Семинар STEP (online)

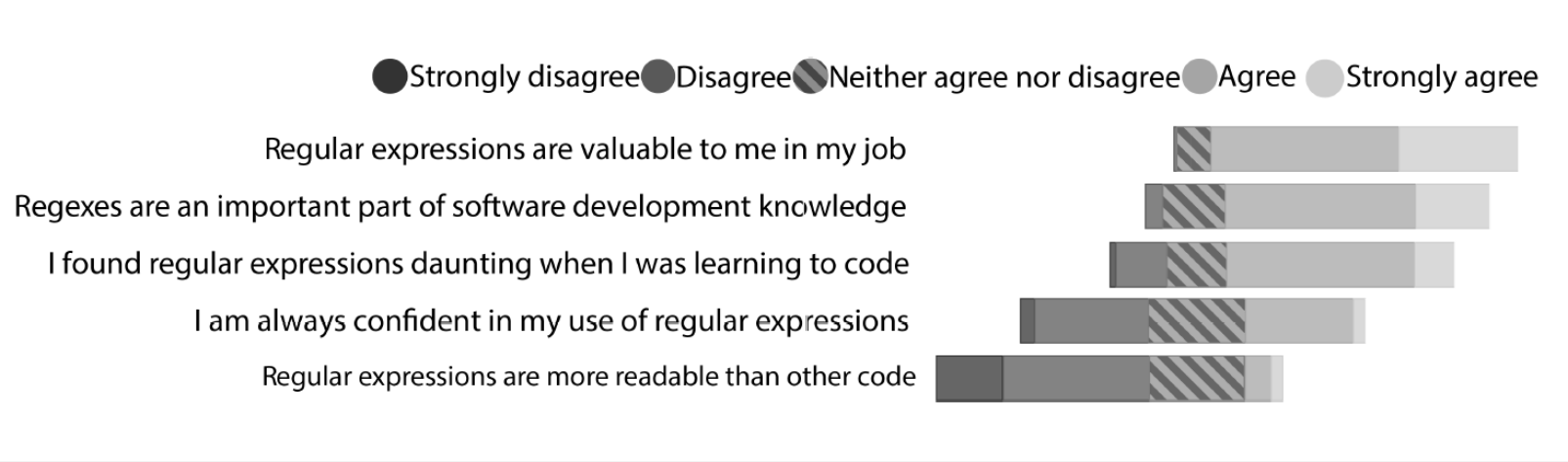
День Проблем, 30 декабря, 2024

Неуловимые формальные языки расширенных регексов

А. Непейвода, a_nevod@mail.ru

Круг проблем

Regexes are Hard: Decision-making, Difficulties, and Risks in Programming Regular Expressions, 2023



- Библиотеки регексов (RegexLib) содержат слишком мало примеров.
- Очень мало паттернов и анти-паттернов проектирования регексов, особенно в расширенном случае.

Расширенные регексы

- опережающие и ретроспективные проверки:
  - экспоненциально экономичнее в неэкзотических случаях: `(?= [^c]*a)(?=[^b]*b)[^c]*c.*`
  - релевантны (в ≈10% проектах выборки [EREUCP])
  - в общем виде не всегда поддерживаются из-за «невозможности безвозвратного разбора» [DBEREM].
- обратные ссылки на строку:
  - выражают контекстно-зависимые свойства, не выразимые в рамках контекстно-свободных языков: `(([a-z])+)\s.*\1.*`
  - релевантны (в ≈5% проектах выборки [EREUCP])
  - задача сопоставления NP-полна даже в ограниченном случае.
- обратные ссылки на выражение: `((a(?1)b)?)` — аналог оператора минимальной неподвижной точки, экзотика.

Языки проекций vs языки выражений

Извлечение содержимого группы захвата порождает язык, распознаваемый не всем выражением, а только одной группой.

- Выразительная сила проекций и выражений в случае расширенных контекстно-зависимыми операциями регулярных выражений не совпадает.
- Игнорируются в (рассмотренных) теоретических работах.

К вопросу о применимости теории полугрупп

Утверждение: Проблема пустоты языка регулярных выражений с обратными ссылками на выражения и строки и опережающими проверками неразрешима.

