SRS. Иерархия Хомского. Регулярные грамматики

Теория формальных языков

2021 г.

String RS, или semi-Thue systems

Частный случай TRS — SRS (плоские TRS).

- Множество данных — строки (слова) в алфавите $\mathfrak{A}$.
- Правила переписывания имеют вид $u \to v$, где u, v — строки из $\mathfrak{A}^*$.
- Правило $u \to v$ применимо к строке Φ, если Φ содержит хотя бы одну подстроку u.
- В общем случае применение не детерминированно.

String RS, или semi-Thue systems

Частный случай TRS — SRS (плоские TRS).

- Множество данных — строки (слова) в алфавите $\mathfrak{A}$.
- Правила переписывания имеют вид $u \to v$, где u, v — строки из $\mathfrak{A}^*$.
- Правило $u \to v$ применимо к строке Φ, если Φ содержит хотя бы одну подстроку u.
- В общем случае применение не детерминированно.

Каково множество нормализованных строк в алфавите $\{\mathbf{A}, \mathbf{B}\}$ относительно системы правил $\mathbf{AB} \to \varepsilon$, $\mathbf{AA} \to \mathbf{A}$? относительно только правила $\mathbf{AB} \to \varepsilon$?

Выразительная сила SRS

Теорема

SRS позволяют выразить любую рекурсивную функцию.

Доказательство: алгорифмы Маркова.

Определение

Нормальный алгоритм (алгорифм) Маркова (НАМ) — это SRS с детерминированным поведением:

- top-down сопоставление (от верхних правил к нижним);
- выбор самой левой подстроки;
- существование терминальных правил.

Грамматики

Определение

Грамматика — это четвёрка $G = \langle N, \Sigma, P, S\rangle$, где:

- N — алфавит нетерминалов;
- Σ — алфавит терминалов;
- P — множество правил переписывания $\alpha \to \beta$ типа $\langle (N \cup \Sigma)^+ \times (N \cup \Sigma)^* \rangle$;
- $S \in N$ — начальный символ.

$\alpha \Rightarrow \beta$, если $\alpha = \gamma_1 \alpha' \gamma_2$, $\beta = \gamma_1 \beta' \gamma_2$, и $\alpha' \to \beta' \in P$.
$\Rightarrow^*$ — рефлексивное транзитивное замыкание $\Rightarrow$.

Определение

Язык $L(G)$, порождаемый G — множество $\{u \mid u \in \Sigma^* \; \& \; S \Rightarrow^* u\}$.

α-преобразование

По-разному воспринимают переименовку:

- Переменные vs конструкторы в TRS;
- Нетерминалы vs терминалы в грамматиках.

α-преобразование

По-разному воспринимают переименовку:

- Переменные vs конструкторы в TRS;
- Нетерминалы vs терминалы в грамматиках.

Для любой инъективной σ применение σ к правилам грамматики/trs для переменных и нетерминалов также называется α-преобразованием.

- α-преобразование не меняет терминальный язык;
- обычно термы различаются с точностью до α-преобразования.

α-преобразование

По-разному воспринимают переименовку:

- Переменные vs конструкторы в TRS;
- Нетерминалы vs терминалы в грамматиках.

Для любой инъективной σ применение σ к правилам грамматики/trs для переменных и нетерминалов также называется α-преобразованием.

- α-преобразование не меняет терминальный язык;
- обычно термы различаются с точностью до α-преобразования.

Неформально: контейнеры определяются не именем, а содержимым (см. экстенсиональность в логике).

Иерархия Хомского без ε-правил

$A, B \in N,\ a \in \Sigma^*,\ \alpha, \beta \in (N \cup \Sigma)^*,\ \gamma \in (N \cup \Sigma)^+$

Иерархия грамматик

Тип 0	Рекурсивно-перечислимые	$\forall$
Тип 1	Контекстно-зависимые	$\alpha A \beta \to \alpha\gamma\beta,\ \gamma \neq \varepsilon$
Тип 2	Контекстно-свободные	$A \to \gamma$
Тип 3	Праволинейные (регулярные)	$A \to a,\ A \to aB$

Иерархия Хомского без ε-правил

$A, B \in N,\ a \in \Sigma^*,\ \alpha, \beta \in (N \cup \Sigma)^*,\ \gamma \in (N \cup \Sigma)^+$

Иерархия грамматик

Тип 0	Рекурсивно-перечислимые	$\forall$
Тип 1	Контекстно-зависимые	$\alpha A \beta \to \alpha\gamma\beta,\ \gamma \neq \varepsilon$
Тип 2	Контекстно-свободные	$A \to \gamma$
Тип 3	Праволинейные (регулярные)	$A \to a,\ A \to aB$

Примеры языков

Тип 0	$\{u \mid L(u) = L(r)\}$, r — фикс. regex, u — regex;
Тип 1	$\{ww \mid w \in \Sigma^+\}$
Тип 2	непустые палиндромы в алфавите $\{a, b\}$
Тип 3	$\{w \mid w = aw_1\ \&\ (w = a^{2k} \vee w = a^{3k} \vee w \neq a^{5k})\}$

Иерархия Хомского с ε-правилами

$A, B \in N,\ a \in \Sigma^*,\ \alpha, \beta \in (N \cup \Sigma)^*,\ \gamma \in (N \cup \Sigma)^+$.

