

# Unifying the Interpretation of Redundant Information

Rocchi Paolo, IBM, via Shangai 53, Roma Italy, paolorocchi@it.ibm.com

**Keywords:** Redundancy, control, information theory, reliability theory

**Received:** November 2001

*This paper discusses the possibility of interpreting redundant information beyond the particular views emerging in specialist sectors. We introduce a theoretical framework that aims at unifying and calculating the main features of redundant information. This theoretical layout has been introduced in professional tuition.*

## 1 Introduction

People tried to handle redundant information from immemorial times. For example, copyists introduced several abbreviations and writing simplification in order to reduce the language redundancy.

Specialists did not tackle redundancy by rigorous methods until the early twentieth century when telegraph and telephone networks, radio emitters began to connect towns, then nations and continents. Infrastructures for telecommunication involved heavy investments and economic pressures drove engineers to optimize the use of these facilities. H. Nyquist and others started to search for optimal transmission and finally C. Shannon established the fundamental laws of data compression and marked the birth of the information theory [1].

These authors accomplished their purposes and aided the progress of technology, but the thorough comprehension of redundancy still remains an open question. Writers brought this problem to light nearly fifty years ago [2] [3] [4], although a formal theory on redundancy does not seem to attract mathematicians' attention so far. The debate remains on the philosophical plane, for example see the initial "Theory of Redundancy" and the next "Deflationary Theory of Truth" [5]. Modern advances in computer science, especially in the Internet, press toward the rigorous comprehension of the different forms of redundancy. We select three essential points from the queries that thinkers have raised.

a) Redundancy regards any kind of information and we question whether results pertaining to the binary technology may be extended to other forms of information [6]. The evidence should prove the contrary.

b) The entropy and the redundancy factor quantify redundancy of digital information. As they are logically disparate, we should integrate them into a comprehensive notion expressed by the mathematical language.

c) Redundancy increases the reliability of data during transmission and storing and also improves the machinery reliability. The relationship between redundant codes and redundant systems should be fully clarified in order to achieve the general and exhaustive knowledge of redundancy.

I was persuaded that these ample themes should be handled within a unifying logic and have driven a theoretical research for years. This paper sets out some results and tries to answer the above points.

## 2 Redundancy

Let the set  $\{\varepsilon\}$  include the entities  $\varepsilon_1, \varepsilon_2, \varepsilon_3, \dots, \varepsilon_n$ . We assume that two pieces of information are the entities  $\varepsilon_i$  and  $\varepsilon_j$  that have the property of being distinct

$$\varepsilon_i \neq \varepsilon_j \quad i, j = 1, 2, \dots, n \quad (2.1)$$

The item of information  $\varepsilon_x$  stands for something and we assume that the meaning of information is the function  $\mu$  of representing  $\alpha$ .

$$\varepsilon_x \bullet \text{-----} \bullet \alpha \quad \mu \quad x = 1, 2, \dots, n \quad (2.2)$$

We could say that  $\mu$  is the main job of  $\varepsilon_x$  or, in other terms,  $\varepsilon_x$  works as a model.

The statements (2.1) and (2.2) formalize two ideas universally shared in current literature. Notably they establish that information is distinct and has semantic

properties. Discrete formalism is usual in information technology (IT) and the pair (2.1) and (2.2) follows this vein.

The word “redundant” derives from Latin and hints something abundant and repetitive with respect to its use. I put forward the following definition of the redundancy in accordance to this naïf idea.

