Prof. Dr. Hans-Jörg Kreowski, Dr. Sabine Kuske
Studiengang Informatik
Linzer Str. 9a
OAS 3001, 3005
Tel.: 2956, 2335, 3697 (Sekr.), Fax: 4322
E-Mail: {kreo,kuske}@informatik.uni-bremen.de
www.informatik.uni-bremen.de/theorie

Juni 2004

Theoretische Informatik 2

Das Cocke-Kasami-Younger-Verfahren

Das Verfahren von Cocke, Kasami und Younger ist eine Lösung des Wortproblems kontextfreier Grammatiken in Chomsky-Normalform. Es beruht auf dem Kontextfreiheitslemma und hat kubischen Aufwand.
Sei $G = (N, T, P, S)$ eine kontextfreie Grammatik in *Chomsky-Normalform*,[1] d.h. für jede Produktion $A ::= r$ ist $r \in T$ oder $r \in N^2$. Will man wissen, ob ein Wort $w \in T^*$ aus einem nichtterminalen Zeichen $A \in N$ ableitbar ist, ob also $A \stackrel{*}{\longrightarrow} w$ gilt, gibt es nur zwei relevante Fälle, in denen die Antwort positiv ausfällt:

1. $|w| = 1$ und $A ::= w \in P$ oder
2. $|w| > 1$ und es existieren $A ::= BC \in P$ sowie $B \stackrel{*}{\longrightarrow} u$, $C \stackrel{*}{\longrightarrow} v$ mit $w = uv$.

Wegen der Chomsky-Normalform leitet kein nichtterminales Zeichen das leere Wort ab. Außerdem hat die Ableitung $A \stackrel{*}{\longrightarrow} w$ immer mindestens einen ersten Schritt $A \longrightarrow r$ für $A ::= r \in P$. Ist $r \in T$, muss die restliche Ableitung $r \stackrel{*}{\longrightarrow} w$ die Länge 0 haben, d.h. $r = w$. Ist $r = BC$ für $B, C \in N$, dann induziert die restliche Ableitung $r = BC \stackrel{*}{\longrightarrow} w$ nach dem Kontextfreiheitslemma zwei Ableitungen $B \stackrel{*}{\longrightarrow} u$ und $C \stackrel{*}{\longrightarrow} v$ mit $w = uv$. Außerdem hat damit w mindestens die Länge 2. Das ergibt insgesamt die beiden genannten Fälle, wenn man beachtet, dass die jeweiligen Rückrichtungen offensichtlich sind.
Um also für terminale Wörter der Länge 1 die Ableitbarkeit aus einem nichtterminalen Zeichen zu prüfen, muss man nur die terminierenden Regeln anschauen. Um sie für längere Wörter zu prüfen, muss man die nichtterminalen Regeln anschauen und das Wort in zwei Teile teilen. Das Anfangsstück muss aus dem ersten, das Endstück aus dem zweiten nichtterminalen Zeichen der rechten Regelseite ableitbar sein. Das ist dieselbe Frage, aber für kürzere Wörter, so dass diese Rekursion nach endlich vielen Schritten abbricht. Beachtet man noch, dass bei weiteren Zerlegungen der Wörter beliebige Teilwörter des ursprünglichen Wortes entstehen können, erhält man folgende Formulierung der obigen beiden Fälle, wobei als Gesamtwort $x_1 \cdots x_n$ (mit $x_l \in T$ für $l = 1, \ldots, n$) und als Teilwort $x_i \cdots x_{i+j-1}$ für $i = 1, \ldots, n$ und $j = 1, \ldots, n - i + 1$ betrachtet werden:

[1] Zu jeder kontextfreien Grammatik, die nicht das leere Wort erzeugt, kann eine kontextfreie Grammatik in dieser Form konstruiert werden, die dieselbe Sprache erzeugt. Genaueres lässt sich z.B. in [HMU02, Kapitel 7] oder [EP00, Abschnitt 6.2 und 6.3] nachlesen.

$A \xrightarrow{*} x_i \cdots x_{i+j-1}$ gdw.

- $j = 1$ und $A ::= x_i \in P$ oder
- $j > 1$, $A ::= BC \in P$ und es existiert k mit $1 \leq k < j$ derart, dass $B \xrightarrow{*} x_i \cdots x_{i+k-1}$ und $C \xrightarrow{*} x_{i+k} \cdots x_{i+j-1}$.

Das liefert ein Verfahren, um die nichtterminalen Zeichen zu bestimmen, aus denen sich Teilwörter von $x_1 \cdots x_n$ ableiten lassen.
Seien für $i = 1, \ldots, n$ und $j = 1, \ldots, n - i + 1$

$$CELL_{i,j} = \{A \in N \mid A \xrightarrow{*} x_i \cdots x_{i+j-1}\}.$$

Dann können diese "Zellen" nach der obigen Überlegung folgendermaßen berechnet werden:

- Für $i = 1, \ldots, n$: $CELL_{i,1} = \{A \in N \mid A ::= x_i \in P\}$;
- für $j = 2, \ldots, n$ und $i = 1, \ldots, n - j + 1$:

$$CELL_{i,j} = \bigcup_{k=1}^{j-1} \{A \in N \mid A ::= BC \in P,\ B \in CELL_{i,k},\ C \in CELL_{i+k,j-k}\}.$$

Damit ist auch das Wortproblem für G gelöst, denn es gilt:

$$x_1 \cdots x_n \in L(G) \text{ gdw. } S \xrightarrow{*} x_1 \cdots x_n \text{ gdw. } S \in CELL_{1,n}.$$

Da es für jedes $j = 1, \ldots, n$ jeweils $n - j + 1$ Zellen mit j als zweitem Index gibt, müssen insgesamt

$$\sum_{j=1}^{n} (n - j + 1) = \frac{n \cdot (n + 1)}{2}$$

Zellen berechnet werden. Für die Zellen $CELL_{i,1}$ geht das in konstanter Zeit, weil nur die gegebenen terminierenden Produktionen inspiziert werden müssen (höchstens $\#N \cdot \#T$ viele). Um eine Zelle $CELL_{i,j}$ mit $j > 1$ zu bilden, muss man für $k = 1, \ldots, j - 1$ auf die Zellenpaare $CELL_{i,k}$ und $CELL_{i+k,j-k}$ zugreifen. Man kann annehmen, dass die bereits berechnet sind, weil ihre zweiten Indizes kleiner als das aktuelle j sind. Jede dieser Zellen enthält eine beschränkte Zahl von nichtterminalen Zeichen (höchstens $\#N$ viele). Aus den $j - 1$ Zellenpaaren sind die beschränkt vielen Elementpaare zu bilden (höchstens $\#N^2$ viele) und mit den rechten Seiten der nichtterminalen Produktionen (höchstens $\#N^3$ viele) zu vergleichen. Da $j \leq n$ ist, sind für die $O(n^2)$ vielen Zellen höchstens $O(n)$ viele konstante Aktionen erforderlich, was einen Gesamtaufwand von $O(n^3)$ ergibt.

Literatur

[EP00] Katrin Erk and Lutz Priese. *Theoretische Informatik.* Springer, Berlin Heidelberg, 2000.

[HMU02] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. *Einführung in die Automatentheorie, Formale Sprachen und Komplexitätstheorie.* Pearson Studium, 2002.