ՀԱՅԿԱԿԱՆ ՍՍՀ ԳԻՏՈՒԹՅՈՒՆՆԵՐԻ ԱԿԱԴԵՄԻԱՅԻ ԶԵԿՈՒՅՑՆԵՐ
ДОКЛАДЫ АКАДЕМИИ НАУК АРМЯНСКОЙ ССР

LXIX 1979 1

УДК 681.1422

МАТЕМАТИКА

А. А. Ордян, Г. В. Джулакян

Алгоритм синтаксического анализа для индексных языков

(Представлено чл.-корр. АН Армянской ССР Р. Р. Варшамовым 3/IV 1979)

Для описания синтаксиса языков программирования широко используются формальные грамматики, введенные Хомским ([1]). Для класса контекстно-свободных (КС) формальных грамматик разработаны достаточно эффективные алгоритмы синтаксического анализа, успешно применяемые в современных синтаксически-ориентированных трансляторах. Однако с помощью КС-грамматик невозможно полностью описать синтаксис языков программирования высокого уровня (таких как АЛГОЛ-60, ФОРТРАН, ПЛ-1 и т. д.), так как в синтаксических конструкциях последних существенно используются так называемые контекстные условия.

Известно ([2]), что контекстные условия можно моделировать (описать) с помощью контекстно-чувствительных формальных грамматик, класс которых полностью включает в себя класс КС-грамматик.

Настоящая работа посвящена описанию нового алгоритма синтаксического анализа для индексных языков, порождаемых индексными грамматиками ([3]). Класс индексных грамматик (языков) шире класса КС-грамматик (языков), но является собственным подклассом класса контекстно-чувствительных грамматик (языков).

Разработанный алгоритм использует автомат нового типа — D-автомат, переходы которого осуществляются с помощью некоторых бинарных отношений — отношений предшествования.

Определение некоторых понятий, не приведенных в этой работе, можно найти в ([3–5]).

Индексной грамматикой называется пятерка $G=(N, T, F, P, S)$, где

N — конечное множество терминальных символов;

T — конечное непустое множество нетерминальных символов;

F — конечное множество символов, называемых индексами или посредниками;

P—конечное множество упорядоченных пар вида (X,χ), где либо $X\in N$, $\chi\in(NF^*\cup T)^*$, либо $X\in NF$, $\chi\in(N\cup T)^*$.

Пара $(X,\chi)\in P$ называется правилом грамматики G и обозначается $X\to\chi$, а P называется множеством правил грамматики G; $S\in N$—выделенный (начальный) символ грамматики G.

В этом определении предполагается, что $N\cap T\cap F=\varnothing$ и для каждого индекса $f\in F$ имеется хотя бы одно правило вида $Af\to X_1\ldots X_k$.

Пусть α, $\beta\in(NF^*\cup T)^*$. Говорят, что в G из α непосредственно порождается β, $\alpha\Rightarrow\beta$, если либо

1. $\alpha=\gamma A\xi\delta$, $(A\to X_1\eta_1\ldots X_k\eta_k)\in P$,
$\beta=\gamma X_1\theta_1\ldots X_k\theta_k\delta$, где $\theta_i=\eta_i\xi$ для $X_i\in N$, $\theta_i=\varepsilon$ для $X_i\in T$, $1\leqslant i\leqslant k$, либо
2. $\alpha=\gamma Af\xi\delta$, $(Af\to X_1X_2\ldots X_k)\in P$,
$\beta=\gamma X_1\theta_1\ldots X_k\theta_k\delta$, где $\theta_i=\xi$ для $X_i\in N$, $\theta_i=\varepsilon$ для $X_i\in T$, $1\leqslant i\leqslant k$.

В этих обозначениях ε—пустое слово, $A\in N$, $f\in F$, $\gamma,\delta\in(NF^*\cup T)^*$, $\xi,\eta_i,\theta_i\in F^*$.

Пусть α, β, $\gamma\in(NF^*\cup T)^*$.

Рекурсивно определяется n-ая степень ($n\geqslant 0$) отношения $\Rightarrow$, а именно

1. $\alpha\Rightarrow^0\alpha$;
2. $\alpha\Rightarrow^{n+1}\beta$, если $\alpha\Rightarrow^n\gamma$, $\gamma\Rightarrow\beta$ для некоторого γ.

Транзитивное замыкание отношения $\Rightarrow$ обозначается через $\Rightarrow^+$, а транзитивное и рефлексивное замыкание—через $\Rightarrow^*$. Ясно, что $\alpha\Rightarrow^*\beta$ тогда и только тогда, когда $\alpha\Rightarrow^i\beta$ для некоторого $i\geqslant 0$, и $\alpha\Rightarrow^+\beta$ тогда и только тогда, когда $\alpha\Rightarrow^i\beta$ для некоторого $i\geqslant 1$.

Говорят, что β выводится из α, если $\alpha\Rightarrow^*\beta$. Если $\alpha\in(NF^*\cup T)^*$ и $S\Rightarrow^*\alpha$, то α называется сентенциальной формой.

Слово из символов вида $A\xi$, где $A\in N$, $\xi\in F^+$, называется индексированным нетерминалом. Пусть G—индексная грамматика. Множество $L(G)=\{w\mid w\in T^*,\ S\Rightarrow^+w\}$ называется индексным языком.

Для языка $L(G)$ задача синтаксического анализа заключается в следующем: дать алгоритм, определяющий для любого $w\in T^*$ истинность логического выражения $w\in L(G)$. В случае, если действительно $w\in L(G)$, алгоритм дает последовательность выполнения правил, с помощью которых порождается w в грамматике G. Для однозначных грамматик, а именно такие мы рассматриваем, выходом алгоритма синтаксического анализа при работе над словом w является последовательность $\alpha_n,\alpha_{n-1},\ldots,\alpha_1$, где $\alpha_n=w$, $\alpha_1=S$. Переход от α_{i+1} к α_i ($1\leqslant i\leqslant n-1$) называется редукцией. Отношение редукции обозначим через $\to_R$.

Из определения непосредственного вывода для индексных грамматик следует:

$\beta\to_R\alpha$, если либо

1. $\beta=\gamma X_1\theta_1\ldots X_k\theta_k\delta$, где $\theta_i=\eta_i\xi$ для $X_i\in N$, $\theta_i=\varepsilon$ для $X_i\in T$

$\alpha=\gamma A\xi\delta$ и имеется правило $A\to X_1\eta_1\dots X_k\eta_k$, либо

2. $\beta=\gamma X_1\theta_1\dots X_k\Theta_k\delta$, где $\theta_i=\xi$ для $X_i\in N$, $\Theta_i=\varepsilon$ для $X_i\in T$, $\alpha=\gamma Af\xi\delta$ и имеется правило $Af\to X_1\dots X_k$.

В частности, если $\beta=aB\eta_1\xi C\theta\xi b$ и имеется правило $A\to aB\eta C\theta b$, то $\beta\Longrightarrow_R\alpha$, где $\alpha=A\xi$.

Если $\beta=aB\xi C\xi$ и имеется правило $Af\to aBC$, то $\beta\Longrightarrow_R\alpha$ где $\alpha=Af\xi$.

Пусть Z—некоторая цепочка. X назовем префиксом (суффиксом) Z, если $Z=XY$ ($Z=YX$), где Y некоторая (возможно пустая) цепочка.

Пусть дана грамматика $G=(N, T, F, P, S)$[1].

В последующих определениях предполагается, что $x, y\in(NF^*\cup T)$, $f\in F$, $\alpha, \beta, \gamma, \gamma_1\in(NF^*\cup T)^*$, $\xi, \xi'\in F^*$, $A, B, C\in N\cup NF$, $Z\in NF^*$ $\beta', \gamma'\in(F^*\cup NF^*\cup T)^*$.

Определим следующие бинарные отношения над множеством $T\cup NF^*\cup F$:

1. $X\doteq Y$, если существует некоторое правило $(A\to\alpha XY\beta)\in P$,

2. $X<Y$, если существуют некоторое правило $(A\to\alpha ZB\beta)\in P$ и вывод $B\Rightarrow^+ Y\gamma$, где X—любой префикс Z.

