Об усилении хигманова вложения при построении регулярных приближений циклов

Антонина Н. Непейвода
Институт программных систем РАН
г. Переславль-Залесский

Объединенный семинар ИЦСА и ИЦПУ
26 февраля 2013 г., Переславль–Залесский

Семантическое дерево программы — дерево путей исполнения программы на возможных входных данных.

Пример

Рекурсивное определение факториала в арифметике Пеано

```
f(0)=1;

f(x+1)=m(x+1,f(x));

m(0,y)=0;

m(x+1,y)=a(y, m(x,y));

a(0,y)=y;

a(x+1,y)=a(x,y)+1;
```

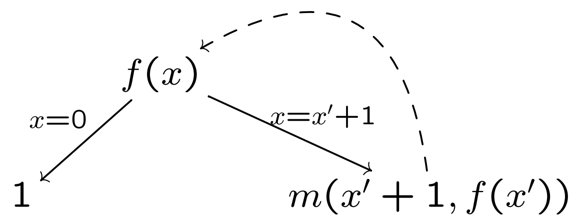
Начало развертки семантического дерева исполнения `f(x)`

$$
\begin{tikzcd}
 & f(x) \arrow[dl, "x=0"'] \arrow[dr, "x=x'+1"] & & \\
1 & & m(x'+1, f(x')) \arrow[dl, "x'=0"'] \arrow[dd, "x'=x''+1"] & \\
 & m(1,1) \arrow[d] & & \\
 & a(1, m(0,1)) & m(x''+1, m(x''+1, f(x''')) \arrow[dl, "x''=0"'] \arrow[dr, "x''=x'''+1"] & \\
 & \ldots & & \ldots
\end{tikzcd}
$$

Это дерево может содержать в себе пути сколь угодно большой длины (зависящей от x). Такие пути семантически соответствуют циклам.

Имеет смысл задача нахождения критериев для построения приближений циклов в программах. В частности, для семантических деревьев такой критерий должен сравнивать два узла $Node_1$, $Node_2$, таких, что $Node_1$ — предок $Node_2$, и строить аппроксимацию цикла, если структура $Node_2$ некоторым образом повторяет структуру $Node_1$.

Пример



$f(x)$ — подтерм $m(x'+1, f(x'))$ с точностью до переименования переменных. На основании этого наблюдения ветвь дерева сворачивается в граф.

Отношение R называется *почти полным (almost well)* на множестве последовательностей строк $S \subset \Sigma^*$ в некотором алфавите Σ, если всякая бесконечная последовательность строк $\{a_n\}$ содержит a_i, a_j, такие, что $i < j$ и $R\,(a_i, a_j)$.

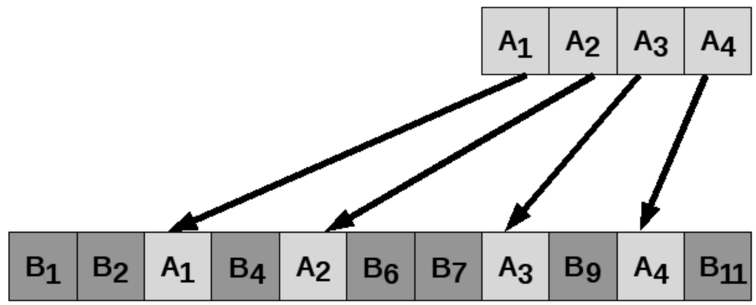
Это свойство гарантирует, что аппроксимация цикла будет построена на всех бесконечных ветвях семантического дерева программы.

Последовательность $\{a_n\}$, обладающая свойством $\forall i, j(i < j \Rightarrow (a_i, a_j) \notin R)$, называется *плохой последовательностью*.

Если отношение транзитивно и почти полно, оно называется *полным квазипорядком* (wqo).

Пример

Пусть дан конечный алфавит Σ. $A = a_1 a_2 \ldots a_m$, $B = b_1 b_2 \ldots b_n$, $\forall i, j (i \geq 1\ \&)$. Если $\forall i (i < m \Rightarrow \exists j, k (b_j = a_i\ \&\ b_k = a_{i+1}\ \&\ k > j))$, то A вкладывается в B в смысле Хигмана ($A \trianglelefteq B$).



