Original paper completed in 1987 to support TEMPEST class.

Basic Information Theory and the Use of IR's

Bruce C. Gabrielson, PhD and George C. Caldwell

Security Engineering Services
5005 Bayside Road
Chesapeake Beach, MD 20732

Introduction

There is a considerable amount of confusion existing in the TEMPEST industry related to information theory, information ratios, and the application of ratios to practical analysis problems. This paper attempts to help clarify the confusion in an unclassified format by explaining the requisite background information regarding the theory before developing an unclassified figure of merit (an IR) for breaking down detected signals into useful information. Other figure of merits may have been developed for specific purposed, but the information and techniques presented herein are generic, and will help in the understanding of other similar values.

Information

Information is a quantitative term, measured by the degree to which is clarifies that which is unknown; a totally predictable event contains no information. In general, information has the property of reducing the uncertainty of a situation. The uncertainty is called entropy (H), and entropy exists to the extent that information is lacking (Information + Entropy = 100% or I + E = 1). If the entropy of a situation is small, only a small amount of information is required to clarify it. If the entropy is large, then much more information will be required before the uncertainty can be replaced by an acceptable degree of clarity.

If the probability of an event E occurring is P(E), then the information obtained when E occurs is:

$$I(E) = \log \frac{1}{P(E)}$$

with the choice of base for the logarithm corresponding to the unit of information (i.e. base 2 for bits, base 10 for Hartleys), where 1 Hartley is equal to 3.22 bits. In this case, there is a clear distinction between a bit as a unit of information or a binary number, and a baud as a unit of signaling speed.

Baud is characterized by the presence or absence of a pulse in a channel, and is a measure of the maximum rate of pulses (code elements) per second in the system. Baud rate is found by taking the reciprocal of the length (in seconds) of the shortest pulse
used in creating a character. As an example, the length of a pulse of baudot code as used with 60 wpm teletypewriters is 0.022 seconds (22 ms). The baud rate is the reciprocal (1/0.22 =45.45 bauds), which is the highest possible baud rate of this signal. Figure 1 describes the condition. In TEMPEST work, bauds are not considered.

Bit is the abbreviation for binary digit, with both binary states called bits, since both states carry the same amount of information. The number of bits required to identify any particular selection from a group of N possible selections is:

$$I = \log_2 N$$

provided that all N selections have equal opportunity of being chosen.

Original paper completed in 1987 to support TEMPEST class.

The term "byte" is used to describe a group of consecutive bits that are treated as a unit. Most computers are designed to use byte-sized characters of eight bits. The transmission of both bits and characters are usually measured in terms of so many per second, whereas words are measured in so many per minute. Figure 2 provides conversions between the various ways to describe information transfers.

Data is transferred across communications channels using various baseband coding techniques, depending on the number of lines available and the application. Tables 1 and 2 describe ASCII and other common codes.

Messages

A message, in information theory, is merely the output of some information source. The quantitative value of a message is based on several factors. First, it must be established how much was known about the contents of the message before it was received. Second, it must be known how many messages were in the set from which the message was chosen. Third, to be precise, there should be some way of knowing the probability of each event that the message could describe.

A few examples may help clarify what is being described. If the source is a telephone transmitter, the message would be the analog voltages impressed on the telephone line. If the source is a teletypewriter, the message could be a character, one of the bits making up a character, or the message could be the entire word. The composition of a message can therefore be a number of things depending on how it is defined.

A further breakdown relates to language. If the message is in commonly spoken English, the message consists of a reduced set of words, each made up of letters of the English alphabet. If the message is in computer language, another significant reduction exists in the number of probabilities possible.

Assuming that the probability of occurrence of a symbol in the source language does not depend on the symbol(s) preceding it (a definition of a zero-memory source), each symbol gives information equal to:

$$I(s_i) = \log \frac{1}{P(s_i)}$$

and the average information per symbol is:

$$H(S) = \Sigma_i P(s_i) I(s_i) \text{ ; H(S) is the entropy of the source}$$

Obviously, a zero-memory source is not realistic in a communications context. In English, the probability of a "h" occurring is greater is the preceding symbol is a "t" rather than an "l". Therefore, the sources considered here will be ergodic Markov sources, sources where the occurrence of a symbol depends on a finite number (the order of the Markov source) of preceding symbols, and where if the source input is long enough, a typical sequence <u>will</u> occur.

From various empirical sources, the entropy of English (26 letters and a space) is:

(1)	As a zero memory source	4.03 bits/symbol
(2)	As a first order Markov source	3.32 bits/symbol
(3)	As a third order Markov source	3.10 bits/symbol

Shannon, the father of information theory, has by theoretical methods estimated the entropy of actual English to be between .06 and 1.3 bits/symbol.

Codes

Original paper completed in 1987 to support TEMPEST class.

Regardless of code used (ASCII or even COMSEC), the intent of reducing uncertainty is the decoding a of symbol sequence. Even though English is a code for thought, the information above is not adequate to vigorously describe an encoded information channel. In this regard, a better description of code requirements is necessary.

The codes we will be considering are Instantaneous, Uniquely Decodable, and Non-Singular Black Code. These codes are described below.

Block Code - A code where every source is represented by a finite fixed sequence of code symbols, called a code word.

Non-Singular - All code words are unique.

Uniquely Decodable - For every finite sequence of the source alphabet, the encoded sequence is a Block Code. For example, binary numbers with leading zero suppressed are not uniquely decodable (100 could be 4 or 1,0,0 or 2,0).

Instantaneous - Can be decoded without decoding succeeding sequences.

Based on the preceding definitions, a Kraft inequality can be formed for a binary code which states:

$$\sum_{i=1}^{q} 2^{-l_i} \le 1$$

where q = number of code words
2 = number of symbols in code alphabet
l_i = length of i(th) code word

The proof of the Kraft inequality is involved, and left to those who are interested. For our purposes, the average code length is the important consideration. From the Kraft inequality, for source symbols with probabilities $p_1, ...p_q$ and code word lengths $l_i, ... l_q$:

the average code length $L = \sum_{i=1}^{q} p_i l_i$.

This is valid for uniquely decodable code from either zero-memory or Markov sources. Going back to the definition of entropy,

$$H(S) = \sum_{i=1}^{q} p_i \log p_i \qquad \log p_i = \frac{1}{p_i}$$

we can see that: $H(S) \le L$

Without going into an involved proof, it is intuitive that the longer the code required for a source, the less information per code symbol. The equation states the lower bound of the problem.

With a lower bound, it is possible to choose a code word length l_i such that summed over the code

$$H(S) \le L \langle H(S) + 1$$

If the above is applied to a Markov source, n symbols at a time, knowing the entropy of Sn = n *S,

$H(S) \le Ln/n \langle H(S) + 1/n$ Ln/n is the average number of code symbols, coding this source in groups of n

The above expression leads to Shannon's first Theorem, the noiseless coding theorem:

$$\lim_{n\to\infty} Ln/n = H(S)$$

or the number of code symbols per source symbol can be no smaller than the entropy, or information, of the source.

Based on the above determination, the redundancy of a code can be defined. If the efficiency is the ratio of information to code length H(S)/L, then:

$$\text{redundancy} = 1 - \text{efficiency}$$

$$= L - \frac{H(S)}{L}$$

An Information Channel and Shannon Capacity

The mathematics involved to this point is nice, but much too theoretical to be of use to a TEMPEST engineer. We need to now direct our discussions to a more realistic realm, the definition of an information channel. An information channel consists of an input alphabet A = {a(i)} where i = 1,2, ...r, an output alphabet B = {b(i)}i where i = 1, 2, ...r, and a set of conditional probabilities p[b(j)/a(i)] such that b(j) will be received in a(i) is sent for all j and i. This allows the introduction of noise into a communication.

Let's first look at channel noise power. As signal and noise approach the same power level, with constant channel bandwidth, the signal must exist for longer periods of time in each discrete state in order to be detected. The theoretical maximum bit rate C, through a channel of bandwidth (BW) and signal-to-random-noise power ratio S/N is:

$$C = BW \, Log_2 (1 + S/N)$$

The S/N power ratio indicates the relative strength of the signal to that of channel noise.

$$Log_2 X = Log_2 10 \, Log_{10} X$$

$$Log_2 10 = 3.32193 = 1/Log_{10} 2$$

In the presence of noise, a binary signal is obviously more easily detected than one using several bits per code element. The required minimum S/N power ratio from a known bit rate and bandwidth is:

C/BW

Original paper completed in 1987 to support TEMPEST class.

$$S/N = 2$$

where C is the maximum bit rate through a channel.

As the bit content (number of levels) of a code element is increased, a corresponding increase in the S/N power ratio must be made to maintain equal detection capability relative to a binary signal. For a common binary (digital) channel, S/N = 3, or signals 4.8 dB above the noise level are directly detectable.

Noise uncertainty leads also to effects on probability calculations. Using arguments parallel to those for zero-memory and Markov sources, we can arrive at the channel equivocation:

$$\frac{Ln/n}{\lim n \to \infty} = H(A/B)$$

It must be stated that channel equivocation applies only to instantaneous uniquely decodable codes. Channel equivocation is the average knowledge gained from observing an output symbol in B produced by an input symbol for A. Conversely, H(A/B) bits are required to produce an output symbol in B from an input symbol in A. Since it takes H(A) bits to define an input symbol in A, the average information gained from receiving a symbol in B is:

$I(A,B) = H(A) - H(A/B)$ the mutual information channel.

The mutual information of a channel contains less information than the source because of the noise uncertainties. This decrease in information from the observed signal reduces the efficiency of communication transmission in a manner similar to code efficiency.

$$\frac{\text{Channel information}}{\text{source code information}} = \frac{H(A) - H(A/B)}{H(A)}$$

The transmission efficiency is less than 1, and mitigates against communication. If a redundancy is built into the source code, either by some error checking protocol, or by semantics, as in the case of natural languages, a communicability factor effects efficiency.

$K^* \; \dfrac{1}{\text{redundancy}}$ *transmission efficiency

$K \; \dfrac{L}{[L-H(A)]} \; \dfrac{[H(A) - H(A/B)]}{H(A)}$ arises where K is an empirical, context dependent factor.

This communicability factor is predicated upon observing the output signal. If an unobserved channel is postulated, as would be the case in encryption schemes, communicability becomes:

$$K\left\{\frac{l}{L-H(A)}\right\}\left\{\frac{H(A) - H(A/B)}{H(A)}\right\} - H(K)$$

The Gabrielson/ Caldwell IR (GCIR)

and can be reduced to any level. The above equation can be considered an information ratio figure of merit for information detectibility. Negative values of IR would indicate a "red herring" strategy. Cascaded channel cases have not been addressed for the sake of simplicity of presentation. Also, since mutual information is additive, signals can be expressed directly by the above communicability factor.

Conclusion

Original paper completed in 1987 to support TEMPEST class.

The use of information ratios, IR's, provides a statistical measure of the information content of a recovered signal. One important factor is the redundancy of the language being used. English is approximately 75% redundant. Other factors include the probability that certain characters are part of the massage.

Rigerously applying an IR to a practical problem is not an easy task, since a solution involves the use of a much higher level of mathematics than used thus far in this derivation. The basic intent was to show that information might exists in many formats, coded emissions being but one such media. The mear detection of a coded signal in a media is not sufficient to determine if enough information is present to identify the message being sent.