

Grundbegriffe der Informationstheorie

Zur Geschichte

Die Informationstheorie wurde Ende der vierziger Jahre des vergangenen Jahrhunderts von Claude E. Shannon begründet. Erstmals hatte damit der Versuch Erfolg, einen Formalismus bereitzustellen, der die flüchtige Größe "Information" zu quantisieren hilft.

Die nachgerade klassische Informationstheorie ist seit Jahrzehnten bei der Beschreibung und Bewertung technischer Datenübermittlungsvorgänge unerlässlich. Darüberhinaus kann sie wohl beanspruchen, als eines der Kerngebiete der *Informatik*, der Lehre der Informationsverarbeitung,¹ zu gelten, wie ja nicht zuletzt der Name dieser immer noch jungen Disziplin andeutet.

In der ursprünglichen Form hatte Shannon für die zu messende "Information" den Begriff der *Entropie* geprägt, der auch heute noch oft benutzt wird.

Ein Maß für Information

Die gesamte Informationstheorie basiert auf stochastischen Modellen. So sinkt beispielsweise der Informationswert eines "Guten Abend allerseits!", wenn es aus dem Mund von Heribert Faßbender kommt, aufgrund seiner Hochfrequenz ab. Ebenso ähnelt sich der Neuigkeits-, sprich Informationswert der Floskel "Das Wetter." aus dem Munde eines Ulrich Wickert seinem ohnehin schon geringen Unterhaltungswert an.

Allgemein wird daher definiert, daß der **Informationswert** $I(x)$ eines Signals oder Zeichens x , das eine bestimmte Quelle X absondert, direkt von seiner Auftrittswahrscheinlichkeit $p(x) = P(X = x)$ abhängt:²

$$I(x) = \log \frac{1}{p(x)} = -\log p(x)$$

Durch die logarithmisch reziproke Form der Beziehung wird gewährleistet, daß der Informationswert selbst eine nichtnegative Größe ist und mit schrumpfendem $p(x)$

¹Mathematiker behaupten gerne, die Informatik sei eine Anwendung der Mathematik. Diese Ansicht muß man nicht teilen; man kann ebenso darauf insistieren, daß die Mathematik erst eine (obgleich fortgeschrittene) Anwendung grundlegendster formaler Symbol-, also Informationsverarbeitung, also Informatik, ist. Es ist müßig, darüber zu streiten, ob dieses Fundament, die formale Logik, zuallererst Mathematik oder Informatik ist. Wo beide Disziplinen nichttrivial werden, läßt sich bald zeigen, daß keine der anderen strukturell vorgelagert ist.

²Statt $P(X = x)$ wäre korrekterweise $P("X = x")$ zu schreiben, um anzudeuten, daß "X = x" nicht der auszuwertende boolesche Ausdruck $X = x$ ist – sonst gäbe es nur zwei Wahrscheinlichkeiten: $P(\text{TRUE})$ und $P(\text{FALSE})$ –, sondern ein *Bezeichner* für das Ereignis, daß X den Wert x annimmt.

wächst.³

In der stochastischen Terminologie wäre X hier eine formale Zufallsvariable, die diskrete Werte x_i annimmt, jeweils mit Wahrscheinlichkeit $p_i = p(x_i) = P(X = x_i)$.

In der Nomenklatur der formalen Sprachen hingegen ist X ein Alphabet aus Zeichen x_i ; die stochastische Verteilung p läßt sich in eine solche Notation $x_i \in X$ nicht integrieren.

Die Entropie, der **Informationsgehalt**, die “Information” einer Informationsquelle X ist nun der *durchschnittliche* Informationswert der abgegebenen Signale bzw. Zeichen x . Dies ist aber genau der Erwartungswert an Information $I(x)$ bzgl. der Wahrscheinlichkeitsverteilung p ,

$$\begin{aligned} H(X) &= \mathcal{E}(I(x)) \\ &= \sum_{i=1}^n p_i \cdot I(x_i) \\ &= - \sum_{i=1}^n p_i \cdot \log p_i, \end{aligned}$$

also der jeweilige Informationswert gewichtet mit der Wahrscheinlichkeit seines Auftretens. Für $\sum_i p_i = 1$ wird dieser elegante und kompakte Ausdruck auch als $H(p_1, p_2, \dots, p_n) := H(X)$ bezeichnet.

Werden mehrere Zufallsvariable X und Y simultan betrachtet, so lassen sich Abhängigkeiten zwischen ihren Informationsgehalten genau analog der stochastischen Notation angeben:

- Ebenso, wie die Wahrscheinlichkeit, daß Y im Falle eines festen X einen bestimmten Wert annimmt, von $P(Y)$ verschieden sein kann, nämlich $P(Y|X)$, so kann auch der Informationsgehalt $H(Y|X)$ von $H(Y)$ abweichen,⁴ wenn X bekannt ist.

Die bedingte Wahrscheinlichkeit des Verhaltens von Y macht dann den Informationsgehalt ebenfalls zu einem bedingten.

- Die Wahrscheinlichkeit, daß X und Y jeweils feste Werte annehmen, wird mit der Verbundwahrscheinlichkeit $P(X, Y)$ beschrieben; analog ist von einer Verbundinformation $H(X, Y)$, die beide Variablen zusammen tragen, die Rede.

Für Wahrscheinlichkeiten wie auch Informationsgehalte gilt in ähnlicher Weise:

$$P(Y|X) = \frac{P(X, Y)}{P(X)} \geq P(X, Y) \quad H(Y|X) = H(X, Y) - H(X) \leq H(X, Y)$$

Hier kann den Formeln für Entropien H noch ihre Herkunft aus Wahrscheinlichkeiten P qua Logarithmierung angesehen werden. Wo sich Wahrscheinlichkeiten multiplizieren – nämlich bei unkorrelierten (stochastisch unabhängigen) Ereignissen –,

$$P(X, Y) = P(X) \cdot P(Y),$$

ist eine Summation der Entropien

$$H(X, Y) = H(X) + H(Y)$$

³Meist wird der Logarithmus zur Basis Zwei (*logarithmus dualis*) gewählt; dann wird die Information in der Einheit bit gemessen. 1 bit, die Antwort auf eine Ja/Nein-Frage, ist gewissermaßen das Basisquantum für die Meßgröße Information.

⁴Im Gegensatz zur Stochastik, wo $P(Y|X)$ größer oder kleiner oder gleich $P(Y)$ sein kann, ist stets $H(Y|X) \leq H(Y)$! Gemäß Shannons Theorie kann man also, anders als im realen Leben, durch beliebige Information (über X) nicht “dümmer” werden als ohne sie ...

zu beobachten, was der Isomorphie

$$\text{Multiplikation positiver Zahlen} \cong \text{Addition ihrer Logarithmen}$$

Rechnung trägt. Lediglich die Ordnungsrelationen sind einander invers, da die Wahrscheinlichkeiten P stets ≤ 1 , ihre Logarithmen demnach nichtpositiv sind.

Beispiel: Beim Würfeln seien zwei Ereignisse, die also jeweils eintreten oder nicht (=“¬”) eintreten können, definiert:

x = ”eine 4, eine 5 oder eine 6 gewürfelt”;

y = ”eine 6 gewürfelt”.

Die Wahrscheinlichkeiten bestimmen sich wie folgt:

$$p_1 := P(x) = \frac{1}{2} \quad p_2 := P(y) = \frac{1}{6}$$

Wegen $y \Rightarrow x$ gilt für die Verbundwahrscheinlichkeit $P(x, y) = P(y)$, und die bedingten Wahrscheinlichkeiten sind somit

$$P(x|y) = \frac{P(x, y)}{P(y)} = \frac{1/6}{1/6} = 1$$

$$P(y|x) = \frac{P(x, y)}{P(x)} = \frac{1/6}{1/2} = \frac{1}{3}$$

Der Informationsgehalt der Zufallsprozesse X und Y ist:

$$H(X) = H(P(x), P(\neg x)) = H(p_1, 1 - p_1) = 1 \text{ bit}$$

$$H(Y) = H(P(y), P(\neg y)) = H(p_2, 1 - p_2) \approx 0,650 \text{ bit}$$

$$H(X, Y) = - \overbrace{P(x, y) \log P(x, y)}^{P(y) \log P(y)} - \overbrace{P(\neg x, y) \log P(\neg x, y)}^0 - \underbrace{P(x, \neg y) \log P(x, \neg y)}_{\frac{2}{6} \log \frac{2}{6}} - \underbrace{P(\neg x, \neg y) \log P(\neg x, \neg y)}_{P(\neg x) \log P(\neg x)}$$

$$\approx 0,431 \text{ bit} + 0 + 0,528 \text{ bit} + \frac{1}{2} \text{ bit} = 1,459 \text{ bit}$$