$\trianglelefteq$ — почти полное отношение на произвольных последовательностях слов в конечном алфавите (Хигман, 1952).

В 1995 году М.Г. Серенсен предложил использовать хигманово вложение термов в узлах деревьев для построения аппроксимаций циклов в семантическом дереве программы.

Ранее, в 1988 году, В.Ф. Турчин предложил использовать для построения аппроксимаций циклов отношение над стеками функциональных вызовов, сходное с хигмановым вложением.

Набор $\langle \mathbf{\Sigma}, \mathbf{R}, \Gamma_0 \rangle$, где $\mathbf{\Sigma}$ — алфавит $\Gamma_0 \in \mathbf{\Sigma}^+$ — начальное слово и $\mathbf{R} \subset \mathbf{\Sigma}^+ \to \mathbf{\Sigma}^*$ — правила переписывания, называется *системой переписывания префиксов* (СПП), если правило $R : R_l \to R_r$ может применимо лишь к словам вида $R_l\Phi$ и порождает слова вида $R_r\Phi$.

Пример

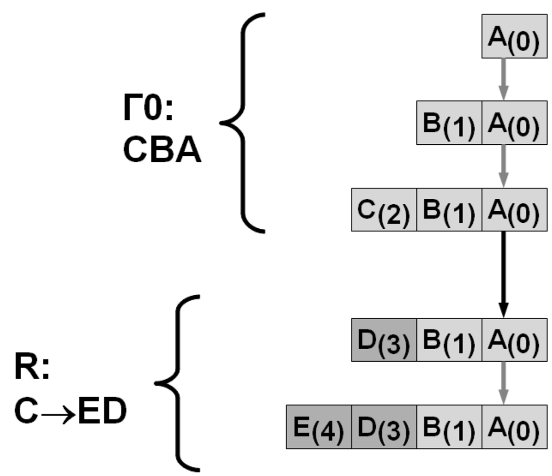
$aabab$ после применения правила $a \to bb$ превращается в $bbabab$, но не в $abbbab$ или $aabbbb$.

Если $|R_l| = 1$ для всех $R : R_l \to R_r$, то система переписывания префиксов называется *алфавитной* (АСПП).

Пусть последовательность слов $\{\Phi_i\}_{i=1}^{n}$ порождена СПП $\mathbf{G} = \langle \mathbf{\Sigma}, \mathbf{R}, \Gamma_0 \rangle$. Пометим все Φ_i *индексами времени*:

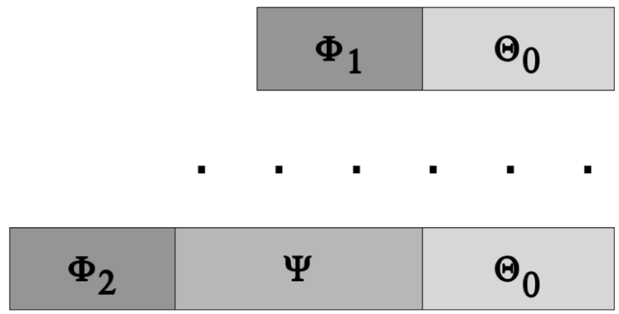
1. $\Phi_1[i]$ помечается индексом $|\Gamma_0| - i$;

2. Пусть наибольший индекс времени в отрезке последовательности $\{\Phi_i\}_{i=1}^{k}$ $(k < n)$ равен M и Φ_{k+1} порождено из Φ_k правилом $R : R_l \to R_r$. Тогда $\Phi_{k+1}[i]$ $(i \le |R_r|)$ получает временной индекс $M + |R_r| - i + 1$. Остальные буквы в Φ_{k+1} сохраняют прежние временные индексы.



Определение отношения Турчина

$\Gamma \preceq \Delta \Leftrightarrow \Gamma = \Phi_1\Theta_0 \,\&\, \Delta = \Phi_2\Psi\Theta_0 \,\&\, \Phi_1 \approx \Phi_2$



Теорема Турчина

$\preceq$ — почти полное отношение на последовательностях, порожденных АСПП над конечным алфавитом.