Обоснование: Теорема 1.3.4 из доклада В.Г. Дурнева на PSSV-2024.

Внутри формально регулярных языков

В теории регулярные выражения с пересечениями и дополнением были исследованы ещё в 90-е годы.

- Модели машины — переключающиеся автоматы (alternating finite automata).
- Комбинаторный взрыв при переходе от AFA к ДКА (увеличение числа состояний до 2^{2^n} раз).
- Альтернативные техники разбора — производные Антимирова для булевских регулярных выражений (проблема mintermization).

На практике: усовершенствованные теоретические методы — [DBEREM], 2022–2023.

За рамками регулярности

Многократные формализации регексов с обратными ссылками:

- без переиспользований групп захвата и без неинициализированных ссылок (Campeanu–Salomaa–Yu, 2003).
- без переиспользований групп захвата и в ε-семантике.
- с переиспользованием групп захвата (Freydenberger, Schmid, 2012–2013).

Систематизация формализмов и сравнение выразительной силы — в [RREwB].

Конфликты формализаций

- CSY≠Schmid: конфуз с использованием в [CRRRL] контрпримера, сформулированного для более слабого класса языков.

- CSY≠Schmid: конфуз с циклическим переопределением ссылок и контрпримером в [LREB].

Проблема: трудно отделить синтаксические ограничения, ведущие и не ведущие к семантическим изменениям.
Пример: $[_1a]_1[_2\backslash 1]_2[_1\backslash 2a]_1$ не выводит из класса CSY, $[_1a]_1([_2\backslash 1]_2[_1\backslash 2a]_1)^*$ выводит из класса CSY.

В [И-Н] — попытка разрешения проблемы путём определения ACREG-класса, допускающего переименовки в CSY-регексы.

Методы анализа расширенных регексов

- Лемма о накачке для CSY-формализации: конечное число итерируемых классов эквивалентности по суффиксной конгруэнции — слишком узкое для Schmid-формализации $\Rightarrow$ недоразумения с применением из-за почти идентичных названий разных классов языков.

- Утверждение из работы [EPREB]: регексы в Schmid-формализации являются подклассом индексных языков — адекватный формализм, но вложение не «плотное»: индексные языки замкнуты относительно обращения, языки Schmid-а не замкнуты.

WiP по анализу Schmid-регексов

- Доказательство незамкнутости относительно реверсирования Schmid-языков в [И-Н].
- Передоказанное утверждение из [LREB] о меньшей выразительной силе Schmid-языков без опережающих проверок по сравнению с Schmid+LA-языками.

Общие черты: находится инвариант, нарушаемый в рамках итерации в префиксе слова.

Определим $h_c(x) = \begin{cases} c, x = c \\ \$, x \neq c \end{cases}$, $compress_c(x)$ — слово, полученное заменой максимальных блоков из букв c на (одну и ту же) свежую букву c'.

Если $\exists n, c(\forall w \in \mathscr{L}(|w| \geq n \Rightarrow w = w_1 w_2 \ \& \ w_1 = xyz \ \& \ |xy| \leq n \ \& \ |xyz|_c = 0 \forall i \exists w_3 (xy^i z w_3 \in \mathscr{L} \ \& \ |w_3|_c > 0) \Rightarrow$
множество $\{compress_\$(h_c(w_3)) \mid xy^i z w_3 \in \mathscr{L}\}$ образует конечное число классов эквивалентности относительно $xy^i z))$

