Практическое занятие №6. LL(1)-анализ

Д.Ю. Чалый

24 декабря 2010 г.

1 Построение множеств $FIRST$ и $FOLLOW$

Ключевой задачей при использовании метода LL(1)-анализа является корректное построение множеств $FIRST$ и $FOLLOW$.

Рассмотрим следующую грамматику:

$$\begin{array}{lcl} S & \to & ABa \mid DE \\ A & \to & EC \mid AC \\ B & \to & b \mid \varepsilon \\ C & \to & c \\ D & \to & d \mid \varepsilon \\ E & \to & BAB \mid c \end{array}$$

Построим множества $FIRST(X)$, где $X \in N$. Физический смысл множеств $FIRST(X)$ состоит в том, что это множество терминальных символов, с которых могут начинаться строки, выводимые из X. Построение происходит по алгоритму, который имеет итеративный характер, т.е. мы постепенно пополняем множества $FIRST$ для нетерминальных символов грамматики. При этом пользуются следующими правилами:

1. Если в грамматике есть продукция вида $X \to a\alpha$, где $a \in T$, $\alpha \in (N \cup T)^*$, то в множество $FIRST(X)$ помещаем a. Таким образом, после использования этого правила мы получим следующий набор множеств $FIRST$:

$$\begin{array}{lcl} FIRST(S) & = & \varnothing \\ FIRST(A) & = & \varnothing \\ FIRST(B) & = & \{b, \varepsilon\} \\ FIRST(C) & = & \{c\} \\ FIRST(D) & = & \{d, \varepsilon\} \\ FIRST(E) & = & \{c\} \end{array}$$

2. Если в грамматике есть продукция вида $X \to A\alpha$, где $A \in N$, $\alpha \in (N \cup T)^*$, то в множество $FIRST(X)$ включаем $FIRST(A) \setminus \{\varepsilon\}$. Таким образом, после использования этого правила мы получим следующий набор множеств $FIRST$:

$$
\begin{array}{lcl}
FIRST(S) & = & \varnothing \cup (FIRST(D) \setminus \{\varepsilon\}) = \{d\} \\
FIRST(A) & = & \varnothing \cup (FIRST(E) \setminus \{\varepsilon\}) = \{c\} \\
FIRST(B) & = & \{b, \varepsilon\} \\
FIRST(C) & = & \{c\} \\
FIRST(D) & = & \{d, \varepsilon\} \\
FIRST(E) & = & \{c\} \cup (FIRST(B) \setminus \{\varepsilon\}) = \{c, b\}
\end{array}
$$

3. Для грамматик с ε-продукциями необходимо учитывать следующее. Если в грамматике есть продукция вида $X \to A_1 \ldots A_n$, где $A_i \in N \cup T$, причем $\forall i < j \colon \varepsilon \in FIRST(A_i)$ и $\varepsilon \notin FIRST(A_j)$, то в множество $FIRST(X)$ включаем $\bigcup_{i=1}^{j} FIRST(A_i) \setminus \{\varepsilon\}$. Таким образом, после использования этого правила мы получим следующий набор множеств $FIRST$:

$$
\begin{array}{lcl}
FIRST(S) & = & \{d\} \cup (FIRST(E) \setminus \{\varepsilon\}) = \{b, c, d\} \\
FIRST(A) & = & \{c\} \\
FIRST(B) & = & \{b, \varepsilon\} \\
FIRST(C) & = & \{c\} \\
FIRST(D) & = & \{d, \varepsilon\} \\
FIRST(E) & = & \{c, b\} \cup ((FIRST(B) \cup FIRST(A)) \setminus \{\varepsilon\}) = \{b, c\}
\end{array}
$$

4. Еще одно правило для грамматик с ε-продукциями. Если в грамматике есть продукция вида $X \to A_1 \ldots A_n$, где $A_i \in N \cup T$, то мы включаем в $FIRST(X)$ элемент ε тогда и только тогда, когда $\forall i \colon \varepsilon \in FIRST(A_i)$. В нашей грамматике таких продукций нет.

Для построения итоговых множеств $FIRST$ необходимо применять вышеописанные правила до тех пор, пока возможно пополнение хотя бы одного множества $FIRST$. Проделайте это самостоятельно, а в итоге должно получиться:

$$
\begin{array}{lcl}
FIRST(S) & = & \{b, c, d\} \\
FIRST(A) & = & \{b, c\} \\
FIRST(B) & = & \{b, \varepsilon\} \\
FIRST(C) & = & \{c\} \\
FIRST(D) & = & \{d, \varepsilon\} \\
FIRST(E) & = & \{b, c\}
\end{array}
$$

Множества $FOLLOW$ строятся на основе рассчитанных множеств $FIRST$. Физический смысл множества $FOLLOW(X)$, где $X \in N$ состоит в том, что это множество терминальных символов, которые могут находиться непосредственно за нетерминалом X в выводимой сентенциальной форме. Необходимо заметить, что в множество $FOLLOW$ не может входить ε, однако может входить специальный знак $\sharp$, обозначающий, что нетерминал может быть последним в выводимой сентенциальной форме. Очевидно, что знак $\sharp$ входит в множество $FOLLOW(S)$, где S — начальный символ грамматики, так как мы начинаем вывод именно с него. Сам алгоритм основывается на следующих правилах:

1. Помещаем в множество $FOLLOW(S)$ знак $\sharp$.

2. Пусть $X \to \alpha A a \beta$, где $\alpha, \beta \in (N \cup T)^*$, $A \in N$, $a \in T$, то помещаем в множество $FOLLOW(A)$ символ a.

3. Пусть $X \to \alpha A B \beta$, где $\alpha, \beta \in (N \cup T)^*$, $A, B \in N$, то помещаем в множество $FOLLOW(A)$ элементы множества $FIRST(B)$.

4. Пусть $X \to \alpha A$, где $\alpha \in (N \cup T)^*$, $A \in N$, то помещаем в множество $FOLLOW(A)$ элементы множества $FOLLOW(X)$.

5. Если в грамматике есть ε-продукции, то необходимо это учитывать. Пусть $X \to \alpha A X_1 X_2 \ldots X_n \beta$, где $\alpha, \beta \in (N \cup T)^*$, $A, X_1, X_2 \ldots X_n \in N$, причем $\forall i < n \colon \varepsilon \in FIRST(X_i)$, тогда помещаем в $FOLLOW(A)$ множества $FIRST(X_1), \ldots, FIRST(X_n)$.

6. Еще одно правило для того чтобы учитывать наличие ε-продукций. Пусть $X \to \alpha X_1 X_2 \ldots X_n$, где $\alpha \in (N \cup T)^*$, $X_1, X_2 \ldots X_n \in N$, причем $\forall 1 < i \leqslant n \colon \varepsilon \in FIRST(X_i)$, тогда помещаем в $FOLLOW(X_1), \ldots FOLLOW(X_n)$ элементы множества $FOLLOW(X)$.

7. Применяем вышеприведенные правила до тех пор, пока можем пополнять хоть одно множество $FOLLOW$.

Рассмотрим применение этих правил на примере нашей грамматики. Согласно правил 1 и 2, начальное содержимое множеств будет таким:

$$\begin{aligned}
FOLLOW(S) &= \{\sharp\} \\
FOLLOW(A) &= \varnothing \\
FOLLOW(B) &= \{a\} \\
FOLLOW(C) &= \varnothing \\
FOLLOW(D) &= \varnothing \\
FOLLOW(E) &= \varnothing
\end{aligned}$$

Применяем правила 3 и 5, получаем:

$$\begin{aligned}
FOLLOW(S) &= \{\sharp\} \\
FOLLOW(A) &= \varnothing \cup ((FIRST(B) \cup \{a\} \cup FIRST(C)) \setminus \{\varepsilon\}) = \{a, b, c\} \\
FOLLOW(B) &= \{a\} \cup (FIRST(A) \setminus \{\varepsilon\}) = \{a, b, c\} \\
FOLLOW(C) &= \varnothing \\
FOLLOW(D) &= \varnothing \cup (FIRST(E) \setminus \{\varepsilon\}) = \{b, c\} \\
FOLLOW(E) &= \varnothing \cup (FIRST(C) \setminus \{\varepsilon\}) = \{c\}
\end{aligned}$$

Теперь применим правила 4 и 6:

$$\begin{aligned}
FOLLOW(S) &= \{\sharp\} \\
FOLLOW(A) &= \{a, b, c\} \cup FOLLOW(E) = \{a, b, c\} \\
FOLLOW(B) &= \{a, b, c\} \cup FOLLOW(E) = \{a, b, c\} \\
FOLLOW(C) &= \varnothing \cup FOLLOW(A) = \{a, b, c\} \\
FOLLOW(D) &= \{b, c\} \\
FOLLOW(E) &= \{c\} \cup FOLLOW(S) = \{c, \sharp\}
\end{aligned}$$

...и еще несколько раз повторим применение правил и получим в итоге следующее

$$\begin{array}{lcl} FOLLOW(S) & = & \{\sharp\} \\ FOLLOW(A) & = & \{a, b, c, \sharp\} \\ FOLLOW(B) & = & \{a, b, c, \sharp\} \\ FOLLOW(C) & = & \{a, b, c, \sharp\} \\ FOLLOW(D) & = & \{b, c\} \\ FOLLOW(E) & = & \{c, \sharp\} \end{array}$$

2 Критерий для LL(1)-грамматик

Для того чтобы убедиться, что грамматика действительно является LL(1)-грамматикой, существует критерий, состоящий из двух частей:

1. Для любых двух продукций $A \to \alpha$ и $A \to \beta$, $FIRST(\alpha) \cap FIRST(\beta) = \varnothing$.

2. Если существует продукция $A \to \varepsilon$, то тогда для любой продукции $A \to \alpha$ должно выполняться $FOLLOW(A) \cap FIRST(\alpha) = \varnothing$.

Грамматика, рассмотренная ранее, не является LL(1)-грамматикой, из-за того, что не выполняется первый пункт критерия. Например, если рассмотреть продукции для нетерминала S, получим $FIRST(ABa) \cap FIRST(DE) = (FIRST(A)) \cap (FIRST(D) \cup FIRST(E)) = \{b, c\} \cap (\{d, \varepsilon\} \cup \{b, c\}) = \{b, c\} \cap \{\varepsilon, b, c, d\} = \{b, c\} \neq \varnothing$.

Вторая часть критерия не выполняется если рассмотреть продукции для нетерминала B. Действительно, $FOLLOW(B) \cap FIRST(b) \neq \varnothing$.

3 Метод LL(1)-анализа

Метод LL(1)-анализа является эффективным методом синтаксического анализа, который позволяет строить вывод входной строки за линейное время. Его можно промоделировать при помощи построения управляющей таблицы LL(1)-анализатора. Рассмотрим следующую LL(1)-грамматику:

$$\begin{array}{lcl} S & \to & AB \mid DE \\ A & \to & aA \mid \varepsilon \\ B & \to & b \mid dB \\ D & \to & cA \\ E & \to & e \mid \varepsilon \end{array}$$

Множества $FIRST$ и $FOLLOW$ для символов грамматики получаются такими (убедитесь в этом!):

$$
\begin{aligned}
FIRST(S) &= \{a, b, c, d\} \\
FIRST(A) &= \{\varepsilon, a\} \\
FIRST(B) &= \{b, d\} \\
FIRST(D) &= \{c\} \\
FIRST(E) &= \{\varepsilon, e\}
\end{aligned}
$$

$$
\begin{aligned}
FOLLOW(S) &= \{\sharp\} \\
FOLLOW(A) &= \{b, d, e, \sharp\} \\
FOLLOW(B) &= \{\sharp\} \\
FOLLOW(D) &= \{e, \sharp\} \\
FOLLOW(E) &= \{\sharp\}
\end{aligned}
$$

Управляющая таблица представляет собой структуру, в которой столбцы поименованы терминалами (плюс знак конца строки $\sharp$), строки нетерминалами, а в ячейках таблицы располагаются продукции. Сначала удобно расположить все продукции, не являющиеся ε-продукциями. Они располагаются в таблице по принципу если $A \to \alpha$ – продукция в исходной грамматике, то поместим ее в строку, отмеченную A и в столбцы, отмеченные символами из $FIRST(\alpha)$. Если действовать по такому принципу, то для нашей грамматики таблица заполнится так:

	a	b	c	d	e	$\sharp$
S	$S \to AB$	$S \to AB$	$S \to DE$	$S \to AB$		
A	$A \to aA$					
B		$B \to b$		$B \to dB$		
D			$D \to cA$			
E					$E \to e$	

Продукции вида $X \to \varepsilon$ помещаются в столбцы таблицы, помеченные символами из множества $FOLLOW(X)$. Для нашей грамматики мы получим такую таблицу:

	a	b	c	d	e	$\sharp$
S	$S \to AB$	$S \to AB$	$S \to DE$	$S \to AB$		
A	$A \to aA$	$A \to \varepsilon$		$A \to \varepsilon$	$A \to \varepsilon$	$A \to \varepsilon$
B		$B \to b$		$B \to dB$		
D			$D \to cA$			
E					$E \to e$	$E \to \varepsilon$

Посмотрим теперь, как это работает. Пусть на вход подана строка ce. Мы начинаем вывод с S. Смотрим, какая продукция находится в строке S и столбце c. Это продукция $S \to DE$, значит вывод надо начинать с нее:

$$S \to DE$$

Теперь нам надо раскрыть нетерминал D, заглядываем в строку, отмеченную им, и в столбец c, там находится продукция $D \to cA$, получаем

$$S \to DE \to cAE$$

Первым в сентенциальной форме cAE стоит терминал, который мы успешно сравниваем с первым символом входной строки, перейдем ко второму символу, e. Смотрим в строку A таблицы и столбец e. Там находится продукция $A \to \varepsilon$. получаем:

$$S \to DE \to cAE \to cE$$

Далее видим в строке E и столбце e находится продукция $E \to e$, применяем ее:

$$S \to DE \to cAE \to cE \to ce$$

Вот так мы вывели входную строку, пользуясь управляющей таблицей.