Quellencodierung I: Redundanzreduktion, redundanzsparende Codes

1. Redundanzreduktion

1.1 Definition der Redundanz

Definition: Die **Codewortlänge** eines Codewortes ist die Anzahl der Binärzeichen (0 oder 1), die bei der Codierung des Codewortes benutzt werden.

Definition: Die **mittlere Codewortlänge m** ist der gewichtete Mittelwert (d.h. der Erwartungswert) der Codewortlängen aller n Zeichen:

$$m := \sum_{i=1}^{n} p(x_i) \cdot m_i$$

mit $p(x_i)$: Wahrscheinlichkeit, daß das i-te Zeichen auftritt,

m_i : Codewortlänge des i-ten Zeichens.

Definition: Der **mittlere Informationsgehalt H** von n Zeichen ist der gewichtete Mittelwert des Informationsgehaltes I_i der einzelnen Zeichen:

$$H := \sum_{i=1}^{n} p(x_i) \cdot I_i$$

mit $I_i = \log_2\left(\frac{1}{p(x_i)}\right)$: Informationsgehalt des i-ten Zeichens,

$p(x_i)$: Wahrscheinlichkeit, daß das i-te Zeichen auftritt.

Definition: Die **Redundanz** eines Codes ist die Differenz zwischen der mittleren Codewortlänge und dem mittleren Informationsgehalt: $R := m - H$.

Die **relative Redundanz** gibt an, wieviel Prozent der Codierung redundant sind: $r := \frac{m - H}{m}$.

1.2 allgemeine Redundanzreduktion

Die Redundanz wird reduziert, indem statt Blockcodes Codewörter unterschiedlicher Länge benutzt werden. Zeichen, die mit hoher Wahrscheinlichkeit auftreten, erhalten ein kurzes Codewort, Zeichen, die selten auftreten, ein langes.
Bei der Codierung wird die mittlere Codewortlänge kleiner als die Länge eines Blockcodes.

Fano-Bedingung (Präfix-Eigenschaft):

Kein Codewort aus einem Code bildet den Anfang eines anderen Codewortes.
Wenn diese Bedingung erfüllt ist, muß die Codewortlänge nicht übertragen werden, und die Decodierung ist eindeutig.

Shannonsches Codierungstheorem:

Die Codierung eines Zeichenvorrats kann immer so vorgenommen werden, daß die Redundanz minimal wird.

2. redundanzsparende Codes

2.1 Codierung nach Shannon

Voraussetzungen:

- Die n verschiedenen Zeichen $x_1,...,x_n$ treten mit den Wahrscheinlichkeiten $p(x_1), p(x_2),..., p(x_n)$ auf.
- Die Zeichen sind nach fallender Wahrscheinlichkeit geordnet: $p(x_1) \geq p(x_2) \geq ... \geq p(x_n)$.
- Für jedes Zeichen x_i ist P_i die kumulative Wahrscheinlichkeit aller vorigen Zeichen. Das bedeutet
  $P_1 = 0$, $P_2 = p(x_1)$, $P_3 = p(x_1) + p(x_2)$, $P_4 = p(x_1) + p(x_2) + p(x_3)$, ...,
  $P_n = p(x_1) + p(x_2) + ... + p(x_{n-1})$.
  Allgemein: $P_{i+1} = P_i + p(x_i)$ für $i = 1,..., n-1$.

Berechnung der Codewortlänge:

Die zu benutzende Codewortlänge des i-ten Zeichens wird berechnet als $m_i = \lceil I_i \rceil = \left\lceil \log_2\left(\frac{1}{p(x_i)}\right) \right\rceil$.

Sie ist also die kleinste natürliche Zahl, die größer oder gleich dem Informationsgehalt des Zeichens x_i ist.

Berechnung der Codewörter:

Die kumulative Wahrscheinlichkeit P_i, die zum Zeichen x_i gehört, wird in eine Dualzahl umgerechnet.
Dabei gelten folgende Regeln:

- Die Dualzahl muß möglichst groß, aber kleiner als P_i sein.
- Die Vorkommastelle wird ignoriert, weil sie sowieso immer 0 ist (sozusagen ein „hidden bit“); es werden also nur die Nachkommastellen umgerechnet.
- Die Dualzahl wird das gesuchte Codewort.
- Die Dualzahl wird nach der m_i -ten Stelle abgebrochen, denn das Codewort soll genau m_i Stellen lang sein.

Man erhält also die Binärzeichen $b_1, b_2,..., b_{m_i}$, die das gesuchte Codewort ergeben.

Beispiel:

Zeichen x_i	$p(x_i)$	P_i	m_i	**Codewort**	Z_i
A	**0,4**	0	2	**00**	0
B	**0,2**	0,4	3	**011**	0,375
C	**0,15**	0,6	3	**100**	0,5
D	**0,15**	0,75	3	**110**	0,75
E	**0,05**	0,9	5	**11100**	0,875
F	**0,05**	0,95	5	**11110**	0,9375

Die mittlere Codewortlänge ist m=2,8 Bit; der mittlere Informationsgehalt ist H=2,246 Bit.
Daraus ergeben sich die Redundanz R=0,554 Bit und die relative Redundanz r=19,77%.

*

2.2 Codierung nach Fano

Voraussetzungen:

- Die n verschiedenen Zeichen $x_1,...,x_n$ treten mit den Wahrscheinlichkeiten $p(x_1), p(x_2),..., p(x_n)$ auf.
- Die Zeichen sind nach fallender Wahrscheinlichkeit geordnet: $p(x_1) \geq p(x_2) \geq ... \geq p(x_n)$.

Codierungsvorschrift:

1. Die Menge der Zeichen wird in 2 Teile aufgeteilt, die möglichst gleichwahrscheinlich sind.
2. Der einen Teilmenge wird die Null als erste Codewortstelle zugeordnet, der anderen die Eins.

3. Für beide Teilmengen werden diese Schritte rekursiv durchgeführt, bis Teilmengen entstehen, die nur ein Zeichen enthalten.

Beispiel:

Zeichen x_i	$p(x_i)$	Codewort
A	0,4	00
B	0,2	01
C	0,15	10
D	0,15	110
E	0,05	1110
F	0,05	1111

Die mittlere Codewortlänge ist m=2,35 Bit; der mittlere Informationsgehalt ist H=2,246 Bit.
Daraus ergeben sich die Redundanz R=0,104 Bit und die relative Redundanz r=4,41%.

*

2.3 Codierung nach Huffman

Voraussetzungen:

- Die n verschiedenen Zeichen $x_1,...,x_n$ treten mit den Wahrscheinlichkeiten $p(x_1), p(x_2),..., p(x_n)$ auf.
- Die Zeichen sind nach fallender Wahrscheinlichkeit geordnet: $p(x_1) \geq p(x_2) \geq ... \geq p(x_n)$.

Codierungsvorschrift:

1. Zwei Zeichen der kleinsten Wahrscheinlichkeit werden herausgesucht.
2. Dem einen Zeichen wird die Null als (links anzufügende) Codewortstelle zugeordnet, dem anderen die Eins.
3. Die beiden Zeichen werden zu einem einzigen zusammengefaßt (Addition der Wahrscheinlichkeiten).
4. Das Verfahren wird wiederholt, bis alle Zeichen zusammengefaßt sind.

Die Huffman-Codes führen zu den kleinstmöglichen mittleren Codewortlängen.

Beispiel:

Zeichen x_i	$p(x_i)$	Codewort
A	0,4	0
B	0,2	100
C	0,15	101
D	0,15	110
E	0,05	1110
F	0,05	1111

Die mittlere Codewortlänge ist m=2,3 Bit; der mittlere Informationsgehalt ist H=2,246 Bit.
Daraus ergeben sich die Redundanz R=0,054 Bit und die relative Redundanz r=2,33%.

*

2.4 Codeumschaltung

Bei der Codeumschaltung werden zwei Codetabelle benutzt, zwischen denen mittels spezieller Codewörter umgeschaltet werden kann.
Damit werden statistische Abhängigkeiten ausgenutzt, um die Redundanz zu verringern.

Quellen: T. Grams: Codierungsverfahren, BI-Wissenschaftsverlag, 1986
Bärbel Mertsching: Grundzüge der Informatik B1, Uni HH, 1997