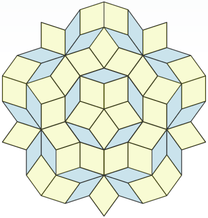


Структуры над словами: образцы и уравнения

Летняя практика, Переславль-Залесский

4–6 июля, 2022 г.

Проектирование структур с образцами

- Вопрос достижимости образца:

```
f (A : x) = Expr1
f [] = Expr2
f [A] = Expr3
```

- Вопрос накрытия образцами:

```
f ((x : y) : z) = Expr1
f [] = Expr2
```

- Вопрос перестановочности образцов:

```
f (x : (A : y)) = Expr1
f (A : (y : z)) = Expr2
```

...а с отказом от свободы и единственности вхождений переменных эти вопросы становятся намного сложнее.

Проектирование структур с образцами

- Вопрос достижимости образца:

```
f {x t t y = Expr1 }
f {x 'A' t z1 'A' y = Expr2 }
f {x 'A' y 'A' z = Expr3 }
```

- Вопрос накрытия образцами:

```
f {x1 (z) x2 t x3  = Expr1}
f {x (z)  = Expr2}
f {t x  = Expr3}
```

- Вопрос перестановочности образцов:

```
f {x, x 'A' : 'A' x = Expr1}
f (x, x 'AB' : 'BA' x = Expr2}
```

Необходимо определить выразительную силу образцов — языки, которые они описывают, и свойства этих языков.

Базовые определения

$V_{\mathcal{T}}$ — множество переменных типа $\mathcal{T}$, $V = \bigcup^{\mathcal{T}} V_{\mathcal{T}}$.
Рассматриваем e-переменные (типа строка/выражение) и t-переменные (типа терм). Т.е. $V = V_{\mathtt{e}} \cup V_{\mathtt{t}}$.

Кратность терма T в образце P обозначаем $|P|_T$.

Σ — по умолчанию **неограниченный** алфавит констант. $\mathcal{B}[S]$ — множество скобочных структур над строками из S.

Плоский образец P — строка в алфавите $V_{\mathtt{e}} \cup \mathcal{B}[\Sigma \cup V_{\mathtt{t}}]$. Образец P линеен, если $\forall x \in V_{\mathtt{e}}\ (|P|_x = 1)$. Подстановка в образец — гомоморфизм, сохраняющий константы (т.е. для всех $\mathbf{A} \in \Sigma\ \mathbf{A}\sigma = \mathbf{A}$).

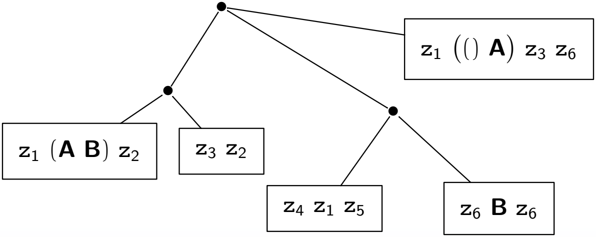
Образец допускает плоское разбиение, если он плоский, либо имеет вид $(P_1)\ (P_2)\ldots\ (P_n)\ P_{n+1}$, где все P_i допускают плоское разбиение. Максимальные плоские подобразцы такого образца называем фрагментами плоского разбиения (ФПР).

Плоские разбиения и деревья

Рассмотрим следующий образец:

$$\Big(\big(z_1\ (\mathbf{A}\ \mathbf{B})\ z_2\big)\ z_3\ z_2\Big)\Big(\big(z_4\ z_1\ z_5\big)\ z_6\ \mathbf{B}\ z_6\Big)\ z_1\ \big(()\ \mathbf{A}\big)\ z_3\ z_6$$

Структура его ФПР приведена ниже.



Поскольку скобочные структуры могут возникнуть только сразу справа от открывающей скобки, то ФПР образуют древесные структуры, аналогичные АТД.