*Definition 2.1: The set  $\{\varepsilon\}$  is minimal when the number of informational entities is equal to the number of the objects to be represented*

$$n = n_\alpha \tag{2.3}$$

*It is insufficient when*

$$n < n_\alpha \tag{2.3 bis}$$

*When*

$$n > n_\alpha \tag{2.3 tris}$$

*The set  $\{\varepsilon\}$  is redundant, notably the information surplus provides the redundancy of  $\{\varepsilon\}$*

$$R = n - n_\alpha \tag{2.4}$$

The more  $R$  is high and the more  $\{\varepsilon\}$  is redundant. Redundancy is null if (2.3) is true and is negative if (2.3bis) is true. In substance  $R$  gives the excess (when positive) and the lack (when negative) of the models  $\varepsilon$ . Redundancy is null when the representations are just enough. For ease, the picture of the car in the web page and three phrases invite the web-visitor to buy the car. Four pieces of information have the same promoting significance and make the message redundant

$$R = 4 - 1 \tag{2.5}$$

Any item of information may be modeled as an algebraic entity and definition (2.4) begins to respond to the question b).

### 3 Methods

Engineers follow two major approaches in order to ensure the reliability of a system. The former provides remedies after the failure has occurred. The latter method is precautionary and precedes the damage

*Method (1): Repairs the failure.*  
*Method (2): Prevents the failure.*

Redundancy is a precautionary solution against information failures and falls into (2) [7]. IT specialists

preliminarily take care of errors, noise and random irregularities, which will injure transmission and storage. When the information set  $\{\varepsilon\}$  is redundant, (2.3ter) yields two possibilities

*Method (2.1): All the items of  $\{\varepsilon\}$  are used as models.*  
*Method (2.2):  $R$  items are not used as models.*

We detail these either-or ways.

*Method (2.1) -* When (2.3ter) is true and all the items of information are used, at least one object  $\alpha$  has  $k$  models

$$k \geq 2 \tag{3.1}$$

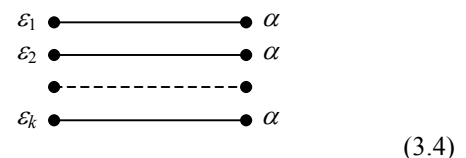
Let  $P$  the probability of altering one piece of information, the probability of altering  $k$  items of information is

$$P_k = P^k \tag{3.2}$$

As  $P$  is lower than unit,  $k$  pieces of information, which stand for one entity  $\alpha$ , are more reliable than only one item of information

$$P_k < P \tag{3.3}$$

This graph, derived from (2.2), visually evidences how  $k$  items of information representing the same object are similar to  $k$  units working in parallel



Expressions from (3.1) to (3.3) are formally symmetrical to the formulas that calculate  $k$  machines in parallel and they reach the same conclusions [8]. In short, *Method (2.1)* answers point c). The present theory shrinks the gap between the information theory and the reliability theory.

*Method (2.2) -* When (2.3ter) is true and  $R$  pieces of information are unused,  $\{\varepsilon\}$  splits into two separate subsets

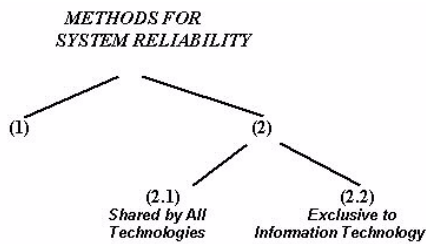
$$\{\varepsilon_u\} \cap \{\varepsilon_z\} = \emptyset \tag{3.5}$$

The subset  $\{\varepsilon_u\}$  includes  $n_\alpha$  pieces of information with precise significance. The subset  $\{\varepsilon_z\}$  has  $R$  meaningless items; hence the entities differ from the semantic viewpoint

$$\varepsilon_u \neq \varepsilon_z \tag{3.6}$$

They apply (2.1) thus engineers can follow a special method, which is exclusive to IT and cannot be compared elsewhere. They prepare the subset  $\{\varepsilon_u\}$  which is meaningful and  $\{\varepsilon_z\}$  meaningless. If the control

detects the unmeaning piece  $\varepsilon_z$ , it reveals an error. This technique, based on the inequality (2.1), is exclusive to information, while *Method* (2.1) is universal.



This framework clarifies the relations existing between the information sector and other engineering fields, and elucidates point c).

