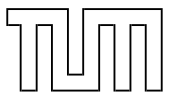
Grundlagen: Algorithmen und Datenstrukturen

Prof. Dr. Hanjo Täubig

Lehrstuhl für Effiziente Algorithmen
(Prof. Dr. Ernst W. Mayr)
Institut für Informatik
Technische Universität München

Sommersemester 2010



Übersicht

Datenkompression

Problem:

- Dateien enthalten oft viel Redundanz (z.B. Wiederholungen) und nehmen mehr Speicherplatz ein als erforderlich

$\Rightarrow$ mit Wissen über die Struktur der Daten und Informationen über die Häufigkeit von Zeichen bzw. Wörtern kann man die Datei so kodieren, dass sie weniger Platz benötigt (Kompression)

Präfixcodes

Definition

Ein Präfixcode (auch *präfixfreier Code*) ist ein Code, bei dem kein Codewort ein Präfix eines anderen Codeworts ist
(kein Codewort taucht als Anfang eines anderen Codeworts auf).

Vorteil:

- $\Rightarrow$ wenn man den codierten Text von vorn abläuft, merkt man sofort, wenn das aktuelle Codewort zu Ende ist
- bei einem Code, der die Präfix-Eigenschaft nicht erfüllt, wird u.U. erst an einer späteren Position klar, welches Codewort weiter vorn im Text gemeint war oder evt. ist die Dekodierung mehrdeutig

Präfixcodes

Beispiele:

- Der Code $\{a \mapsto 0, b \mapsto 01, c \mapsto 10\}$ ist kein Präfixcode, weil das Codewort für a als Präfix des Codeworts für b auftaucht.

  So wäre z.B. unklar, ob 010 für ac oder für ba steht.

  Für 0110 wäre zwar am Ende des Codes klar, dass dieser nur für bc stehen kann, allerdings wäre nach dem Ablaufen der ersten 0 noch nicht klar, ob diese für a steht, oder den Anfang des Codes für b darstellt. Das sieht man erst, nach dem man die folgenden 11 gesehen hat.

- Der Code $\{a \mapsto 0, b \mapsto 10, c \mapsto 11\}$ ist ein Präfixcode, weil kein Codewort als Präfix eines anderen Codeworts auftaucht.

Optimimale Kodierung

Eingabe:

- Wahrscheinlichkeitsverteilung p auf Alphabet A

Ausgabe:

- optimaler Präfixcode

$$f : A \mapsto \{0,1\}^*$$

für die Kodierung von A bei Verteilung p, d.h.
minimale erwartete Codelänge pro Eingabezeichen:

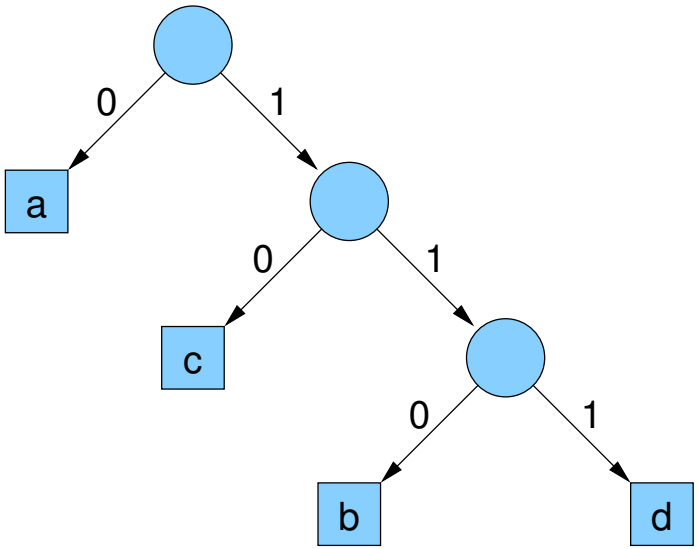
$$\sum_{x \in A} |f(a)| \cdot p(a)$$

Baumdarstellung

Beobachtung:

- Präfixcodes lassen sich als Baum darstellen
- Baumkanten sind mit Zeichen des Codes beschriftet (hier Bits 0 und 1, also Binärbaum)
- an den Blättern stehen die kodierten Zeichen aus A

Beispiel:



Alphabet: $A = \{a, b, c, d\}$

Kodierung:

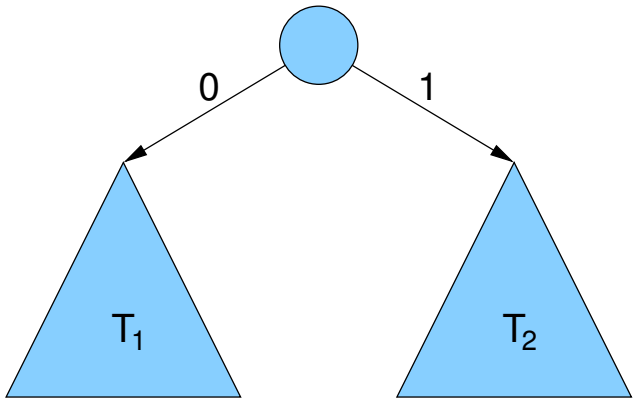
- $f(a) = 0$
- $f(b) = 110$
- $f(c) = 10$
- $f(d) = 111$

Huffman Code

Huffman Code: optimale Kodierung

Strategie:

- anfangs ist jedes Zeichen in A ein Baum für sich (also Wald aus $|A|$ Bäumen)
- Wiederhole bis nur noch ein Baum übrig ist
  - bestimme 2 Bäume T_1 und T_2 mit kleinster Summe ihrer Zeichenwahrscheinlichkeiten $\sum_{a \in T_{1/2}} p(a)$
  - verbinde T_1 und T_2 zu neuem Baum



Huffman Code / Beispiel

Zeichen $x \in A$	a	b	c	d	_
Wahrscheinlichkeit $p(x)$	0,35	0,1	0,2	0,2	0,15