

DISS ETH 6102

A STUDY ON SIMPLE AND EFFICIENT
BINARY DATA COMPRESSION SYSTEMS

A B H A N D L U N G

zur Erlangung

des Titels eines Doktors der technischen Wissenschaften

der

E I D G E N O E S S I S C H E N T E C H N I S C H E N
H O C H S C H U L E Z U E R I C H

vorgelegt von

P a u l K e e - S h e k W A H

Dipl. El.-Ing. ETHZ

geboren am 15. August 1942

von Taiwan, China

Angenommen auf Antrag von

Prof. Dr. F. E. Borgnis, Referent
Prof. Dr. P. E. Leuthold, Korreferent

1978

A B S T R A C T

In this work the theoretical description and practical realization of data compression techniques are treated in detail. Both reversible and irreversible source encoding systems are studied. In addition to the computer simulations of different well known coding schemes, the major effort of this work consists in formulating different schemes mathematically so that they can be judged objectively. As a figure of merit, source encoding efficiency W_r (for redundancy reduction) and W_e (for entropy reduction) are introduced. Closed formulas of these efficiencies are found for almost all the discussed algorithms, allowing a quick comparison of the real system performances in different situations. Simple and efficient source encoding systems are then developed. Essential results are as follows: For one-dimensional redundancy reduction of both dependent and independent sequences, the run-length code with bit-overflow (BRLC) seems to be the best. However, if extremely high efficiency of compressing high order Markov chains is required, the standard decorrelation encoding (SDE) appears to be very suitable. For two-dimensional redundancy reduction, the consecutive decorrelation encoding (CDE) promises to be both practical and efficient. If transmission errors are encountered, effective strategy is also proposed. So far as entropy reduction of binary symmetric source is concerned, linear block codes with block length $n \leq 63$ are recommended, achieving a moderately high efficiency W_e of about 90 %. If the distortion d prescribed by the receiver is lower than 0.125, as in cases of practical interest, these linear block codes appear to be more advantageous than the involved tree-coding. For an asymmetric source, the transmission rate obtained by the constant block distortion encoding (CBDE) is very close to the Shannon bound $R(d)$, especially if the appearance probability p is smaller than 0.2.

Z U S A M M E N F A S S U N G

Der Hauptvorteil digitaler Kommunikationssysteme besteht neben der hohen Arbeitsgeschwindigkeit der integrierten Schaltkreise, hoher Systemzuverlässigkeit, geringem Raumbedarf, Kompatibilität mit verschiedenen Datenträgern und Datenarten sowie der Möglichkeit einer Zeitmultiplex-Uebertragung und -Vermittlung darin, dass sich die diskreten Nachrichtensignale längs des Uebertragungswegs vollständig regenerieren lassen. Angeregt unter anderem durch die Fortschritte in der Technologie der Large-Scale-Integration (LSI) werden heute immer mehr digitale Nachrichtenübertragungssysteme entwickelt, was zu einem immer grösseren Anfall von Daten führt. Die Verarbeitung, Speicherung und Uebertragung immer grösserer Datenmengen bringt neue Probleme mit sich, die gelöst werden müssen. Eine Reduktion der Daten durch gezielte Auswahl der wichtigen oder relevanten Informationen ist eine erste Massnahme. Sie gründet sich auf den Umstand, dass ein beträchtlicher Teil der digitalisierten Signale aus den Nachrichtenquellen wie Fernsehen, Sprache oder Faksimile unwesentliche Information enthält, weshalb bei einer zweckmässiger Eliminierung Qualität und Verständlichkeit der übertragenen Signale nur wenig beeinträchtigt werden. Auch im Hinblick auf die Zuverlässigkeit und Wirtschaftlichkeit der Datenübertragung spielt diese Datenreduktion eine bedeutende Rolle. Einerseits wird die Störfestigkeit der Daten bekanntlich sehr stark geschwächt, wenn man sich auf die Uebertragung des wesentlichen Teils einer Nachricht beschränkt, andererseits schützt aber die inherente Redundanz die Nachricht nur unzureichend. Deshalb besteht die optimale Strategie im Sinne der Informationstheorie darin, die überflüssige Information (Redundanz und Teile der vernachlässigbaren Entropie, manchmal auch Irrelevanz genannt) zunächst möglichst vollständig zu eliminieren und erst dann wieder eine "zweckmässige Redundanz" zuzusetzen. Diese Redundanz kann dann sehr gut an die Eigenschaften und Gegebenheiten des Uebertragungskanal angepasst werden, sodass ein optimaler Schutz gegen eventuelle Störungen im Uebertragungsweg gewährleistet ist. Ein solches Hinzufügen von

Redundanz ist Gegenstand der Kanalcodierung und wird heute weitgehend beherrscht. Die vorangehende Redundanz- und Entropie-Reduktion ist dagegen ein Problem der Quellencodierung. Es ist bis heute noch nicht vollständig befriedigend gelöst worden. Es wird hier hauptsächlich versucht, einen Beitrag zu diesem Themenkreis zu leisten.

Unter Quellencodierung versteht man eine Abbildung zwischen Quellen- und Codesymbolen mit dem Ziel, möglichst wenig Codesymbole pro Quellensymbol zu verwenden, um Speicherplatz, Kanalbandbreite oder Sendeleistung einzusparen. Man unterscheidet zwischen zwei Arten von Quellencodierung: der reversiblen und der irreversiblen Codierung. Die erste Art dient zur Redundanzreduktion einer Nachrichtenquelle, wobei die ursprüngliche Information vollständig aus den komprimierten Daten zurückgewonnen werden kann (abgesehen von den seltenen Störungen bei Buffer-Overflow). Dies setzt jedoch voraus, dass die Quelle Redundanz aufweist. Hingegen wird bei einer irreversiblen Quellencodierung nicht nur die Quellen-Redundanz reduziert, sondern auch die Entropie teilweise verringert. Selbstverständlich wird nur jener Teil der Entropie weggelassen, der für den Empfänger nicht wesentlich ist. Dieser mit Absicht ignorierte Anteil von Information beeinträchtigt die Güte der Nachrichtenübertragung, die deshalb vom Empfänger vorgeschrieben werden muss. Die Grösse der tolerierbaren Verzerrungen (Distortion d) wird als Qualitätskriterium (Fidelity Criterion) bezeichnet. Die notwendige Übertragungsgeschwindigkeit R wird dann um so kleiner sein, je grösser diese Toleranz ist. Die Beziehung zwischen R und d wird "Rate-Distortion-Funktion $R(d)$ " genannt und wurde von Shannon erstmals schon im Jahre 1959 souverän behandelt. Seither bemüht sich die Fachwelt mit nicht allzuviel Erfolg, diese Erkenntnisse in der Praxis anzuwenden. Viele Fragen sind noch ungelöst, so z. B. der Verlauf von $R(d)$ von Markovschen Ketten oder das praktische Vorgehen bei der Annäherung von $R(d)$. Für die letztere Frage zeichnet sich neuerdings eine Lösung ab, nämlich das sog. "Tree-Coding mit Viterbi-Decodierungsverfahren", das jedoch einen sehr grossen Aufwand erfordert, auch wenn man sich auf binäre Daten ohne innere statistische Abhängigkeiten beschränkt.

Die vorliegende Arbeit gliedert sich hauptsächlich in zwei Teile, wobei zuerst die Reduktion der Redundanz und anschliessend die Reduktion der Entropie untersucht werden. Nach eingehender Behandlung der zugehörigen Theorie folgen ein Bericht über den Stand der Technik und Computer-Simulationen verschiedener Datenkompressionssysteme, die einen Vergleich der hier beschriebenen neuen Verfahren mit bereits bekannten Methoden erlauben. Das Schwergewicht wurde darauf gelegt, geschlossene Formeln für verschiedene Systeme zu finden, um eine objektive Beurteilung zu ermöglichen. Dazu werden zwei Begriffe eingeführt: der Wirkungsgrad W_r für die Redundanzreduktion und der Wirkungsgrad W_e für die Entropiereduktion. Die Untersuchung erstreckt sich auch auf Daten mit inneren statistischen Abhängigkeiten. Grosser Wert wird auf die Einfachheit und Wirksamkeit der vorgeschlagenen Systeme gelegt. Ausserdem wird auch die Puffer-Organisation und die Kanalstörungen-Protektion behandelt.

Im folgenden sollen einige wesentliche Resultate angegeben werden. Für die Redundanzreduktion eindimensionaler Markovscher Sequenzen niedrigerer Ordnung scheint der "Run-Length Code mit Bit-Overflow" (BRLC) am zweckmässigsten zu sein. Für Markovsche Sequenzen höherer Ordnung ist das "Standard Decorrelation Encoding" (SDE) sehr gut geeignet. Im Falle der zweidimensionalen Redundanz-Reduktion ergibt das "Consecutive Decorrelation Encoding" (CDE) oft bessere Resultate. Was die Entropie-Reduktion binärer symmetrischer Quellen anbetrifft, so ist die Verwendung von linearen Gruppencodes mit Block-Längen $n \leq 63$ zu empfehlen. Damit wird schon ein relativ hoher Wirkungsgrad W_e von etwa 90 % erreicht. Falls die vorgeschriebene Verzerrung d kleiner als 0.125 ist, sind diese einfacheren Gruppencodes den komplizierten Tree-Codes vorzuziehen. Für asymmetrische Quellen scheint das "Constant Block Distortion Encoding" (CBDE) sehr gut geeignet zu sein, vor allem, wenn die Auftretenswahrscheinlichkeit p kleiner als 0.2 ist.