Исследовательские вопросы

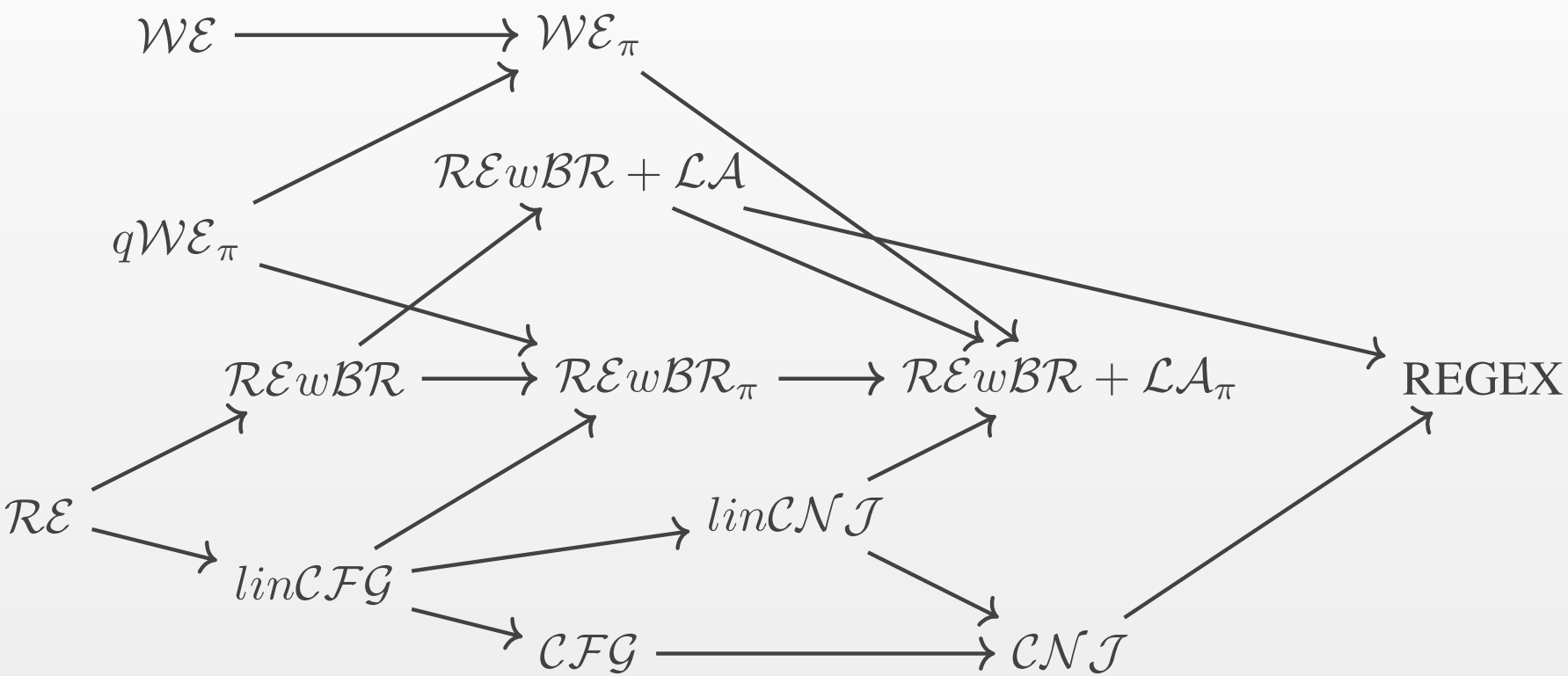
- Алгебраические методы анализа расширенных регексов (аналоги производных?) $\Rightarrow$ более эффективный в среднем парсинг, оптимизации.

- Более точное отделение классов регексов в различных формализмах друг от друга $\Rightarrow$ уменьшение числа ошибок при переносе алгоритмов из одного класса в другой.

- Анализ конгруэнций LTS-структур на базе регексов $\Rightarrow$ нахождение анти-паттернов проектирования расширенных регулярных выражений.

Известные соотношения классов

Более подробно — см. доклад на STEP по ссылке.



Литература

- [DBEREM] **Derivative Base+d Extended Regular Expression Matching Supporting Intersection, Complement and Lookaround**, 2023.
- [EREUCP] **Exploring Regular Expression Usage and Context in Python**, 2016
- [RREwB] **Re-examining regular expressions with backreferences**, 2023
- [CRRRL] **Characterising REGEX by Regular Ref-Languages**, 2016
- [LREB] **On Lookaheads in Regular Expressions with Backreferences**, 2023
- [EPREB] **On the expressive power of regular expressions with backreferences**, 2023
- [DetRW] **Deterministic Regular Expressions with Back-References**, 2019