Пример образца, не разбиваемого на ФПР:

$$\big(x_1\ (\mathbf{A}\ x_2)\ x_1\ x_2\big)\ x_1$$

Языки, распознаваемые образцами

Определение

Языком $\mathscr{L}(\mathsf{P})$***, распознаваемым образцом*** P, *назовем множество элементов* $\Phi \in \mathcal{B}[\Sigma]^*$*, для которых существует подстановка* σ*:* $\mathsf{P}\sigma = \Phi$*. Образец* P_1 *сводится к образцу* P_2*, если* $\mathscr{L}(\mathsf{P}_1) \subseteq \mathscr{L}(\mathsf{P}_2)$*.*

Подстановка $\mathtt{x}\sigma = \varepsilon$ допустима! В терминологии pattern languages — рассматриваются E-pattern languages (EPL, сокращение от Erasing Pattern Languages, языки стирающих образцов).

- Язык, распознаваемый образцом-строкой $\mathsf{P} \in \Sigma^*$, есть $\{\mathsf{P}\}$.
- Язык, распознаваемый образцом $\mathsf{P} = \mathtt{x}_1\ \mathtt{x}_2\ \mathtt{x}_1$, есть всё множество $\mathcal{B}[\Sigma]^*$.

Языки, распознаваемые образцами

Определение

Языком $\mathscr{L}(\mathsf{P})$*, распознаваемым образцом* P*, назовем множество элементов* $\Phi \in \mathcal{B}[\Sigma]^*$*, для которых существует подстановка* σ*:* $\mathsf{P}\sigma = \Phi$*. Образец* P_1 *сводится к образцу* P_2*, если* $\mathscr{L}(\mathsf{P}_1) \subseteq \mathscr{L}(\mathsf{P}_2)$*.*

- Язык, распознаваемый образцом-строкой $\mathsf{P} \in \Sigma^*$, есть $\{\mathsf{P}\}$.
- Язык, распознаваемый образцом $\mathsf{P} = \mathtt{x}_1\ \mathtt{x}_2\ \mathtt{x}_1$, есть всё множество $\mathcal{B}[\Sigma]^*$.

С точки зрения семантики сопоставления, образец $\mathtt{x}_1\ \mathtt{x}_2\ \mathtt{x}_1$ также неудачный: $\mathtt{x}_1$ всегда успешно сопоставляется с ε. Бывает и иначе: хотя $\mathscr{L}(\mathtt{z}_1\ \mathtt{z}_2\ \mathtt{z}_2) = \mathscr{L}(\mathtt{x}_1\ \mathtt{x}_2\ \mathtt{x}_1) = \mathcal{B}[\Sigma]^*$ из-за существования тривиальной подстановки $\mathtt{z}_2 := \varepsilon$, но ленивое сопоставление строки **ABB** с $\mathtt{z}_1\ \mathtt{z}_2\ \mathtt{z}_2$ построит подстановку $\mathtt{z}_2 := \mathbf{B}$, а вовсе не $\mathtt{z}_2 := \varepsilon$.

Сводимость и эквивалентность

Если P_1, P_2 оба из $(V_e \cup \mathcal{B}[\Sigma])^*$, то:

- P_1 сводится к P_2 $\Leftrightarrow$ существует подстановка σ такая, что $P_2\sigma = P_1$;
- если P_2 линеен, тогда вычислительная сложность проверки сводимости образца P_1 к образцу P_2 линейна от суммы длин P_1 и P_2.

Из-за того, что образцы стирающие (определяют EPL), двухсторонняя сводимость не эквивалентна наличию переименовки: вспомним те же $\mathtt{x}_1\ \mathtt{x}_2\ \mathtt{x}_1$ и $\mathtt{z}_1\ \mathtt{z}_2\ \mathtt{z}_2$.

Сводимость и эквивалентность

Если P_1, P_2 оба из $(V_e \cup \mathcal{B}[\Sigma])^*$, то:

- P_1 сводится к P_2 $\Leftrightarrow$ существует подстановка σ такая, что $P_2\sigma = P_1$;
- если P_2 линеен, тогда вычислительная сложность проверки сводимости образца P_1 к образцу P_2 линейна от суммы длин P_1 и P_2.