Пример

$$\mathbf{G_F}:$$
$$\begin{array}{lll} R_1 : f \to \Lambda & R_3 : m \to \Lambda & R_5 : a \to \Lambda \\ R_2 : f \to fm & R_4 : m \to ma & R_6 : a \to a \end{array}$$

Начальное слово Γ_0 — fa. Начальный отрезок порождаемой последовательности может выглядеть следующим образом

$$\begin{array}{c} \Gamma_0 : \mathbf{f}_{(1)} a_{(0)} \\ \downarrow{\scriptstyle R_2} \\ \Gamma_1 : \mathbf{f}_{(3)} m_{(2)} a_{(0)} \\ \downarrow{\scriptstyle R_2} \\ \Gamma_2 : \mathbf{f}_{(5)} m_{(4)} m_{(2)} a_{(0)} \\ \downarrow{\scriptstyle R_1} \\ \Gamma_3 : \mathbf{m}_{(4)} m_{(2)} a_{(0)} \\ \downarrow{\scriptstyle R_4} \\ \Gamma_4 : \mathbf{m}_{(7)} a_{(6)} m_{(2)} a_{(0)} \end{array}$$

$\Gamma_0 \preceq \Gamma_1$, $\Gamma_0 \preceq \Gamma_2$, $\Gamma_1 \preceq \Gamma_2$, и $\Gamma_3 \preceq \Gamma_4$.

Верхняя оценка на длину слова в плохой последовательности, порожденной АСПП

$$|\Gamma_i| \leq |\Gamma_0| + \sum(|(R_i)_r| - 1)$$

$\sum(|(R_i)_r| - 1)$ пробегает все правила с $|R_r| \neq \Lambda$.

Обозначим Γ^- слово Γ без первой буквы. Правило R неукорачивающее, если $R_r \neq \Lambda$.

Пусть R — неукорачивающее правило, $R(\Theta_0[1])\Theta_0^-$ предшествует $R(\Theta_1[1])\Theta_1^-$ и $\exists i(\Theta_1^-[i] = \Theta_0^-[1])$, тогда $R(\Theta_0[1])\Theta_0^- \preceq R(\Theta_1[1])\Theta_1^-$.

Ограниченность длины слова в плохой последовательности доказывает почти полноту отношения $\preceq$ в конечном алфавите.

Грубая верхняя оценка на длину плохой последовательности относительно $\preceq$

$$C_{Max} = card(\Sigma)^{|\Gamma_0| + \sum(|(R_i)_r| - 1)}.$$

АСПП $\mathbf{G}$=$\langle\mathbf{\Sigma}, \mathbf{R} \subset \mathbf{\Sigma} \rightarrow \mathbf{\Sigma}^*, \Gamma_0\rangle$ называется *обогащенной* или $\mathfrak{G}$-*системой*, если

1. Для всяких двух правил $R : a \rightarrow R_r$, R': $b \rightarrow R'_r$ $\exists i, j(R_r[i] \approx R'_r[j]) \Rightarrow i = j \;\&\; |R| = |R'| \;\&\; \forall i(i < |R| \Rightarrow R_r[i] \approx R'_r[i])$;

2. Если $R^a \in \mathbf{R}$, $R^a : a \rightarrow R_r$ и $b \in \mathbf{\Sigma}$, то $R^b : b \rightarrow R_r \in \mathbf{R}$.

Рассмотрим следующее преобразование АСПП $\mathbf{G}$ в обогащенную $\mathbf{G}'$.

1. Пусть $a = R_r[i]$, $a \in \Sigma$. Заменим a на пару $(a, 2^r * 3^{i-1})$. Начальное слово считаем правилом номер 0.

2. К каждой паре правил $a_1 \to \Phi$, $a_2 \to \Psi$ добавим правила $a_1 \to \Psi$ и $a_2 \to \Phi$ $(a_i \in \Sigma)$.