Иерархия грамматик

Тип 0	Рекурсивно-перечислимые	$\forall$
Тип 1	Контекстно-зависимые	$\alpha A \beta \to \alpha \gamma \beta,\ \gamma \neq \varepsilon$ $\vee S \to \varepsilon \ \& \ \forall p : \alpha \to \beta \in P \forall \beta_1, \beta_2 (\beta \neq \beta_1 S \beta_2)$
Тип 2	Контекстно-свободные	$A \to \alpha$
Тип 3	Регулярные	$A \to a,\ A \to aB,\ A \to \varepsilon$

Академические регулярные выражения $\mathcal{RE}$

Допустимые операции

- A^* — замыкание Клини — ноль или больше итераций A;
- A^+ — одна или больше итерация A;
- $A?$ — 0 или 1 вхождение A;
- $A|B$ — альтернатива (вхождение либо A, либо B).

Академические регулярные выражения $\mathcal{RE}$

Допустимые операции

- A^* — замыкание Клини — ноль или больше итераций A;
- A^+ — одна или больше итерация A;
- $A?$ — 0 или 1 вхождение A;
- $A|B$ — альтернатива (вхождение либо A, либо B).

Следствия

Если r_1, r_2 — $\mathcal{RE}$, тогда

- $r_1 \,|\, r_2$ — $\mathcal{RE}$;
- $r_1 r_2$ — $\mathcal{RE}$;
- r_1^*, r_2^+ — $\mathcal{RE}$.

Операции в регулярных грамматиках

Объединение

Дано: G_1 и G_2 — праволинейные. Построить $G : L(G) = L(G_1) \cup L(G_2)$.

1. Переименовать нетерминалы из N_1 и N_2, чтобы стало $N_1 \cap N_2 = \varnothing$ (сделать α-преобразование). Применить переименовку к правилам G_1 и G_2.
2. Объявить стартовым символом свежий нетерминал S и для всех правил G_1 вида $S_1 \to \alpha$ и правил G_2 вида $S_2 \to \beta$, добавить правила $S \to \alpha$, $S \to \beta$ в правила G.
3. Добавить в правила G остальные правила из G_1 и G_2.

Операции в регулярных грамматиках

Конкатенация

Дано: G_1 и G_2 — праволинейные. Построить $G : L(G) = L(G_1)L(G_2)$.

1. Переименовать нетерминалы из N_1 и N_2, чтобы стало $N_1 \cap N_2 = \varnothing$ (сделать α-преобразование).
2. Построить из G_1 её вариант без ε-правил (см. ниже).
3. По всякому правилу из G_1 вида $A \to a$ строим правило G вида $A \to aS_2$, где S_2 — стартовый нетерминал G_2.
4. Добавить в правила G остальные правила из G_1 и G_2. Объявить S_1 стартовым.
5. Если $\varepsilon \in L(G_1)$ (до шага 2), то по всем $S_2 \to \beta$ добавить правило $S_1 \to \beta$.

Операции в регулярных грамматиках

Положительная итерация Клини

Дано: G_1 — праволинейная. Построить $G : L(G) = L(G_1)^+$.

1. Построить из G_1 её вариант без ε-правил.
2. По всякому правилу из G_1 вида $A \to a$ строим правило G вида $A \to aS_1$, где S_1 — стартовый нетерминал G_1.
3. Добавить в правила G все (включая вида $A \to a$) правила из G_1. Объявить S_1 стартовым.
4. Если $\varepsilon \in L(G_1)$ (до шага 2), добавить правило $S_1 \to \varepsilon$ и вывести S_1 из рекурсии.

Построение грамматики без ε-правил

Дано: G — праволинейная. Построить G' без правил вида $A \to \varepsilon$ такую, что $\mathrm{L}(\mathrm{G}') = \mathrm{L}(\mathrm{G})$ или $\mathrm{L}(\mathrm{G}') \cup \{\varepsilon\} = \mathrm{L}(\mathrm{G})$.

1. Перенести в G' все правила G, не имеющие вид $A \to \varepsilon$.
2. Если существует правило $A \to \varepsilon$, то по всем правилам вида $\mathrm{B} \to \mathrm{a}A$ дополнительно строим правила $\mathrm{B} \to \mathrm{a}$.

Пересечение регулярных грамматик

Дано: G_1, G_2 — праволинейные. Построить G' такую, что $L(G') = L(G_1) \cap L(G_2)$.

1. Построить стартовый символ G' — пару $\langle S_1, S_2 \rangle$, где S_i — стартовый символ грамматики G_i.
2. Поместить $\langle S_1, S_2 \rangle$ в множество U неразобранных нетерминалов. Множество T разобранных нетерминалов объявить пустым.
3. Для каждого очередного нетерминала $\langle A_1, A_2 \rangle \in U$:
   3. если $A_1 \to a \in G_1$, $A_2 \to a \in G_2$, тогда добавить в G' правило $\langle A_1, A_2 \rangle \to a$;
   3. если $A_1 \to aA_3 \in G_1$, $A_2 \to aA_4 \in G_2$, тогда добавить в G' правило $\langle A_1, A_2 \rangle \to a\langle A_3, A_4 \rangle$, а в U — нетерминал $\langle A_3, A_4 \rangle$, если его ещё нет в множестве T;
   3. если все пары правил, указанные выше, были обработаны, тогда переместить $\langle A_1, A_2 \rangle$ из U в T.
4. Повторять шаг 3, пока множество U не пусто.
5. Если $\varepsilon \in L(G_1)$ & $\varepsilon \in L(G_2)$, тогда добавить в G' правило $\langle S_1, S_2 \rangle \to \varepsilon$.

От $\mathcal{RE}$ к грамматике

Теорема

Если $\mathrm{E} \in \mathcal{RE}$, то существует праволинейная регулярная грамматика G такая, что $\mathrm{L}(\mathrm{G}) = \mathrm{L}(\mathrm{E})$

От $\mathcal{RE}$ к грамматике

Теорема

Если $E \in \mathcal{RE}$, то существует праволинейная регулярная грамматика G такая, что $L(G) = L(E)$

Для каждой константы a_i в E построим правило $S_i \to a_i$. Объявим грамматику с одним этим правилом G_i. Последовательно соберём из таких грамматик грамматику для E, используя вышеописанные операции итерации, конкатенации, объединения.

Построим регулярную грамматику для $(a|(ab))^*b^+$.

- Объявим исходные правила: $S_1 \to a$, $S_2 \to ab$, $S_3 \to b$ (для краткости сразу для ab).
- Создадим грамматику G_4 для $G_1 \cup G_2$:
  $S_4 \to a \quad S_4 \to ab$
- По G_4 построим $G_5 = (G_4)^*$:
  $$\begin{array}{lll} S_5 \to aT & S_5 \to abT & S_5 \to \varepsilon \\ S_5 \to a & S_5 \to ab & T \to a \\ T \to aT & T \to abT & T \to ab \end{array}$$
- По G_3 построим $G_6 = (G_3)^+$: $S_6 \to bS_6$, $S_6 \to b$.
- Осталось построить $G_7 = G_5G_6$. Удаляем ε-правило:
  $$\begin{array}{llll} S_5 \to aT & S_5 \to abT & S_5 \to a & S_5 \to ab \\ T \to a & T \to aT & T \to abT & T \to ab \end{array}$$
  Проводим конкатенацию и возвращаем ε-правило:
  $$\begin{array}{llll} S_5 \to aT & S_5 \to abT & S_5 \to aS_6 & S_5 \to abS_6 \\ T \to aS_6 & T \to aT & T \to abT & T \to abS_6 \\ S_5 \to b & S_5 \to bS_6 & S_6 \to bS_6 & S_6 \to b \end{array}$$

Неподвижная точка $\mathcal{RE}$

Лемма Ардена

Пусть $X = (AX) \,|\, B$, где X — неизвестное $\mathcal{RE}$, а A, B — известные, причём $\varepsilon \notin L(A)$. Тогда $X = (A)^* B$.

Рассмотрим систему уравнений:

$$
\begin{array}{lcl}
X_1 & = & (A_{11}X_1) \,|\, (A_{12}X_2) \,|\, \ldots \,|\, B_1 \\
X_2 & = & (A_{21}X_1) \,|\, (A_{22}X_2) \,|\, \ldots \,|\, B_2 \\
 & \ldots & \\
X_n & = & (A_{n1}X_1) \,|\, (A_{n2}X_2) \,|\, \ldots \,|\, B_n
\end{array}
$$

Положим $\varepsilon \notin A_{ij}$. Будем последовательно выражать X_1 через $X_2, \ldots, X_n$, X_2 через $X_3 \ldots X_n$ и т.д. Получим регулярное выражение для X_n.

От грамматики к $\mathcal{RE}$

- Объявляем каждый нетерминал переменной и строим для него уравнение:
  - По правилу $A \to aB$ добавляем альтернативу aB;
  - По правилу $A \to b$ добавляем альтернативу без переменных.
  - Правило $S \to \varepsilon$ обрабатываем отдельно, не внося в уравнение: добавляем в язык альтернативу $(\mathcal{RE} \mid \varepsilon)$.
- Решаем систему относительно S.

От грамматики к $\mathcal{RE}$

Пример

Построим $\mathcal{RE}$ по грамматике:

$S \to aT \quad S \to abS$

$T \to aT \quad T \to bT \quad T \to b$

Строим по правилам грамматики систему:

$S = (abS) \mid (aT)$

$T = ((a \mid b)T) \mid b$

Решаем второе уравнение:

$T = (a \mid b)^* b$

Подставляем в первое:

$S = (abS) \mid (a(a \mid b)^* b)$

Получаем ответ:

$S = (ab)^* a(a|b)^* b$