Informatik II - SS 2018

(Algorithmen & Datenstrukturen)

Vorlesung 21 (11.7.2018)

String Matching (Textsuche) II
Greedy Algorithmen I

UNI FREIBURG

Fabian Kuhn
Algorithmen und Komplexität

Textsuche / String Matching

UNI FREIBURG

Gegeben:

- Zwei Zeichenketten (Strings)
- Text T (typischerweise lang)
- Muster P (engl. pattern, typischerweise kurz)

Ziel:

- Finde alle Vorkommen von P in T

Annahmen:

- Länge Text T : $\boldsymbol{n}$, Länge Muster P : $\boldsymbol{m}$

Knuth-Morris-Pratt Algorithmus

UNI FREIBURG

Idee:

- Falls wir beim Testen des Musters P an Stelle t feststellen, dass $P[t]$ nicht mit dem Text an der entsprechenden Stelle übereinstimmt, dann wissen wir, dass die Stellen $P[0 \dots t-1]$ übereingestimmt haben.
- Das können wir bei der weiteren Suche ausnutzen

Beispiel: $P = \mathrm{ABDABLABDABK}$

Knuth-Morris-Pratt Algorithmus

Vorberechnung: Array S der Länge $m+1$

- $S[i]$: Stelle in P, an welcher man die neue Suche beginnt, falls beim Testen der Stelle i im Pattern ein Mismatch auftritt
- $S[0] = -1, \quad S[1] = 0$
- $S[m]$: Stelle in P, an welcher man weitersucht, nachdem P erfolgreich gefunden wurde

Beispiel:

$P =$ `[ A, B, D, A, B, L, A, B, D, A, B, D]`

$S =$ `[-1, 0, 0, 0, 1, 2, 0, 1, 2, 3, 4, 5, 3]`

Initialisierung

Vorberechnung von Array S:

- $P =$ [A, B, D, A, B, L, A, B, D, A, B, D]

  $S =$ [-1, 0, 0, 0, 1, 2, 0, 1, 2, 3, 4, 5, 3]

- An Position in $S[i]$ (für $i \in \{2, \dots, m\}$) steht

$$\boldsymbol{S[i] := \min_{k<i}\{P[i-k \dots i-1] = P[0 \dots k-1]\}}$$

- $S[i]$: Länge des längsten echten Teilstückes von $P[0 \dots i-1]$, welches an Stelle $i-1$ endet, und welches auch Anfangsstück von P ist

Knuth-Morris-Pratt Algorithmus

UNI FREIBURG

$t \coloneqq 0;\ \ p \coloneqq 0$ *// t: Position in Text, p: Position im Pattern*

while $t < n$ **do**

  if $T[t] = P[p]$ **then** *// characters match*

    if $p = m - 1$ **then** *// pattern found*

      `pattern found at position` $t - m + 1$

      $p \coloneqq S[m];\ \ t \coloneqq t + 1$

    else

      $p \coloneqq p + 1;\ \ t \coloneqq t + 1$

  else *// characters don't match*

    if $p = 0$ **then** *// mismatch at first character*

      $t \coloneqq t + 1$

    else

      $p \coloneqq S[p]$

Knuth-Morris-Pratt Alg.: Beispiel

Pattern: ABCABC $\quad S = [-1,0,0,0,1,2,3]$

Text:

A	D	A	B	C		D	A	B	C	A	G	A	B	V	A	B	C	A	B	C	A	B	C	

Knuth-Morris-Pratt Alg.: Laufzeit

UNI FREIBURG

Laufzeit ohne Initialisierung des Arrays S:

