Корректность алгоритма разбора LC(k)-грамматик

Шпилевой Денис Б05-153
Руководитель - Павел Ахтямов

Цель исследований

1. Формализация LC-разбора.

2. Как часть глобальной задачи - исследования возможности определения класса языков по некоторым данным, которые разбираются за линейное время.

План презентации

- Формальные языки, иерархия Хомского
- КС-грамматики
- Разбор
- LL(k)-грамматики
- LC(k)-грамматики
- План работы на семестр
- Источники

Формальные языки. Иерархия Хомского

Опр. Формальный язык - множество конечных слов над конечным алфавитом.

Иерархия языков Хомского:

- Регулярные (праволинейные)	$A \to aB$	$A, B \in N, a \in \Sigma$
- Контекстно-свободные	$A \to X$	$A \in N, X \in (N \cup \Sigma)^*$
- Контекстно-зависимые	$XAY \to XZY$	$A \in N, X, Y, Z \in (N \cup \Sigma)^*$
- Порождающие	$X \to Y$	$X, Y \in (N \cup \Sigma)^*$

КС-грамматики. Парсинг

Опр. G = (N, Σ, P, S) - КС-грамматика, если все ее правила имеют вид:

$$A \to X, \text{ где } A \in N, X \in (N \cup \Sigma)^*$$

Парсинг (синтаксический анализ) - процесс сопоставления линейной слова формального языка с его формальной грамматикой. Результатом обычно является дерево разбора.

Алгоритм разбора - алгоритм, определенный на множестве слов некоторого языка, и биективно сопоставляющий каждому слову его разбор (последовательность примененных правил).

Алгоритм Эрли (1968)

Алгоритм Эрли - алгоритм синтаксического анализа цепочки по контекстно-свободной грамматике, основанный на методе динамического программирования.

Разбирает все КС-грамматики, но асимптотика - от квадрата до куба в зависимости от коэффициента неоднозначности грамматики.

Рассмотрим примеры алгоритмов, разбирающих не на все КС-грамматики, но выполняющих разбор за линейное время от длины слова.

LL(k)-грамматики (1968)

LL(k)-разбор - левый разбор по известным k символам после головки.

Опр. КС-грамматику G = (N, Σ, P, S) называют LL(k)-грамматикой, если из существования двух левых выводов:

- $S \Rightarrow wAa \Rightarrow_1 wba \Rightarrow wx$
- $S \Rightarrow wAa \Rightarrow_1 wca \Rightarrow wy$ $\quad w, x, y \in \Sigma^*, A \in N, a, b, c \in (N \cup \Sigma)^*$
- при этом $FIRST_k(x) = FIRST_k(y)$, тогда b = c.

LL(k)-грамматики (1968)

Критерий LL(k)-грамматики. КС-грамматика G = (N, Σ, P, S) является LL(k)-грамматикой тогда и только тогда, когда для двух различных правил A => x, A => y пересечение $FIRST_k(xa)$ и $FIRST_k(ya)$ пусто при всех таких wAa, что S => wAa, где w ϵ Σ*, A ϵ N, x, y ϵ (N ∪ Σ)*.

Цель критерия - показать, что для определения того, какое правило необходимо применить, не нужно помнить всю цепочку w до рассматриваемых k символов после головки.

LL(k)-грамматики (1968)

Алгоритм разбора LL(k)-грамматик. Построение управляющей таблицы разбора LL(k)-грамматик.

Вход: LL(k)-грамматика G = (N, Σ, P, S) и множество её LL(k)-таблиц Г.

Выход: Корректная управляющая таблица разбора M.

Метод: Управляющая таблица M определяется (Г ∪ Σ ∪ {$}) x Σ*k так:

1. Если A -> $x_0B_1x_1B_2x_2\ldots B_mx_m$ - правило с номером i, и $T_{A,L}$ є Г, то для всех u, для которых $T_{A,L}(u)$ = (A -> $x_0B_1x_1B_2x_2\ldots B_mx_m$, <Y1, … Ym>) полагаем, что $M(T_{A,L}, u) = (x_0 T_{B_1,Y_1} x_1 T_{B_2,Y_2} x_2 \ldots T_{B_m,Y_m} x_m, i)$.
2. M(a, av) = **выброс**, для всех v є Σ*(k - 1).
3. M($, e) = **допуск.**
4. В остальных случаях M(X, u) = **ошибка.**
5. $T_{\$, e}$ - начальная таблица.

LC(k)-грамматики (1970)

LC(k)-разбор - левый разбор по известным k символам после головки и известен вывод из левого участка.

Опр. КС-грамматику G = (N, Σ, P, S) называют LC(k)-грамматикой, если она удовлетворяет таким условиям:
Пусть S =>* wAx, w є Σ*, A є N, x, y є (N ∪ Σ)*. Тогда для каждой цепочки v є Σ* и вывода A =>* By существует не более одного такого правила B => a, что:

1. а) Если a = Cb, где a, b є (N ∪ Σ)*, C є N, то первые k букв u є $FIRST_k$(byx)

   б) Если C = A, то u не є $FIRST_k$(x).

2. Если a начинается терминалом, то u є $FIRST_k$(ayx)

LC(k)-грамматики

Алгоритм разбора LC(1)-грамматик. Построение управляющей таблицы разбора LC(1)-грамматик.

Вход: LC(1)-грамматика G = (N, Σ, P, S).

Выход: Корректная управляющая таблица разбора T.

1. Пусть B -> x - правило с номером i.

   а) Если x = Cy, где C - нетерминал, то T([A, C], a) = (y[A, B], i) для всех A ϵ N, x ϵ $FIRST_1$(yde), таких, что S => wAe и A => Bd. x, y, d, e ϵ (N ∪ Σ)*, a ϵ N.

   б) Если x не начинается нетерминалом, то T(A, a) = (x[A, B], i) для всех A ϵ N, x ϵ $FIRST_k$(xde), таких, что S => wAe и A => Bd. x, d, e ϵ (N ∪ Σ)*, a ϵ N.

2. T([A, A], a) = (e, e), для всех A ϵ N, a ϵ $FIRST_k$(e), таких, что S => wAe.
3. T(a, av) = **выброс**, для всех a ϵ Σ.
4. T($, e) = **допуск.**
5. Иначе T(X, a) = **ошибка.**

План работы на семестр

1. Описать алгоритм разбора LC(k)-грамматик.
2. Ввести эквивалентный критерий.
3. Доказать корректность алгоритма.
4. Доказать линейную сложность алгоритма.

Источники и литература

- Ахо А. Ульман Дж., Теория синтаксического анализа, перевода и компиляции, том 1, Синтаксический анализ.
- 11th Annual Symposium on Switching and Automata Theory (swat 1970), deterministic left corner parsing, pages 139-152.