### 4 Redundant codes

We restrict our attention to *Method* (2.2) and in particular examine the redundancy of codes. We assume the length  $L$  is fixed for the sake of simplicity. The combinatorial calculus provides the ensuing result

$$n = B^L \tag{4.1}$$

Where  $B$  is the base of the code  $\{\varepsilon\}$ . Now we consider the *minimal code*  $\{\varepsilon_\alpha\}$  that, in accordance to (2.3), is the set of codewords just sufficient to represent  $n_\alpha$  objects. The base  $B$  and the minimal length  $L_\alpha$  allow us to calculate  $n_\alpha$ . This quantity along with (4.1) makes explicit the redundancy (2.4)

$$R = B^L - B^{L_\alpha} \tag{4.2}$$

As  $B$  exceeds the unit,  $\{\varepsilon\}$  is redundant if and only if  $L$  is larger than the minimal length  $L_\alpha$

$$R > 0 \iff L > L_\alpha \quad B \geq 2 \tag{4.3}$$

This result proves that the redundancy of a digital code relies on its length.

1] Expression (4.3) suggests to calculate the *digital redundancy*  $R_D$  of  $\{\varepsilon\}$  by the difference of lengths

$$R_D = L - L_\alpha \tag{4.4}$$

2] The *redundancy factor*  $R_C$  of the code  $\{\varepsilon\}$ , already in use, relates the length  $L$  with respect to the minimal length

$$R_C = \frac{L}{L_\alpha} \tag{4.5}$$

We make explicit  $L$  and  $L_\alpha$  with combinatorial calculus and we put them into (4.5)

$$R_C = \frac{L}{L_\alpha} = \frac{\log_B(n)}{\log_B(n_\alpha)} = \log_{n_\alpha}(n) \tag{4.6}$$

This result evidences that  $R_C$  depend on  $n$  and  $n_\alpha$ . This property also regards  $R_D$  that has the same variables of  $R_C$ . Both of them are coherent with (2.4) in point of mathematics. They differ on the practical plane:  $R_D$  and  $R_C$  regard digital information instead  $R$  has general usage. This theory brings to light the relations existing between various measurements of redundancy and clarifies point b).

The base  $B$  and  $n_\alpha$  objects are usually given in the professional environment, hence the length  $L_\alpha$  is the essential reference for digital calculations. Combinatorial calculus provides this result

$$L_\alpha = \log_B(n_\alpha) \tag{4.7}$$

That although neglects the frequency of the codeword. Shannon has the merit of discovering the accurate value of  $L_\alpha$  and calculating it by means of the entropy

$$H = -C \sum_i^{n_\alpha} P_i \cdot \log_B(P_i) \tag{4.8}$$

Where  $C$  is a positive constant and

$$\sum_i^{n_\alpha} P_i = 1 \tag{4.9}$$

$H$  provides the rigorous minimal length and brings evidence of the roughness of (4.7). In fact, if the codewords of the code  $\{\varepsilon\}$  are equiprobable

$$P_i = 1/n_\alpha \tag{4.10}$$

The entropy equals to (4.7) up to the constant  $C$

$$\begin{aligned} H &= - \sum_i^{n_\alpha} 1/n_\alpha \log_B(1/n_\alpha) = \\ &= \log_B(n_\alpha) \end{aligned} \tag{4.11}$$

May be proved that this result is the maximum of  $H$ , when  $n_\alpha$  and  $B$  are given [1]. In short the Shannon entropy provides  $L_\alpha$  in general, while (4.7) is true only if the codewords are equiprobable.

### 5 Technical refinements

Let codewords be duplicated, tripled etc.

$$R_C \geq 2 \tag{5.1}$$

High redundancy entails high reliability because the detection of errors is immediate. Per contra volumes are bulky and this range confines the problem

$$R_C < 2 \quad (5.2)$$

The small number of characters hinders the detection of errors and engineers refine *Method (2.2)* by means of the algorithm, which enhances the control.

This solution although does not rule out the possibility of an input unrecognizable by the algorithm. The generic codeword may be modified during the transmission or the storing to the extent that it could neither belong to  $\{\varepsilon_u\}$  nor to  $\{\varepsilon_z\}$ . In this case, the *Method (2.2)* flops.

