Informationsebenen

A) Statistik (Zeichenmenge, Auftrittswahrscheinlichkeit, Häufigkeit):
Der statistische Informationsgehalt einer Zeichenkette
(gemessen in Bits: BIT=binary digit)

B) Lexik (griech. lexikón (biblíon) = Wörterbuch):
Zeichensatz/Wortschatz einer Sprache zur Informationsdarstellung
(Alphabet/Wörter=Lexeme/Zeichencodierung)
Lexikologie = Wortlehre, Wortkunde, Wortschatzuntersuchung

C) Syntax (griech. syntaksis=Zusammen-Ordnung):
Regeln der Zusammenstellung von Zeichen und Wörtern / Beziehungen
der Zeichen untereinander (Sinn-Sätze als Träger semantischer
Information/Grammatik)

D) Semantik: (griech: σημαίνειν sēmainein „bezeichnen“)
Bedeutung der Zeichen/Wörter für ...? (Bedeutungslehre)
(Aussage, Sinn, Botschaft)

E) Pragmatik (griech. pragmatike =Handlungsaspekt, die Kunst richtig zu handeln):
beabsichtigte/ausgeführte/ausgelöste Handlung bzw. Tat

F) Apobetik (griech. αποβαίνοντα = Ergebnis, Erfolg):
beabsichtigtes/erreichtes Ziel bzw. Ergebnis

siehe auch: Semiotik, Syntaktik, Sigmatik

Informationskriterien

- Effektivität (Relevanz und Angemessenheit, Konsistenz, Richtigkeit, Verwendbarkeit)
- Effizient (Bereitstellung mit wirtschaftlichen Ressourcen)
- Vertraulichkeit
- Integrität
- Verfügbarkeit
- Complianz (Einhaltung externer und interner Regeln)
- Verlässlichkeit

Information ist zusätzliches zweckorientiertes Wissen.

Information hängt vom Wissensstand einer Person ab.

Informationstheorie

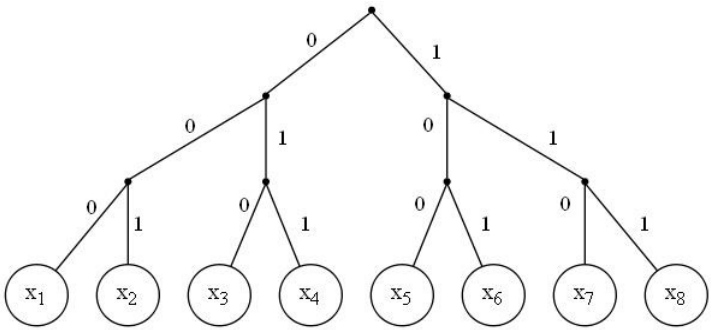
Begründer: Claude Shannon (1916-2001) in 1948

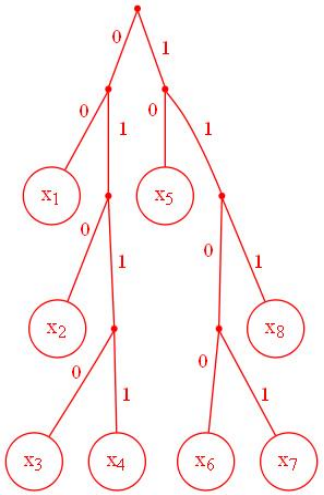
Ziel: quantitative Beschreibung des Informationsgehaltes einer Nachricht

Nachricht: besteht aus Zeichen $x_i \in X$ eines Alphabetes X

Die Wahrscheinlichkeit des Auftretens eines Zeichens x_i in der Nachricht des Senders sei $p_i=p(x_i)$ (Auftrittswahrscheinlichkeit)

Codebaum: mit Zeichen in den Blättern des Baumes





Präfix-Eigenschaft (Links-Fano-Bedingung): keine Codefolge ist selbst Anfang (Präfix) einer anderen Codefolge

010101101111... ?

Def.: Die Anzahl der Entscheidungsschritte H im Codebaum zur eindeutigen Identifizierung eines x_i wird als Entscheidungsgehalt des Zeichens x_i bezeichnet (H in Bit).

