СПИСОК ЛИТЕРАТУРЫ

1. OMG UML Version 2.3 [Электронный ресурс]: <http://www.omg.org/spec/UML/2.3/>, 2010.

2. *Бураков В. В.* Управление качеством программных средств. СПб: СПбГУАП, 2009. 287 с.

Сведения об авторе

Вадим Витальевич Бураков — д-р техн. наук, профессор; СПИИРАН, лаборатория информационных технологий в системном анализе и моделировании; E-mail: Burakov@eureca.ru

Рекомендована СПИИРАН

Поступила в редакцию 10.06.14 г.

УДК 681.3.062

Л. Н. ФЕДОРЧЕНКО

МЕТОД РЕГУЛЯРИЗАЦИИ ГРАММАТИК В СИСТЕМАХ ПОСТРОЕНИЯ ЯЗЫКОВЫХ ПРОЦЕССОРОВ

Описывается алгоритм регуляризации приведенных контекстно-свободных грамматик, основанный на эквивалентных преобразованиях, который совместно с алгоритмом устранения рекурсий редуцирует грамматику к единственному регулярному выражению.

Ключевые слова: *контекстно-свободная грамматика, эквивалентное преобразование грамматики.*

Регулярные множества и контекстно-свободные языки с различными ограничениями и расширениями успешно применяются в технологии построения трансляторов. При создании разного вида трансляторов языков программирования используется множество технологических средств построения анализаторов формальных языков. Как правило, эти технологические средства обеспечивают лишь проверку предъявляемых к грамматике требований и выдачу диагностических сообщений об их нарушениях. Для получения эквивалентной грамматики, удовлетворяющей условиям алгоритма анализа, существуют способы эквивалентных преобразований контекстно-свободных (КС) грамматик. Эти преобразования могут быть выполнены автоматически. Такие эквивалентные преобразования реализованы в программном средстве SynGT (Syntax Graph Transformations), разработанном в СПИИРАН.

В настоящей статье рассматривается алгоритм редукции приведенной КС-грамматики к одному регулярному выражению с помощью эквивалентных преобразований.

Определим отношение R зависимости между нетерминалами КС-грамматики следующим образом.

О п р е д е л е н и е 1. Пусть $G = (V_N, V_T, P, S)$ — КС-грамматика, где V_N — алфавит нетерминалов, V_T — алфавит терминалов, P — множество правил грамматики, $S \in V_N$ — начальный символ грамматики. Будем считать, что нетерминал $A \in V_N$ зависит от нетерминала $B \in V_N$, если существует правило вида $A \to \alpha B \beta \in P$, где $\alpha,\ \beta \in V^*,\ V = V_N \cup V_T$. Этот факт будем записывать как $(A, B) \in \mathbb{D}$, а множество всех таких пар $\mathbb{D}$ будем называть *отношением зависимости между нетерминалами* КС-грамматики. Другими словами, $\mathbb{D} \subseteq V_N \times V_N$. При $A = B$ считаем нетерминал A *самозависимым (рекурсивным).*

О п р е д е л е н и е 2. Нетерминал $A \in V_N$ назовем *абсолютно независимым*, если $\neg\exists B : (A, B) \in \mathbb{D}$.

Другими словами, абсолютно независимые нетерминалы определяются правилами, в правых частях которых нет ни одного нетерминала.

Рассмотрим схему эквивалентного преобразования КС-грамматики в одно регулярное выражение.

1. Множество всех нетерминалов V_N данной приведенной КС-грамматики разбивается на непересекающиеся подмножества (уровни) l_i, $0 \le i \le k < |N|$, в соответствии с уровнем зависимости нетерминала. Нетерминалы, правила для которых содержат только терминальные символы в правых частях, принадлежат к самому нижнему нулевому уровню l_0, т.е. такие нетерминалы $A \in V_N$, значения (регулярные множества) $R(A)$ которых содержатся в правой части соответствующего A-правила с терминальными символами.

