Теория автоматов и формальных языков Контекстно-свободные языки

Лектор: Екатерина Вербицкая

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»

23 октября 2020

В предыдущей серии

- Регулярные выражения, регулярные грамматики и конечные автоматы задают класс регулярных языков
- Класс регулярных языков замкнут относительно теоретико-множественных операций, конкатенации, итерации, гомоморфизма цепочек
- Определение принадлежности слова языку осуществляется за O(n) операций
- Однако класс регулярных языков достаточно узок, ни один используемый в промышленности язык программирования не является регулярным
 - ▶ Лемма о накачке для доказательства нерегулярности языка
 - Язык правильных скобочных последовательностей, язык палиндромов не являются регулярными

Контекстно-свободная грамматика

Четверка
$$\langle V_T, V_N, P, S \rangle$$

- V_T алфавит терминальных символов (терминалов)
- V_N алфавит нетерминальных символов (нетерминалов)
 - $V_T \cap V_N = \emptyset$
 - $V ::= V_T \cup V_N$
- ullet Р конечное множество правил вида A o lpha
 - $ightharpoonup A \in V_N$
 - $\alpha \in V^*$
- ullet S начальный нетерминал грамматики, $S \in V_N$

Пример: арифметические выражения

$$E \rightarrow E + E \mid E * E \mid N$$

$$N \rightarrow 0 \mid 1 \mid \dots \mid 9$$

Вывод в грамматике

• Отношение выводимости:

$$\forall \alpha, \gamma, \delta \in V^*, A \in V_N : A \to \alpha \in P : \gamma A \delta \Rightarrow \gamma \alpha \delta$$

- Вывод транизитивное, рефлексивное замыкание отношения выводимости $(\stackrel{*}{\Rightarrow}, \stackrel{+}{\Rightarrow}, \stackrel{k}{\Rightarrow})$
- **Левосторонний (правосторонний) вывод** на каждом шаге заменяем самый левый (правый) нетерминал
 - ▶ Если не специфицируется, подразумевается левосторонний вывод
- По сути, правила грамматики рассматриваются как правила переписывания

Пример вывода

Построим левосторонний вывод цепочки 2 + 3 * 4 в грамматике

$$\langle \{0, 1, \dots, 9, +, *\}, \{E, N\}, P, E \rangle$$

$$E \rightarrow E + N \mid E * N \mid N$$
$$N \rightarrow 0 \mid 1 \mid \cdots \mid 9$$

$$\underline{E} \Rightarrow \underline{E} * N \Rightarrow \underline{E} + N * N \Rightarrow \underline{N} + N * N \Rightarrow 2 + N * N \stackrel{2}{\Rightarrow} 2 + 3 * 4$$

Существование левостороннего вывода

Теорема

Если для цепочки ω существует некоторый вывод $S \stackrel{*}{\Rightarrow} \omega$, то существует и левосторонний вывод для этой цепочки $S \stackrel{*}{\Rightarrow} \omega$

Доказательство.

Докажем более общее утверждение: если существует $A\stackrel{*}{\Rightarrow}\omega$, то существует $A\stackrel{*}{\Rightarrow}\omega$, где $A\in V_N$.

Доказываем по индукции по длине вывода k

 $k=1:A\Rightarrow\omega$ — тривиально.

 $k \mapsto k+1 : \triangleleft A \Rightarrow \alpha \stackrel{*}{\Rightarrow} \omega.$

Обозначим $\alpha = B_1 B_2 \dots B_m \stackrel{*}{\Rightarrow} \omega_1 \omega_2 \dots \omega_m = \omega; \forall i : B_i \stackrel{l_i}{\Rightarrow} \omega_i, l_i \leq k$ По индукционному предположению $\forall i : B_i \stackrel{*}{\Rightarrow} \omega_i$

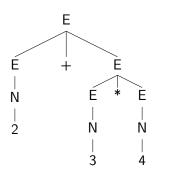
 $\Rightarrow: A \Rightarrow B_1 B_2 \dots B_m \stackrel{*}{\Rightarrow} \omega_1 B_2 \dots B_m \stackrel{*}{\Rightarrow} \omega$ — левосторонний вывод

