Seminar “Implementierung von Stringalgorithmen”

Leiß/Goller, CIS

14.5.2004

4 Anwendungen von Suffix-Tries

Aufgabe 4.1 Sei $T(t)$ der Suffix-Trie des Texts $t \in \Sigma$.

(i) Schreibe eine Funktion, die die Anzahl der verschiedenen Teilwörter (Substrings) w von t der Länge n,

$$\mathit{faktoren}(t)(n) := |\{w \in \Sigma^* \mid w \text{ ist Teilwort von } t \text{ mit } |w| = n\}|,$$

mit Hilfe von $T(t)$ berechnet.

(ii) Schreibe eine Funktion, die die Menge der Anfangspositionen von w in t aus dem Suffix-Trie $T(t)$ und $n = |t|$ berechnet. Die Anfangspositionen der Vorkommen von w in t kann man aus den Längen der Suffixe wv von t und n berechnen, also an den Längen der Wege vom durch w bestimmten Knoten zu den Blättern von $T(t)$ erkennen.

(iii) Schreibe eine Funktion, die einen Trie so ändert, das die Anzahl der Blätter unterhalb eines Knotens am Knoten vermerkt werden. Konstruiere damit aus dem Suffix-Trie $T(t)$ einen anders markierten Suffix-Trie $T'(t)$. Benutze $T'(t)$, um die Anzahl der Vorkommen von w in t in $|w|$ Schritten zu berechnen.

Aufgabe 4.2 Der *gemeinsame* Suffix-Trie mehrer Texte $t_1, \ldots, t_k \in \Sigma^*$ soll ein Σ-Trie sein, der die Vereinigung der Suffixmengen $\mathit{Suf}(t_1) \cup \cdots \cup \mathit{Suf}(t_k)$ repräsentiert, wobei aus einer Marke eines wesentlichen Knotens hervorgeht, zu welchen der Texte das durch den Knoten repräsentierte Wort ein Suffix ist. (Die Marke kann als Bitvektor der Länge k oder als Liste von Zahlen aus $\{1, \ldots, k\}$ dargestellt werden).

Man kann die bisherige Konstruktion so modifizieren, daß man nach der Konstruktion des Suffix-Tries $T(t_1)$ von t_1 (mit vorläufiger Markierung der wesentlichen Knoten) dessen Wurzel als Suffix-Trie $T(\epsilon)$ des leeren Präfixes von t_2 benutzt und damit die Konstruktion von $T(t_2)$ beginnt, usw., sodaß man die Bäume von den Wurzeln aus überlagert. (Dabei muß man an den Knoten die Marken anpassen, d.h. ggf. das betreffende Bit für „wesentlich bzgl. Text t_2“ setzen, und darf die Suffixzeiger aus $T(t_1)$ nicht irrtümlich als Suffixzeiger in der Konstruktion von $T(t_2)$ lesen.)