Регулярные грамматики и выражения. Теорема Клини

Теория формальных языков

2023 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.

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

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

То есть во всех сентенциальных формах либо нет нетерминалов, либо он единствен и расположен строго справа от терминальных символов.
Каждый нетерминал N описывает собственный язык $\mathscr{L}(N)$ относительно G — язык слов, которые выводятся из N за конечное число применений правил грамматики G.

Регулярные грамматики и НКА

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

То есть во всех сентенциальных формах либо нет нетерминалов, либо он единствен и расположен строго справа от терминальных символов.

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

Пусть n — число нетерминалов в регулярной грамматике G для языка $\mathscr{L}$.

Рассмотрим слово $w \in \mathscr{L}(G)$, $|w| \geqslant n+1$. Оно получается применением цепочки из $n+1$ правил $\Rightarrow$ после применения хотя бы двух из них нетерминал в сентенциальной форме результата повторится.

$$\underbrace{S \to \cdots \to \Phi\ A \to \cdots \to \Phi\ \Psi\ A}_{\text{не больше } n+1 \text{ шага}} \to \cdots \to \Phi\ \Psi\ \Theta$$

$$\Downarrow$$

$$|\Phi| + |\Psi| \leqslant n+1$$

По построению, $\Theta \in \mathscr{L}(A)$ (поскольку A в конечном счёте раскрывается в Θ), и также $\Psi\Theta \in \mathscr{L}(A)$, причём $|\Psi| > 0$. Кроме того, $\Phi\mathscr{L}(A) \subseteq \mathscr{L}(G)$, поскольку $S \to^* \Phi A$.

Лемма о накачке

Рассмотрим слово $w \in \mathscr{L}(G)$, $|w| \geqslant n + 1$. Оно получается применением цепочки из $n + 1$ правил $\Rightarrow$ после применения хотя бы двух из них нетерминал в сентенциальной форме результата повторится.

Известно, что $|\Phi| + |\Psi| \leqslant n + 1$.

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

Поскольку $A \to^* \Psi A$, то $\forall k(A \to^* \Psi^k A)$ (достаточно повторить k раз вывод ρ_2). Значит, $\forall k(\Phi\ \Psi^k\ \Theta \in \mathscr{L}(G))$.

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 $|\Phi| + |\Psi| \leqslant n + 1$.

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

Поскольку $A \to^* \Psi A$, то $\forall k(A \to^* \Psi^k A)$ (достаточно повторить k раз вывод ρ_2). Значит, $\forall k(\Phi\ \Psi^k\ \Theta \in \mathscr{L}(G))$.

Ещё раз о структуре накачек

Утверждение

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

- n — длина накачки;
- Φ — префикс накачки;
- Ψ — накачиваемый фрагмент (или просто «накачка»);
- Θ — суффикс накачки;
- $\Phi\Psi$ — область накачки;
- слово $\Phi\Theta$ (случай $k = 0$) — результат «пустой накачки» или «отрицательной накачки»;
- слова $\Phi\Psi^k\Theta$, где $k \geqslant 2$ — результаты «положительной накачки».

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 $\mathscr{L} = \left\{w\,w^R \mid w \in \Sigma^+\right\}$.

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.

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

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 $k \leqslant n\ \exists \nu(n + k \cdot \nu = n + n!)$. 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 $a^{n+k\cdot i} b^n$ не принадлежат $\mathscr{L}$.

Анализ на достаточность

Является ли лемма о накачке достаточной характеристикой регулярных языков? Существуют ли языки, которые «накачиваются» согласно её формулировке, но не регулярны?

Гипотеза

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

Анализ на достаточность

Гипотеза

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

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

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

Анализ на достаточность

Гипотеза

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

Мы нашли длину накачки для $\left\{ w\, w^R\, z \mid w \in \Sigma^+ \;\&\; z \in \Sigma^+ \right\}$ (она равна 4), но язык регулярным не является. Следовательно, лемма о накачке — только необходимое, но не достаточное условие регулярности.

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 $\mathscr{L}$.

Покажем, что у некоторых регулярных языков длина накачки действительно меньше, чем размер минимального НКА (или минимальной регулярной грамматики).

Смысл леммы о накачке

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

Таким образом, в грамматике для $\mathscr{L}$ должно быть больше двух нетерминалов (можно обойтись тремя).