2. Для каждого следующего уровня l_i и для всех нетерминалов этого уровня осуществляется замещение (подстановка) нетерминалов уровня l_{i-1} их значениями. Самый последний уровень l_k содержит только один элемент — начальный символ S грамматики, который замещается регулярным выражением на последнем шаге преобразования.

Все подстановки осуществляются в соответствии с иерархией на нетерминалах, определяемой отношением зависимости $\mathbb{D}$, т.е. если правая часть A_j-правила содержит вхождение нетерминала B_j, то пара (A_j, B_j) принадлежит отношению зависимости $\mathbb{D}$ согласно определению 1.

Задача состоит в том, чтобы разбить все нерекурсивные нетерминалы на непересекающие множества s_0, s_1, s_2, ... , s_m , $m \le n$, где $n = |V_N|$, которые должны обладать следующими свойствами.

1) Все нетерминалы $A \in s_0$ абсолютно независимы, т.е. регулярные выражения для таких нетерминалов *априори* определены A-правилами.

2) Нетерминалы любого множества s_i, $1 \le i \le m$, *независимы* друг от друга, т.е. для любой пары нетерминалов $(A, B) \in s_i$ выполняется условие $(A, B) \notin \mathbb{D}$.

3) Нетерминалы множества s_i, $1 \le i \le m$, *непосредственно* вычислимы по регулярным значениям нетерминалов из множеств s_j, $1 \le j \le i-1$, предыдущих уровней. Другими словами, если $A \in s_i$, то в правой части A-правила все вхождения нетерминалов замещаются регулярными значениями вычисленных к этому моменту нетерминалов из множества s_j, $1 \le j \le i-1$.

Примем, что нетерминалы из множества s_l относятся к уровню l. Очевидно, что на максимальном уровне (m) всегда располагается начальный нетерминал S грамматики, и только он, если он не встречается в правых частях правил. Это условие задается предварительно [1].

На уровне „0“ находятся нетерминалы, регулярные значения которых уже определены: их можно назвать опорными. Далее, при параллельном продвижении по уровням, „вычисляются“ регулярные значения всех других нетерминалов путем замещения в соответствии с вышеуказанным свойством 3. В результате такого процесса формируется регулярное выражение для начального нетерминала S грамматики на уровне m. Именно оно и есть искомый результат эквивалентных преобразований исходной КС-грамматики.

Лево- и/или *правосторонние* и *центральная* рекурсии в преобразуемых правилах, в случае их обнаружения, исключаются согласно приведенному в работах [2, 3] алгоритму с использованием операций итерации (бинарной или унарной).

В качестве примера будем использовать грамматику с трактовкой правил как регулярных формул с операциями объединения, конкатенации и обобщенной итерации [3]. В процессе преобразований правил грамматики могут появляться и другие регулярные операции: два вида замыкания — *рефлексивно-транзитивное*, обозначаемое „звездочкой“ Клини, и *транзитивное*, обозначаемое „плюсом“ Клини, а также обобщенная итерация, обозначаемая как „#“ (см. определение в работах [2, 3]).

Пример 1. Дана КС-грамматика $G = (V_N, V_T, P, S)$, где

$$V_T = \{'d', '.', '\backslash', 'e', '+' ; '-'\};$$

$$V_N = \{ A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9, A_{10}, A_{11}, A_{12}, A_{13}, A_{14}, A_{15}\}; S = A_{15};$$

P состоит из 15 правил.

Список правил грамматики приведен на рис. 1.

A_1: '+' ; '–'.	A_2: A_1 ; ε.	A_3: A_2 , A_{14}.	A_4: '\' ; 'e'.
A_5: A_4 , A_3.	A_6: A_{14} ; A_{11}.	A_7: A_6 , A_5.	A_8: '.' , A_{14}.
A_9: A_{14}.	A_{10}: A_9 ; ε.	A_{11}: A_{10} , A_8.	A_{12}: 'd'.
A_{13}: A_{12} ; A_{13} , A_{12}.	A_{14}: A_{13}.	A_{15}: A_{14} ; A_{11} ; A_7.	