3. $X>Y$, если существуют некоторое правило $(A\to\alpha B\xi C\beta)\in P$ и выводы $B\Rightarrow^+\gamma Z$ и $C\Rightarrow^+ Y\gamma_1$, где X—любой префикс Z.

4. $X\boxdot f$, если существует вывод $S\Rightarrow^+\alpha Xf\beta'$ такой, что в сентенциальной форме $\alpha Xf\beta'$ элементы X и f не порождаются одновременно с помощью некоторого правила вида $(B\to\gamma Xf\gamma')\in P$.

Отношения 1—4 назовем отношениями предшествования. Пусть $\alpha, \beta\in(NF^*\cup T)^*$, $X_1, X_2\in(NF^*\cup T)$, $\xi_1, \xi_2\in F^*$.

Теорема 1. *Если $\alpha=\alpha X_1\xi_1X_2\xi_2\beta$—некоторая сентенциальная форма, то между X_1 и X_2 имеет место хотя бы одно из отношений 1—3. Если $X_1=X_2$—единственное отношение и $X_1, X_2\in T$ или $X_1, X_2\in NF^*$ одновременно, то $\xi_1=\xi_2$.*

Грамматику $G=(N, T, F, P, S)$ назовем приведенной, если для G выполняются следующие условия:

1. Если $(A\to\alpha)\in P$, то $\alpha\neq\varepsilon$, где ε—пустое слово;

2. Если $\gamma_1A\gamma_2$ сентенциальная форма, то в G существует вывод $A\Rightarrow^+ w$, где $w\in T^+$;

3. Для всех $A\in NF\cup N$ не имеет места $A\to^+A$.

Приведенную грамматику назовем грамматикой предшествования, если выполняются следующие условия:

1. Правые части всех правил различны;

2. $|\lessdot|\cap|\gtrdot|=\varnothing$;

3. Если $(|\lessdot|\cap|\doteq|)\cup(\{\doteq|\cap|\gtrdot\})=|(X_i, Y_i)|$, где $i=1, 2\dots$, $X_i, Y_i\in NF^*$, то для всех сентенциальных форм $\rho_i=\alpha X_i\xi_i, Y_i\xi_i\beta$ имеет место $\xi_i, \xi_i\neq\varepsilon$. Если $\xi_{i_1}=\xi_{i_2}$, то в ρ_i элементы X_i и Y_i появились

---

[1] Здесь и далее, если нет специальной оговорки, под грамматикой будем понимать индексную грамматику.

одновременно применением некоторого правила вида $A \to \alpha_i X_i Y_i \beta_i$;

4) если $X \lfloor \doteq \rfloor f$, то P не содержит правила вида $A \to \alpha X f \beta$, но содержит хотя бы одно правило вида $B \to \alpha_1 X \beta_1$.

Пусть для грамматики предшествования G все правила вывода пронумерованы натуральными числами, образующими множество Q.

Ниже приводится определение D—автомата, реализующего синтаксический анализ цепочек, порождаемых грамматикой предшествования.

D—автоматом называется пятерка $D=(V, R, Q, L_H, L_k)$, где

$V=N \cup T \cup F$—множество входных символов автомата;

R—отображение множества $E=(\perp(V \cup \{\Diamond\})^* \times F^* \times V^* \perp \times (Q \Diamond \Diamond)^*)$ на множество $E \cup \{ОШИБКА\}$;

$L_H=|\perp, \varepsilon, w \perp, \varepsilon|$—начальная конфигурация автомата, где w—анализируемая цепочка;

$L_k=|\perp, \varepsilon, S \perp, \pi|$—заключительная конфигурация автомата, где $\pi \in (Q \Diamond)^*$—слово, указывающее последовательность редукций.

Автомат останавливается либо при достижении состояния ОШИБКА, либо на конфигурации L_k. В последнем случае $w \in L(G)$ и выводится в G с помощью правил $P_1, P_2, \ldots P_n$, где $P_1 \Diamond P_2 \Diamond \ldots P_n \Diamond = \pi$. Вспомогательные символы $\perp$ и $\Diamond$ не принадлежат множеству V.

Включение $(a, b) \in R$ обозначим $a \vdash_R b$.

