Регулярные грамматики и выражения. Теорема Клини

Теория формальных языков

2022 г.

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 $\Rightarrow$.

Язык $\mathscr{L}(G)$, порождаемый G — множество $\{u \mid u \in \Sigma^* \ \&\ S \Rightarrow^* u\}$. Сентенциальная форма — элемент множества $\{u \mid u \in (N \cup \Sigma)^* \ \&\ S \Rightarrow^* u\}$.

Регулярные грамматики и НКА

Регулярная грамматика имеет правила вида $S \to \varepsilon$ (причём S не встречается в правых частях никаких правил), $T_i \to a_i$, $T_i \to a_i T_j$.

НКА (неформально) определяется списком правил перехода и финальными состояниями.

- $T_i \to a_i T_j$ соответствует переходу $\langle T_i, a_i, T_j \rangle$;
- $T_i \to a_i$ соответствует переходу $\langle T_i, a_i, F \rangle$, где F — уникальное финальное состояние;
- $S \to \varepsilon$ соответствует объявлению S финальным.

Лемма о накачке

Рассмотрим слово $w \in \mathscr{L}(G)$, $|w| \geqslant n + 1$. Оно получается применением не меньше, чем $n + 1$ правил $\Rightarrow$ после применения хотя бы двух из них в сентенциальной форме справа будет стоять один и тот же нетерминал A.

$$\underbrace{S \to \dots \to \Phi\ A \to \dots \to \Phi\ \Psi\ A}_{\text{не больше } n \text{ шагов}} \to \dots \to \Phi\ \Psi\ \Theta$$

$$\Downarrow$$

$$|\Phi| + |\Psi| \leqslant n$$

Лемма о накачке

Рассмотрим слово $w \in \mathscr{L}(G)$, $|w| \geqslant n + 1$. Оно получается применением не меньше, чем $n + 1$ правил $\Rightarrow$ после применения хотя бы двух из них в сентенциальной форме справа будет стоять один и тот же нетерминал A.

Известно, что $|\Phi| + |\Psi| \leqslant n$.

$$\underbrace{S \to \cdots \to \Phi}_{\rho_1:\,\text{вывод } \Phi A \text{ из } S}\ \overbrace{A \to \cdots \to \Phi\ \Psi}^{\rho_2:\,\text{вывод } \Psi A \text{ из } A}\ \underbrace{A \to \cdots \to \Phi\ \Psi\ \Theta}_{\rho_3:\,\text{вывод } \Theta \text{ из } A}$$

Все выводы вида $\rho_1\,(\rho_2)^*\,\rho_3$ допустимы в $G \Rightarrow$ $\forall k(\Phi\ \Psi^k\ \Theta \in \mathscr{L}(G))$.

Лемма о накачке

Утверждение

Если G — регулярная, то существует такое $n \in \mathbb{N}$, что $\forall w\big(w \in \mathscr{L}(G)\ \&\ |w| > n \Rightarrow \exists w_1, w_2, w_3\big(|w_2| > 0\ \&\ |w_1| + |w_2| \leqslant n\ \&\ w = w_1\ w_2\ w_3\ \&\ \forall k(k \geqslant 0 \Rightarrow w_1\ {w_2}^k\ w_3 \in \mathscr{L}(G))\big)\big)$.

Известно, что $|\Phi| + |\Psi| \leqslant n$.

$$\underbrace{S \to \cdots \to \Phi}_{\rho_1:\,\text{вывод } \Phi A \text{ из } S}\ \overbrace{A \to \cdots \to \Phi\ \Psi}^{\rho_2:\,\text{вывод } \Psi A \text{ из } A}\ \underbrace{A \to \cdots \to \Phi\ \Psi}_{\rho_3:\,\text{вывод } \Theta \text{ из } A}\ \Theta$$

Все выводы вида $\rho_1\ (\rho_2)^*\ \rho_3$ допустимы в $G \Rightarrow$ $\forall k(\Phi\ \Psi^k\ \Theta \in \mathscr{L}(G))$.

Примеры применения леммы о накачке

Обозначим обращение (reversal) слова w как w^{R}.
Рассмотрим язык $\mathscr{L} = \left\{w\, w^{R} \mid w \in \Sigma^{+}\right\}$.

Пусть длина накачки — n. Рассмотрим слово $b^{n+1} a\, a\, b^{n+1} \in \mathscr{L}$. Поскольку $|\Phi| + |\Psi| \leqslant n$, то $\Psi = b^{k},\ k \geqslant 1$. Но $b^{m} a\, a\, b^{n} \notin \mathscr{L}$, если $m \neq n$. Поэтому $\mathscr{L}$ — не регулярный.

