Общие ненаправленные алгоритмы анализа

Метод Ангера. CYK. НФХ.

В. С. Полозов

Кафедра системого программирования СПбГУ

Теория автоматов и формальных языков

Метод Ангера без ε-правил и циклов

Автор: S.H. Unger
Рассмотрим грамматики без ε-правил и циклов.
Принцип. Возьмём правило:

$$S \to ABC|DE|F$$

и вход $pqrs$.
Для каждой правой части всеми способами разобьем вход на соответствующее количество непустых подстрок - «корзин».
Проверим. Можно использовать DFS или BFS. В авторской статье используется DFS.

Метод Ангера без ε-правил и циклов

Автор: S.H. Unger
Рассмотрим грамматики без ε-правил и циклов.
Принцип. Возьмём правило:

$$S \to ABC|DE|F$$

и вход $pqrs$.
Для каждой правой части всеми способами разобьем вход на соответствующее количество непустых подстрок - «корзин».
Проверим. Можно использовать DFS или BFS. В авторской статье используется DFS.

Пример

Рассмотрим грамматику:

$$\begin{aligned} Expr &\rightarrow Expr + Term \mid Term \\ Term &\rightarrow Term \times Factor \mid Factor \\ Factor &\rightarrow (Expr) \mid i \end{aligned}$$

И вход

$$(i + i) \times i$$

Метод Ангера с ε-правилами

Рассмотрим правила:

$$\begin{aligned} S &\to A\ B\ C \\ B &\to S\ A \end{aligned}$$

и вход pqr.

Пример

Рассмотрим правила:

$$\begin{aligned} S &\to L\ S\ D \mid \varepsilon \\ L &\to \varepsilon \\ D &\to d \end{aligned}$$

и пару входов: d и dd.

Метод CYK

Назван по именам создателей: J. Cocke, D.H. Younger, и T. Kasami.

Две фазы

1. Строится таблица соответствия выводам подстрок входного предложения нетерминалам
2. Используя построенную таблицу строим все возможные выводы

Пример

$$
\begin{aligned}
\mathit{Numbers} &\rightarrow \mathit{Integer} \mid \mathit{Real} \\
\mathit{Integer} &\rightarrow \mathit{Digit} \mid \mathit{Integer}\ \mathit{Digit} \\
\mathit{Real} &\rightarrow \mathit{Integer}\ \mathit{Fraction}\ \mathit{Scale} \\
\mathit{Fraction} &\rightarrow .\ \mathit{Integer} \\
\mathit{Scale} &\rightarrow e\ \mathit{Sign}\ \mathit{Integer} \mid \mathit{Empty} \\
\mathit{Digit} &\rightarrow 0|1|2|3|4|5|6|7|8|9 \\
\mathit{Empty} &\rightarrow \varepsilon \\
\mathit{Sign} &\rightarrow +|-
\end{aligned}
$$

и вход – 23.4e+5

- Что делать с ε?

Рассмотрим вход – 43.1

Пример

$$
\begin{aligned}
\mathit{Numbers} &\rightarrow \mathit{Integer} \mid \mathit{Real} \\
\mathit{Integer} &\rightarrow \mathit{Digit} \mid \mathit{Integer}\ \mathit{Digit} \\
\mathit{Real} &\rightarrow \mathit{Integer}\ \mathit{Fraction}\ \mathit{Scale} \\
\mathit{Fraction} &\rightarrow .\ \mathit{Integer} \\
\mathit{Scale} &\rightarrow e\ \mathit{Sign}\ \mathit{Integer} \mid \mathit{Empty} \\
\mathit{Digit} &\rightarrow 0|1|2|3|4|5|6|7|8|9 \\
\mathit{Empty} &\rightarrow \varepsilon \\
\mathit{Sign} &\rightarrow +|-
\end{aligned}
$$

и вход – 23.4e+5

- Что делать с ε?