$t \coloneqq 0;\ \ p \coloneqq 0$

while $t < n$ **do**

    if $T[t] = P[p]$ **then**

        if $p = m - 1$ **then**

            pattern found

            $p \coloneqq S[m];\ \ t \coloneqq t + 1$

        else

            $p \coloneqq p + 1;\ \ t \coloneqq t + 1$

    else

        if $p = 0$ **then**

            $t \coloneqq t + 1$

        else

            $p \coloneqq S[p]$

Berechnung von $S[i]$

Berechnung von $S[i]$

Berechnung von $S[i]$: Beispiel

$h \coloneqq S[i-1]$

`while` $h \geq 0$ `do`

    `if` $P[i-1] = P[h]$ `then`

        $S[i] \coloneqq h+1;\ \ h \coloneqq -2$

    `else`

        $h \coloneqq S[h]$

`if` $h = -1$ `then` $S[i] = 0$

Beispiel: $P = [\texttt{A},\texttt{B},\texttt{D},\texttt{A},\texttt{B},\texttt{L},\texttt{A},\texttt{B},\texttt{D},\texttt{A},\texttt{B},\texttt{D},\texttt{X}]$

Berechnung von $S[i]$: Laufzeit

$h \coloneqq S[i-1]$
while $h \geq 0$ do
    if $P[i-1] = P[h]$ then
        $S[i] \coloneqq h+1$; $h \coloneqq -2$
    else
        $h \coloneqq S[h]$
if $h = -1$ then $S[i] = 0$

Beobachtung:

$$S[i] \leq S[i-1] + 1$$

Falls $\boldsymbol{S[i] = S[i-1] + 1}$: 1 Schleifendurchlauf

Falls $\boldsymbol{S[i] < S[i-1]}$:

- Wert von h nimmt in jedem Schleifendurchlauf ab
- Am Schluss ist $S[i] = h + 1$
- Anzahl Schleifendurchläufe $\leq \Delta h + 1 = S[i-1] - S[i] + 2$

Berechnung von $S[i]$: Laufzeit

UNI FREIBURG

Falls $\boldsymbol{S[i] = S[i-1] + 1}$:

- Anzahl Schleifendurchläufe $= 1 = S[i-1] - S[i] + 2$

Falls $\boldsymbol{S[i] < S[i-1]}$:

- Anzahl Schleifendurchläufe $\leq \Delta h + 1 = S[i-1] - S[i] + 2$

Gesamtlaufzeit:

Knuth-Morris-Pratt Alg.: Zusammenfassung

UNI FREIBURG

Knuth-Morris-Pratt Algorithmus:

- Berechnet zuerst in Zeit $O(m)$ das Array S der Länge m
  - hängt nur vom Pattern P ab
  - beschreibt an jeder Position im Pattern, wo (im Pattern) man bei einem Mismatch weitersuchen muss
- Mit Hilfe von S werden dann alle Vorkommen von P in T in Zeit $O(n)$ gefunden
  - In jedem Schritt kann man entweder die aktuelle Suchposition in T oder die Position des Suchfensters in T um mindestens 1 nach rechts verschieben

Gesamtlaufzeit: $\boldsymbol{O(m+n) = O(n)}$

Greedy Algorithmen

UNI FREIBURG

- Greedy Algorithmen sind eine Algorithmenmethode, um Optimierungsprobleme zu lösen.
  - wie auch z.B. dynamische Programmierung

- Greedy Algorithmen funktionieren oft nicht gut, aber wenn Sie funktionieren, dann sind sie besonders einfach und oft effizient

- Wir haben schon Beispiele gesehen und werden noch ein weiteres wichtiges Beispiel anschauen

- Vielen Dank an Christian Scheideler (U. Paderborn) für die Folien

Greedy Algorithmen

UNI FREIBURG

“Gierige” Strategie für Optimierungsprobleme:

- Aufbau einer Lösung in „kleinen“ Schritten
- In jedem dieser Schritte wird entsprechend eines definierten Optimierungskriteriums eine irreversible Entscheidung getroffen

Frage 1

Wann kann eine solche Strategie zu einer optimalen Lösung führen?

Frage 2

Wie beweist man, dass ein Greedy Algorithmus eine optimale Lösung liefert?

- Wir werden das an einem Beispiel sehen:
  zusätzliche Beispiele und verschiedene Beweisstrategien werden in der Algorithmentheorie besprochen

Greedy Algorithmen

1. Greedy Algorithmen bestimmen Lösung durch sukzessives Er-weitern von bereits gefundenen Teillösungen.

2. Erweitern geschieht durch lokal optimale Wahlen.

3. Analyse muss dann zeigen, dass lokal optimale Wahlen zu global optimalen Lösungen führt.

Greedy Algorithmen – Prim's MST Alg.

