Практическое занятие №5. Синтаксический анализ с использованием магазинных автоматов

Д.Ю. Чалый

24 декабря 2010 г.

1 Построение эквивалентного магазинного автомата для КС-грамматики

Рассмотрим следующую грамматику:

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

Тогда эквивалентный этой грамматике МП-автомат будет таким: $M = (\{q\}, \{a, b, c, d\}, \{S, A, B, C, D, E\} \cup \{a, b, c, d\}, \delta, q, S, \{q\})$, где:

- $\delta(q, \varepsilon, S) = \{(q, ABa), (q, DE)\}$;
- $\delta(q, \varepsilon, A) = \{(q, EC), (q, AC)\}$;
- $\delta(q, \varepsilon, B) = \{(q, b), (q, \varepsilon)\}$;
- $\delta(q, \varepsilon, C) = \{(q, c)\}$;
- $\delta(q, \varepsilon, D) = \{(q, d), (q, \varepsilon)\}$;
- $\delta(q, \varepsilon, E) = \{(q, BAB), (q, c)\}$;
- $\delta(q, a, a) = \{(q, \varepsilon)\}$;
- $\delta(q, b, b) = \{(q, \varepsilon)\}$;
- $\delta(q, c, c) = \{(q, \varepsilon)\}$;
- $\delta(q, d, d) = \{(q, \varepsilon)\}$;

В общем-то построение автомата является простой задачей, а в динамике он ведет себя следующим образом. Если в стеке наверху находится нетерминал, то он заменяется на одну из правых частей продукций для этого нетерминала (так как МП-автомат является недетерминированным вычислительным устройством, то нельзя сказать на какую), а если сверху стека находится терминал, то происходит его сравнение с очередным символом входной строки. Если сравнение успешно, то символ изымается из стека, в противном случае автомат останавливается.

Проблема состоит в том, что МП-автомат является недетрминированным по своей природе и мы должны как-то промоделировать его детерминированными средствами (все алгоритмы в своей основе построены на положении, что на основе текущего состояния мы однозначным, детерминированным образом определяем как изменить это состояние).

Для МП-автоматов есть два алгоритма — поиск в ширину и поиск в глубину. Однако они применимы только когда исходная грамматика не является леворекурсивной. Данная грамматика имеет левую рекурсию. После преобразования (проделайте это сами и не забудьте про скрытую левую рекурсию) получится следующая грамматика[1]:

$$
\begin{array}{rcl}
S & \rightarrow & ABa \mid DE \\
A & \rightarrow & EC \mid ECA' \\
A' & \rightarrow & C \mid CA' \\
B & \rightarrow & b \mid \varepsilon \\
C & \rightarrow & c \\
D & \rightarrow & d \mid \varepsilon \\
E & \rightarrow & bAB \mid c \mid bABE' \mid cE' \\
E' & \rightarrow & CA'B \mid CB \mid CBE' \mid CA'BE'
\end{array}
$$

Перед рассмотрением алгоритмов необходимо поименовать продукции. Будем их именовать символом, который находится в голове продукции, а индекс будет задавать номер продукции по порядку вышеприведенной грамматики. Например, продукция $S \rightarrow ABa$ будем иметь имя S_1, а продукция $A' \rightarrow CA' - A'_2$.

Пусть на вход подана строка $\alpha = dccb$. Необходимо найти вывод этой строки в данной грамматике. Кстати, вообще говоря, нет необходимости явно строить МП-автомат для грамматики, так как он имеет простую структуру, и работать непосредственно с грамматикой.

2 Алгоритм поиска в ширину

Алгоритм поиска ширину начинает свою работу со нижеследующего состояния, которое соответствует начальной конфигурации магазинного автомата:

[1]Обратите внимание что приведенная грамматика имеет ε-продукции. В общем случае перед применением алгоритма удаления левой рекурсии от них необходимо избавляться, но для данной конкретной грамматики алгоритм применим напрямую. Наличие ε-продукций позволяет сократить объем грамматики.

1

	$dccb\sharp$
	$S\sharp$

Состояние удобно задавать в виде таблицы, где в правой части записывается сверху остаток входной строки, а снизу стек МП-автомата. Алгоритм моделирования в ширину состоит из применения двух видов шагов: раскрытие нетерминала и сравнение терминалов. Сначала совершаем раскрытие согласно тех продукций, которые имеются для нетерминала сверху стека (верхний символ располагает слева, в данном случае это символ S, второй символ, символ $\sharp$, находится под ним):

	$dccb\sharp$
S_1	$ABa\sharp$
S_2	$DE\sharp$

Так как для S имеются две продукции, то мы исходную строчку заменили на две, а слева записали имена продукций, которые мы применили. Шаг раскрытия нетерминала повторяется до тех пор, пока в правой части таблицы стеки начинаются с нетерминалов. Следующий шаг преобразует состояние следующим образом:

	$dccb\sharp$
S_1A_1	$ECBa\sharp$
S_1A_2	$ECA'Ba\sharp$
S_2D_1	$dE\sharp$
S_2D_2	$E\sharp$

