Der Suffix-Baum eines Texts t

- entsteht durch Pfad- und Markenkomprimierung aus dem Suffix-Trie,
- hat Größe $O(|t|)$ statt $O(|t|^2)$,
- kann in $O(|t|)$ Schritten konstruiert werden.

Pfadkomprimierung

Liegen auf einem Pfad im Suffix-Trie, nur *nicht verzweigende* Knoten, so werden diese Knoten weggelassen und die Markierungen zu einer Folge verkettet. (Der Suffix-Baum hat also Zeichen*reihen* als Marken.)

Prop. 2 Der Suffixbaum $S(t)$ von t hat $O(|t|)$ Knoten.

Bew: Die inneren Knoten von $S(t)$ –außer der Wurzel– haben ≥ 2 direkte Nachfolger. Ein solcher Baum hat (nach Induktion über die Baumhöhe) nur $O(n)$, Knoten, wenn er n Blätter hat. $S(t)$ hat $\leq |t|$ Blätter, denen ja Suffixe von t entsprechen.

Markierungskomprimierung

Alle Markierungen $w = a_i \cdots a_k$ an Pfaden werden durch Zahlen (i, j) auf Positionen von Teilworten $w = t[i, j]$ von t ersetzt.

Da $S(t)$ nur $O(|t|)$ Knoten, also auch Kanten hat, und $|(i, j)| = O(1)$ ist, braucht $S(t)$ nur $O(|t|)$ Platz.

Suffixbaum

Der *Suffixbaum* $S(t)$ des Texts $t \in \Sigma^*$ ist ein Baum,

(i) dessen Kanten mit Anfangs- und Endpositionen (i, j) von Teilwörtern $t[i, j]$ von t markiert sind,

(ii) an dessen Verzweigungsknoten nur Äste abgehen, deren Markierungen (i, j) Wörtern $t[i, j]$ mit paarweise verschiedenen Anfangsbuchstaben entsprechen,

(iii) dessen Knoten in wesentliche und unwesentliche eingeteilt sind, so daß die dargestellte Wortmenge $L(S(t))$

$$= \{w \in \Sigma^* \mid w \text{ entspricht einem wesentlichen Knoten}\}$$

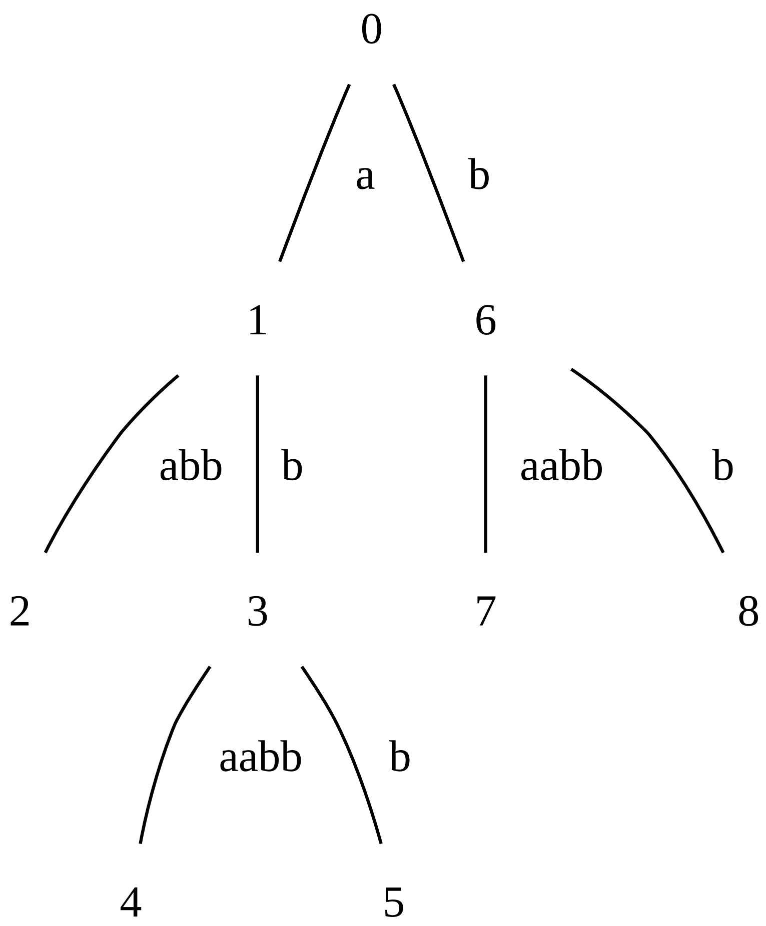
die Menge der Suffixe von t ist,

(iv) dessen innere unwesentliche Knoten mindestens zwei Nachfolger haben.

Einem Knoten *entspricht* das Wort, das sich aus der Verkettung der (expandierten) Kantenmarkierungen $t[i, j]$ auf dem Weg von der Wurzel zum Knoten ergibt.