Engineers cure the problem and state that the subsets  $\{\varepsilon_u\}$  and  $\{\varepsilon_z\}$  be mutually exclusive. Using the set theory we write the following constraint

$$\{\varepsilon\} = [\{\varepsilon_z\} \cap \{\varepsilon_{uC}\}] \cup [\{\varepsilon_u\} \cap \{\varepsilon_{zC}\}] \quad (5.3)$$

Where  $\{\varepsilon_{zC}\}$  and  $\{\varepsilon_{uC}\}$  are the complementary subsets of  $\{\varepsilon_z\}$  and  $\{\varepsilon_u\}$ . As (3.5) is true we have

$$\begin{aligned} \{\varepsilon_u\} \cap \{\varepsilon_{zC}\} &= \{\varepsilon_u\} \\ \{\varepsilon_z\} \cap \{\varepsilon_{uC}\} &= \{\varepsilon_z\} \end{aligned} \quad (5.4)$$

And finally we get

$$\{\varepsilon_u\} \cup \{\varepsilon_z\} = \{\varepsilon\} \quad (5.5)$$

This equation along with (3.5) establishes the *Excluded Middle Principle*. The bits and the binary codewords comply with this constraint. A binary word is necessary included either in the subset  $\{\varepsilon_u\}$  or in  $\{\varepsilon_z\}$ . Specialists elaborate the most advanced control techniques thanks to this special property [9] which provides the answer to question a).

## 6 Conclusions

Some authors pursue complex studies about redundancy on the philosophical plane but the conclusions appear generic to engineers. I have searched for the replies to the initial queries by means of the mathematical language and believe that this feature may be appreciated.

These pages stage the examination of redundancy and progressively come from the most ample themes to the specialist ones. In detail:

- This paper proposes the redundancy definition (2.4) which relates the physical nature of information to the semantics. This unitary approach gives an answer to point b).
- This work puts forward the redundancy  $R_D$  of digital words which is verbally expressed so far, and relates it to  $R_C$  and to  $R$ . These discourses try to clarify point b).

- Equation (3.3) and scheme (3.4) elucidate the links between the informational redundancy and the reliability theory as point c) demands.
- The *Excluded Middle Principle* (5.5) and (3.5) explain why technicians can develop very sophisticate binary solutions as point a) presumes.

The cultural meaning of the present work has proved to possess valid educational qualities. They have been partially taught in high schools and in basic training in IBM.

An ample theory on information, systems and control includes the equations presented in this paper [10] and this is the last feature, which I aim at highlighting.

## References

- [1] Shannon C., Weaver W. (1949) - *The Mathematical Theory of Communication* - Univ. Illinois Press, Urbana.
- [2] Calderbank R. ed. (1996) - *Different Aspect of Coding Theory* - American Mathematical Society.
- [3] Kriebel H.C. (1965) - *A Resume of Mathematical Research on Information Systems*. - Carnegie Institute of Technology, Pittsburg.
- [4] Cherry C. (1996) - *On Human Communication: a Review and a Criticism* - MIT Press, Cambridge.
- [5] Field H. (1986) - The Deflationary Conception of Truth - in MacDonald G and Wright C. (eds.) *Fact, Science and Morality*, Blackwell, Oxford.
- [6] Marin L. (1994) - *De la Representation* - Gallimard, Paris.
- [7] Ramakur R. (1993) - *Reliability Engineering: Fundamentals and Applications* - Prentice Hall, N.Y.
- [8] Shen K., Xie M. (1990) - On the Increase of System Reliability by Parallel Redundancy - *IEEE Transactions on Reliability* vol 39, n.5.
- [9] Wakerly J. (1978) - *Error Detecting Codes, Self-checking Circuits and Applications* - North-Holland, Amsterdam.
- [10] Rocchi P. (2000) - *Technology + Culture = Software* - IOS Press, Amsterdam.