Если рассматривать только образцы без идущих подряд переменных из V_e, тогда уже для образцов без переменных из V_t выполняется утверждение

$$\mathscr{L}(P_1) = \mathscr{L}(P_2) \Leftrightarrow \exists\sigma(P_1\sigma = P_2 \;\&\; \forall x \in V_e(x\sigma \in V_e))$$

Краткие и избыточные образцы

Определение

Образец P_1 *называется кратким, если любой образец* P_2 *такой, что* $\mathscr{L}(\mathsf{P}_1) = \mathscr{L}(\mathsf{P}_2)$*, имеет длину, не меньшую, чем* P_1*.*
Иначе P_1 *называется избыточным.*

Пример

Образец $\boxed{\mathtt{x}_1\ \mathtt{x}_2}$ **A** $\mathtt{x}_3$ **B** $\boxed{\mathtt{x}_1\ \mathtt{x}_2}$ избыточен.

Образец $\mathtt{x}_1\ \mathtt{x}_2\ \mathtt{x}_2\ \mathtt{x}_1$ является кратким.

Алгебраисты также говорят, что избыточные образцы определяются нетривиальными неподвижными точками морфизмов над образцами (т.е. существует нетривиальная подстановка, переводящая избыточный образец в себя).

Критерий избыточности образца (Reidenbach, 2004)

Образец P избыточен, если существует представление
$P = Q_0 \boxed{R_1} Q_1 \dots \boxed{R_n} Q_n$, $Q_i \in \{\mathcal{B}[\Sigma] \cup V_e\}^*$, $R_i \in {V_e}^+ {V_e}^+$,
такое, что:

- множества переменных образцов Q_i и R_j не пересекаются;
- в каждом слове R_i найдется имеющая единственное вхождение в R_i переменная x_i *(выделенная)* такая, что

$$\forall j(| \boxed{R_j} |_{x_i} > 0 \Rightarrow \boxed{R_j} = \boxed{R_i}).$$

Этот критерий является необходимым и достаточным условием при рассмотрении плоских образцов в $\{\mathcal{B}[\Sigma] \cup V_e\}^*$[a].

[a] У Рейденбаха он доказан для образцов в V_e^*. Для образцов над $\mathcal{B}[\Sigma]$ доказательство где-то в моих старых тетрадях — здесь существенно, что скобки порождают бесконечный «алфавит констант».

Критерий Рейденбаха под лупой

Пусть искомое разбиение образца P существует. Тогда по каждому блоку R_i построим подстановку σ_{fix} так: $x_i\sigma_{\mathrm{fix}} = R_i$, а образы прочих переменных из R_i равны ε (они могут встречаться и в сочетании с другой выделенной переменной x_k в прочих R-блоках). Очевидно, $P\sigma_{\mathrm{fix}} = P$.

Образец $P \in V_e^*$ допускает неоднозначную нестирающую подстановку $\Leftrightarrow$ образец P избыточен по Рейденбаху.

Критерий Рейденбаха под лупой

Образец $P \in V_e^*$ допускает неоднозначную нестирающую подстановку $\Leftrightarrow$ образец P избыточен по Рейденбаху.

- Для образцов, содержащих константные фрагменты, это не верно: $\mathtt{x}_1$ **A** $\mathtt{x}_2$ допускает много подстановок в $\mathbf{A}^n$, хотя является кратким. Однако это верно для фрагментов таких образцов, не содержащих констант.
- Стирающих подстановок может быть и несколько: например, $\mathtt{x}_1\ \mathtt{x}_2\ \mathtt{x}_2\ \mathtt{x}_1$ допускает две подстановки в **A B A B**. Однако поиск возможных подстановок в такой формулировке может быть сделан экспоненциально быстрее, чем без знания о критерии Рейденбаха.
- Иногда структура слова слишком однородна, чтобы критерий Рейденбаха гарантировал единственность подстановки: см. $\mathbf{A}^n$ и любой краткий образец без констант, содержащий как минимум две различные переменные.

Добавление переменных типа терм

За увеличение выразительной силы образцов приходится платить усложнением теоретических конструкций.

- $\mathscr{L}(\mathsf{P}_1) \subseteq \mathscr{L}(\mathsf{P}_2)$ уже не определяется подстановкой.

  $\mathsf{P}_1 = \mathbf{A}\ \mathtt{x}_1$, $\mathsf{P}_2 = \mathtt{x}_2\ \mathtt{t}$. Язык P_1 вкладывается в язык P_2, а подстановки нет.