Пусть $H_{00}=|\perp X_1 \ldots X_r, \varepsilon, X_{r+1} \ldots X_n \perp, \delta|$,

$H_{01}=|\perp X_1 \ldots X_r X_{r+1}, \varepsilon, X_{r+2} \ldots X_n \perp, \delta|$,

$H_{02}=|\perp X_1 \ldots X_r \Diamond, Y, X_{r+2} \ldots X_n \perp, \delta|$,

$H_{03}=|\perp \beta, \varepsilon, \beta_1 A X_{r+1} \ldots X_n \perp, q \Diamond \delta|$,

$H_{10}=|\perp X_1 \ldots X_r \Diamond \ldots X_r, Y, X_{r+1} \ldots X_n \perp, \delta|$,

$H_{11}=|\perp X_1 \ldots X_r \Diamond \ldots X_r X_{r+1}, Y, X_{r+2} \ldots X_n \perp, \delta|$,

$H_{12}=|\perp X_1 \ldots X_r \Diamond \ldots X_r X_{r+1}, Y', X_{r+3} \ldots X_n \perp, \delta|$,

$H_{13}=|\perp X_1 \ldots X_r Z_r \ldots X_r Z_r X_{r+1} \Diamond, Y'', X_{r+3} \ldots X_n \perp, \delta|$,

$H_{14}=|\perp X_1 \ldots X_r Z_r \ldots X_r Z_r X_{r+1}, \varepsilon, X_{r+2} \ldots X_n \perp, \delta|$.

Пусть для всех $X \in N F^* \cup T$ имеет место $\perp \lessdot X$ и $X \gtrdot \perp$.

Отображение R определяется следующим образом.

I. $H_{00} \vdash_R H_0$, причем

$H_0=H_{01}$, если $(X_r \lessdot X_{r+1}) \vee (X_r \doteq X_{r+1})$;

$H_0=H_{02}$, если $(X_r \lfloor \doteq \rfloor X_{r+1})$, где $Y=f_1 \ldots f_r$—слово максимальной длины из F^+, следующее за X_r, при этом $f_1=X_{r+1}$;

$H_0=H_{03}$, если $(X_r \gtrdot X_{r+1}) \wedge (X_k \lessdot X_{k+1}) \wedge (X_j \doteq X_{j+1})$, где $k+1 \leqslant j \leqslant r-1$ и $(A \to X_{k+1} \ldots X_r) \in P$— правило с номером q, β—терминальный префикс максимальной длины цепочки $X_1 \ldots X_k$;

H_0—ОШИБКА, если либо X_r и X_{r+1} не связаны никакими отношениями предшествования, либо нет правила с правой частью $X_{k+1} \ldots X_r$.

II. $H_{10} \vdash_R H_1$, причем

$H_1=H_{11}$, если $(X_r \doteq X_{r+1}) \wedge (X_{r+1} \in T)$;

$H_1 = H_{12}$, если $(X_r \doteq X_{r+1}) \wedge (X_{r+1} | \dot{-} | X_{r+2}) \wedge (Y \doteq Y')$, где $Y' = f_1 \ldots f_e$ —слово максимальной длины из F^+, следующее за X_{r+1}, при этом $f_1 = X_{r+2}$;

$H_1 = H_{13}$, если $(X_r \doteq X_{r+1}) \wedge (X_r \lessdot X_{r+1}) \wedge (X_{r+1} | \dot{-} | X_{r+2}) \wedge (Y \neq Y'')$, где $Z_i = Y$ для $X_i \in NF^+$, $Z_i = \varepsilon$ для $X_i \in T$, $t \leqslant i \leqslant r$, $y'' = f_1 \ldots f_l$ —слово максимальной длины из F^+, следующее за X_{r+1}, при этом $f_1 = X_{r+2}$;

$H_1 = H_{14}$, если $X_r \lessdot X_{r+1}$, где $Z_i = Y$ для $X_i \in NF^*$, $z_i = \varepsilon$ для $X_i \in T$, $t \leqslant i \leqslant r$;

