Information Processing & Management Vol. 17, No. 6, pp. 341–347, 1981
Printed in Great Britain

0306-4573/81/060341-07$02.00/0
Pergamon Press Ltd.

INFORMATION THEORY AND INFORMATION SCIENCE†

PRANAS ZUNDE
School of Information and Computer Science, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A.

(*Received for publication* 7 *April* 1981)

Abstract—The empirical import of Shannon's Information Theory and its impact on information science are discussed. It is argued that extension of the scope of Information Theory as well as development of new theories of information science presupposes better understanding of relevant empirical regularities and laws. Possibilities of broadening the empirical foundation of Information Theory by introduction of appropriate least effort criteria are discussed.

INTRODUCTION

The subject of study of information science is empirical phenomena associated with various information processes such as information generation, transmission, transformation, compression, storage and retrieval. The ultimate purpose is to gain a better understanding of the nature of information. Starting—as does every other empirical discipline—with a description of phenomena in the domain of its interest, information science seeks to establish general principles by means of which the observed phenomena can be explained, accounted for, and predicted.

A *theory* of an empirical science is, in one sense, a representation or statement of a body of knowledge, founded on empirically verified relationships and laws, and knitted together by general principles, which may not necessarily be immediately apparent. It is in this sense that one refers to theory of light in physics or theory of valence in biology.

The term *theory* is also used in a different sense, in the sense of a systematic representation of formal rules and principles, which have no direct empirical import. Thus a theory of equations in mathematics is a theory in the latter sense of the word. It is common to refer to such theory as calculus, i.e. calculus of variations. The essential difference between these two senses of the term *theory* is that in the first case the statements of a theory are interpreted and are empirically verified, at least to a certain extent, to "hold" for observed data. In the second case, a theory is not interpreted and, as far as its postulates and rules go, does not need an interpretation, i.e, does not need to "hold" for any particular kind of observed phenomena. Rather, it is an abstract or formal theory.

On the other hand, it is not at all uncommon that an originally formal or abstract theory is later discovered to have fruitful interpretation within certain empirical contexts. This then becomes somewhat of an unexpected windfall for the empirical science in the domain of which a formal theory finds such a meaningful interpretation. Think, for instance, of the formal theories of logic and, in particular, of Boolean algebras, which though known for many years in their abstract, uninterpreted form, recently found important applications in modelling electronic circuits and computational devices.

Information science is a yound discipline and neither its empirical laws nor its theories are sufficiently well developed. To some, Shannon's Information Theory is the *only* theory in this subject field. The purpose of this paper is, first, to discuss the Information Theory as originally developed by Shannon and then to evaluate its potential significance as a theory of broader empirical foundation of information science.

CALCULUS OF INFORMATION THEORY

Shannon's Information Theory is first and foremost a formal theory, built around the paradigm of entropy. The entropy equation is derived from certain axioms, which have no

†This research was partially supported by the National Science Foundation grant IST-7917567.

empirical import. Since the understanding of the formal nature of Information Theory is essential for the exposition of its role in information science, we shall beiefly review these axioms (for the discrete case)[1, 9].

Let X be a discrete random variable which can take on a finite number of values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$. Let $h(\cdot)$ be a function which assigns a real number $h(p_i)$ to each event $\{X = x_i\}$, whose probability is p_i, $i = 1, 2, \ldots, n$. Let $H(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} p_i h(p_i)$ and denote $H[(1/n), \ldots (1/n)] = f(n)$.

Now, we postulate the following axioms, i.e. require that the function $h(\cdot)$ satisfy the following formal conditions:

(1) $f(n)$ is a monotonically increasing function, i.e. $n > n'$ implies $f(n) > f(n')$;
(2) $f(m \cdot n) = f(m) + f(n)$;
(3) $H(p_1, p_2, \ldots, p_{n-1}, q_1, \ldots, q_m) = H(p_1, \ldots, p_{n-1}, p_n) + p_n H\left(\frac{q_1}{p_n}, \ldots, \frac{q_m}{p_n}\right)$; where $q_1 + q_2 + \cdots + q_m = p_n$;
(4) $H(p, 1-p)$ is a continuous function of p.

It has been shown that the only function $h(\cdot)$ which satisfies all of the above requirements (axioms) is, up to a constant multiplier,

$$h(p_i) = -\log_a p_i, \quad i = 1, 2, \ldots, n$$

and thus

$$H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_a p_i.$$

The above four axioms which have been postulated as requirements for the functions $h(\cdot)$ and $H(\cdot)$, are completely formal in nature and require no interpretation. In particular, there was no need to introduce the notion of "information" in connection with the above axioms, nor any other information-related notion such as "uncertainty" or "message." In fact, all of the calculus of Shannon's information theory can be developed without any reference to such constructs or notions. It is not even necessary to interpret the numbers $p_1, p_2, \ldots, p_n$ in the above axioms as "probabilities" to avoid indirect empirical import which the term "probability," depending on its interpretation, may imply. We can simply say that $p_1, \ldots, p_n$ are non-negative numbers which sum to 1 and are values of some measure on x. In particular, the entropy function $H(p_1, p_2, \ldots, p_n)$ could be formally interpreted as a measure of spread of the values of a measure function $p(x_i)$, $i = 1, 2, \ldots, n$.