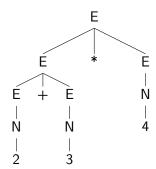
Единственность вывода

Не всегда (левосторонний) вывод единственен: 2 вывода строки 2+3*4

$$E \rightarrow E + E \mid E * E \mid N$$

$$N \rightarrow 0 \mid 1 \mid \cdots \mid 9$$





Однозначность грамматики

- Грамматика называется **однозначной**, если для *любого* слова языка существует *единственный* (левосторонний) вывод
- Грамматика называется **неоднозначной**, если *существует* слово языка, такое что для него *существует несколько* (левосторонних) выводов

Однозначность грамматики

- Грамматика называется **однозначной**, если для *любого* слова языка существует *единственный* (левосторонний) вывод
- Грамматика называется неоднозначной, если существует слово языка, такое что для него существует несколько (левосторонних) выводов
- По однозначной грамматике можно тривиальным образом построить неоднозначную: продублировать правило

$$S o A$$
 $A o a$ $S o A \mid B$ $A o a$ $B o a$

 Не существует общего алгоритма преобразования неоднозначной грамматики в однозначную

Примеры однозначной и неоднозначной грамматики

Неоднозначная грамматика

$$E \rightarrow E + E \mid E * E \mid N$$

$$N \rightarrow 0 \mid 1 \mid \dots \mid 9$$

Однозначная грамматика

$$\begin{split} E \rightarrow E + N \mid E * N \mid N \\ N \rightarrow 0 \mid 1 \mid \cdots \mid 9 \end{split}$$

Проверка однозначности грамматики неразрешима

Теорема

He существует алгоритма, определяющего по произвольной грамматике, что она является неоднозначной

Доказательство.

Сведение решения проблемы соответствий Поста к нашей задаче

Проблема соответствий Поста

Даны списки
$$A=(a_1,\ldots,a_n)$$
 и $B=(b_1,\ldots,b_n)$, где $\forall i:a_i,b_i\in\Sigma^*$. Существует ли непустая последовательность (i_1,\ldots,i_k) , удовлетворяющая условию $a_{i_1}\ldots a_{i_k}=b_{i_1}\ldots b_{i_k}$, где $\forall j:1\leq i_j\leq n$?

Проверка однозначности грамматики неразрешима

Теорема

Не существует алгоритма, определяющего по произвольной грамматике, что она является неоднозначной

Доказательство.

$$riangle$$
 алфавит $\Sigma = \{a_1, \ldots, a_n, b_1, \ldots, b_n\} \sqcup \{z_1, \ldots, z_n\}$, где a_i, b_i из ПСП

$$S \rightarrow A \mid B$$

$$A \rightarrow a_i A z_i \mid \varepsilon$$

$$B \rightarrow b_i B z_i \mid \varepsilon$$

Если грамматика неоднозначна, то существует выводимое слово вида:

$$a_{i_1}a_{i_2}\ldots a_{i_k}z_{i_k}z_{i_{k-1}}\ldots z_{i_1}=\omega=b_{i_1}b_{i_2}\ldots b_{i_k}z_{i_k}z_{i_{k-1}}\ldots z_{i_1}$$

Если бы умели решать ПСП, то мы могли бы проверить грамматику на однозначность, но ПСП неразрешима, а значит и проверка на однозначность неразрешима

Контекстно-свободный язык

- Язык называется контекстно-свободным, если для него существует контекстно-свободная грамматика
- Язык, задаваемый КС грамматикой $\langle V_T, V_N, P, S \rangle$: $\{\omega \in V_T^* \mid S \stackrel{*}{\Rightarrow} \omega\}$
- КС язык называется существенно неоднозначным, если для него не существует однозначной грамматики
 - ▶ $\{0^a1^b2^c\mid a=b$ либо $b=c\}$ доказательство в книге Ахо Ульмана

Пустота КС языка

Теорема

Существует алгоритм, определяющий, является ли язык, порождаемый КС грамматикой, пустым

Доказательство.

Для доказательства потребуется следующая лемма



Лемма

Теорема

Если в данной грамматике выводится некоторая цепочка, то существует цепочка, дерево вывода которой не содержит ветвей длиннее m, где m — количество нетерминалов грамматики

Доказательство.