3. Пусть $(a_i, n_i) = R'_r[i]$. Для каждого $a_i \to \Phi$ добавим в $\mathbf{R}$ правила $(a_i, n_i) \to \Phi$.

Если в $\mathfrak{G}$-системе $\exists R(R : (a, n) \to R_r)$, то для всякого (b, m) $\exists R(R : (b, m) \to R_r)$, поэтому правила переписывания в $\mathfrak{G}$-системах будут записываться в общем виде: $R : x \to R_r$.

Пример

Переведем АСПП $\mathbf{G_F}$ в $\mathfrak{G}$-форму.

$$
\begin{array}{ll}
\mathbf{G'_F}: & \\
R_0 : x \to f^1 a^3 & R_2 : x \to m^4 a^{12} \\
R_1 : x \to f^2 m^6 & R_3 : x \to a^8 \\
R_4 : x \to \Lambda &
\end{array}
$$

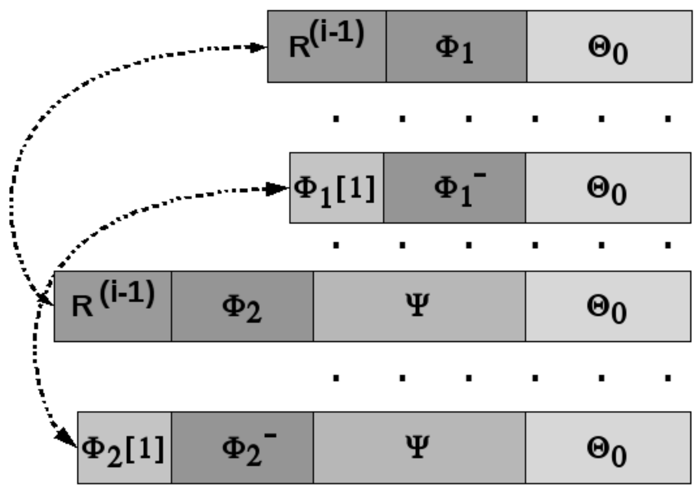
Порожденная последовательность после преобразования:

$$
\begin{array}{c}
\Gamma_0 : \mathbf{f}^1_{(1)} a^3_{(0)} \\
\big\downarrow R_1 \\
\Gamma_1 : \mathbf{f}^2_{(3)} m^6_{(2)} a^3_{(0)} \\
\big\downarrow R_1 \\
\Gamma_2 : \mathbf{f}^2_{(5)} m^6_{(4)} m^6_{(2)} a^3_{(0)} \\
\big\downarrow R_4 \\
\Gamma_3 : \mathbf{m}^6_{(4)} m^6_{(2)} a^3_{(0)} \\
\big\downarrow R_2 \\
\Gamma_4 : \mathbf{m}^4_{(7)} a^{12}_{(6)} m^6_{(2)} a^3_{(0)}
\end{array}
$$

Теперь $\Gamma_0 \not\preceq \Gamma_1$.

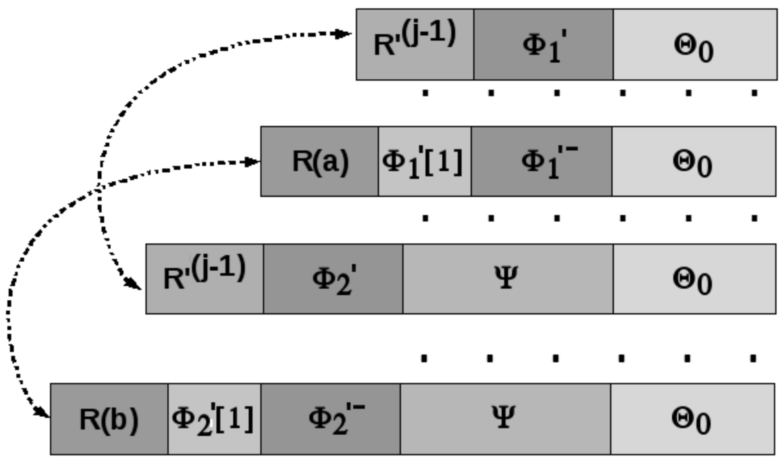
Самая первая пара вложенных по Турчину слов в последовательности, порожденной $\mathfrak{G}$-системой, имеет вид $R(a)\Theta_0$, $R(b)\Psi\Theta_0$.