- Нет (пока ещё) исследованного понятия избыточного и краткого образца. Более того, образцы для одного и того же языка не образуют нижнюю полурешётку.

  Образы $\mathtt{x}\ \mathtt{t}$ и $\mathtt{t}\ \mathtt{x}$ оба краткие.

Плавающие t-переменные

Определение

*Назовем переменную t_i в плоском линейном образце P **якорной**, если*

- t_i *имеет кратность, не меньшую 2;*
- *или в P существует подслово α, не содержащее переменных из V_e, такое, что $\alpha = \alpha_1\ t_i\ \alpha_2$, причем α_1 и α_2 оба содержат хотя бы один символ или t-переменную, имеющую кратность не меньше 2.*

*В противном случае назовем t_i **плавающей**.*

Пример

Рассмотрим образец $t_1\ t_2\ x_1\ t_3\ t_4\ t_2\ x_2\ t_5$. Якорными переменными являются t_2 и t_1.

Плавающие переменные и языки образцов

Плавающая переменная в образце — указатель на то, что в соответствующий фрагмент образца нельзя подставить пустое слово. Аналог «нестираемых» фрагментов.

Плавающий сегмент линейного образца P — максимальное подслово P, содержащее только плавающие t-переменные и переменные из V_e.

Образец, в котором все e-переменные входят в плавающие сегменты — аналог нестирающего (non-erasing) образца. Хуже всего, если есть и стирающие, и нестирающие фрагменты.

Плавающие переменные и языки образцов

Плавающий сегмент линейного образца P — максимальное подслово P, содержащее только плавающие t-переменные и переменные из $V_{\mathtt{e}}$.

Образец, в котором все e-переменные входят в плавающие сегменты — аналог нестирающего (non-erasing) образца. Хуже всего, если есть и стирающие, и нестирающие фрагменты.

Язык образца $P_1 = \textbf{BBA}\,\mathtt{x}\,\textbf{ABCDA}$ вкладывается в язык образца $P_2 = \mathtt{z}_0\,\mathtt{t}\,\mathtt{t}\,\mathtt{z}_1\,\mathtt{t}_1\,\mathtt{t}_2\,\mathtt{t}_3\,\mathtt{t}\,\mathtt{z}_2$. Чтобы это доказать, приходится перебирать два случая: пустоты и непустоты подставляемого в x значения.

Multi-pattern languages (Kari, Salomaa)

Определение

Языком $\mathscr{L}(P)$, ***распознаваемым множеством образцов*** P_i *(англ. — multi-pattern language, сокращенно MPL), назовем множество элементов* $\Phi \in \mathcal{B}[\Sigma]^*$*, для которых существует* $i \in \mathbb{N}$ *и подстановка* σ*:* $\sigma(P_i) = \Phi$*.*

Множество MPL–объединений стирающих образцов совпадает с множеством MPL–объединений нестирающих образцов. Образец с плавающими t-переменными тоже определяет мультиобразец, и здесь уже смешивание стирающих и нестирающих фрагментов может быть разрешено.

Однако переход от стирающих образцов к нестирающим порождает экспоненциальное разрастание описания MPL.

Пример «плавающего» MPL

Пусть $\mathsf{P}_1 = \mathtt{x}_1$ **A A C** $\mathtt{x}_2$ **C A B** $\mathtt{x}_3$ **B B C**,
$\mathsf{P}_2 = \mathtt{z}_1\ \mathtt{t}_n\ \mathtt{t}_n\ \mathtt{z}_2\ \mathtt{t}_{m1}\ \mathtt{z}_3\ \mathtt{t}_{m2}\ \mathtt{z}_4\ \mathtt{t}_{m3}\ \mathtt{z}_5\ \mathtt{t}_n\ \mathtt{z}_6$.

Множество нестирающих образцов, порождающих P_2:

P_2^1	$\mathtt{t}_n\ \mathtt{t}_n\ \mathtt{z}_1\ \mathtt{z}_2\ \mathtt{z}_3\ \mathtt{t}_n$
P_2^2	$\mathtt{t}_n\ \mathtt{t}_n\ \mathtt{z}_1\ \mathtt{z}_2\ \mathtt{z}_3\ \mathtt{t}_n\ \mathtt{z}_4$
P_2^3	$\mathtt{z}_0\ \mathtt{t}_n\ \mathtt{t}_n\ \mathtt{z}_1\ \mathtt{z}_2\ \mathtt{z}_3\ \mathtt{t}_n$
P_2^4	$\mathtt{z}_0\ \mathtt{t}_n\ \mathtt{t}_n\ \mathtt{z}_1\ \mathtt{z}_2\ \mathtt{z}_3\ \mathtt{t}_n\ \mathtt{z}_4$

Множество нестирающих образцов, порождающих P_1, и обобщающие их подобразцы из P_2:

A A C C A B B B C	$\mathsf{P}_2^3\sigma_1,\ \mathtt{t}_n\sigma_1 = \mathbf{C}$
A A C C A B $\mathtt{x}_3$ **B B C**	$\mathsf{P}_2^3\sigma_2,\ \mathtt{t}_n\sigma_2 = \mathbf{C}$
A A C $\mathtt{x}_2$ **C A B B B C**	$\mathsf{P}_2^2\sigma_3,\ \mathtt{t}_n\sigma_3 = \mathbf{A}$
A A C $\mathtt{x}_2$ **C A B** $\mathtt{x}_3$ **B B C**	$\mathsf{P}_2^2\sigma_4,\ \mathtt{t}_n\sigma_4 = \mathbf{A}$
$\mathtt{x}_1$ **A A C C A B B B C**	$\mathsf{P}_2^3\sigma_5,\ \mathtt{t}_n\sigma_5 = \mathbf{C}$
$\mathtt{x}_1$ **A A C C A B** $\mathtt{x}_3$ **B B C**	$\mathsf{P}_2^3\sigma_6,\ \mathtt{t}_n\sigma_6 = \mathbf{C}$
$\mathtt{x}_1$ **A A C** $\mathtt{x}_2$ **C A B B B C**	$\mathsf{P}_2^4\sigma_7,\ \mathtt{t}_n\sigma_7 = \mathbf{A}$
$\mathtt{x}_1$**A A C** $\mathtt{x}_2$ **C A B** $\mathtt{x}_3$ **B B C**	$\mathsf{P}_2^4\sigma_8,\ \mathtt{t}_n\sigma_8 = \mathbf{A}$

Размер алфавита

Все хорошие свойства образцов, позволяющие работать с ними обычными методами (поиск подстановки, разбиение Рейденбаха) — следствие того, что мы подразумеваем $|\Sigma| = O(|\Sigma_{\mathtt{Prog}}|^2)$, где Σ — алфавит входных данных, $\Sigma_{\mathtt{Prog}}$ — множество символов, явно входящих в образцы. Допущение реалистичное, учитывая, что «буквами» выступают и константные деревья.

Языки образцов $\mathtt{x}$ **A B** $\mathtt{y}$ **A** $\mathtt{z}$ и $\mathtt{x}$ **A** $\mathtt{y}$ **B A** $\mathtt{z}$ в алфавите {**A**, **B**, **C**} очевидно не сравнимы: первый распознаёт слово **ABCA**, второй распознаёт **ACBA**. А в алфавите {**A**, **B**} эти образцы описывают один и тот же язык[a].

[a] И поэтому, если алфавит входных данных явно присутствует в образцах, нужны другие способы проверки подстановок на однозначность.

О распознаваемых словах

Предположим, мы рассматриваем краткий образец $\mathsf{P} = \mathrm{x}_1\ \mathrm{x}_2\ \mathrm{x}_2\ \mathrm{x}_1$. Если он сопоставляется со словом $\mathbf{A}^n$, то, как уже было видно, сопоставлений может быть много. С какими ещё словами происходит такая же ситуация?

Чтобы ответить на указанный вопрос, предположим, что слово сопоставилось с P двумя разными способами. То есть нашлись x_1, x_2, z_1, z_2 такие, что

$$x_1\ x_2\ x_2\ x_1 = z_1\ z_2\ z_2\ z_1$$

Что нам даёт такое равенство и как его упрощать? Ответы на этот вопрос потребуют краткое введение в теорию уравнений в словах.