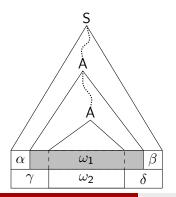
Рассмотрим дерево вывода цепочки ω . Если в нем есть 2 узла, соответствующих одному нетерминалу A, обозначим их n_1 и n_2 . Предположим, n_1 расположен ближе к корню дерева, чем n_2 ; $S \stackrel{*}{\Rightarrow} \alpha A_{n_1} \beta \stackrel{*}{\Rightarrow} \alpha \omega_1 \beta$; $S \stackrel{*}{\Rightarrow} \gamma A_{n_2} \delta \stackrel{*}{\Rightarrow} \gamma \omega_2 \delta$; ω_2 является подцепочкой ω_1 . Заменим в изначальном дереве узел n_1 на n_2 . Полученное дерево является деревом вывода $\alpha \omega_2 \beta$. Повторяем процесс замены одинаковых нетерминалов до тех пор, пока в дереве не останутся только уникальные нетерминалы.

В полученном дереве не может быть ветвей длины большей, чем *т.* По постороению оно является деревом вывода.

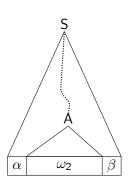
Лемма

Теорема

Если в данной грамматике выводится некоторая цепочка, то существует цепочка, дерево вывода которой не содержит ветвей длиннее m, где m — количество нетерминалов грамматики







Алгоритм проверки пустоты КС языка

Доказательство.

Строим коллекцию деревьев, представляющих вывод в грамматике.

- Инициализируем коллекцию деревом из одного узла S
- Добавляем в коллекцию дерево, полученное применением единственного правила грамматики из какого-нибудь дерева из коллекции, если его в нем еще нет, и самая длинная ветвь не длиннее m
- Всли после окончания построения коллекции в ней существует дерево, являющееся деревом вывода некоторой цепочки терминалов, значит, язык не пуст



Упрощение КС грамматики: удаление непродуктивных нетерминалов

Продуктивный нетерминал: нетерминал, для которого существует цепочка терминалов, выводимая из него $(\exists \omega \in V_T^* : A \stackrel{*}{\Rightarrow} \omega)$

Непродуктивный нетерминал: нетерминал, не являющийся продуктивным

Упрощение КС грамматики: удаление непродуктивных нетерминалов

Теорема

Для любой КС грамматики $G = \langle V_T, V_N, P, S \rangle$: $L(G) \neq \varnothing$ можно построить эквивалентную грамматику, каждый нетерминал которой продуктивен

Доказательство.

Удаляем из грамматики все нетерминалы $A:L(A)=\varnothing$, а также правила, использующие их.

Полученную грамматику обозначаем G_1 .

Докажем, что $L(G) = L(G_1)$.

Очевидно, $L(G_1) \subseteq L(G)$.

Докажем от противного, что $L(G) \subseteq L(G_1)$.

Предположим, что $\exists \omega \in L(G)$, но $\omega \notin L(G_1)$. Тогда $S \stackrel{*}{\Rightarrow} \alpha_1 A \alpha_2 \stackrel{*}{\Rightarrow} \omega$, где $A \in V_N \setminus V_{N_1}$, но тогда $\exists \gamma \in V_T^* : A \stackrel{*}{\Rightarrow} \gamma$. Противоречие

Упрощение КС грамматики: приведение

Теорема

Для любой КС грамматики, порождающей непустой язык, можно постороить эквивалентную, для каждого нетерминала A которой существует вывод вида $S \stackrel{*}{\Rightarrow} \omega_1 A \omega_3 \stackrel{*}{\Rightarrow} \omega_1 \omega_2 \omega_3, \omega_i \in V_T^*$

Доказательство.

Будем рассматривать грамматику без непродуктивных нетерминалов $G_1 = \langle V_{N_1}, V_T, P_1, S \rangle$.