Как видно, слева формируются строки, которые хранят историю применяемых продукций. Применяем шаг раскрытия продукций до тех пор, пока есть хоть одна строка, в левой части которой стек начинается с нетерминала. В итоге получим:

	$dccb\sharp$
$S_1A_1E_1$	$bABCBa\sharp$
$S_1A_1E_2$	$cCBa\sharp$
$S_1A_1E_3$	$bABE'CBa\sharp$
$S_1A_1E_4$	$cE'CBa\sharp$
$S_1A_2E_1$	$bABCA'Ba\sharp$
$S_1A_2E_2$	$cCA'Ba\sharp$
$S_1A_2E_3$	$bABECA'Ba\sharp$
$S_1A_2E_4$	$cE'CA'Ba\sharp$
S_2D_1	$dE\sharp$
$S_2D_2E_1$	$bAB\sharp$
$S_2D_2E_2$	$c\sharp$
$S_2D_2E_3$	$bABE'\sharp$
$S_2D_2E_4$	$cE'\sharp$

Теперь все строки в правой части таблицы начинаются с терминального символа. Производим сравнение этого символа с очередным символом на входе, символом d. Если символ сверху стека не совпадает с ним, то вычеркиваем строку. Если же символ совпадает, то моделируем его успешное сравнение, перенося из правой части таблицы в левую. Непосредственно после этого действия получим:

d	$ccb\sharp$
S_2D_1d	$E\sharp$

У нас снова в правой части таблицы появились строчки, которые начинаются с нетерминалов. Раскрываем нетерминалы так же, как и ранее:

d	$ccb\sharp$
$S_2D_1dE_1$	$bAB\sharp$
$S_2D_1dE_2$	$c\sharp$
$S_2D_1dE_3$	$bABE'\sharp$
$S_2D_1dE_4$	$cE'\sharp$

Снова все строки начинаются с терминальных символов, производим сравнение и получаем:

dc	$cb\sharp$
$S_2D_1dE_2c$	$\sharp$
$S_2D_1dE_4c$	$E'\sharp$

В таблице появилась строка, начинающаяся с нетерминала, раскрываем его:

dc	$cb\sharp$
$S_2D_1dE_2c$	$\sharp$
$S_2D_1dE_4cE'_1$	$CA'B\sharp$
$S_2D_1dE_4cE'_2$	$CB\sharp$
$S_2D_1dE_4cE'_3$	$CBE'\sharp$
$S_2D_1dE_4cE'_4$	$CA'BE'\sharp$

Следующий шаг:

dc	$cb\sharp$
$S_2D_1dE_2c$	$\sharp$
$S_2D_1dE_4cE'_1C_1$	$cA'B\sharp$
$S_2D_1dE_4cE'_2C_1$	$cB\sharp$
$S_2D_1dE_4cE'_3C_1$	$cBE'\sharp$
$S_2D_1dE_4cE'_4C_1$	$cA'BE'\sharp$

Производим сравнение:

dcc	$b\sharp$
$S_2D_1dE_4cE_1'C_1c$	$A'B\sharp$
$S_2D_1dE_4cE_2'C_1c$	$B\sharp$
$S_2D_1dE_4cE_3'C_1c$	$BE'\sharp$
$S_2D_1dE_4cE_4'C_1c$	$A'BE'\sharp$

После раскрытия нетерминалов получим:

dcc	$b\sharp$
$S_2D_1dE_4cE_1'C_1cA_1'C_1$	$cB\sharp$
$S_2D_1dE_4cE_1'C_1cA_2'C_1$	$cA'B\sharp$
$S_2D_1dE_4cE_2'C_1cB_1$	$b\sharp$
$S_2D_1dE_4cE_2'C_1cB_2$	$\sharp$
$S_2D_1dE_4cE_3'C_1cB_1$	$bE'\sharp$
$S_2D_1dE_4cE_3'C_1cB_2E_1'C_1$	$cA'B\sharp$
$S_2D_1dE_4cE_3'C_1cB_2E_1'C_1$	$cB\sharp$
$S_2D_1dE_4cE_3'C_1cB_2E_1'C_1$	$cBE'\sharp$
$S_2D_1dE_4cE_3'C_1cB_2E_1'C_1$	$cA'BE'\sharp$
$S_2D_1dE_4cE_4'C_1cA_1'C_1$	$cBE'\sharp$
$S_2D_1dE_4cE_4'C_1cA_2'C_1$	$cA'BE'\sharp$
$S_2D_1dE_4cE_4'C_1cA_1'C_1$	$cBE'\sharp$
$S_2D_1dE_4cE_4'C_1cA_2'C_1$	$cA'BE'\sharp$

Производим сравнение:

$dccb$	$\sharp$
$S_2D_1dE_4cE_2'C_1cB_1$	$\sharp$
$S_2D_1dE_4cE_3'C_1cB_1$	$E'\sharp$

Снова раскрываем нетерминалы:

$dccb$	$\sharp$
$S_2D_1dE_4cE_2'C_1cB_1$	$\sharp$
$S_2D_1dE_4cE_3'C_1cB_1E_1'C_1$	$cA'B\sharp$
$S_2D_1dE_4cE_3'C_1cB_1E_2'C_1$	$cB\sharp$
$S_2D_1dE_4cE_3'C_1cB_1E_3'C_1$	$cBE'\sharp$
$S_2D_1dE_4cE_3'C_1cB_1E_4'C_1$	$cA'BE'\sharp$