Рассмотрим вход – 43.1

Обозначения

Введем обозначения:

- $R_\varepsilon = \{Empty,\ Scale\}$ - множество нетерминалов A таких, что $A \to \varepsilon$
- $s_{i,l}$ - подстрока $z_i z_{i+1} ... z_{i+l-1}$ строки $z = z_1 z_2 ... z_n$ начиная с позиции i и длиной l.
- $R_{s_{i,l}}$ - множество нетерминалов из которых выводима строка $s_{i,l}$.
- Для строк длины 0 положим: $s_{i,0} = \varepsilon$ и $R_{s_{i,0}} = R_\varepsilon$

Замечание

Основная переборная сложность CYK связана с особенностями грамматики:

- Цепные правила: $A \to B \to ... \to C \overset{*}{\to} s_{i,l}$
- ε-правила

Обозначения

Введем обозначения:

- $R_\varepsilon = \{Empty,\ Scale\}$ - множество нетерминалов A таких, что $A \to \varepsilon$
- $s_{i,l}$ - подстрока $z_i z_{i+1} ... z_{i+l-1}$ строки $z = z_1 z_2 ... z_n$ начиная с позиции i и длиной l.
- $R_{s_{i,l}}$ - множество нетерминалов из которых выводима строка $s_{i,l}$.
- Для строк длины 0 положим: $s_{i,0} = \varepsilon$ и $R_{s_{i,0}} = R_\varepsilon$

Замечание

Основная переборная сложность CYK связана с особенностями грамматики:

- Цепные правила: $A \to B \to ... \to C \overset{*}{\to} s_{i,l}$
- ε-правила

CYK: НФХ

Определение

Нормальная форма Хомского (CNF): все правила имеют вид $A \to a$ или $A \to BC$.

Получаем:

- $R_\varepsilon = \emptyset$
- $R_{s_{i,1}}$ вычисляется непосредственно из правил $A \to a$.
- Для правила $A \to BC$
  из B должно выводиться $s_{i,k}$, т.е. $B \in R_{s_{i,k}}$,
  из C должно выводиться $s_{i+k,l-k}$, т.е. $C \in R_{s_{i+k,l-k}}$

CYK: НФХ

Т.е. можно организовать *таблицу распознавания* (рис.).

$R_{s_{1,n}}$				
$R_{s_{1,n-1}}$	$R_{s_{2,n-1}}$			
...	...	...		
$R_{s_{1,2}}$	...	...	$R_{s_{n-1,2}}$	
$R_{s_{1,1}}$	$R_{s_{2,1}}$	...	$R_{s_{n-1,1}}$	$R_{s_{n,1}}$
s_1	s_2	...	s_{n-1}	s_n

$R_{i,j} = ?$

Соответственно, можно изменить порядок обхода:

- off-line order
- on-line order

Сложность?

CYK: НФХ

Т.е. можно организовать *таблицу распознавания* (рис.).

$R_{s_{1,n}}$				
$R_{s_{1,n-1}}$	$R_{s_{2,n-1}}$			
...	...	...		
$R_{s_{1,2}}$	...	...	$R_{s_{n-1,2}}$	
$R_{s_{1,1}}$	$R_{s_{2,1}}$	...	$R_{s_{n-1,1}}$	$R_{s_{n,1}}$
s_1	s_2	...	s_{n-1}	s_n

$R_{i,j} = ?$

Соответственно, можно изменить порядок обхода:

- off-line order
- on-line order

Сложность?

CYK: НФХ

Т.е. можно организовать *таблицу распознавания* (рис.).

$R_{s_{1,n}}$				
$R_{s_{1,n-1}}$	$R_{s_{2,n-1}}$			
...	...	...		
$R_{s_{1,2}}$	...	...	$R_{s_{n-1,2}}$	
$R_{s_{1,1}}$	$R_{s_{2,1}}$	...	$R_{s_{n-1,1}}$	$R_{s_{n,1}}$
s_1	s_2	...	s_{n-1}	s_n

$R_{i,j} = ?$

Соответственно, можно изменить порядок обхода:

- off-line order
- on-line order

Сложность?

CYK: НФХ

Т.е. можно организовать *таблицу распознавания* (рис.).

$R_{s_{1,n}}$				
$R_{s_{1,n-1}}$	$R_{s_{2,n-1}}$			
...	...	...		
$R_{s_{1,2}}$	...	...	$R_{s_{n-1,2}}$	
$R_{s_{1,1}}$	$R_{s_{2,1}}$	...	$R_{s_{n-1,1}}$	$R_{s_{n,1}}$
s_1	s_2	...	s_{n-1}	s_n

$R_{i,j} = ?$

Соответственно, можно изменить порядок обхода:

- off-line order
- on-line order

Сложность?