Верно: если существует $S \stackrel{*}{\Rightarrow} \alpha_1 A \alpha_3, \alpha_i \in V^*$, то $S \stackrel{*}{\Rightarrow} \alpha_1 A \alpha_3 \stackrel{*}{\Rightarrow} \omega_1 A \omega_3 \stackrel{*}{\Rightarrow} \omega_1 \omega_2 \omega_3, \omega_i \in V_T^*$

Строим множество нетерминалов, встречающихся в выводах: добавляем сначала S, потом добавляем нетерминалы, встречающиеся в правой части правил для нетерминалов из множества. Завершаем процесс, когда больше ничего не добавить. Обозначаем полученное множество V_{N_2} , удаляем все правила грамматики, содержащие нетерминалы из $V_{N_1} \setminus V_{N_2}$

Упрощение КС грамматики: приведение

Доказательство.

Получили грамматику $G_2 = \langle V_{N_2}, V_T, P_2, S \rangle$.

Докажем: $L(G_2) = L(G_1)$

 $L(G_2)\subseteq L(G_1)$, так как $P_2\subseteq P_1$

Докажем: $L(G_1)\subseteq L(G_2)$. Пусть $S\overset{*}{\underset{G_1}{\Longrightarrow}}\omega$. Все нетерминалы,

встречающиеся в этом выводе содержатся в V_{N_2} , соответственно используются только правила из $P_2 \Rightarrow S \stackrel{*}{\underset{G_2}{\longrightarrow}} \omega$

Так как все нетерминалы V_{N_2} продуктивны, то

$$S \stackrel{*}{\Rightarrow} \omega_1 A \omega_3 \stackrel{*}{\Rightarrow} \omega_1 \omega_2 \omega_3, \omega_i \in V_T^*$$

Грамматика G_2 называется **приведенной**, ее нетерминалы — **достижимыми**

Недостижимые и непродуктивные нетерминалы называются **бесполезными**

Упрощение КС грамматики: удаление цепных правил

Правило называется **цепным**, если оно имеет вид A o B; $A, B \in V_N$.

Теорема

Для любой КС грамматики $G = \langle V_N, V_T, P, S \rangle$ можно построить эквивалентную, не содержащую цепных правил

Доказательство.

Строим новое множество правил P_1 .

Включаем в него все нецепные правила P.

Затем добавляем в P_1 правила вида $A\stackrel{*}{\Rightarrow} \alpha$, если $A\stackrel{*}{\Rightarrow} B$, где $A,B\in V_N$ и $B\to \alpha$ — нецепное правило из P.

Замечание: достаточно проверять только цепные выводы длины меньшей, чем V_{N}

Обозначим полученную грамматику за $G_1=\langle V_N,V_T,P_1,S
angle$, докажем $L(G_1)=L(G)$

Упрощение КС грамматики: удаление цепных правил

Доказательство.

Очевидно $L(G_1) \subseteq L(G)$ Покажем $L(G) \subseteq L(G_1)$. Пусть $\omega \in L(G)$. Рассмотрим левосторонний вывод $S \Rightarrow \alpha_0 \Rightarrow \alpha_1 \Rightarrow \ldots \Rightarrow \alpha_n = \omega$. Предположим $\alpha_i \Rightarrow \alpha_{i+1}$ — первый шаг, выполняемый посредством цепного правила в выводе; $\forall k \in [i..j]: \alpha_k \Rightarrow \alpha_{k+1}$ — посредством цепного правила; $\alpha_j \stackrel{\Rightarrow}{\underset{G}{\Rightarrow}} \alpha_{j+1}$ — посредством нецепного правила Тогда $|\alpha_i| = |\alpha_{i+1}| = \cdots = |\alpha_i|$, и на каждом шаге заменяется один и тот же нетерминал. Тогда $\alpha_i \Longrightarrow \alpha_{j+1}$ посредством правила из $P_1 \setminus P$, следовательно $\omega \in L(G_1)$

Нормальная форма Хомского

КС грамматика находится в **нормальной форме Хомского**, если все ее правила имеют вид:

$$A o BC$$
 $A,B,C\in V_N$ $A o a$ $A\in V_N,a\in V_T$ $S o arepsilon$ $S o au$ Стартовый нетерминал

Нормальная форма Хомского

Теорема

Для любой КС грамматики можно построить эквивалентную в нормальной форме Хомского