1. Greedy Algorithmen bestimmen Lösung durch sukzessives Er-weitern von bereits gefundenen Teillösungen.

2. Erweitern geschieht durch lokal optimale Wahlen.

3. Analyse muss dann zeigen, dass lokal optimale Wahlen zu global optimalen Lösungen führt.

1. Algorithmus von Prim bestimmt minimalen Spannbaum durch sukzessives Hinzufügen von Kanten.

2. Prims Algorithmus wählt möglichst leichte Kante, die isolierten Knoten mit Teilbaum verbindet.

3. Analyse über Schnitte und die Definition von leichten Kanten über einen Schnitt

Greedy Algorithmen – Kruskal's MST Alg.

UNI FREIBURG

1. Greedy Algorithmen bestimmen Lösung durch sukzessives Er-weitern von bereits gefundenen Teillösungen.

2. Erweitern geschieht durch lokal optimale Wahlen.

3. Analyse muss dann zeigen, dass lokal optimale Wahlen zu global optimalen Lösungen führt.

1. Algorithmus von Kruskal bestimmt minimalen Spannbaum durch sukzessives Hinzufügen von Kanten.

2. Kruskals Algorithmus wählt möglichst leichte Kante, die Zusammenhangskomp. verbindet.

3. Analyse über Schnitte und die Definition von leichten Kanten über einen Schnitt

Greedy Algorithmen – Datenkompression

UNI FREIBURG

Datenkompression

- Reduziert Größen von Files
- Viele Verfahren für unterschiedliche Anwendungen: MP3, MPEG, JPEG, …
- Wie funktioniert Datenkompression?

Zwei Typen von Kompression:

- Verlustbehaftete Kompression (Bilder, Musik, Filme,…)
- Verlustfreie Kompression (Programme, Texte, …)

Kodierung:

- Computer arbeiten auf Bits (Symbole 0 und 1), nutzen also das Alphabet {0,1}
- Menschen nutzen umfangreichere Alphabete (z.B. Alphabete von Sprachen)
- Darstellung auf Rechner erfordert Umwandlung in Bitfolgen

UNI FREIBURG

Beispiel:

- Alphabet Σ={a,b,c,d,…,x,y,z, ,.,:,!,?,&} (32 Zeichen)
- 5 Bits pro Symbol: $2^5 = 32$ Möglichkeiten

a	b	…	z		.	:	!	?	&
00000	00001		11001	11010	11011	11100	11101	11110	11111

Fragen?

- Sind 4 Bits pro Symbol nicht genug ?
- Müssen wir im Durchschnitt 5 Bits für jedes Vorkommen eines Symbols in langen Texten verwenden?

Greedy Algorithmen – Datenkompression

UNI FREIBURG

Beobachtung:

- Nicht jeder Buchstabe kommt gleich häufig vor
- z. B. kommen x, y und z in der deutschen Sprache viel seltener vor als e, n oder r

Idee:

- Benutze kurze Bitstrings für Symbole die häufig vorkommen

Effekt:

- Gesamtlänge der Kodierung einer Symbolfolge (eines Textes) wird reduziert

UNI FREIBURG

Grundlegendes Problem:

- Eingabe: Text über dem Alphabet Σ
- Gesucht: Eine binäre Kodierung von Σ, so dass die Länge des Textes in dieser Kodierung minimiert wird

Beispiel:

- $\Sigma=\{0,1,2,\ldots,9\}$
- Text = 00125590004356789 (17 Zeichen)

Länge der Kodierung:
$4 \cdot 17 = 68$ Bits

0	1	2	3	4	5	6	7	8	9
0000	0001	0010	0011	0100	0101	0110	0111	1000	1001

Greedy Algorithmen – Datenkompression

Grundlegendes Problem:

- Eingabe: Text über dem Alphabet Σ
- Gesucht: Eine binäre Kodierung von Σ, so dass die Länge des Textes in dieser Kodierung minimiert wird

Beispiel:

- $\Sigma=\{0,1,2,\ldots,9\}$
- Text = 0012559004356789 (17 Zeichen)