Преобразование к НФХ

Преобразование к НФХ:

- удаление ε-правил
- удаление цепных правил $(A \to B)$
- удаление непродуктивных нетерминалов
- удаление недостижимых нетерминалов
- оставшаяся грамматика преобразуется добавлением правил, пока все правила не будут вида $A \to a$ или $A \to BC$.

Удаление ε-правил

Грамматика $G = (V_T, V_N, P, S)$

1. Для правила $A \to \varepsilon$
   - Правила вида $B \to \alpha A \beta$ заменим на $B \to \alpha A' \beta$ и $B \to \alpha \beta$
   - Повторить, в т.ч. и для новых правил
2. Повторить для всех ε-правил
3. Добавим нетерминал A' для всех нетерминалов A:
   $A' \to \alpha$, если $A \to \alpha \in P$ и $\alpha \neq \varepsilon$

Удаление цепных правил

Грамматика $G = (V_T, V_N, P, S)$

1. Для правила $A \to B$, где
   $B \to \alpha_1|\alpha_2|...|\alpha_n$
   $A \to \alpha_1|\alpha_2|...|\alpha_n$
2. Повторить, кроме правил $A \to A$

Удаление непродуктивных нетерминалов

Идея: непродуктивными являются нетерминалы, которые не являются продуктивными.

Удаление недостижимых нетерминалов

Идея: недостижимыми являются нетерминалы, которые не являются достижимыми.

Замечание

Удаление недостижимых нетерминалов не добавляет непродуктивных нетерминалов. (Обратное неверно)

Преобразование к CNF

Грамматика $G = (V_T, V_N, P, S)$, без ε-правил, цепных правил, все нетерминалы достижимы и продуктивны.
Осталось два типа правил:

1. $A \to a$ – уже в хорошей форме
2. $A \to X_1 X_2 ... X_m, m \geq 2$
   - Для каждого терминала вводим новый нетерминал $T_b \to b$
   - Заменить вхождения
   - Повторить
3. $A \to X_1 X_2 ... X_m, m \geq 3, X_i \in V_N$
   - Заменить правило на
     $A \to A_1 X_3 ... X_m$
     $A_1 \to X_1 X_2$
   - Повторить

Пример

$$
\begin{aligned}
\mathit{Numbers} &\rightarrow 0|1|2|3|4|5|6|7|8|9 \\
\mathit{Numbers} &\rightarrow \mathit{Integer}\ \mathit{Digit} \\
\mathit{Numbers} &\rightarrow \mathit{N1}\ \mathit{Scale}' \mid \mathit{Integer}\ \mathit{Fraction} \\
\mathit{N1} &\rightarrow \mathit{Integer}\ \mathit{Fraction} \\
\mathit{Integer} &\rightarrow 0|1|2|3|4|5|6|7|8|9 \\
\mathit{Integer} &\rightarrow \mathit{Integer}\ \mathit{Digit} \\
\mathit{Fraction} &\rightarrow \mathit{T1}\ \mathit{Integer} \\
\mathit{T1} &\rightarrow . \\
\mathit{Scale}' &\rightarrow \mathit{N2}\ \mathit{Integer} \\
\mathit{N2} &\rightarrow \mathit{T2}\ \mathit{Sign} \\
\mathit{T2} &\rightarrow e \\
\mathit{Digit} &\rightarrow 0|1|2|3|4|5|6|7|8|9 \\
\mathit{Sign} &\rightarrow +|-
\end{aligned}
$$

Откат преобразования к CNF

При преобразовании к CNF могут потеряться нетерминалы:

- при удаление ε-правил (В примере: Scale, Real, Empty)
- удаление цепных правил (В примере: Real)

Дополним грамматику правилом:

$$Real \rightarrow N1\ Scale' \mid Integer\ \ Fraction$$

Откат преобразования к CNF

При преобразовании к CNF могут потеряться нетерминалы:

- при удаление ε-правил (В примере: Scale, Real, Empty)
- удаление цепных правил (В примере: Real)

Дополним грамматику правилом:

$$Real \rightarrow N1\ Scale' \mid Integer\ \ Fraction$$

Вопросы