- Удаляем непродуктивные нетерминалы
- $oldsymbol{0}$ Удаляем цепные правила. Теперь $orall A o B:B\in V_T$
- $oldsymbol{2}$ Заменяем каждое правило $A o B_1B_2\dots B_n$ на $A o C_1C_2\dots C_n$
 - ▶ Если $B_i \in V_N$, $C_i = B_i$,
 - ▶ Если $B_i \in V_T$, добавляем также правило $C_i \to B_i$,
- $oldsymbol{3}$ Заменяем правило $A o C_1 C_2 \dots C_n$ на множество правил:

$$A \rightarrow C_1D_1$$

$$D_1 \rightarrow C_2D_2$$

$$\dots$$

$$D_{n-3} \rightarrow C_{n-2}D_{n-2}$$

$$D_{n-2} \rightarrow C_{n-1}C_n$$

Полученная грамматика находится в НФХ и эквивалентна данной

Пример приведения в НФХ

$$G = \langle \{S,A,B\}, \{a,b\},P,S
angle$$
, где P :

$$S \rightarrow bA \mid aB$$

 $A \rightarrow a \mid aS \mid bAA$
 $B \rightarrow b \mid bS \mid aBB$

- $S \rightarrow bA \Rightarrow S \rightarrow C_1A; C_1 \rightarrow b$
- $S \rightarrow aB \Rightarrow S \rightarrow C_2B$; $C_2 \rightarrow a$
- $A \rightarrow aS \Rightarrow A \rightarrow C_3S$; $C_3 \rightarrow a$
- $A \rightarrow bAA \Rightarrow A \rightarrow C_4D_1$; $C_4 \rightarrow b$; $D_1 \rightarrow AA$
- $B \rightarrow bS \Rightarrow B \rightarrow C_5S$; $C_5 \rightarrow b$
- $B \rightarrow aBB \Rightarrow B \rightarrow C_6D_2$; $C_6 \rightarrow a$; $D_2 \rightarrow BB$

Еще немного упростим

$$S \rightarrow bA \mid aB$$

 $A \rightarrow a \mid aS \mid bAA$
 $B \rightarrow b \mid bS \mid aBB$

$$S
ightarrow C_b A \mid C_a B$$

 $A
ightarrow a \mid C_a S \mid C_b D_1$
 $B
ightarrow b \mid C_b S \mid C_a D_2$
 $D_1
ightarrow AA$
 $D_2
ightarrow BB$
 $C_a
ightarrow a$
 $C_b
ightarrow b$

Алгоритм приведения в НФХ

- 1 Удалить стартовый нетерминал из правых частей правил
 - добавляется новое правило $S_0 o S, S_0 \notin V_N, S_0$ делается новым стартовым
- 2 Избавиться от неодиночных терминалов в правых частях
 - ▶ новое правило $C_c \rightarrow c$
- 3 Удалить длинные правила (длины больше 2)
- 4 Удалить непродуктивные правила (ε -правила)
 - ▶ Если $A \rightarrow \varepsilon$, то $A \varepsilon$ -правило
 - ▶ Если $A \to X_1 X_2 \dots X_n, \forall i : X_i \varepsilon$ -правило, то $A \varepsilon$ -правило
 - Заменяем $A \to X_1 X_2 \dots X_n$ на множество правил, где каждое из ε -правил опущено во всех возможных комбинациях, удаляем те, которые выводят ε
- 5 Удалить цепные правила
 - ▶ Для каждой пары правил $A \to B$; $B \to X_1 X_2 \dots X_n$ добавить правило $A \to X_1 X_2 \dots X_n$, цепное правило удалить

Порядок действий при приведении в НФХ

Порядок важен!

- 1 Удалить стартовый нетерминал из правых частей правил
- 2 Избавиться от неодиночных терминалов в правых частях
- Удалить длинные правила (длины больше 2)
- $oldsymbol{4}$ Удалить непродуктивные правила (arepsilon-правила)
- 5 Удалить цепные правила
- 1 шаг порождает новые цепные правила, поэтому его нельзя выполнять после 5 шага
- Если выполнить 4 шаг перед 3 шагом, то произойдет экспоненциальный взрыв грамматики
- 5 шаг приводит к квадратичному возрастанию размера грамматики
- Наиболее эффективны порядки 1, 2, 3, 4, 5 и 1, 3, 4, 5, 2

Увеличение размера грамматики при нормализации

Порядок важен!