0	1	2	3	4	5	6	7	8	9
0	11000	11001	11010	11011	10	11100	11101	11110	11111

Greedy Algorithmen – Datenkompression

UNI FREIBURG

MorseCode:

- Elektrische Pulse über Kabel
- Punkte (kurze Pulse)
- Striche(Lange Pulse)

Beispiele aus dem MorseCode:

- e ist 0 (ein einzelner Punkt)
- t ist 1 (ein einzelner Strich)
- a ist 01 (Punkt – Strich)

Problem:

- Ist 0101 eta, aa, etet, oder aet ?

Problem Mehrdeutigkeit:

- Ist die Kodierung eines Buchstabens ein Präfix der Kodierung eines anderen Buchstabens, dann kann es passieren, dass die Kodierung nicht eindeutig ist.

Beispiel:

- e = 0, a = 01
- 0 ist Präfix von 01

UNI FREIBURG

Präfix-Kodierung:

Eine Präfix-Kodierung für ein Alphabet Σ ist eine Funktion γ, die jeden Buchstaben $x \in \Sigma$ auf eine endliche Sequenz von 0 und 1 abbildet, so dass für $x,y \in \Sigma$, $x \neq y$, die Sequenz $\gamma(x)$ *kein* Präfix der Sequenz $\gamma(y)$ ist.

Beispiel (Präfix-Kodierung):

$x \in \Sigma$	0	1	2	3	4	5	6	7	8	9
$\gamma(x)$	00	0100	0110	0111	1001	1010	1011	1101	1110	1111

Präfix-Code: Dekodierung

- Wie dekodiert man eine gegebenen mit einem Präfix-Code kodierte Zeichenkette?

Beispiel

$x \in \Sigma$	0	1	2	3	4	5	6	7	8	9
$\gamma(x)$	00	1000	1001	1010	010	011	1011	1101	1110	1111

Kodierte Zeichenkette

0100010100110011111011000101101100010000100010100100

Präfix-Code als Binärbaum

- Wie dekodiert man eine gegebenen mit einem Präfix-Code kodierte Zeichenkette?

Beispiel

$x \in \Sigma$	0	1	2	3	4	5	6	7	8	9
$\gamma(x)$	00	1000	1001	1010	010	011	1011	1101	1110	1111

UNI FREIBURG

Definition (Frequenz)

- Die Frequenz f[x] eines Buchstaben $x \in \Sigma$ bezeichnet den Bruchteil der Buchstaben im Text, die x sind.

Beispiel:

- $\Sigma = \{0,1,2\}$
- Text =„0010022001“ (10 Zeichen)
- f[0] = 3/5
- f[1] = 1/5
- f[2] = 1/5

UNI FREIBURG

Definition (Kodierungslänge)

Die Kodierungslänge eines Textes mit n Zeichen bzgl. einer Kodierung γ ist definert als

$$\text{Kodierungslänge} = \sum_{x \in X} n \cdot f[x] \cdot |\gamma(x)|$$

Beispiel:

- Σ = {a,b,c,d}
- γ(a) = 0; γ(b) =101; γ(c)= 110; γ(d)=111
- Text = „aacdaabb“
- Kodierungslänge = 16

UNI FREIBURG

Definition (durchschn. Kodierungslänge)

Die durchschnittliche Kodierungslänge eines Buchstabens in einem Text mit n Zeichen und bzgl. einer Kodierung γ ist definiert als

$$\mathbf{ABL}(\gamma) = \sum_{x \in \Sigma} f[x] \cdot |\gamma(x)|$$

Beispiel:

- Σ = {a,b,c,d}
- γ(a) = 0; γ(b) =101; γ(c)= 110; γ(d)=111
- Text = „aacdaabb“
- Durchschnittliche Kodierungslänge = 16/8 = 2

UNI FREIBURG

Problem einer optimalen Präfix-Kodierung:

- **Eingabe:**
  Alphabet Σ und für jedes $x \in \Sigma$ seine Frequenz f[x]

- **Ausgabe:**
  Eine Kodierung γ, die ABL(γ) minimiert