
Internet-Algorithmik

Abgabetermin: 05.05.2004 vor der Vorlesung

Besorgen Sie sich und lesen Sie den Artikel

Alfred V. Aho and Margaret J. Corasick. **Efficient string matching: an aid to bibliographic search.** *Communications of the ACM*, 18(6):333–340, 1975.

entweder online unter <http://doi.acm.org/10.1145/360825.360855>¹ oder aus der Bibliothek. Bearbeiten Sie anschließend folgende Aufgaben.

Aufgabe 1

Vergleichen Sie den Algorithmus von Aho-Corasick mit dem KMP-Algorithmus aus der Vorlesung bezüglich ihrer Funktionsweisen und Struktur. Wo gibt es Äquivalenzen oder Analogien?

Aufgabe 2

- (i) Konstruieren Sie die Goto-, Failure-, und Output-Funktion für folgende Wort-Menge: {aab, aabba, abaab, baa, abba}. Nummerieren Sie Ihre Zustände in lexikographischer Ordnung, d.h. ein Zustand, der ein lexikographisch kleineres Wort repräsentiert, hat eine kleinere Nummer².
- (ii) Konstruieren Sie aus der vorherigen Teilaufgabe den zugehörigen endlichen deterministischen Automaten (DFA) mit vollständiger Angabe alle Übergänge.

Aufgabe 3

Benutzen Sie den String baaabaabbaabaabbbab als Eingabe für Ihren Automaten aus Aufgabe 2. Legen Sie dazu eine Tabelle mit 21 Spalten und 3 Zeilen an. In der ersten Zeile soll in der Spalte i der i -te Buchstabe des Strings stehen. In der zweiten Zeile sollen in der i -ten Spalte die Zustände eingetragen werden, die die „Pattern Matching Machine“ aus 2(i) vor dem Lesen des i -ten Zeichens durchläuft. In der letzten Spalte steht dann der Zustand, mit dem die Maschine endet. In der dritten Zeile soll in der i -ten Spalte der Zustand des Automaten aus 2(ii) vor dem Lesen des i -ten Zeichens eingetragen werden. (Vergleiche auch Abbildung 2 aus oben genannten Artikel).

Aufgabe 4

Modifizieren Sie den Algorithmus von Aho-Corasick so, dass es nicht die Fundstellen ausgibt, sondern für jedes Wort der Suchmenge die Anzahl der Vorkommen berechnet.

¹Der Proxy muss unbedingt richtig konfiguriert sein und der Zugriff kann nur aus dem Universitätsnetzwerk (oder über VPN) erfolgen, z.B. durch automatische Konfiguration über <http://pac.lrz-muenchen.de>.

²Die lexikographische Ordnung über $\Sigma = \{a, b\}$ ist $\epsilon < a < b < aa < ab < ba < bb < aaa < \dots$