Достаточный вариант леммы о накачке

Видно, что проблемы с языком $\{w w^{\mathsf{R}} z \mid w \in \Sigma^+ \ \& \ z \in \Sigma^+\}$ возникают из-за того, что у него очень удачный префикс: любая степень буквы, большая первой, начинается с палиндрома. Однако, если бы мы потребовали, чтобы слово из $\mathscr{L}$ начиналось с палиндрома хотя бы длины 4, подобное рассуждение уже не привело бы к успеху.

Достаточный вариант леммы о накачке

Мы можем искать не первый повтор нетерминала в пути разбора по грамматике, а любой, если осталось разобрать ещё достаточно длинный суффикс.

$$\underbrace{S \to \cdots \to \Phi\ A_0}_{\substack{\text{Произвольное}\\ \text{число шагов}}} \to \underbrace{\Phi\ \Psi'\ A \to \cdots \to \Phi\ \Psi'\ \Psi\ A}_{\substack{\text{Не более } m \text{ шагов}\\ \text{до повтора нетерминала}}} \to \cdots \to \Phi\ \Psi'\ \Psi\ \Theta$$

$\mathscr{L}$ регулярный $\Leftrightarrow$ существует универсальная длина накачки m такая, что $w \in \mathscr{L}$ $(|w| \geqslant m)$ для любого $i \leqslant |w| - m$ может быть представлено как $\Phi\Psi'\Psi\Theta$, где $|\Phi| = i$, $1 \geqslant |\Psi| \leqslant m$, $|\Psi'| + |\Psi| \leqslant m$, причём $\forall k(\Phi\Psi'\Psi^k\Theta \in \mathscr{L})$.

Академические регулярные выражения $\mathcal{RE}$

- $A \mid B$ — альтернатива (вхождение слова или из A, или из B);
- $A\,B$ — конкатенация (множество слов с префиксами из A и суффиксами из B);
- A^* — итерация Клини (0 или более конкатенаций A с собой).

- A^+ — положительная итерация (синтаксический сахар для выражения $A\,A^*$);
- $A?$ — опция (синтаксический сахар для выражения $(A \mid \varepsilon)$).

И менее очевидные синтаксические конструкции, такие как отрицание, положительные и отрицательные «ретроспективные» и «опережающие» проверки (моделирующие в т.ч. пересечения), сохраняющие выразительную силу регулярных языков.

Академические регулярные выражения $\mathcal{RE}$

- $A \mid B$ — альтернатива (вхождение слова или из A, или из B);
- $A\,B$ — конкатенация (множество слов с префиксами из A и суффиксами из B);
- A^* — итерация Клини (0 или более конкатенаций A с собой).

Приоритет операций: итерация $>$ конкатенация $>$ альтернатива, то есть $ab^* \mid c^*d$ — то же, что $\big(a(b^*)\big) \mid \big((c^*)d\big)$.

Определим $r_1 = r_2 \Leftrightarrow \mathscr{L}(r_1) = \mathscr{L}(r_2)$. Для всех $r_1, r_2, r_3 \in \mathcal{RE}$:

- операции конкатенации и альтернативы ассоциативны;
- $r_1 \mid r_2 = r_2 \mid r_1$;
- $r_1(r_2 \mid r_3) = r_1r_2 \mid r_1r_3$;
- $(r_1 \mid r_2)r_3 = r_1r_3 \mid r_2r_3$.

Полукольца

- Коммутативный моноид по сложению;
- Полугруппа по умножению;
- $\mathtt{id}$ по сложению — ноль по умножению;
- левая и правая дистрибутивности.

Регулярные выражения — идемпотентное полукольцо.

Алгебра Клини

- $x^*x + 1 = x^* = 1 + xx^*$
- (формализация Клини): $p + qx = x \Rightarrow x = q^*p$ (где q не распознаёт ε)
- (формализация Козена): $q + pr \leqslant r \Rightarrow p^*q \leqslant r$; $q + rp \leqslant r \Rightarrow qp^* \leqslant r$, где $x \leqslant y \Leftrightarrow x + y = y$.

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:

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

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.

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 $B \to \alpha A$ дополнительно строим правила $B \to \alpha$.

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 $\langle A_1, A_2 \rangle \in U$:
   - если $A_1 \to a \in G_1$, $A_2 \to a \in G_2$, тогда добавить в G' правило $\langle A_1, A_2 \rangle \to a$;
   - 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;
   - 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 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:

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