И, наконец, производим заключительное сравнение:

$dccb\sharp$	
$S_2D_1dE_4cE_2'C_1cB_1\sharp$	

В итоге у нас опустошился стек и одновременно мы прочитали входную строку целиком. Следовательно, эта строка является выводимой в данной грамматике. В левой части таблицы у нас образовалась цепочка символов $S_2D_1dE_4cE_2'C_1cB_1\sharp$, называемая текущим выводом. При помощи нее можно восстановить вывод входной строки. Уберем из текущего вывода терминальные символы, получим $S_2D_1E_4E_2'C_1B_1$. Методы моделирования магазинных автоматов строят левосторонний вывод исходной строки. Вывод начинается с начального символа грамматики S. Последовательно применяем продукции $S_2, D_1, E_4, E_2', C_1, B_1$ в левостороннем выводе: $S \to DE \to dE \to dcE' \to dcCB \to dccB \to dccb$.

3 Алгоритм поиска в глубину

Как было видно, при поиске в ширину таблица может значительно расти в объеме. Поиск в глубину позволяет избавиться от этого, однако алгоритм поиска содержит большее количество шагов.

Начинается поиск в глубину с того же самого состояния, что и поиск в ширину:

1

	$dccb\sharp$
	$S\sharp$

При поиске в глубину мы проходим два этапа. Первый этап состоит в раскрытии нетерминалов, можно назвать его «прямым» ходом. Выбираем первую продукцию для нетерминала S (продукции выбираются по порядку, алгоритм не делает никаких предположений о том, какую продукцию выбрать):

	$dccb\sharp$
S_1	$ABa\sharp$

Прямой ход алгоритма продолжается до тех пор, пока в правой части таблицы стек начинается с нетерминала:

	$dccb\sharp$
S_1A_1	$ECBa\sharp$

$\longrightarrow$

	$dccb\sharp$
$S_1A_1E_1$	$bABCBa\sharp$

Как справа мы встретили терминал, в данном случае b, необходимо произвести сравнение очередного символа на входе (символ d). Сравниваемые символы неравны и алгоритм переходит режим возврата. Начинается режим с попытки перебора продукций для последнего рассмотренного нетерминала. Последним мы рассматривали нетерминал E и продукцию E_1, которая не подошла. Значит необходимо попробовать следующую продукцию, E_2:

	$dccb\sharp$
$S_1A_1E_1$	$bABCBa\sharp$

$\longrightarrow$

	$dccb\sharp$
$S_1A_1E_2$	$cCBa\sharp$

Снова неудача при сравнении и берем следующую продукцию:

	$dccb\sharp$
$S_1A_1E_2$	$cCBa\sharp$

$\longrightarrow$

	$dccb\sharp$
$S_1A_1E_3$	$bABE'CBa\sharp$

и опять...:

Снова неудача при сравнении и берем следующую продукцию:

	$dccb\sharp$
$S_1A_1E_3$	$bABE'CBa\sharp$

$\longrightarrow$

	$dccb\sharp$
$S_1A_1E_4$	$cE'CBa\sharp$

Все, продукции для E закончились, значит необходимо осуществить возврат по нетерминалу. Мы убираем из стека тело последней рассмотренной продукции для E и на ее место перемещаем сам нетерминал из левой части таблицы в правую:

	$dccb\sharp$
$S_1A_1E_4$	$cE'CBa\sharp$

$\longrightarrow$

	$dccb\sharp$
S_1A_1	$ECBa\sharp$

Теперь у нас в левой части таблицы находится продукция A_1, которая, как показала работа алгоритма, не подошла. Следовательно, надо попробовать следующую продукцию, A_2:

	$dccb\sharp$
S_1A_1	$ECBa\sharp$

$\longrightarrow$

	$dccb\sharp$
S_1A_2	$ECA'Ba\sharp$

Чтобы не загромождать изложение, можно сразу увидеть, что далее рассмотрение продукций для E не приведет к успеху, мы так же безуспешно просмотрим все продукции для E и совершим возврат. Следовательно, необходимо совершить возврат по A:

	$dccb\sharp$
S_1A_1	$ECA'Ba\sharp$

$\longrightarrow$

	$dccb\sharp$
S_1	$ABa\sharp$

Снова идем вперед, и берем следующую продукцию для S:

	$dccb\sharp$
S_1	$ABa\sharp$

$\longrightarrow$

	$dccb\sharp$
S_2	$DE\sharp$

Раскрываем нетерминал D:

	$dccb\sharp$
S_2	$DE\sharp$

$\longrightarrow$

	$dccb\sharp$
S_2D_1	$dE\sharp$

Производим успешное сравнение:

	$dccb\sharp$
S_2D_1	$dE\sharp$

$\longrightarrow$

d	$ccb\sharp$
S_2D_1d	$E\sharp$

Далее без комментариев дальнейшая работа алгоритма будет такой: