Dirk Fox

Entropie

Das Maß für den Informationsgehalt von Nachrichten spielt eine wichtige Rolle in der Informationssicherheit – und wird, besonders bei Passworten, häufig intuitiv völlig falsch geschätzt.

Hintergrund

Für viele Anwendungen der Informatik ist die Bestimmung des Informationsgehalts eines Wertes, Textes oder Datensatzes von erheblicher Bedeutung. Diese Messgröße dient in der Codierungstheorie dazu, optimale Darstellungen (Codierungen) zu finden, um Informationen mit möglichst geringem Speicherbedarf abzulegen – und möglichst effizient auszuwerten.

So ist intuitiv klar, dass die Information „Geschlecht" in einem Datensatz nur zwei Werte annehmen kann. Man kann sie als natürlichsprachlichen Text („Mann" bzw. „Frau") speichern; oder auch verkürzt als „Hr." bzw. „Fr.", noch kürzer als „M" bzw. „F", oder – optimal codiert – als den Zahlenwert 0 bzw. 1. Der Informationsgehalt aller dieser Darstellungen ist identisch (genau ein bit), der Speicherbedarf liegt im ersten Fall beim 32fachen der optimalen Codierung.

Aber auch in der Kryptographie ist die Kenntnis des Informationsgehalts in vielerlei Hinsicht hilfreich: So kann aus dem Informationsgehalt einer verschlüsselten Nachricht manchmal auf das verwendete Verschlüsselungsverfahren geschlossen werden.

Entropie-Bestimmung

Der Begriff der Entropie wurde von Claude E. Shannon (1916-2001) im Jahr 1948 in „A Mathematical Theory of Communication" eingeführt. Der Begriff prägt seitdem die moderne Informationstheorie.

Sein Modell zur Bestimmung des Informationsgehalts ist leicht zu verstehen: Er betrachtet zunächst eine beliebige Zeichenfolge, die sich aus endlich vielen verschiedenen Zeichen, dem Alphabet – das auch aus anderen Zeichen als den uns bekannten 26 Buchstaben bestehen kann – zusammensetzt. Diese Zeichenfolge entspringt, Zeichen für Zeichen, einer Quelle – dem Mund eines Sprechers, einer Nachrichtenübermittlung oder einer Tastatur. Vereinfachend nimmt Shannon an, dass die Quelle kein „Gedächtnis" hat, d.h. die ausgegebenen Zeichen alle (stochastisch) unabhängig von einander sind.

Jedes Zeichen *z* hat nun eine bestimmte Auftrittswahrscheinlichkeit *p(z)*, die sich messen lässt. Aus dieser Wahrscheinlichkeit berechnet Shannon den Informationsgehalt *I(z)* wie folgt:

$$I(z) = -\log_2(p(z))\ bit$$

Das Ergebnis leuchtet intuitiv ein. Hat ein Zeichen – wie die 0 aus einer gleichverteilten 01-Folge, beispielsweise einer Zufallsquelle – die Auftrittswahrscheinlichkeit 0,5 (also 50%), dann liegt der Informationsgehalt der 0 bei einem bit. Liefert die Quelle hingegen reine 1er-Folgen, dann ist der Informationsgehalt der 1 null bit – denn wir wissen schon vorher, dass die Quelle eine 1 ausgeben wird, durch den Empfang der 1 haben wir keine Information hinzugewonnen.

Ausgedrückt in den Begriffen der Wahrscheinlichkeitstheorie ist die Entropie eines Zeichens, die „Ungewissheit", welches Zeichen als nächstes folgt, also der (durchschnittliche) Erwartungswert dessen Informationsgehalts. Je kleiner die Auftrittswahrscheinlichkeit eines Zeichens, desto höher ist sein Informationsgehalt; umgekehrt ist der Informationsgehalt eines Zeichens gering, wenn es oft auftritt.

Die Entropie ist damit zugleich ein Maß für die Informationsdichte eines Zeichens oder einer Nachricht. Im optimalen Fall entspricht die Informationsdichte (der erwartete Informationsgehalt), also die Entropie in bit, der Anzahl an Bits, die für die Darstellung (Codierung) der Information aufgewendet werden. Will man die Entropie einer Zeichenquelle bestimmen, wird der Informationsgehalt jedes einzelnen Zeichens berechnet und alle diese Werte summiert.

Entropie natürlicher Sprachen

Auch für natürliche Sprachen kann die Entropie bestimmt werden. Tatsächlich sind die Buchstabenhäufigkeiten sehr ungleichmäßig verteilt: das „E" kommt wesentlich häufiger vor als alle anderen Buchstaben des Alphabets. Über lange, charakteristische Texte kann der Wert der Entropie angenähert werden. Danach liegt die Entropie der deutschen Sprache bei 4,0629 bit pro Zeichen. Die Verwendung von 26 Zeichen im Alphabet sorgt damit für eine Redundanz von 0,637 bit/Zeichen – dieselbe Entropie könnte man also theoretisch mit einem um vier Zeichen gekürzten Alphabet erreichen.

Noch genauer lässt sich die Entropie einer natürlichen Sprache bestimmen, wenn auch die Auftrittswahrscheinlichkeiten von Buchstabenkombinationen berücksichtigt werden (Paare: Bigramme, Triple: Trigramme, ...). Karl Küpfmüller bestimmte die Entropie der deutschen Sprache im Jahr 1954 auf 1,5 bit je Buchstabe – damit liegt die Redundanz der deutschen Sprache bei etwa 3,2 bit je Buchstabe. Anders ausgedrückt: etwa 68% der Zeichen eines deutschen Textes sind – aus der Sicht des Informationsgehalts der Nachricht – überflüssig.

Entropie eines Passworts

Tatsächlich kommt es insbesondere bei der Wahl eines Passworts auf einen hohen Informationsgehalt an – sonst nutzen auch große Passwortlängen wenig. Wird ein deutsches Wort als Passwort gewählt, liegt der Informationsgehalt nur bei ca. 1,5 bit je Buchstabe. Ein zehn Zeichen langes Passwort hat also einen Informationsgehalt von nur 15 bit. Ohne Sonderzeichen, Ziffern und eine zufällige Passwortwahl reduziert sich die Zahl möglicher Passworte damit auf nur etwa 33.000 – ein Kinderspiel für Passwortcracker.