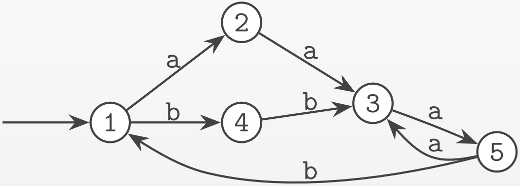
Если $\mathscr{L}$ — конечноавтоматный, то $\exists n \in \mathbb{N}. \forall w \big(w \in \mathscr{L} \mathbin{\&} |w| > n \Rightarrow \exists w_1, w_2, w_3 \big(|w_2| > 0 \mathbin{\&} |w_1| + |w_2| \leq n \mathbin{\&} w = w_1\, w_2\, w_3 \mathbin{\&} \forall k\, (k \geq 0 \Rightarrow w_1\, {w_2}^k\, w_3 \in \mathscr{L})\big)\big)$.

Универсальная лемма о накачке

$\mathscr{L}$ конечноавтоматный $\Leftrightarrow \exists m \in \mathbb{N}. \forall w \in \mathscr{L}. \forall i \in \mathbb{N}(|w| \geq m \mathbin{\&} (i \leq |w| - m) \Rightarrow \exists w_1, w_2, w_3, w_4 (w = w_1 w_2 w_3 w_4 \mathbin{\&} |w_1| = i \mathbin{\&} 1 \geq |w_3| \leq m \mathbin{\&} |w_2| + |w_3| \leq m \mathbin{\&} \forall k\, (w_1 w_2 w_3^k w_4 \in \mathscr{L})))$.

Другой взгляд на конечные системы переходов

- $(1), \ldots, (5)$ — это нульместные функции;
- функция (k) только читает символ с ленты и передаёт управление другой функции.



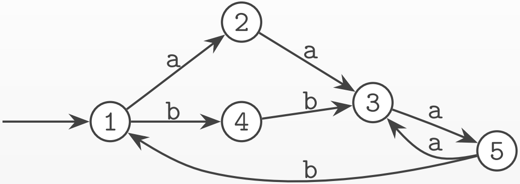
- получается система переписывания термов, фиксирующая возможные поведения стека.

$$(1) \xrightarrow{a} (2) \quad (1) \xrightarrow{b} (4) \quad (2) \xrightarrow{a} (3)$$

$$(4) \xrightarrow{b} (3) \quad (3) \xrightarrow{a} (5) \quad (5) \xrightarrow{a} (3)$$

$$(5) \xrightarrow{b} (1)$$

Другой взгляд на конечные системы переходов



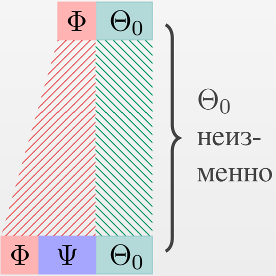
- получается система переписывания термов, фиксирующая возможные поведения стека.

$$(1) \xrightarrow{a} (2) \quad (1) \xrightarrow{b} (4) \quad (2) \xrightarrow{a} (3)$$
$$(4) \xrightarrow{b} (3) \quad (3) \xrightarrow{a} (5) \quad (5) \xrightarrow{a} (3)$$
$$(5) \xrightarrow{b} (1)$$

- Очевидно, что в любом отрезке вычисления длины больше N минимум две конфигурации стека (а значит, и конфигурации вычисления вообще) повторятся.

Теорема Турчина о регулярном поведении стеков

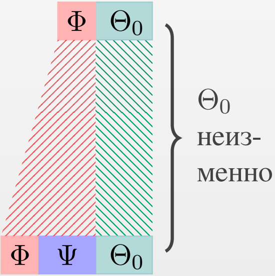
Если вдоль пути вычислений встречаются состояния стека: $\rho_1 : \Phi\Theta_0$, $\rho_2 : \Phi\Psi\Theta_0$, то будем говорить, что ρ_1 и ρ_2 образуют турчинскую пару ($\rho_1 \preceq \rho_2$), если значение Θ_0 неизменно на отрезке вычислений начиная от ρ_1 и до ρ_2.



Если Φ порождает бесконечный цикл с состояниями стека $\Phi\Psi^n\Theta_0$, то $\rho_1 \preceq \rho_2$. Если все функции нульместны, то условие $\rho_1 \preceq \rho_2$ необходимо и достаточно для существования бесконечного цикла.

Теорема Турчина о регулярном поведении стеков

Если вдоль пути вычислений встречаются состояния стека: $\rho_1 : \Phi\Theta_0$, $\rho_2 : \Phi\Psi\Theta_0$, то будем говорить, что ρ_1 и ρ_2 образуют турчинскую пару ($\rho_1 \preceq \rho_2$), если значение Θ_0 неизменно на отрезке вычислений начиная от ρ_1 и до ρ_2.

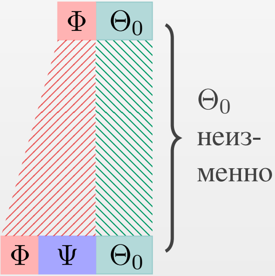


Плохая последовательность относительно $\preceq$: отрезок вычислений, не содержащий пар, связанных $\preceq$.

Длина наибольшей п.п. относительно $\preceq$ ограничена числом правил в программе P.

Теорема Турчина о регулярном поведении стеков

Если вдоль пути вычислений встречаются состояния стека: $\rho_1 : \Phi\Theta_0$, $\rho_2 : \Phi\Psi\Theta_0$, то будем говорить, что ρ_1 и ρ_2 образуют турчинскую пару ($\rho_1 \preceq \rho_2$), если значение Θ_0 неизменно на отрезке вычислений начиная от ρ_1 и до ρ_2.



Длина наибольшей п.п. относительно $\preceq$ ограничена числом правил в программе P.

Выполняется также для размера вершины стека $|\Phi| = 1$.

Нульместное переписывание со стеком

- Пусть теперь нульместные функции так же читают одну букву, но вызывают не обязательно не больше одной другой функции.
- Правила переписывания стека примут вид $N_i \xrightarrow{a} M_1 \dots M_n$.

Пусть $\mathscr{L}$ — КС-язык. Тогда он может быть порождён грамматикой G с правилами вида $N_i \mapsto \gamma_i$ и $N_i \mapsto \gamma_i M_{1,i} \dots M_{k,i}$, где $\gamma_i \in \Sigma$, $N_i, M_j \in \mathcal{N}$.

Классическая лемма о накачке

Пусть $\mathscr{L}$ — КС-язык. Тогда существует длина накачки $p \in \mathbb{N}$ такая что для всех $w \in \mathscr{L}$, $|w| \leq p$ выполняется условие:

$$\exists x_i, y_i, z\big(w = x_1 y_1 z\, y_2 x_2 \ \&\ |y_1 y_2| \geq 1 \ \&\ |y_1 z y_2| \leq p \\ \&\ \forall k \in \mathbb{N}(x_1 y_1^k z y_2^k x_2 \in \mathscr{L})\big)$$

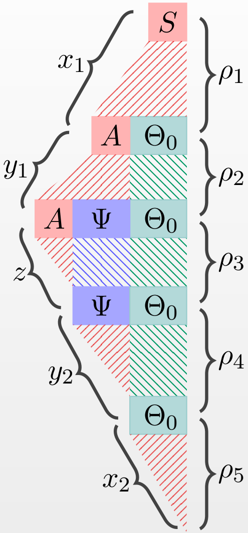
- Можно выбрать заведомо накачиваемые позиции (лемма Огдена);
- Можно выбрать запрещённые позиции (теорема Бадера–Маура);
- Или множественные накачки (Multiple Pumping Lemma)...