Bem. Hat t ein besonderes Endsymbol \$, so haben alle innere Knoten mindestens zwei direkte Nachfolger (da es dann keine wesentlichen inneren Knoten gibt).

Der Suffixbaum von *abaabb*
(Kantenmarken expandiert)



Wesentliche Knoten sind die Blätter und die Knoten 6 und 0.

Konstruktion des Suffixbaums $S(t)$
Ukkonen 1992

Die Konstruktion von $S(t)$ erfolgt analog der online-Konstruktion des Tries $T(t)$.

Kantenmarken (i, j) identifizieren wir oft mit ihren Expansionen $\alpha = t[i, j] \in \Sigma$.

Ein *impliziter* Knoten (v, α) von $S(t)$ ist ein Knoten v mit einem echten Präfix α der expandierten Kantenmarke zu einem Nachfolger von v. Ist $\alpha = \epsilon$, so heißt (v, α) *echt* oder *real*.

Unterscheide bei der Konstruktion des Suffix-Tries T_i aus T_{i-1}:

Fall 1 Ein *Blatt* v von T_{i-1} erhält einen a_i-Nachfolger.

Dann ist v auch ein Blatt in S_{i-1}, und es muß nur die Marke (k, l) an der Kante nach v zu $(k, l + 1)$ verlängert werden.

Das kann man sich sparen, wenn man nur Marken (k, ∞) an den Kanten zu Blättern von S_{i-1} verwendet. Erst bei S_n ersetzt man ∞ durch n.

Fall 2 Ein *innerer* Knoten von T_{i-1} erhält einen a_i-Nachfolger.

Die inneren Knoten der Suffixkette von T_{i-1}, die in T_i neue a_i-Nachfolger erhalten, bilden die *abgekürzte* Suffixkette von T_{i-1}, der in S_{i-1} gewisse implizite Knoten

$$(v_{k-1}, \alpha_{k-1}), \ldots, (v_l, \alpha_l) \tag{2}$$

entsprechen. Sei (v, α) ein Knoten dieser Kette. Dem Einfügen seines a_i-Nachfolgers $(v, \alpha)' = (v, \alpha a_i)$ in T_i entspricht in der Konstruktion von S_i:

(i) falls (v, α) nicht real und $(v, \alpha\beta)$ der nächste reale Nachkomme in S_{i-1} ist, bei (v, α) einen realen Knoten von S_i mit Nachfolger $(v, \alpha\beta)$ einfügen, und die beiden Kantenmarken anpassen,

(ii) an den realen (inneren) Knoten (v, α) einen Zweig mit Kantenmarke (i, ∞) zum neuen Blatt $(v, \alpha)' = (v, \alpha a_i)$ von S_i anlegen,

(iii) einen Suffixzeiger vom ggf. zuletzt angelegten a_i-Nachfolger $(u, \gamma)'$ eines (u, γ) aus (2) nach $(v, \alpha a_i)$ anlegen.

Am Ende (v_l, α_l) muß ein weiterer Suffixzeiger angelegt werden:

(iv) Falls in T_{i-1} der Suffixzeiger von (v_l, α_l) auf einen Knoten (v_{l-1}, α_{l-1}) zeigt, der schon einen a_i-Nachfolger in T_{i-1} hatte, muß man in S_i von $(v_l, \alpha_l)'$ darauf zeigen, sonst auf die Wurzel.

Dies ist gleichzeitig ein Repräsentant des Anfangsknotens der Kette (2) für S_i. In S_0 und S_1 ist es die Wurzel.

Zu klären bleibt:

Q: Wie folgt man in S_{i-1} der Suffixkette (2) von T_{i-1}?

Sei (v_{k-1}, α_{k-1}) *kanonisch* repräsentiert, d.h. v_{k-1} sein nächster realer Vorfahre. (v_{k-1}, α_{k-1}) stellt das $(k-1)$-Suffix von $a_1 \cdots a_{i-1}$ dar, also (v_{k-2}, α_{k-2}) das $(k-2)$-Suffix. Man findet also $(v_{k-2}, \alpha_{k-2}) = (suf(v_{k-1}), \alpha_{k-1})$ über den Suffixzeiger von v_{k-1} und $\alpha_{k-1} = t[r, i-1]$.

Q: Wie erkennt man das Ende der kurzen Suffixkette (2)?

Wie oben finden wir (v_{l-1}, α_{l-1}) aus einem kanonischen Repräsentanten von (v_l, α_l). Daran, daß $(v_{l-1}, \alpha_{l-1}a_i)$ ein vorhandener impliziter Knoten ist, erkennt man, daß die abgekürzte Suffixkette verlassen wurde.

Q: Wie berechnet man den kanonischen Repräsentanten von (v, α)?

Man folgt vom realen Knoten v dem durch die Buchstaben von α definierten Weg, merkt sich jeweils den letzten durchlaufenen realen Knoten und die erste nicht benutzte Position r „in“ α. Mit ihr und der Endposition stellt man α' dar.