Рассмотрим первую пару $\Phi_1\Theta_0$, $\Phi_2\Psi\Theta_0$, такую, что $\Phi_1\Theta_0 \preceq \Phi_2\Psi\Theta_0$ ($\Phi_1 \approx \Phi_2$), обрывающую плохую последовательность. $\Phi_1[1]$ и $\Phi_2[1]$ должны порождаться различными применениями R, и если $\Phi_1[1] \approx R_r[i]$, то $\Phi_2[1] \approx R_r[i]$. Обозначим слово $R_r[1]R_r[2]...R_r[i-1]$ как $R^{(i-1)}$ и рассмотрим моменты применения R. Первый имеет вид $R^{(i-1)}_{(k_1)}\Phi_1\Theta_0$, второй — $R^{(i-1)}_{(k_2)}\Phi_2\Psi\Theta_0$.



Они образуют турчинскую пару, следовательно, совпадают с $\Phi_1\Theta_0$ и $\Phi_2\Psi\Theta_0$.

Итак, $\Phi_1 = R(a)\Phi_1'$, $\Phi_2 = R(b)\Phi_2'$ ($\Phi_1' \approx \Phi_2'$). Пусть $\Phi_1' \neq \Lambda$. Тогда $\exists R', j(\Phi_1'[1] \approx R_r'[j]$ & $\Phi_2'[1] \approx R_r'[j])$ и $\Phi_1'[1] \neq \Phi_2'[1]$. Обозначим слово $R'[1]R'[2]...R'[j-1]$ как $R'^{(j-1)}$ и рассмотрим моменты применения R' к предшественникам $\Phi_1\Theta_0$ и $\Phi_2\Psi\Theta_0$.



Они выглядят как $R'^{(j-1)}_{(l_1)}\Phi_1'\Theta_0$ и $R'^{(j-1)}_{(l_2)}\Phi_2'\Psi\Theta_0$ и образуют турчинскую пару. Следовательно, $\Phi_1\Theta_0 = R(a)\Theta_0$ и $\Phi_2\Psi\Theta_0 = R(b)\Psi\Theta_0$.

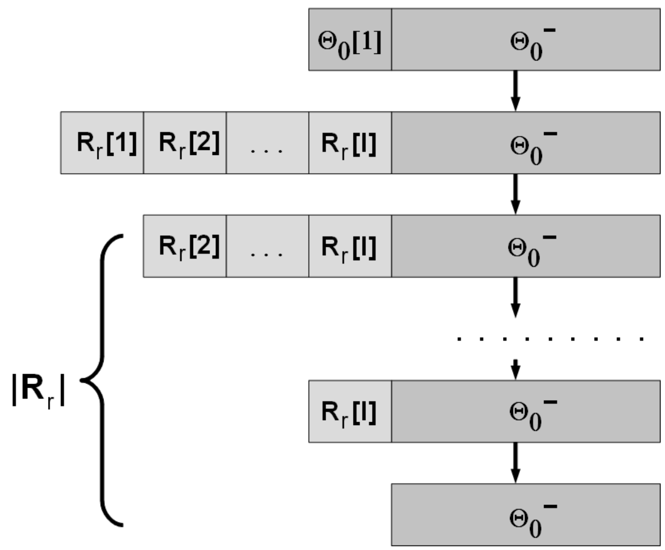
Применение правила $R(\Delta[1])$ *отменено* в $\Delta_2 \Leftrightarrow \Delta_2[1] = \Delta_1^-[1] \vee \forall i(\Delta_2[i] \neq \Delta_1^-[1])$.

Если некоторое неукорачивающее правило R никогда не применялось при порождении слова Г либо все его применения были отменены (назовем такие правила *свободными*), то применение R позволит построить более длинную плохую последовательность, чем стирание первой буквы Г.