In einem balancierten Codebaum gilt: $H \leq \lceil ld\, n \rceil$ (n = Anzahl der Zeichen in X)

Def.: Entscheidungsgehalt einer Nachricht aus N Zeichen $H_N = N * ld\ n$

Def.: Informationsgehalt eines Zeichens $x_i \in X$

$$I(x_i) = ld\left(\frac{1}{p(x_i)}\right) = -ld\left(p(x_i)\right)$$ (in Bits pro Zeichen)

also: eine sichere Auftrittswahrscheinlichkeit bedeutet Null Information !

kleine $p_i = p(x_i)$ bedeutet viel Information!

Merke: Information = Grad der Unsicherheit

Def.: Der mittlere Informationsgehalt (= Entropie) einer Quelle wird definiert als:

$$H(x) = \sum_{i=1}^{n}\left(p(x_i) * I(x_i)\right) = \sum_{i=1}^{n}\left(p(x_i) * ld\left(\frac{1}{p(x_i)}\right)\right) = -\sum_{i=1}^{n}\left(p(x_i) * ld\left(p(x_i)\right)\right)$$

Def.: Sei l_i die Länge eines Codes des i-ten Zeichens der Quelle, dann ist die mittlere Codelänge L der Quelle definiert als:

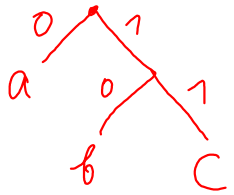
$$L = \sum_{i=1}^{n}(p(x_i) * l_i)$$

Bsp:

X	p_i	p_i	$I(x_i)$	$H(x)$	p_i	$I(x_i)$	$H(x)$	p_i
a	50%	$=\frac{1}{2}=2^{-1}$	1	$\frac{1}{2}\cdot 1+\frac{1}{2}\cdot 1$	25% $=\frac{1}{4}$	2	0,81	100%
b	50%	$=\frac{1}{2}=2^{-1}$	1	$=$ 1	75% $=\frac{3}{4}$	2-ld3		0% ...

	p_i	$I(x_i)$	$H(x)$	p_i	$I(x_i)$	$H(x)$	Code	l_i
a	33%	ld3	$\frac{1}{3}\cdot$ld3	50%	1		0	1
b	33%	ld3	$\frac{1}{3}$ld3	25%	2	1,5	10	2
c	33%	ld3	$\frac{1}{3}$ld3	25%	2		01	2
			$=$ld3					

$$1{,}5 = H(x) = \sum p_i \cdot l_i = 1{,}5 = L$$



x(i)	p(i)	I(x(i))	H(x)	Code	I(i)	L
a	50%	1	0,5*1+0,5*1=1	0	1	0,5*1+0,5*1=1
b	50%	1		1	1	

a	25%	2	0,25*2+0,75(2-ld 3)=0,81	0	1	0,25*1+0,75*1=1
b	75%	2-ld3		1	1	

a	33%	ld 3	0,333*ld 3+0,333*ld 3*0,333*ld 3=ld 3	0	1	0,33*1+0,33*2+0,33*2=1,66666
b	33%	ld 3		10	2	
c	33%	ld 3		11	2	

a	25%	2	0,25*2+0,25*2+0,5*1=1,5	11	2	0,25*2+0,25*2+0,5*1=1,5
b	25%	2		10	2	
c	50%	1		0	1	

a	25%	2	0,25*2+0,25*2+0,25*2+0,25*2=2	00	2	0,25*2+0,25*2+0,25*2+0,25*2=2
b	25%	2		01	2	
c	25%	2		10	2	
d	25%	2		11	2	

a	12,5%	3	0,125*3+1,125*3+0,25*2+0,5*1=1,75	111	3	0,125*3+1,125*3+0,25*2+0,5*1=1,75
b	12,5%	3		110	3	
c	25%	2		10	2	
d	50%	1		0	1	

a	20%	ld 5	0,2*ld 5+0,2*ld 5+0,2*ld 5+0,2*ld 5+0,2*ld 5=ld 5=2,32193	111	3	0,2*3+0,2*3+0,2*3+0,2*3+0,2*1=2,6
b	20%	ld 5		110	3	
c	20%	ld 5		101	3	
d	20%	ld 5		100	3	
e	20%	ld 5		0	1	

a	6,25%	4	0,0625*4+0,0625*4+0,125*3+0,25*2+0,5*1=1,875	1111	4	0,0625*4+0,0625*4+0,125*3+0,25*2+0,5*1=1,875
b	6,25%	4		1110	4	
c	12,5%	3		110	3	
d	25%	2		10	2	
e	50%	1		0	1	

Nach Shannon gilt: $H(x) \leq L \leq H(x) + 1$

d.h. die mittlere Codelänge einer optimalen Codierung
ist stets $\geq$ der Entropie der zu codierenden Nachricht.

Problem: Kann evtl. eine (minimale) Codierung gefunden werden mit H(x) = L ?

Codierungen:

https://de.wikipedia.org/wiki/Shannon-Fano-Kodierung

https://de.wikipedia.org/wiki/Arithmetisches_Kodieren

https://de.wikipedia.org/wiki/Tunstall-Kodierung

https://en.wikipedia.org/wiki/Tunstall_coding

Lösung: Huffman Codierung von David A.Huffman (1929-99/USA) in 1952

https://de.wikipedia.org/wiki/Huffman-Kodierung

Coderedundanz:= R(X) = L(X) - H(X)

z.B. n=10, x_i mit je 4 Bit codiert

$(\ell_i = 4)$

→ R(X) = 4 - ld 10 ~ 0,68

3,32

relative Coderedundanz:=

R(X) = [L(X) - H(X)] / L(X) = 1 - H(X)/L(X)

$= 1 - \frac{3{,}32}{4} \sim 17\%$

Dt. Schriftsprache 26 Buchstaben

a) Gleichverteilung der Buchstabenhäufigkeit $\Rightarrow$ Entropie = 4,7 Bit/Sym

b) unter Berücksichtigung der realen Buchstabenhäufigkeit
$\Rightarrow$ Entropie = 4,1 Bit/symbol

c) bei Beachtung von Silben mit einer mittleren Symbollänge von 3
$\Rightarrow$ Entropie = 2,8 Bit/symbol

d) Grenzwert der Entropie deutscher Texte $\sim$ 1,3 Bit/symbol
(siehe Küpfmüller) (aber Taschenbuch der TK S.41) 1,6 Bit/symbole

Küpfmüller et.al.:Einführung in die Theoretische Elektrotechnik. Springer

absolute Redundanz: $R = L - E$ = mittlere Codelänge - Entropie z.B. $L = \mathrm{ld}\, N$

relative Redundanz: $r_{rel} = \frac{L-E}{L} = 1 - \frac{E}{L}$

r_{rel} (dt. Sprache) $= 1 - \frac{1,3}{4,7} = \frac{3,4}{4,7} = 73\%$

$(= 0,66 = 66\%)$ nach Taschenbuch der TK

N: Anzahl der Quellsymbole

Algorithmus: Huffman Codierung
(buttom up)

- Zeichenhäufigkeiten bestimmen und von groß nach klein sortieren
- Ordne jeweils denjenigen Zeichen einen Vaterknoten zu, die jeweils die minimalen Häufigkeiten von Zeichen bzw. der vorherig entstandenen, den Vaterknoten zugeordneten, aggregierte Häufigkeiten (also die Summe der Häufigkeiten der beiden Sohnknoten) aufweisen
- Dadurch entsteht buttom-up ein binärer Codebaum mit den Zeichen in den Blätter (Prefix-Eigenschaft ist automatisch erfüllt)

Algorithmus: Shannon Fano Codierung

(top down)

- Zeichenhäufigkeiten bestimmen und von groß nach klein sortieren
- Teile die Menge der auftretenden Zeichen bzgl. der Summe der Häufigkeiten in zwei möglichst gleichgroße Teil-Mengen (Partitionierung). Diese Partitionierung erfolgt gemäß der Reihenfolge der sortierten Häufigkeiten, d.h. es werden beginnend bei dem am häufigsten auftetenden Zeichen solange Zeichen in die erste der beiden Teilmengen aufgenommen bis die aggregierte Zeichenhäufigkeit 50% der Gesamtzeichenhäufigkeit erreicht bzw. überschreitet. Die restlichen Zeichen werden jeweils in die zweite Teilmenge aufgenommen.
- Solange noch keine einelementige Menge resultiert unterteile diese Teilmengen weiter nach den gleichen Kriterien
- Dadurch entsteht ein Codebaum mit den Zeichen in den Blätter (Prefix-Eigenschaft ist automatisch erfüllt)

weitere Entropie-Codierungen

Golomb-Rice-Codes: https://de.wikipedia.org/wiki/Golomb-Code

CABAC oder CAVLC - lernfähige kontextabhängige Entropie-Kodierung

CABAC=Context Adaptive Binary Arithmetic Coding

Teil von MPEG-4/Part10 (H.264/AVC)