Н.Ф. Грейбах & Поведение стека

Пусть $\mathscr{L}$ — КС-язык. Тогда он описывается грамматикой G с правилами вида $N_i(\gamma_i) \mapsto \varepsilon$ и $N_i(\gamma_i) \mapsto M_{1,i} \dots M_{k,i}$, где $\gamma_i \in \Sigma$, $N_i, M_j \in \mathcal{N}$.

- Поведение стека P описывается алфавитной префиксной грамматикой.
- (Алфавитные) префиксные грамматики определяют регулярные языки.
- ...и удовлетворяют их комбинаторным свойствам (например, универсальной лемме о накачке).

Теорема Турчина — это лемма о накачке



- Можно выбрать любой достаточно длинный сегмент, не являющийся плохой последовательностью;
- Можно выбрать любое конечное число запрещённых позиций;
- Можно рассуждать о накачках рекурсивно.

Специализация теоремы Турчина

- Выберем последнюю турчинскую пару на пути вычисления и применим ограничения из теоремы Турчина:

Пусть $\mathscr{L}$ — КС. Тогда существует длина накачки $p \in \mathbb{N}$ такая что для всех $w \in \mathscr{L}$, $|w| \leq p$ выполняется условие:

$$\exists x_i, y_i, z\Big(w = x_1 y_1 z\, y_2 x_2 \;\&\; |y_1| \geq 1 \;\&\; |y_2| \geq 1 \;\&\; |z| \geq 1 \;\&\; |y_1 z y_2| \leq p \;\&\; \frac{|x_2|}{|x_1|} \leq p \;\&\; \forall k \in \mathbb{N}(x_1 y_1^k z y_2^k x_2 \in \mathscr{L})\Big)$$

Специализация теоремы Турчина

- А теперь выберем самую первую турчинскую пару на пути вычислений:

Пусть $\mathscr{L}$ — КС. Тогда существует длина накачки $p \in \mathbb{N}$ такая что для всех $w \in \mathscr{L}$, $|w| \leq p$ выполняется условие:

$$\exists x_i, y_i, z\big(w = x_1 y_1 z\, y_2 x_2 \;\&\; |y_1| \geq 1 \;\&\; |y_2| \geq 1 \;\&\; |z| \geq 1 \\ \&\; |x_1 y_1| \leq p \;\&\; \forall k \in \mathbb{N}(x_1 y_1^k z y_2^k x_2 \in \mathscr{L})\big)$$

Специализация теоремы Турчина

- Применим лемму рекурсивно:

Пусть $\mathscr{L}$ — КС. Тогда существует длина накачки $p \in \mathbb{N}$ такая что для всех $w \in \mathscr{L}$, $|w| \leq p$ выполняется условие:

$\exists x_i, y_i, z\big(w = x_1 y_1 z\, y_2 x_2 \;\&\; |y_1| \geq 1 \;\&\; |y_2| \geq 1 \;\&\; |z| \geq 1 \;\&$
$|x_1 y_1| \leq p \;\&\; \forall k \in \mathbb{N}(x_1 y_1^k z y_2^k x_2 \in \mathscr{L}) \;\&\; (|\xi| \leq p \vee \xi$ содержит независимую область накачки$)\big)$

- Здесь $\xi \in \{x_2, z, y_2\}$.

Too Many Languages Satisfy Ogden's Lemma

Рассмотрим язык

$$\left\{a^n b^m \middle| (n \neq m) \vee (n = m = k^2)\right\}$$

Применим рекурсивно теорему Турчина к слову $a^{p^{p\,2}} b^{p^{p\,2}}$.

- Если существуют хотя бы две области накачки такие, что одна добавляет больше букв a, чем b, а другая наоборот, то контрпример построен.
- Предположим, что все накачки добавляют больше букв a, чем b. Будем вычитать минимальные накачки из слова до тех пор, пока фрагмент из букв a не станет меньше p и накачек в нём уже не останется. При этом фрагмент из букв b будет всё ещё больше, чем p^{p-1} и будет содержать хотя бы одну накачку.