- 1 Удалить стартовый нетерминал из правых частей правил
 - Увеличение на 1
- 2 Избавиться от неодиночных терминалов в правых частях
 - ightharpoonup Увеличение на $|V_T|$ правил
- 3 Удалить длинные правила (длины больше 2)
 - Увеличение не более, чем в 2 раза (для правил длины $k \ge 3$ порождается k-1 новых правил)
- $oldsymbol{4}$ Удалить непродуктивные правила (arepsilon-правила)
 - ▶ Увеличение не более, чем в 3 раза
- 5 Удалить цепные правила
 - ▶ Увеличение не более, чем в $O(n^2)$ (цепных правил не больше n^2 , где n число нетерминалов)

Итого: **полиномиальное** увеличение размеров грамматики при правильном порядке действий

Синтаксический анализ: алгоритм Кока-Янгера-Касами (Cocke-Younger-Kasami algorithm, CYK)

Что значит $A \rightarrow a$?

Синтаксический анализ: алгоритм Кока-Янгера-Касами (Cocke-Younger-Kasami algorithm, CYK)

Что значит $A \rightarrow a$?

$$A \Rightarrow a \stackrel{*}{\Rightarrow} \omega \Leftrightarrow \omega = a$$

Синтаксический анализ: алгоритм Кока-Янгера-Касами (Cocke-Younger-Kasami algorithm)

Что значит $A \rightarrow BC$?

Синтаксический анализ: алгоритм Кока-Янгера-Касами (Cocke-Younger-Kasami algorithm)

Что значит $A \rightarrow BC$?

$$A \Rightarrow BC \stackrel{*}{\Rightarrow} \omega \Leftrightarrow \begin{cases} \exists \omega_1, \omega_2 : \omega = \omega_1 \omega_2 \\ B \stackrel{*}{\Rightarrow} \omega_1 \\ C \stackrel{*}{\Rightarrow} \omega_2 \end{cases}$$

Синтаксический анализ: алгоритм Кока-Янгера-Касами (Cocke-Younger-Kasami algorithm)

Что значит $A \rightarrow BC$?

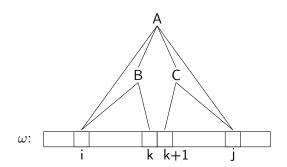
$$A \Rightarrow BC \stackrel{*}{\Rightarrow} \omega \Leftrightarrow \begin{cases} \exists \omega_1, \omega_2 : \omega = \omega_1 \omega_2 \\ B \stackrel{*}{\Rightarrow} \omega_1 \\ C \stackrel{*}{\Rightarrow} \omega_2 \end{cases}$$

Или:

$$A \Rightarrow BC \stackrel{*}{\Rightarrow} \omega \Leftrightarrow \exists k \in [0 \dots |\omega|] : \begin{cases} B \stackrel{*}{\Rightarrow} \omega[0 \dots k] \\ C \stackrel{*}{\Rightarrow} \omega[k+1 \dots |\omega|] \end{cases}$$

Синтаксический анализ: алгоритм Кока-Янгера-Касами (Cocke-Younger-Kasami algorithm, CYK)

- Алгоритм синтаксического анализа, работающий с грамматиками в НФХ
- Динамическое программирование



CYK

- ullet Дано: строка ω длины \emph{n} , грамматика $\emph{G} = \langle \emph{V}_{\emph{T}}, \emph{V}_{\emph{N}}, \emph{P}, \emph{S}
 angle$ в HФX
- Используем трехмерный массив d булевых значений размером $|V_N| \times n \times n, \ d[A][i][j] = true \Leftrightarrow A \stackrel{*}{\Rightarrow} \omega[i \dots j]$
- Инициализация: i = j
 - lacktriangledown d[A][i][i] = true, если в грамматике есть правило $A o \omega[i]$
 - ▶ d[A][i][i] = false, иначе
- Динамика. Предполагаем, d построен для всех нетерминалов и пар $\{(i',j') \mid j'-i' < m\}$
 - $d[A][i][j] = \bigvee_{A \to BC} \bigvee_{k=i}^{j} d[B][i][k] \wedge d[C][k+1][j]$
- В конце работы алгоритма в d[S][1][n] записан ответ, выводится ли ω в данной грамматике