Алгоритм синтаксичного аналізу Кока–Касамі–Янгера для граматик у нормальній формі Хомського

Існує достатньо ефективний метод визначення чи належить ланцюжок мові, що задана граматикою, який оснований на ідеї "динамічного програмування". Цей алгоритм відомий ще як CYK-алгоритм – алгоритм Кока–Янгера–Касамі.

Він використовується *лише для граматик у нормальній формі Хомського.*

На вхід алгоритму подається ланцюжок $w = a_1a_2 \dots a_n$ з T^*.

За час $O(n^3)$ алгоритм будує таблицю, яка говорить, чи належить w мові L.

Алгоритм Кока-Касамі-Янгера для синтаксичного аналізу.

Базис. Обчислюємо перший рядок так. Оскільки ланцюжок, який починається та закінчується в позиції i, являє собою просто термінал ai, а граматика знаходиться в НФХ, єдиний спосіб породити a_i полягає в використаннні продукції вигляду $A \longrightarrow a_i$ граматики G. Отже, X_{ii} є множиною змінних A, для яких $A \longrightarrow a_i$ – продукція G.

Індукція. Нехай потрібно обчислити X_{ij} в $(j - i + 1)$-му рядку, і всі множини X в нижніх рядках вже обчислені, тобто відомі для всіх подланцюжжків, коротших, ніж $a_i a_{i+1} \ldots a_j$, і зокрема, для всіх власних префіксів і суфіксів цього ланцюжка. Можна припустити, що $j - i > 0$, оскільки випадок $j = i$ розглянуто в базисі. Тому будь-який вивід $A \Rightarrow^* a_i a_{i+1} \ldots a_j$ має починатися кроком $A \rightarrow BC$. Тоді B породжує деякий префікс рядка $a_i a_{i+1} \ldots a_j$, скажімо, $B \Rightarrow^* a_i a_{i+1} \ldots a_k$ для деякого $k < j$. Відповідно, C породжує залишок $a_{k+1} a_{k+2} \ldots a_j$, тобто $C \Rightarrow^* a_{k+1} a_{k+2} \ldots a_j$.

Схема індексації таблиці аналізу

X_{15}				
X_{14}	X_{25}			
X_{13}	X_{24}	X_{35}		
X_{12}	X_{23}	X_{34}	X_{45}	
X_{11}	X_{22}	X_{33}	X_{44}	X_{55}
a_1	a_2	a_3	a_4	a_5

Алгоритм Кока-Касамі-Янгера для синтаксичного аналізу.

Доходимо висновку, що для того, щоб A потрапило в X_{ij}, потрібно знайти змінні B і C і ціле k, при яких справедливі такі умови:

1) $i \le k < j$;
2) B належить X_{ik};
3) C належить $X_{k+1,j}$;
4) $A \to BC$ – продукція в G.

Пошук таких змінних А потребує обробки не більше n пар обчислених раніше множин: $(X_{ii}, X_{i+1,j})$, $(X_{i,i+1}, X_{i+2,j})$ і т. д. до $(X_{i,j-1}, X_{jj})$. Отже, ми піднімаємося по колонці, розташованій під X_{ij}, і одночасно спускаємося по діагоналі.