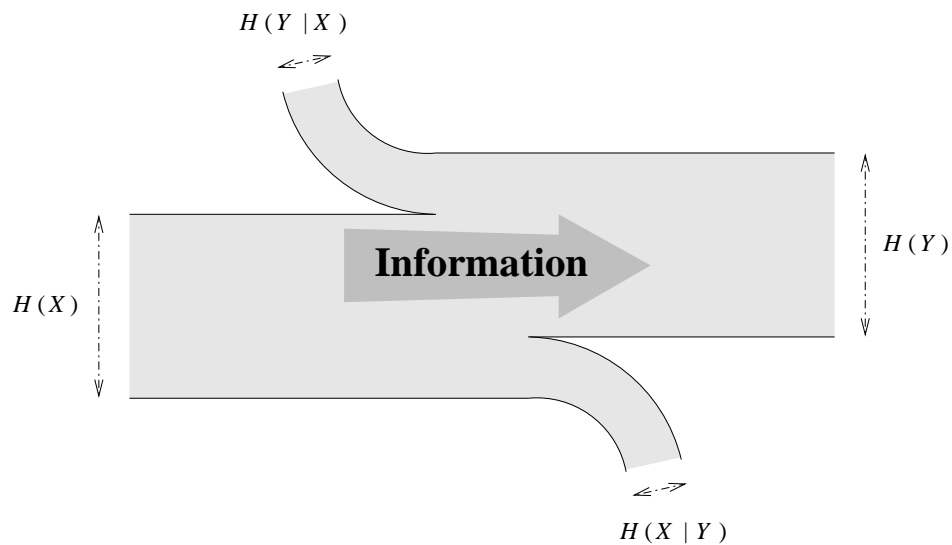
$$H(Y|X) = H(X, Y) - H(X) \approx 1,459 \text{ bit} - 1 \text{ bit} = 0,459 \text{ bit}$$

$$H(X|Y) = H(X, Y) - H(Y) \approx 1,459 \text{ bit} - 0,650 \text{ bit} = 0,809 \text{ bit}$$

Transinformation

Den zentralen Betrachtungen, die die Informationstheorie über die Untiefen der Informationsübermittlung anstellt, liegt nun die Idee zugrunde, diesen Transfer mit zwei stochastisch abhängigen (korrelierten) Zufallsprozessen X und Y – für Informationsquelle und -empfänger – zu modellieren.⁵ Wichtig ist an dieser Stelle nur, daß in interessanten (und leider realistischen) Fällen etwas anderes bei Y ankommt, als X emittiert hat.

Der Vorgang der Übermittlung von Zeichen X zu Empfangszeichen Y – oft sind die Alphabete X und Y (was nicht heißt: die Zufallsvariablen X und Y !) identisch – wird technizistischerweise als “Kanal” bezeichnet und stellt sich im Bergerschen Diagramm schematisch wie folgt dar:



Der Anteil

$$H(Y|X)$$

an der empfangenen Information $H(Y)$ stammt nicht von der Quelle, sondern kommt durch äußere Einflüsse ins Spiel; es ist dies gewissermaßen **Fehlinformation**, die auch Irrelevanz genannt wird. Der verlorengegangene Anteil

$$H(X|Y)$$

an der gesendeten Information $H(X)$ wiederum wird als **Äquivokation** (“Gleichlautendes”) bezeichnet, da er die Mehrdeutigkeit der verbleibenden Information bezüglich der ursprünglich gesendeten bedingt.