Рис. 1

Альтернативы для каждого нетерминала A представлены в виде регулярных выражений с помощью операции объединения и разделены металингвистическом символом „ ; “, а конкатенируемые символы правых частей — символом „ , “; конец A-правила отмечен как „ . “. Таким образом, рассматривается КС-грамматика в регулярной форме, все правила которой имеют следующий вид:

<нетерминал>:<регулярное выражение>.

Строгие определения для обобщенных регулярных выражений в правилах КС-грамматик в регулярной форме приведены в работах [2, 3].

Отношение зависимости между нетерминалами $\mathbb{D}$ легко построить непосредственно по правилам грамматики в виде множества пар $\mathbb{D} \subseteq V_N \times V_N$. В результате получаем

$$\mathbb{D} = \{(A_2, A_1), (A_3, A_2), (A_3, A_{14}), (A_5, A_3), (A_5, A_4), (A_6, A_{11}), (A_6, A_{14}),$$
$$(A_7, A_5), (A_7, A_6), (A_8, A_{14}), (A_9, A_{14}), (A_{10}, A_9), (A_{11}, A_8), (A_{11}, A_{10}), (A_{13}, A_{12}),$$
$$(A_{13}, A_{13}), (A_{14}, A_{13}), (A_{15}, A_7), (A_{15}, A_{11}), (A_{15}, A_{14})\}.$$

Для наглядности отношение зависимости $\mathbb{D}$ можно представить и в виде матрицы (рис. 2). Строки и столбцы этой матрицы — нетерминальные символы грамматики. Непустой элемент матрицы, находящийся i-й строке и j-м столбце, помеченный символом T, означает, что нетерминал A_i зависит от нетерминала A_j (например, A_2 зависит от A_1, A_3 зависит от A_2 и A_{14} и т.д.). Пустые строки матрицы (например, 1, 4, 12-я) означают абсолютную, т. е. полную независимость нетерминалов A_1, A_4, A_{12} от всех других нетерминалов грамматики. Элемент $(A_{13}, A_{13})=T$ означает самозависимость нетерминала A_{13}, т. е. нетерминал A_{13} леворекурсивен. Применяя эквивалентное преобразование, получаем $A_{13} : A_{12}, (A_{12})^*. \equiv (A_{12})\#$.

Рассмотрим метод сортировки нетерминалов. Опишем в абстрактных терминах алгоритм сортировки нетерминалов КС-грамматики, исходя из построенного отношения зависимости. Цель алгоритма — расположить множество нетерминалов грамматики по уровням

таким образом, чтобы все нетерминалы одного уровня l, $1 \le l \le m$, зависели только от нетерминалов уровня $k < l$ в соответствии с отношением $\mathbb{D}$.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}
A_1															
A_2	T														
A_3		T												T	
A_4															
A_5			T	T											
A_6											T				
A_7					T	T									
A_8														T	
A_9														T	
A_{10}									T						
A_{11}								T		T					
A_{12}															
A_{13}												T	T		
A_{14}													T		
A_{15}							T				T			T	

Рис. 2

Алгоритм: *сортировка нетерминалов грамматики.*

Входные данные: $G = (V_N, V_T, P, S)$ — приведенная КС-грамматика, не содержащая рекурсивных нетерминалов; $\mathbb{D} \subseteq V_N \times V_N$ — отношение зависимости нетерминалов.

Выходные данные: $N = \{s_0, s_1, s_2, \ldots, s_m\}$, где $s_k \subseteq V_N$ — подмножество нетерминалов данной грамматики уровня k, $0 \le k \le m$.

Шаг 0. Расположение на уровне $l = 0$ всех абсолютно независимых нетерминалов:

$$l = 0;\ s_0 = \{A \mid \forall\, (A, B \in V_N)\colon (\neg\exists\, (\alpha, \beta \in V^*)\colon A \to \alpha B \beta \in P)\}.$$

Шаг 1. Построение множества нетерминалов следующего уровня:

$$l = l + 1;\ \ s_l = \{A \mid \forall\, (A \in V_N)\colon (\exists\, B \in V_N)\colon (B \in s_{l-1})\ \&\ (A, B) \in R\ \&\ (A \ne B)\}.$$

Шаг 2. Исключение из всех подмножеств s_k, $0 \le k \le l$–1, нетерминалов, содержащихся в подмножестве s_l:

```
for ∀ (A ∈ sl):
do for k from 0 to l–1
     do if  A ∈ sk  then  sk := sk \ {A} od
do;
```

Шаг 3. Определение необходимости продолжения процесса сортировки нетерминалов:

```
if sl ≠ ∅ then goto Шаг 1;
```

Шаг 4. Окончание процесса сортировки:

$$m = l\text{–}1;\ \{\text{максимальный уровень нетерминалов}\}.$$

Результат: $N = \{s_0, s_1, s_2, \ldots, s_m\}$.

Применительно к рассмотренному примеру получен следующий результат:

$$s_0 = \{A_1, A_4, A_{12}\};\ s_1 = \{A_2, A_{13}\};\ s_2 = \{A_{14}\};\ s_3 = \{A_3, A_8, A_9\};\ s_4 = \{A_5, A_{10}\};$$

$$s_5 = \{A_{11}\};\ \ s_6 = \{A_6\};\ s_7 = \{A_7\};\ s_8 = \{\, A_{15}\}.$$

Динамика вычисления уровней зависимости показана на рис. 3, где символ T означает рассматриваемый нетерминал, а символ $\overline{T}$ — удаление его с более низкого уровня отношения

зависимости. Максимальный номер уровня $m = 8$. Заметим, что условие $A \neq B$, используемое на шаге 1 алгоритма, существенно: оно предотвращает бесконечный рост уровней.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}
8															T
7							T								$\not T$
6						T	$\not T$								$\not T$
5						$\not T$	$\not T$				T				$\not T$
4					T		$\not T$			T	$\not T$				
3			T		$\not T$	$\not T$		T	T						$\not T$
2			$\not T$				$\not T$							T	
1		T			$\not T$								T		
0	T			T								T			

Рис. 3

Используя результат сортировки нетерминалов и правила исходной КС-грамматики, построим регулярные выражения для всех нетерминалов, исходя из следующих соображений.

Регулярные значения нетерминалов, представленных регулярными выражениями (см. рис. 1), вычисляются на основе априорных значений тех правил грамматики, правые части которых являются регулярными выражениями над алфавитом терминалов. Такие априорные значения всегда существуют, поскольку грамматика приведенная. Все они относятся к нетерминалам уровня „0“.

В рассматриваемом примере три нетерминала уровня „0“: A_1, A_4 и A_{12}, и соответствующие регулярные выражения для них: $R(A_1) = (\text{'+'} ; \text{'–'})$, $R(A_4) = (\text{'\'} ; \text{'}e\text{'})$ и $R(A_{12}) = (\text{'}d\text{'})$.

Задание регулярных значений для всех нетерминалов уровня „0“ обеспечивает инициализацию процесса построения регулярных значений нетерминалов уровня l, $1 \le l$, по значениям нетерминалов предыдущих уровней. При этом можно утверждать, благодаря выполненной сортировке нетерминалов, что уже существуют регулярные значения, необходимые для вычисления значений нетерминалов уровня l путем замещения (подстановки) вхождений нетерминалов в правила КС-грамматики соответствующими регулярными выражениями, полученными на более низком уровне.