Если неукорачивающие правила применяются к начальному слову длины 1 в порядке $R_0, R_1, ..., R_N$ (то есть R_i всегда приоритетнее, чем R_{i+j}, если оба эти правила применимы без образования турчинской пары), то наибольшая длина плохой последовательности, которая может быть построена этими правилами

$$C'_{Max} = 1 + |(R_0)_r| * (1 + |(R_1)_r| * (... * (1 + |(R_N)_r|)...))$$

Пусть в момент Θ_0 осталось только одно свободное неукорачивающее правило R_0. Длина наибольшего сегмента плохой последовательности, построенном на Θ_0 и заканчивающегося Θ_0^-, равна $|(R_0)_r| + 1$.



Пусть осталось $N+1$ свободных неукорачивающих правил.

Применим самое приоритетное правило R_k, получим $R_k(\Theta_0[1])\Theta_0^-$. На нем построим самую длинную плохую последовательность, сохраняющую $R_k^-(\Theta_0[1])\Theta_0^-$ (длину этого сегмента будем считать известной и обозначим за $C(N)$). Сегмент завершится словом $R_k^-(\Theta_0[1])\Theta_0^-$. На нем построим самую длинную плохую последовательность, сохраняющую $R_k^{--}(\Theta_0[1])\Theta_0^-$, и т.д. до момента Θ_0^-. Каждая буква в $(R_k)_r$ породит $C(N)$ элементов плохой последовательности. Итоговое увеличение длины равно $1+|(R_k)_r|*C(N)$.

Пример

Построим самую длинную плохую последовательность в $\mathfrak{G}$-системе $\mathbf{G}'_{\mathbf{F}}$. Ее длина равна $2*(1+2*(1+2*(1+1))) = 22$.

$$\begin{array}{ll} \mathbf{G}'_{\mathbf{F}}: & \\ R_0 : x \to ab & R_2 : x \to ef \\ R_1 : x \to cd & R_3 : x \to g \\ R_4 : x \to \Lambda & \end{array}$$

$$\begin{array}{llll}
\Gamma_0 : & \mathbf{a}_{(1)}b_{(0)} & \Gamma_{11} : & \mathbf{b}_{(0)} \\
\Gamma_1 : & \mathbf{c}_{(3)}d_{(2)}b_{(0)} & \Gamma_{12} : & \mathbf{c}_{(13)}d_{(12)} \\
\Gamma_2 : & \mathbf{e}_{(5)}f_{(4)}d_{(2)}b_{(0)} & \Gamma_{13} : & \mathbf{e}_{15}f_{(14)}d_{(12)} \\
\Gamma_3 : & \mathbf{g}_{(6)}f_{(4)}d_{(2)}b_{(0)} & \Gamma_{14} : & \mathbf{g}_{16}f_{(14)}d_{(12)} \\
\Gamma_4 : & \mathbf{f}_{(4)}d_{(2)}b_{(0)} & \Gamma_{15} : & \mathbf{f}_{(14)}d_{(12)} \\
\Gamma_5 : & \mathbf{g}_{(7)}d_{(2)}b_{(0)} & \Gamma_{16} : & \mathbf{g}_{(17)}d_{(12)} \\
\Gamma_6 : & \mathbf{d}_{(2)}b_{(0)} & \Gamma_{17} : & \mathbf{d}_{(12)} \\
\Gamma_7 : & \mathbf{e}_{(9)}f_{(8)}b_{(0)} & \Gamma_{18} : & \mathbf{e}_{(19)}f_{(18)} \\
\Gamma_8 : & \mathbf{g}_{(10)}f_{(8)}b_{(0)} & \Gamma_{19} : & \mathbf{g}_{(20)}f_{(18)} \\
\Gamma_9 : & \mathbf{f}_{(8)}b_{(0)} & \Gamma_{20} : & \mathbf{f}_{(18)} \\
\Gamma_{10} : & \mathbf{g}_{(11)}b_{(0)} & \Gamma_{21} : & \mathbf{g}_{(21)}
\end{array}$$

Применение любого из R_1–R_4 к Γ_{21} породит турчинскую пару.

$\preceq$ не является транзитивным, но существует транзитивное отношение T, $T \subset \preceq$, являющееся почти полным на всех последовательностях, порожденных АСПП.