$H_1 = H_{05}$, если [либо $|(X_r \cdot > X_{r+1}) \wedge (\neg (X_r \doteq X_{r+1}))|$, либо $|(X_r \cdot > \cdot > X_{r+1}) \wedge (X_r \doteq X_{r+1}) \wedge (X_{r+1} | \dot{-} | X_{r+2}) \wedge (Y \neq Y')$, где $Y = f_1 \ldots f_e$ — слово максимальной длины из F^+, следующее за X_{r+1}, при этом $f_1 = X_{r+2}|] \wedge |(X_k \lessdot X_{k+1}) \wedge (X_j \doteq X_{j+1})$, где $k+1 \leqslant j \leqslant r|$, при этом $A \to \alpha . X_{k+1} \ldots X_r$ —правило с номером q, β терминальный префикс максимальной длины цепочки $X_1 \ldots X_k$;

H_1=ОШИБКА, если либо X_r и X_{r+1} не связаны никакими отношениями предшествования, либо нет правила с правой частью $X_{k+1} \ldots X_r$, либо ($X_r \doteq X_{r+1}$ единственное отношение между X_r и $X_{r+1}) \wedge (X_{r+1} | \dot{-} | X_{r+2}) \wedge (Y \neq Y')$, где $Y' = f_1 \ldots f_e$ — слово максимальной длины из F^+, следующее за X_{r+1}, при этом $f_1 = X_{r+2}$.

Так как в общем случае отношения предшествования 1--4 определены для цепочек (индексированные нетерминалы), то необходимо решить вопрос однозначного выбора X_{r+1} для указанных отображений.

Пусть текущая конфигурация автомата — $[\perp X_1 \ldots X_r, Y, X_{r+1} \ldots X_n \perp, \tilde{a}]$, где $Y \in F^*$. X_{r+1} выбирается по следующей схеме;

$X_{r+1} = a$, если $a \in T$—первый элемент второго магазина;

$X_{r+1} = f$, если $f \in F$—первый элемент второго магазина;

$X_{r+1} = A\xi$, если $A\xi \in NF^*$ строка символов слева во втором магазине. При этом, если во втором магазине содержится слово $A f_1 \ldots f_k f_{k+1} \ldots f_m X_{r+2} \ldots X_n \perp$, то либо $\xi = f_1 \ldots f_k$, если $A f_1 \ldots f_k | \dot{-} | f_{k+1}$, либо $\xi = f_1 \ldots f_m$, если ни для какого k, $1 \leqslant k \leqslant m$, не имеет место $A f_1 \ldots f_k | \dot{-} | f_{k+1}$.

Для индексных грамматик предшествования указанным способом обеспечивается однозначный выбор X_{r+1}.

Теорема 2. *D-автомат распознает те и только те цепочки, которые порождаются индексными грамматиками предшествования.*

При этом содержание четвертого составляющего заключительной конфигурации показывает порядок применения правил для получения исходной цепочки.

Ереванский комплексный отдел
ВНИИЭГазпрома

Ա. Ա. ՕՐՏՅԱՆ, Գ. Վ. ՋՈՒԼՀԱԿՅԱՆ

Ինդեքսային լեզուների շարահյուսական վերլուծության ալգորիթմ

Հոդվածում ուսումնասիրվում են ևնթատեքստից կախված ֆորմալ լեզուների սնփական ևնթադասի՝ ինդեքսային լեզուների վերլուծության և ճանաչման հարցերը։

Սահմանվում են նախորդման ինդեքսային լեզուները, որոնց համար կառուցվում է շարահյուսական վերլուծության ալգորիթմ։ Ալգորիթմի աշխատանքն էապես հենվում է նոր տիպի պահունակային D-ավտոմատի գործունեության վրա։ Վերջինիս միարժեք անցումները որոշվում են նախորդման բինար հարաբերությունների օգնությամբ։

ЛИТЕРАТУРА—ԳՐԱԿԱՆՈՒԹՅՈՒՆ

[1] *Н. Хомский*, Кибернетический сборник, вып. 2, 1961. [2] *И. Л. Братчиков*, Синтаксис языков программирования, «Наука», М., 1975. [3] *А. Ахо*, в сб. «Языки и автоматы», «Мир», М., 1975. [4] *А. Ахо, Дж. Ульман*, Теория синтаксического анализа, перевода и компиляции, т. I, «Мир», М., 1978.