Wie das obige Diagramm andeutet, ist

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

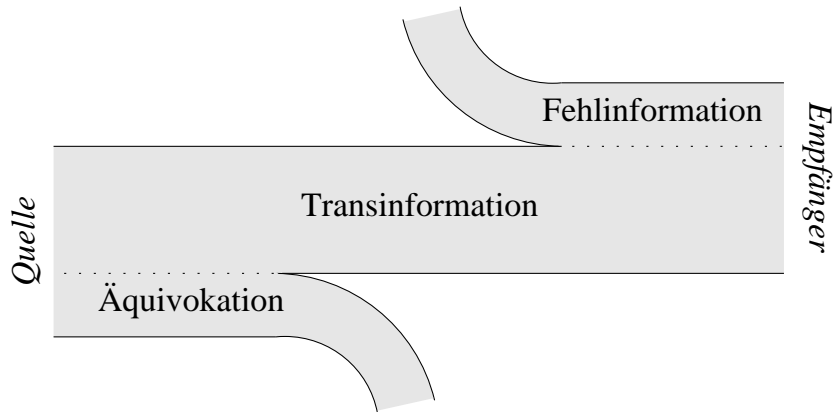
die insgesamt auftretende Gesamt- oder **Totalinformation**, die $H(X) + H(Y)$ nicht überschreiten kann.

Wie ebenfalls dem Diagramm zu entnehmen ist, bezeichnet

$$H(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

⁵Von Spitzfindigkeiten wie der Aufspaltung in Sender, Kodierer, Übertragungskanal, Dekodierer, Empfänger wird in dieser einführenden Abhandlung zugunsten eines ersten Verständnisses abgesehen.

hingegen die tatsächlich übermittelte Information, die sogenannte **Transinformation**.⁶



Ein Übertragungsmodell ist konkret gegeben, wenn alle Auftrittswahrscheinlichkeiten $P(X)$ und alle Übergangswahrscheinlichkeiten $P(Y|X)$ bekannt sind. Zusätzlich zur übrigen stochastischen Größe $P(Y)$ lassen sich hieraus die informatischen Größen

$$H(X), \quad H(Y|X), \quad H(Y), \quad H(X;Y)$$

bestimmen.

In einem solchen Übertragungsmodell, einem "Kanal", kann die Transinformation nicht beliebig hohe Werte annehmen. Sie ist durch die **Kanalkapazität**

$$C = \max_{P(X)} H(X;Y)$$

beschränkt, die durch die Beschaffenheit des Übertragungsmodells sowie durch äußere Einflüsse wie Störwahrscheinlichkeiten (beides geht in $P(Y|X)$ ein) festgelegt ist.

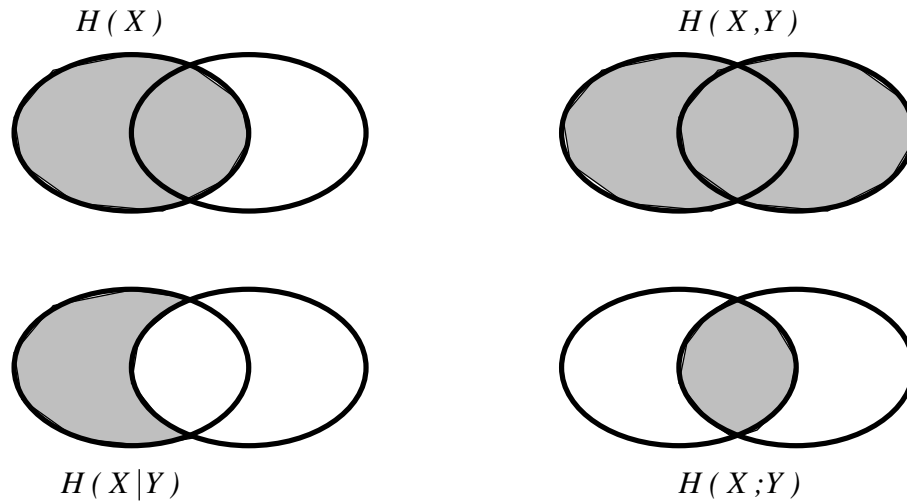
⁶um die Verwirrung komplett zu machen, sei angemerkt, daß sie mancherorts auch als

$$I(X|Y) \quad \text{oder} \quad T(X,Y)$$

notiert wird.

Mengendarstellung der Informationsmaße

Man beachte, daß es sich hier – trotz der übersichtlichen Mengendarstellung – um Informationsmaße, d.h. quantitative Größen handelt; hier wird ihre “Herkunft” bzw. Beschaffenheit klar:



Beispiele für Kanäle:

