

Seqs Zusammenfassung

December 1, 2023

Contents

1	Suffix Arrays	2
1.1	Skew Algorithmus	2
1.2	Longest common prefixes	4
1.3	LCP-Intervalle	5
1.3.1	Range Minimum Queries	6
1.3.2	Kind-Intervalle finden	6
1.4	Bottom-up Konstruktion des LCP-Baums	7
1.5	Suffix Tree	8
1.6	Burrows-Wheeler Transformation	8
1.7	De Bruijn Sequenzen	9

1 Suffix Arrays

Sei $S = c_1c_2\dots c_n$ ein String. der i -te Suffix S_i ist definiert als der String $c_ic_{i+1}\dots c_n$. Ein Suffix Array nimmt die Menge aller n Suffixe und sortiert sie lexikographisch. Dafür ist eine Ordnung $<$ auf dem Alphabet Σ gegeben. Die naive Implementierung, die Substrings anhand ihrer ersten Character vergleicht, hat eine Laufzeit von $\mathcal{O}(n^2)$

1.1 Skew Algorithmus

Dieser ist ein Beispiel für einen linear-Zeit Algorithmus. Dieser wird im Folgenden am Beispiel $S = \text{ctaataatg}$ erklärt.

Schritt 1. Als erstes teilt der Algorithmus den String S in zwei Teile. Sei S_i der Substring beginnend an der Stelle i . Man definiert die Menge $S_L = \{S_i : i \bmod 3 \neq 1\}$ und $S_R = \{S_i : i \bmod 3 = 1\}$. Im Beispiel besteht S_L aus

- taataatg
- aataatg
- atg
- tg
- g

und S_R aus

- ctaataatg
- ataataatg
- atg

Die Suffixmenge von S_L wird mittels Radixsort rekursiv in linearer Zeit sortiert, dabei wird nur bezüglich der ersten drei Stellen aller Suffixe sortiert. Dadurch erhält man

- aat (S_3)
- aat (S_6)
- g\$\$ (S_9)
- taa (S_2)
- taa (S_5)

- $\text{tg\$}(S_8)$

Wären nun alle dieser Tripel echt verschieden, so wäre diese Liste bereits total sortiert, was in diesem Beispiel aber nicht der Fall ist. Die Tripel werden anschließend von oben nach unten nummeriert, wobei gleiche Tripel die gleiche Nummer erhalten. Diese werden als lexikografische Namen bezeichnet. Man definiert \overline{S} als der String der lexikografischen Namen der Substrings S_i mit $i \bmod 3 = 2$ konkateniert mit dem String der lexikografischen Namen der Substrings S_i mit $i \bmod 3 = 0$. Im Beispiel ist $\overline{S} = 334112$. Auf \overline{S} wird der Algorithmus rekursiv angewandt. Von \overline{S} erhalten wir die Suffixe

1. 334112
2. 34112
3. 4112
4. 112
5. 12
6. 2

Diese Liste wird im Sinne eines Suffixarrays sortiert, wodurch man

- 112
- 12
- 2
- 334112
- 34112
- 4112

erhält. Für jeden dieser Suffixe erhält man einen dazugehörigen Suffix vom ursprünglichen String S und seine Position in i in S . In der Tat erhält man die Zuordnungen: $\overline{S}_1 \equiv S_2$, $\overline{S}_2 \equiv s_5$, $\overline{S}_3 \equiv S_8$, ...