Согласно матрице зависимости (см. рис. 2), от регулярного значения нетерминала A_1 зависит значение нетерминала A_2, и только оно, и, следовательно, замещая (в соответствии с рис. 1) в правиле для A_2 вхождение A_1 значением ('+' ; '–'), получаем регулярное выражение (('+' ; '–'), ε) в качестве правой части нового правила для нетерминала A_2. Аналогичные рассуждения приводят к заключению, что регулярное значение $R(A_4) = (\text{'\'} ; \text{'}e\text{'})$ должно передаваться в правую часть правила для A_5, а $R(A_{12}) = (d)$ — для A_{13}. (Для упрощения записи исключим знаки-кавычки „ ' ' “ для обобщенного терминала d.) В результате подстановок получаем регулярные выражения $R(A_5) = ((\text{'\'} ; \text{'}e\text{'}), ((\text{'+'} ; \text{'–'}), \varepsilon), (d)+)$ и $R(A_{13}) = =((d), (d)^*)$ в качестве правых частей новых правил для нетерминалов A_5 и A_{13} уровня „1“.

Далее вычисляем значения нетерминалов уровня „2“, т.е. A_{14}. После подстановки в исходном правиле для A_{14} вместо A_{13} значения $R(A_{13})$ получаем $R(A_{14}) = (((d), (d)^*))$. Подобным же образом вычисляются регулярные значения нетерминалов следующих уровней вплоть до максимального, на котором находится всегда один — начальный — нетерминал грамматики $S = A_{15}$. Его регулярное значение и есть цель преобразований исходной КС-грамматики.

Итак, регуляризованная КС-грамматика G_1, эквивалентная исходной, — есть КС-грамматика в регулярной форме:

$$R(A_{15}) = R(S)\ ((d), (d)^*)\ ;$$
$$(((((d), (d)^*))\ ;\ \varepsilon), \text{'.'}, ((d), (d)^*))\ ;$$
$$(((d), (d)^*), ((((d), (d)^*))\ ;\ \varepsilon), \text{'.'}, ((d), (d)^*))\ ,$$

$$((\text{'\textbackslash'} ; \text{'}e\text{'}), (((((\text{'+'} ; \text{'–'}), \varepsilon), \varepsilon)), ((d), (d)^*)))) \equiv$$
$$\equiv d^+ ; d^*, \text{'.'}, d^+ ; (d^+; d^*, \text{'.'}, d^+), (\text{'\textbackslash'} ; \text{'}e\text{'}), [\text{'+'} ; \text{'–'}], d^+ \equiv$$
$$\equiv (d^+; d^*, \text{'.'}, d^+), [(\text{'\textbackslash'} ; \text{'}e\text{'}), [\text{'+'} ; \text{'–'}], d+].$$

Предложенный алгоритм получения регулярного выражения, эквивалентного приведенной КС-грамматике без ограничений на рекурсии, может быть применен при построении языковых процессоров. Используется метод сортировки нетерминалов КС-грамматики по отношению зависимости [4], реализованный в системе SynGT [2]. Приведенный алгоритм регуляризации позволяет обнаружить любые рекурсии (левая/правая/вложенная) по ходу распределения нетерминалов по уровням. Благодаря применению алгоритма удаления таких рекурсий [2, 3] снимаются ограничения на рекурсивности в исходной грамматике.

СПИСОК ЛИТЕРАТУРЫ

1. Handbook of Formal Languages / Eds.: *G. Rozenberg, A. Salomaa.* Berlin, Heidelberg, New York: Springer-Verlag, 1997. Vol. 2. 527 p.

2. *Fedorchenko L.* Regularization of Context-Free Grammars. Saarbrucken: LAP LAMBERT Academic Publishing, 2011.

3. *Федорченко Л. Н.* О регуляризации контекстно-свободных грамматик // Изв. вузов. Приборостроение. 2006. Т. 49, № 11. С. 50—54.

4. *Мартыненко Б. К.* Регулярные языки и КС-грамматики // Компьютерные инструменты в образовании. 2012. № 1. С. 14—20.

Сведения об авторе

Людмила Николаевна Федорченко — канд. техн. наук; СПИИРАН, ст. научный сотрудник;
E-mail: lnf@iias.spb.su

Рекомендована СПИИРАН

Поступила в редакцию
10.06.14 г.