Примеры применения леммы о накачке

Обозначим обращение (reversal) слова w как w^R.
Рассмотрим язык $\mathscr{L} = \{w\,w^R \mid w \in \Sigma^+\}$.

Пусть длина накачки — n. Рассмотрим слово $b^{n+1}a\,a\,b^{n+1} \in \mathscr{L}$. Поскольку $|\Phi| + |\Psi| \leqslant n$, то $\Psi = b^k,\ k \geqslant 1$. Но $b^m a\,a\,b^n \notin \mathscr{L}$, если $m \neq n$. Поэтому $\mathscr{L}$ — не регулярный.

Рассмотрим язык $\mathscr{L}' = \{a^n b^m \mid n \neq m\}$.

Пусть длина накачки — n. Рассмотрим множество слов $a^n b^{n+n!} \in \mathscr{L}'$. Поскольку $|\Phi| + |\Psi| \leqslant n$, то $\Psi = a^k,\ k \geqslant 1$. Но для всех $k \leqslant n\ \exists v(n + k * v = n + n!)$. Поэтому слово вида $a^{n+n!}b^{n+n!} \in \mathscr{L}'$, что абсурдно. Следовательно, $\mathscr{L}'$ не является регулярным.

Нерегулярные языки

Пусть $\mathscr{L} = \{w \mid |w|_a = |w|_b\}$. Все слова вида $a^k b^k$ принадлежат $\mathscr{L}$. Пусть длина накачки равна n. Рассмотрим слово $a^n b^n$. Поскольку $|\Phi| + |\Psi| \leqslant n$, то $\Psi = a^k$, $k > 0$. Но слова $a^{n+k*i} b^n$ не принадлежат $\mathscr{L}$.

Совпадает ли $\mathscr{L}$ с языком правильных скобочных последовательностей P (язык Дика)? Если да, доказать. Если нет, исследовать язык $L \setminus P$. Регулярен ли он?

Анализ на достаточность

Гипотеза

G — регулярная $\overset{???}{\Longleftrightarrow}$ существует такое $\mathtt{n} \in \mathbb{N}$, что $\forall w\big(w \in \mathscr{L}(\mathsf{G})\ \&\ |w| > \mathtt{n} \Rightarrow \exists w_1, w_2, w_3\big(|w_2| > 0\ \&\ |w_1| + |w_2| \leqslant \mathtt{n}\ \&\ w = w_1\ w_2\ w_3\ \&\ \forall \mathsf{k}(\mathsf{k} \geqslant 0 \Rightarrow w_1\ {w_2}^{\mathsf{k}}\ w_3 \in \mathscr{L}(\mathsf{G}))\big)\big)$.

Анализ на достаточность

Гипотеза

G — регулярная $\overset{???}{\Longleftrightarrow}$ существует такое $\mathfrak{n} \in \mathbb{N}$, что $\forall w \big(w \in \mathscr{L}(\mathsf{G})\ \&\ |w| > \mathfrak{n} \Rightarrow \exists w_1, w_2, w_3 \big(|w_2| > 0\ \&\ |w_1| + |w_2| \leqslant \mathfrak{n}\ \&\ w = w_1\ w_2\ w_3\ \&\ \forall \mathsf{k}(\mathsf{k} \geqslant 0 \Rightarrow w_1\ {w_2}^{\mathsf{k}}\ w_3 \in \mathscr{L}(\mathsf{G})) \big) \big)$.

Рассмотрим язык $\mathscr{L} = \left\{ w\, w^{\mathsf{R}}\, z \mid w \in \Sigma^+\ \&\ z \in \Sigma^+ \right\}$ и $\mathfrak{n} = 4$.

- Если $|w| = 1$, тогда можно разбить слово $w\, w^{\mathsf{R}}\, z$ так: $\Phi = w\, w^{\mathsf{R}}$, $\Psi = z[1]$, $\Theta = z\big[2..|z|\big]$. Тогда для всех k $\Phi\, \Psi^{\mathsf{k}}\, \Theta \in \mathscr{L}$.
- Если $|w| \geqslant 2$, тогда разбиваем так: $\Phi = \varepsilon$, $\Psi = w[1]$, $\Theta = w\big[2..|w|\big]\ w^{\mathsf{R}}\, z$. Слова $w\big[2..|w|\big]\ w^{\mathsf{R}}\, z$ и $w[1]^{\mathsf{k}}\, w\big[2..|w|\big]\ w^{\mathsf{R}}\, z$ при $\mathsf{k} \geqslant 2$ также принадлежат $\mathscr{L}$.

Смысл леммы о накачке

Структура доказательства указывает, что длина накачки n регулярного языка $\mathscr{L}$ не больше (возможно, меньше) числа нетерминалов в минимальной грамматике для $\mathscr{L}$.

Рассмотрим $\mathscr{L} = a \mid b \mid (a\ \{a \mid b\}^* a) | (b\ \{a \mid b\}^* b)$. Если выбрать $n = 2$, то в качестве Ψ можно взять вторую букву слова из $\mathscr{L}$. Пусть G имеет два нетерминала S, T и распознаёт $\mathscr{L}$. Если G содержит правила $S \to aT$ и $S \to bT$ (или $S \to aS$, $S \to bS$), то для некоторого непустого z слова вида az и bz будут либо оба принадлежать $\mathscr{L}$, либо нет, чего не может быть. Значит, G содержит либо пару $S \to aT$, $S \to bS$, либо пару $S \to bT$, $S \to aS$. Рассмотрим первый случай. Тогда для некоторого непустого z имеем $az \in \mathscr{L} \Leftrightarrow b^+ az \in \mathscr{L}$, что абсурдно.

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;
- $A \mid B$ — 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;
- $A \mid B$ — альтернатива (вхождение либо A, либо B).

Следствия

Если r_1, r_2 — $\mathcal{RE}$, тогда

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

Операции в регулярных грамматиках

Объединение

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

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 $G : \mathscr{L}(G) = \mathscr{L}(G_1)\ \mathscr{L}(G_2)$.

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 $\varepsilon \in \mathscr{L}(G_1)$ (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 $G : \mathscr{L}(G) = \mathscr{L}(G_1)^+$.

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 $\varepsilon \in \mathscr{L}(G_1)$ (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 $\mathscr{L}(G') = \mathscr{L}(G)$ или $\mathscr{L}(G') \cup \{\varepsilon\} = \mathscr{L}(G)$.

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 $B \to aA$ дополнительно строим правила $B \to a$.

Пересечение регулярных грамматик

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

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 $\varepsilon \in \mathscr{L}(G_1)$ & $\varepsilon \in \mathscr{L}(G_2)$, тогда добавить в G' правило $\langle S_1, S_2 \rangle \to \varepsilon$.

От $\mathcal{RE}$ к НКА: конструкция Глушкова

Теорема

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

Будем строить сразу же НКА, распознающий то же слово, что и E. Для этого определим следующие множества:

- $\mathsf{First}(\mathsf{E})$ — множество символов, с которых может начинаться слово, распознаваемое E.
- $\mathsf{Last}(\mathsf{E})$ — множество символов, которыми может заканчиваться слово, распознаваемое E.
- $\mathsf{Next}(\mathsf{E})$ — множество пар символов, которые могут идти в словах, распознаваемых E, друг за другом.

От $\mathcal{RE}$ к НКА: конструкция Глушкова

Теорема

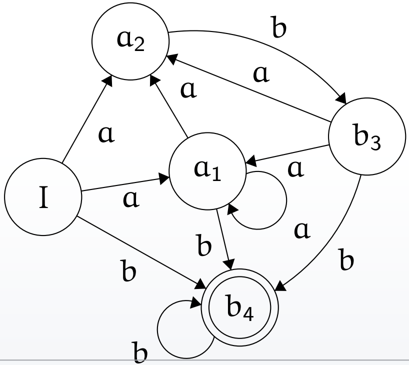
Если $E \in \mathcal{RE}$, то существует праволинейная регулярная грамматика G такая, что $\mathscr{L}(G) = \mathscr{L}(E)$.

В E пронумеруем все символы из Σ разными номерами. Для полученного E' построим $\mathsf{First}(E')$, $\mathsf{Last}(E')$, $\mathsf{Next}(E')$.

- Введём состояния, соответствующие буквам E' (нумерованным), а также входное состояние I.
- Если $\tau \in \mathsf{First}(E')$, тогда порождаем переход из I в τ (по символу τ).
- Если $\tau_1\tau_2 \in \mathsf{Next}(E')$, тогда порождаем переход из τ_1 в τ_2 по символу τ_2.
- Если $\tau \in \mathsf{Last}(E')$, тогда объявляем τ финальным.
- Стираем номера у символов на переходах. НКА, распознающий E, построен.

Построим НКА, распознающий $(a \mid (ab))^*b^+$.

- Линеаризуем: $E' = (a_1 \mid (a_2b_3))^*b_4^+$.
- Порождаем множества First, Last, Next:
  $\mathrm{First}(E') = \{a_1, a_2, b_4\}$
  $\mathrm{Last}(E') = \{b_4\}$
  $\mathrm{Next}(E') = \{a_1a_1, a_1a_2, a_2b_3, b_3a_1, b_3a_2, a_1b_4, b_3b_4, b_4b_4\}$
- Строим конечный автомат:



Производные $\mathcal{RE}$

Множество $a^{-1}U = \{w \mid aw \in U\}$ называется производным Бзрозовски множества U относительно a. Если $\varepsilon \in a^{-1}U$, тогда a распознаётся выражением U.

Λ_E положим равным $\{\varepsilon\}$, если $\varepsilon \in E$, и пустым множеством иначе.

- $a^{-1}\varepsilon = \varnothing$, $a^{-1}\varnothing = \varnothing$;
- $a^{-1}a = \{\varepsilon\}$, $a^{-1}b = \varnothing$;
- $a^{-1}(\Phi \mid \Psi) = a^{-1}(\Phi) \cup a^{-1}(\Psi)$;
- $a^{-1}(\Phi\ \Psi) = a^{-1}(\Phi)\Psi \cup \Lambda_\Phi\, a^{-1}(\Psi)$;
- $a^{-1}(\Phi^*) = a^{-1}(\Phi)\Phi^*$.

С помощью последовательного взятия производных можно свести задачу $w \in \mathscr{L}(R)$ к задаче $\varepsilon \in w^{-1}R$. На этом построен ещё один способ преобразования $\mathcal{RE}$ к автомату.

Пример преобразования

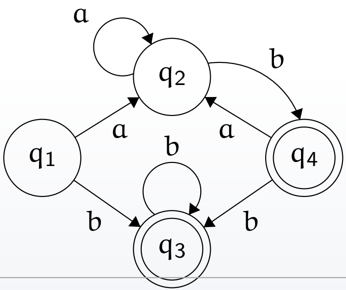
Рассмотрим всё то же выражение $(a \mid (ab))^* b^+$. Построим по нему автомат с помощью производных Брзозовски.

- $a^{-1}(a \mid (ab))^* b^+ = (a^{-1}(a \mid (ab))^*) b^+ \cup (a^{-1} b^+)$, но второе очевидно пусто, поэтому $a^{-1}(a \mid (ab))^* b^+ = (\varepsilon \mid b)\ (a \mid (ab))^* b^+$;
- $b^{-1}(a \mid (ab))^* b^+ = (b^{-1}(a \mid (ab))^*) b^+ \cup (b^{-1} b^+)$, и здесь как раз пусто первое, поэтому производная равна b^*.
- $a^{-1} b^* = \varnothing$; $b^{-1} b^* = b^*$.
- $a^{-1}((\varepsilon \mid b)\ (a \mid (ab))^* b^+)$ вынуждает первую альтернативу в $(\varepsilon|b)$ и порождает само себя.
- $b^{-1}((\varepsilon \mid b)\ (a \mid (ab))^* b^+)$ порождает $(a \mid (ab))^* b^+ \mid b^*$.
- $a^{-1}((a \mid (ab))^* b^+ \mid b^*)$ порождает $(\varepsilon \mid b)\ (a \mid (ab))^* b^+$, $b^{-1}((a \mid (ab))^* b^+ \mid b^*)$ порождает b^*.
- Переходы замкнулись. Осталось собрать производные в состояния автомата.

Пример преобразования

Рассмотрим всё то же выражение $(a \mid (ab))^*b^+$. Построим по нему автомат с помощью производных Брзозовски.

Состояние	Производная
q_1	$(a \mid (ab))^*b^+$
q_2	$(\varepsilon \mid b)\ (a \mid (ab))^*b^+$
q_3	b^*
q_4	$(a \mid (ab))^*b^+ \mid b^*$



Неподвижная точка $\mathcal{RE}$

Неподвижная точка функции $f(x)$ — такое x, что $f(x) = x$.

Лемма Ардена

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

Рассмотрим систему уравнений:

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

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 и НКА к $\mathcal{RE}$

Теорема Клини

По каждому НКА можно построить $\mathcal{RE}$, распознающую тот же язык. Верно и обратное.

Здесь считаем, что в НКА нет ε-переходов.

- Объявляем каждый нетерминал (или состояние НКА) переменной и строим для него уравнение:
  - По правилу $A \to aB$ (или для стрелки из A в B) добавляем альтернативу aB;
  - По правилу $A \to b$ (или для стрелки в финальное состояние) 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:

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