$\mathbf{G_{ABC}}$:

$R_1 : a \to \Lambda \qquad R_3 : d \to \Lambda$

$R_2 : a \to ad \qquad R_4 : b \to fbe$

$R_5 : f \to ad$

$$\begin{array}{c}
\Gamma_0 : \mathbf{a}_{(2)} b_{(1)} c_{(0)} \\
\big\downarrow {\scriptstyle R_2} \\
\Gamma_1 : \mathbf{a}_{(4)} d_{(3)} b_{(1)} c_{(0)} \\
\big\downarrow {\scriptstyle R_1} \\
\Gamma_2 : \mathbf{d}_{(3)} b_{(1)} c_{(0)} \\
\big\downarrow {\scriptstyle R_3} \\
\Gamma_3 : \mathbf{b}_{(1)} c_{(0)} \\
\big\downarrow {\scriptstyle R_4} \\
\Gamma_4 : \mathbf{f}_{(7)} b_{(6)} e_{(5)} c_{(0)} \\
\big\downarrow {\scriptstyle R_5} \\
\Gamma_5 : \mathbf{a}_{(9)} d_{(8)} b_{(6)} e_{(5)} c_{(0)}
\end{array}$$

$\Gamma_0 \preceq \Gamma_1$ и $\Gamma_1 \preceq \Gamma_5$, но $\Gamma_0 \not\preceq \Gamma_5$.

Рассмотрим все порожденные последовательности $\{\Phi_i\}_{i=1}^{\infty}$, такие, что $\exists N\, \forall i\, \exists j (i < j \,\&\, |\Phi_j| \leq N)$.

Для каждой такой последовательности выберем наименьшее N, удовлетворяющее этому свойству. Существует конечное число классов эквивалентности Q' слов длины не больше N, таких, что $\forall i, j (|\Phi_i| \leq N \Rightarrow (\Phi_i \in Q' \,\&\, \Phi_j \in Q' \Leftrightarrow \Phi_i \approx \Phi_j))$. Хотя бы один из них содержит бесконечное число элементов — обозначим его $\{\Phi'_i\}$. Если для некоторых i, j, k $\Phi'_i[k] \neq \Phi'_j[k]$, то моменты, в которые $\Phi'_i[k]$ и $\Phi'_j[k]$ были порождены — результаты применения неукорачивающих правил. Значит, какое-то из них применяется бесконечное число раз.

Все прочие порожденные последовательности удовлетворяют условию $\forall N\,\exists i_N\,\forall j(j > i_N \Rightarrow |\Phi_j| > N)$.

Для каждого N выберем такое первое i_N, что $\forall j(j < i_N \Rightarrow \exists k(k \geq j \ \&\ |\Phi_k| < N))$. Итак, $|\Phi_{i_N-1}| < N$ и $|\Phi_{i_N}| \geq N$, и Φ_{i_N} порождено из предыдущего слова применением неукорачивающего правила R, $|R| \geq 2$: $\Phi_{i_N} = R(\Phi_{i_N-1}[1])\Phi^-_{i_N-1}$. $\Phi^-_{i_N-1}$ неизменно, поскольку $|\Phi_{i_N-1}| < N$. Все элементы последовательности $\{\Phi_{i_N}\}_{i=1}^{\infty}$ начинаются с правой части некоторого неукорачивающего правила, значит, существует бесконечная подпоследовательность $\{\Phi_{i_N}\}_{i=1}^{\infty}$, такая, что все ее элементы начинаются с правой части одного и того же правила.

Следствие

Пусть R — квазипорядок, полный на произвольных последовательностях. $R \cap \preceq$ — почти полное отношение.

Все турчинские пары в T являются результатами применения одного и того же неукорачивающего правила без его отмены. Поэтому турчинское отношение можно сузить до отношения $\preceq' = \{(\Gamma, \Delta) | \Gamma = R(a)\Theta_0 \ \& \ \Delta = R(b)\Psi\Theta_0\}$ без потери почти полноты.