Im Allgemeinen gilt $\overline{S}_{\frac{i+1}{3}} \equiv S_i$ wenn $i = 3 \bmod 2$. Für $i = 0 \bmod 3$ hingegen gilt $\overline{S}_{\frac{n+i}{3}} \equiv S_i$. Dafür wird die Notation

$$\tau(i) = \begin{cases} \frac{i+1}{3}, & i = 2 \bmod 3 \\ \frac{n+i}{3}, & i = 0 \bmod 3 \end{cases}$$

Da wir aber die Richtung von \overline{S} zu S benötigen, brauchen wir die Umkehrabbildung von τ :

$$\tau^{-1}(j) = \begin{cases} 3j - 1, & 1 \leq j \leq \frac{n}{3} \\ 3j - n, & \frac{n}{3} < j \leq \frac{2n}{3} \end{cases}$$

Mittels τ^{-1} können nun die Suffixe von \overline{S} in jene von S umgerechnet werden, die dadurch sortiert sind. Schritt 2: In diesem Teil müssen die Suffixe in S_R sortiert werden. In diesem Schritt wird jeder Suffix von seinem Anfangsbuchstaben getrennt. Man erhält

- $S_1 = cS_2$
- $S_4 = aS_5$
- $S_7 = aS_8$

Wendet man nun Radixsort auf die ersten Buchstaben an, so werden diese Strings automatisch sortiert.

Schritt 3: In diesem Schritt werden die beiden Sublisten wieder zu einem Suffixarray zusammengefügt. Wenn das in der naiven Merge-Sort Implementierung durchgeführt wird, so würde das in quadratischer Laufzeit resultieren. Deswegen wird wieder die Strategie von Schritt 2 verwendet; die Liste S_R wird wieder aufgespalten in erste Buchstaben und Suffixe von S_L . Dadurch sind die beiden Suffixe aber nicht zwingend in der selben Gruppe, weswegen auch noch der zweite Buchstabe abgespalten werden kann. Der Merge Teil kann nun zuerst auf Basis der ersten beiden Buchstaben entschieden werden, außer wenn die beiden Buchstaben der zu vergleichenden Suffixe gleich sind. Dieses Verfahren funktioniert für Substrings S_i aus S_L mit $i \equiv 0 \pmod{3}$. Für Substrings S_i aus S_L mit $i \equiv 2 \pmod{3}$ hingegen würde die abgespaltenen Suffixe wieder in verschiedenen Listen liegen. Deshalb wird in diesem Fall nur ein Buchstabe abgespalten.

Der Vergleich der beiden abgespaltenen Suffixe funktioniert aber nur in linearer Zeit. Deswegen wird hierfür das sogenannte inverse Suffixarray eingeführt.

Definition 1.1.1 (inverses Suffixarray (ISA)). Für ein Suffixarray SA mit $k \rightarrow SA[k]$ ist das inverse Suffixarray definiert durch $ISA[SA[k]] = k$.

Allgemein gilt dann $S_i < S_j \Leftrightarrow ISA[\tau(i)] < ISA[\tau(j)]$.

1.2 Longest common prefixes

Das LCP-Array ist definiert auf der Sortierung des Suffix-Arrays. $LCP[i]$ speichert die Länge des längsten gemeinsamen Präfixes von $S_{SA[i-1]}$ und $S_{SA[i]}$. Insbesondere ist daher $LCP[1]$ undefiniert und wird daher auf -1 gesetzt. Algorithmus:

1. setze $LCP[1] = 1$ und $LCP[n + 1] = -1$
2. konstruiere ISA aus SA (linear)
3. man beginnt mit dem längsten Suffix (bekommt man konstant aus ISA) und vergleicht mit dem Suffix in der Zeile oben
4. dieser Wert l wird eingeschrieben und man geht zum zweitlängsten Suffix
5. Wieder vergleichen mit dem nächst-höheren Suffix im SA. Dort ist die Länge mindestens $l - 1$, daher erst ab dem l -ten Buchstaben vergleichen.
6. iterativ fortsetzen

1.3 LCP-Intervalle

Ein Intervall $[i, i]$ wird als Singleton bezeichnet. Ein Intervall $[i, j]$ in einem LCP Array mit $1 \leq i < j \leq n$ heißt LCP-Intervall mit LCP-Wert l , wenn

1. $LCP[i] \leq l$
2. $LCP[k] \geq l$ für alle k mit $i + 1 \leq k \leq j$
3. $LCP[k] = l$ für mindestens ein $i + 1 \leq k \leq j$
4. $LCP[j + 1] < l$

Wenn man jeden String mit einem Dummy-Wert (z.B. \$), dann ist das Intervall $[1, n]$ immer ein LCP-Intervall.

Lemma 1.3.1. *Zwei LCP-Intervalle können sich nicht überlappen, d.h. es muss einer der folgenden Fälle eingetreten sein:*

1. Intervall $[i, j]$ ist echt in $[p, q]$ enthalten oder
2. Intervalle $[i, j]$ und $[p, q]$ sind disjunkt

Definition 1.3.2. Ein m -Intervall ist in einem l -Intervall enthalten, falls höchstens einer der Ränder gleich ist und $m > l$ gilt.

Das l -Intervall $[i, j]$ ist das umschließende Intervall von $[p, q]$, wenn $[p, q]$ in $[i, j]$ enthalten ist und es kein anderes in $[i, j]$ enthaltenes Intervall gibt, das $[p, q]$ enthält. Diese Intervalle werden Eltern und Kinder-Intervalle genannt.

Ein Intervall der Form $[k, k]$ wird Singleton genannt. Das Elternintervall ist das kleinste LCP-Intervall, dass den Singleton enthält.

1.3.1 Range Minimum Queries

Gegeben ist ein Array A von n ganzen Zahlen. Ein Range Minimum Query gibt für das Array A und Indices i, j ein k zurück, sodass

$$A[k] = \min_{i \leq m \leq j} A[m]$$

So ein Query wird im Folgenden als $RMQ_A(i, j)$ bezeichnet. Man kann zeigen, dass A in linearer Zeit bearbeitet werden kann, sodass das RMQ in konstanter Zeit ausgeführt werden kann.

Definition 1.3.3. Aus der Definition eines LCP-Arrays folgt, dass es einen Index k gibt, an dessen Stelle der LCP-Wert l angenommen wird. All diese Indizes werden in Folge als l -Index bezeichnet.

1.3.2 Kind-Intervalle finden

Lemma 1.3.4. Sei $[i, j]$ ein l -Intervall. Seien $i_1 < i_2 < \dots < i_k$ die l -Indizes in aufsteigender Reihenfolge. Dann sind die Kind-Intervalle gerade

$$[i, i_1 - 1], [i_1, i_2 - 1], \dots, [i_k, i_{k+1} - 1], \dots, [i_k, j]$$

Definition 1.3.5. Für $2 \leq i \leq n$ definiere

$$PSV[i] = \max\{j | 1 \leq j < i \text{ und } LCP[i] < LCP[j]\}$$

und

$$NSV[i] = \min\{j | i < j \leq n + 1 \text{ und } LCP[i] < LCP[j]\}$$

Lemma 1.3.6. Sei $2 \leq k \leq n$ und $LCP[k] = l$. Dann ist $[PSV[k], NSV[k] - 1]$ ein LCP Intervall mit LCP-Wert l

Lemma 1.3.7. Sei $[i, j] \neq [1, n]$ ein LCP-Intervall mit $LCP[i] = p$ und $LCP[j + 1] = q$.

- Ist $p = q$, dann
 - ist das Elternintervall $[PSV[i], NSV[i] - 1]$
 - hat das Elternintervall den LCP-Wert $p = q$
 - sind i und $j + 1$ aufeinander folgende p -Indizes des Elternintervall
- Ist $p > q$, dann

- ist $[PSV[i], j]$ das Elternintervall
- hat das Elternintervall den LCP-Wert p
- ist i der letzte p -Index vom Elternintervall
- Ist $p < q$, dann
 - ist das Elternintervall $[i, NSV[j + 1] - 1]$
 - hat das Elternintervall den LCP-Wert q
 - ist $j + 1$ der letzte q -Index des Elternintervalls

Lemma 1.3.8. *Der LCP-Wert des Elternintervall von $[i, j]$ ist $\max\{LCP[i], LCP[j+1]\}$.*

1.4 Bottom-up Konstruktion des LCP-Baums

Man definiere $\langle p, a, b, A \rangle$ als das Viertupel mit LCP-Wert p , Intervallanfang a , -ende b und die Liste der Kindintervalle A . Algorithmus 5 aus dem Skript funktioniert in linearer Zeit.

Theorem 1.4.1. *Sei top das oberste Intervall auf dem Stack und top_{-1} das darunter für ein k der For-Schleife. Bemerke $top.lcp > top_{-1}.lcp$. Dann, bevor top gepoppt wird, gilt*

- Ist $LCP[k] \leq top_{-1}.lcp$, dann ist top ein Kindintervall von top_{-1} .
- Ist $LCP[k] > top_{-1}.lcp$, dann ist top ein Kindintervall des ersten LCP-Intervall mit LCP-Wert $LCP[k]$, das k enthält. Genauer gesagt, ist top ein Kindintervall von $[top.lb, NSV[k] - 1]$.

Im Folgenden wird eine Anwendung der Bottom-Up Traversierung von LCP-Intervall-Bäumen dargestellt. Diese können sich zum Beispiel hinter der Funktion *process* im LCP-Intervall-Baum Algorithmus verstecken.

Example 1.4.2. Finde alle Teilstrings eines Strings S , die mindestens p und höchstens q mal in S ($2 \leq p \leq q$) vorkommen.

Idee: Folgender Algorithmus

1. *process(lastIntervall)*
2. *for each $\langle l, i, j + 1, [] \rangle$ in lastInterval.childlist*
3. *if $p \leq (j - i + 1)$ and $q \geq (j - i + 1)$*
4. *output (lastInterval.lcp + 1, l, [i, j])*

Die Interpretation der Ausgabe ist zu verstehen als:

Verwende das Intervall $[i, j]$ und betrachte die Präfixe der Länge von $lcp + 1$ bis l . Wenn am Ende die Kindliste auf leer gesetzt wird, dann kann man zusätzlich verhindern, dass der gesamte Baum konstruiert wird.

1.5 Suffix Tree

Sei S ein String der Länge n . Ein Suffix-Baum der Länge $S\$$ ist ein Wurzelbaum $ST(S\$)$ mit den folgenden Eigenschaften

1. ST hat $n + 1$ viele Blätter nummeriert mit 1 bis $n + 1$.
2. Jeder innere Knoten hat mindestens zwei innere Knoten
3. Jede Kante hat als Kantenlabel einen nicht leeren Substring von $S\$$
4. Für jeden Knoten α und jedes $a \in \Sigma$ gibt es nur eine Kante mit $a \Rightarrow \beta$ mit Label av für einen String v und einen Knoten $beginbeta$ in ST
5. Für das Blatt i ist die Konkatenation der Kantenlabels des Pfades zu i genau der Suffix S_i .

1.6 Burrows-Wheeler Transformation

Die Transformation konvertiert einen String der Länge n in den String $BWT[1, n]$ definiert durch $BWT[i] = S[SA[i] - 1]$ für alle i mit $SA[i] \neq 1$ und $BWT[i] = \$$ sonst.

Definition 1.6.1. Ein Substring w von S ist ein Repeat, wenn er mindestens einmal in S vorkommt. Wir definieren die Mengen

$$lc(w) = \{BWT[k] | i \leq k \leq j\}$$

$$rc(w) = \{S[SA[k] + l] | i \leq k \leq j\}$$

Ein Repeat ist supermaximal wenn

$$|lc(w)| = |rc(w)| = occ(w)$$

Mithilfe der bottom-up Traversierung soll nun ein Algorithmus entwickelt werden, der alle supermaximalen Repeats bestimmt. Es fällt auf, dass der rc-Wert für den durch ein LCP-Intervall bestimmten Substring gleich occ ist, wenn das LCP-Intervall kein Kind hat. Der lc Wert kann ganz leicht mit einer Hashmap auf die Buchstaben mit occ verglichen werden.

Definition 1.6.2. Ein Repeat ist links-maximal, wenn $|lc(w)| \geq 2$ und rechts-maximal, wenn $|rc(w)| \geq 2$. Ein links- und rechts-maximaler Repeat ist maximal.

Lemma 1.6.3. *Jeder rechts-maximale Repeat korrespondiert zu einem LCP-Intervall.*

Wenn maximale Repeats gesucht werden, so ist die Rechts-Maximalität bereits gegeben. Die Links-Maximalität ist noch leichter. Um lineare Laufzeit zu gewinnen, muss die Links-Maximalität der Kindintervalle an die Elternintervalle weitergegeben werden. Speichere die größte Position $lastdiff$ mit $1 \leq lastdiff \leq i - 1$ an der ein Buchstabe in $BWT[1...i - 1]$ festgestellt wurde, d.h.

$$BWT[lastdiff - 1] \neq BWT[lastdiff]$$

Lemma 1.6.4 (Berechnung des längsten Repeats). *Um den längsten Repeat zu bestimmen, genügt es, einmal über die LCP Tabelle zu iterieren und den maximalen LCP Wert zu bestimmen. Es wird der String ausgegeben, der an der Stelle des Maximums mit der Länge des LCP Werts steht. Es kann in konstanter Zeit darauf geachtet werden, dass jeder Repeat nur einmal ausgegeben wird.*

1.7 De Bruijn Sequenzen

Definition 1.7.1. Eine De Bruijn Sequenz der Ordnung k ist ein zyklischer String, der jedes String aus Σ^k genau einmal enthält. Zyklisch bedeutet hierbei, dass der Teilstring am Ende beginnen und am Anfang enden kann. Ein zyklischer De Bruijn String kann in einen nicht zyklischen String konvertiert werden, indem die ersten $k - 1$ Zeichen wieder am Ende angehängt werden.

Definition 1.7.2 (De Bruijn Graph). Der De Bruijn Graph für k über das Alphabet Σ enthält einen Knoten für jeden String aus Σ^{k-1} . Eine gerichtete Kante (u, v) existiert genau dann, wenn es einen String $w \in \Sigma^k$ sodass u ein Präfix und v ein Suffix von w ist. Von jedem Knoten gehen genau $|\Sigma|$ Kanten aus.

Eine lineare De Bruijn Sequenz hat Länge $n = |\Sigma|^k + k - 1$ und der maximale LCP Wert m wird an mehreren Stellen in der LCP Tabelle angenommen.

Lemma 1.7.3. *Die gesamte Ausgabelänge des Algorithmus zu Berechnung aller längsten Repeats für eine De Bruijn Sequenz S der Ordnung k ist $\Omega(n \log n)$, wobei $n = 2^k + k - 1$.*

2 komprimierte Ganztext Indizes

2.1 Burrows-Wheeler Transformation

Die BWT für einen String S entsteht aus einer Rotationmatrix des Strings. Dafür wird in Zeile 1 S geschrieben, in Zeile 2 der Suffix S_2 gefolgt von $S[1]$ und so weiter.

Anschließend werden die Zeilen lexikografisch sortiert. Die letzte Spalte L dieser Matrix ist die BWT. Das Problem dieser Konstruktion ist, dass die Laufzeit super-linear ist.

Definition 2.1.1. Sei L wie oben die letzte Spalte und F die erste. Die Funktion LF : $\{1, \dots, n\} \rightarrow \{1, \dots, n\}$ ist wie folgt definiert:

Ist $L[i] = c$ ist die k -te Wiederholung des Characters c , dann wird $LF(i)$ auf das j gesetzt, wo in Spalte F die k -te Wiederholung von c steht.