https://de.wikipedia.org/wiki/Context-Adaptive_Binary_Arithmetic_Coding

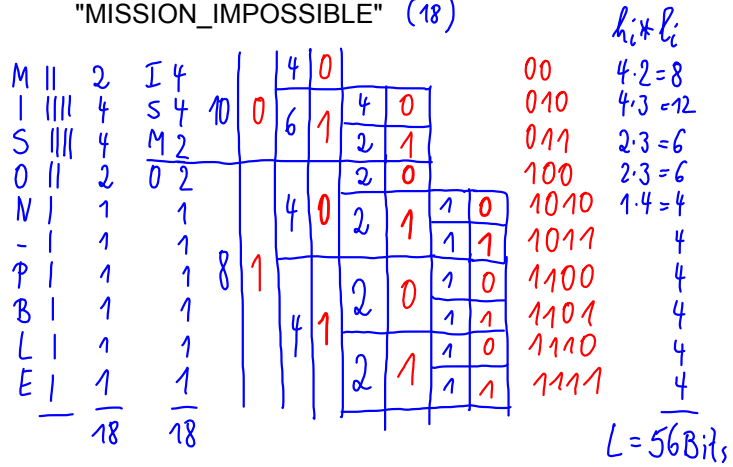
http://iphome.hhi.de/marpe/download/cabac_ieee03.pdf

http://www.xilinx.com/support/documentation/ip_documentation/h264_cabac_ds603.pdf

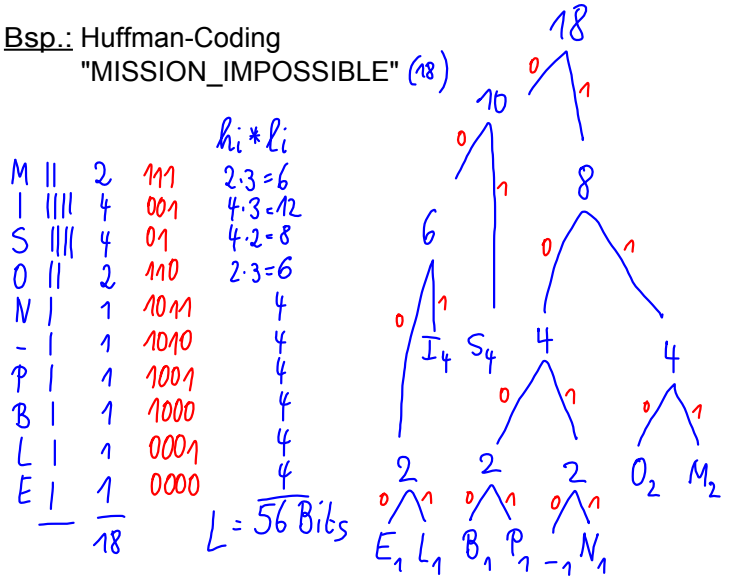
CAVLC=Context Adaptive Variable Length Coding

https://de.wikipedia.org/wiki/Context_Adaptive_Variable_Length_Coding

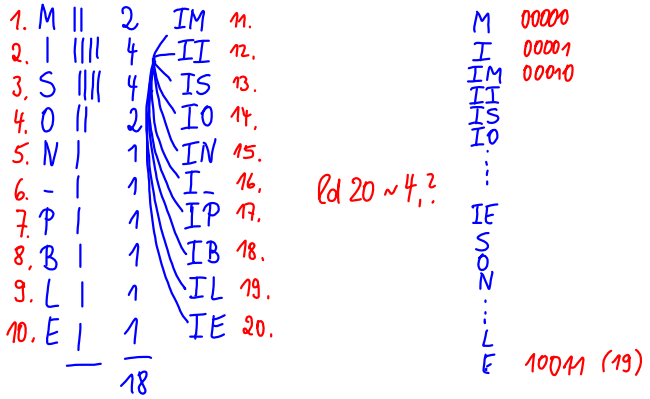
Bsp.: Shannon-Fano-Coding
"MISSION_IMPOSSIBLE" (18)



Bsp.: Huffman-Coding
"MISSION_IMPOSSIBLE" (18)



Bsp.: Tunstall-Coding (äquidistante Codelänge)
"MISSION_IMPOSSIBLE"



Bsp.: arithmetische Coding
"MISSION_IMPOSSIBLE"