Рассмотрим правило $a_1 a_2 \dots a_n \to b_1 b_2 \dots b_m$ неалфавитной СПП. Его действие эквивалентно композиции действий:

$a_1 \to \Lambda$

$\dots$

$a_{n-1} \to \Lambda$

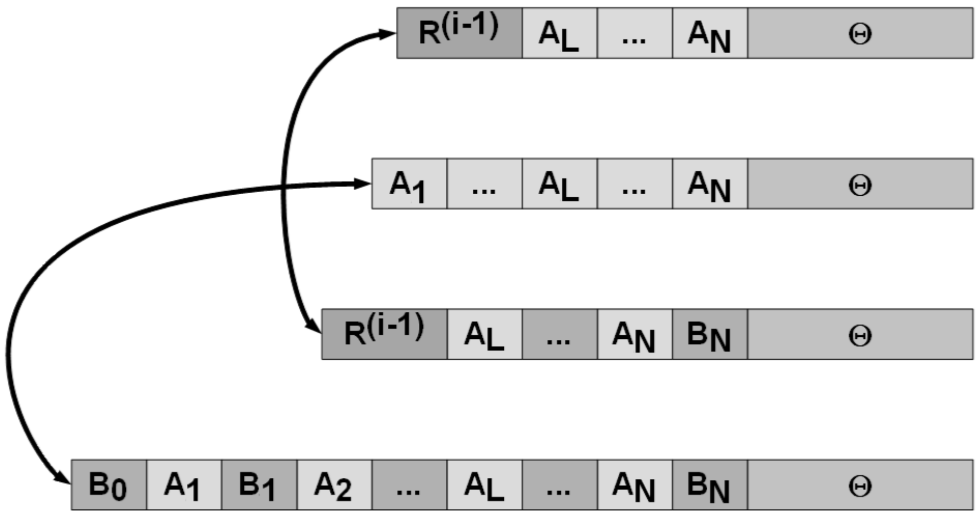
$a_n \to b_1 b_2 \dots b_m$.

Отношение $\preceq'$ почти полно на последовательностях, порожденных АСПП. Значит, $\preceq$ почти полно на последовательностях, порожденных произвольными СПП с конечным набором правил в конечном алфавите.

Первая хигманова пара в последовательности, порожденной $\mathfrak{G}$-системой, является турчинской парой.

Рассмотрим такие Φ_1 и Φ_2, что $\Phi_1 \trianglelefteq \Phi_2$, причем последовательность вплоть до Φ_2 представляет собой плохую последовательность в смысле $\trianglelefteq$.

$\Phi_1 = A_1 A_2 ... A_n \Theta$ и $\Phi_2 = B_0 A'_1 B_2 A'_2 ... B_{n-1} A'_n B_n \Theta$, и $\forall i (i \geq 1 \ \& \ i \leq n \Rightarrow A_i \approx A'_i)$. Вернемся к моментам порождения $A_l[1]$ и $A'_l[1]$ правилом $R : x \to R_r$ (пусть $A_l[1] \approx R_r[i]$).



В эти моменты имеем $R^{(i-1)}_{(k_1)}A_lA_{l+1}...A_n\Theta$ и $R^{(i-1)}_{(k_2)}A'_lB_l...B_{n-1}A_nB_n\Theta$ ($R^{(i-1)}_{k_j}$ обозначает префикс $R_r[1]_{(k_j+i-2)}...R_r[i-1]_{(k_j)}$). Они образуют хигманову пару, следовательно, $l=1=n$. Тогда $\Phi_1 = A_1\Theta$ и $\Phi_2 = A'_1B_n\Theta$, и $\Phi_1 \preceq \Phi_2$.

Выводы

1. $\preceq$ является почти полным над последовательностями, порожденными конечными СПП над конечным алфавитом. Верхняя оценка длины плохой последовательности по $\preceq$ экспоненциальна от размера исходной грамматики.

2. Плохие последовательности по $\trianglelefteq$ и по $\preceq$ совпадают на множестве последовательностей, порожденных обогащенными конечными СПП.

Направление развития

Поиск аналога $\preceq$ для деревьев функциональных вызовов.

Спасибо за внимание!