EMPIRICAL IMPORT OF PROBABILITY

On the other hand, it is precisely the interpretation of the measure function $p(x_i)$ as probability that establishes the first link between information theory and nature. The construct or concept of probability relates information theory to certain empirical observations and permits to view the theory as a *model* of these phenomena. We adopt here the viewpoint of Margenau of probability as a physical quantity which, like most other important constructs of science, is definable both in constitutive and epistemic sense[8]. The constitutive definition is due to Laplace and takes the probability of an event to be the number of favorable cases divided by the total number of "equipossible" cases. It is often called the *a priori* form of probability insofar as the probabilities of events are determined prior to observation and measurement. The epistemic definition of probability was first developed by Ellis, Cournot and others, and takes probability to be the ratio of the actual number of times the event occurs in a series of tests to the total number of events. Consequently, it is often referred to as frequency theory of probability. The latter is obviously a rule of correspondence, linking the theoretical construct "probability" to empirical world or nature. Furthermore, probability now becomes a measurable physical quantity. Consequently, interpreting the numbers $p_1, \ldots, p_n$, which appear

as arguments of the functions $h(\cdot)$ and $H(\cdot)$, as "probabilities" in the epistemic sense, we provide the foundation for the empirical interpretation of the whole theory because of the central role these functions play in it.

Unfortunately, the interpretation of information theory provided by the empirical import of the construct of probability does not automatically provide a link between observable physical phenomena and "information" in any of its empirical senses. Shannon did establish this connection by making certain assumptions about the relationship of probability as a measurable physical quantity on the one hand and the phenomenon of information on the other. Essentially, this connection is based on the assumption that information is not directly observable and thus not directly measurable, but can be observed only in terms of associated phenomena, some of which may be measurable. One such associated observable, Shannon claimed, is the probability of an event: the more improbable an event is, the more information is contained in a message that the event did happen.

Thus, although probability man not be the same thing as information, it might, perhaps, be a sufficiently good *indicant* of it. In a similar vein, the state of a quicksilver column in a thermometer, observed in terms of its extension or height is not the same thing as temperature. But the state of a quicksilver column in an appropriately constructed device is a good indicator or index of temperature, and so we can determine the temperature of an object by bringing it into contact with a thermometer and measuring the height of the quicksilver column in the thermometer. So, the argument runs, why not measure information by measuring the associated probabilities in some appropriate fashion?

Probability as an index of information, which we shall equate, roughly speaking, with the contents of a message, is more specifically interpreted and related to the function $h(p_i)$, called information measure of an event, in the following senses:

(1) The smaller is the probability of an event to occur, the greater is the "amount of information" conveyed by a message that the event did in fact occur, i.e. $p_i > p_j$ implies $h(p_i) < h(p_j)$.

(2) There is no information contained in a message conveying the occurrence of a certain event, i.e. $p_i = 1$ implies $h(p_i) = 0$.

(3) If two events are independent of each other, then the amount of information conveyed by a message that both these events occurred is equal to the sum of information contained in messages conveying the occurrence of these two events separately, i.e. if the event E is a joint event of two independent events E_i and E_j, then $p(E) \times p(E_i)p(E_j)$ implies $h(p) = h(p_i) + h(p_j)$.

At the first glance, there seems to be indeed both a good analogy between probability as an index of information and the height of quicksilver column in a thermometer as an index of temperature, and a good analogy between the methods of associated measurements. Unfortunately, there are significant differences.

First, the relationship between temperature of a body and its extension—in our case the height of quicksilver column in a measuring device—is not just postulated on a speculative basis, but firmly established by a low of physics. There is no empirical law of any kind, relating information associated with an event to its probability.

Secondly, the postulates (1)–(3), although reasonable in some instances, are not at all plausible assumptions for explicating information in its general sense. In particular, the postulate (3) of additivity of information associated with independent events often cannot be justified on empirical grounds, when semantic or pragmatic aspects of information are under consideration. For example, consider two independent events with the same probability of occurrence which carry the same message to a receiver. Clearly, the amount of information transmitted by such two identical messages is not equal twice the amount of information which had been conveyed to him by one message. In fact, the second message, being identical with the first, did not increase at all the amount of information transmitted by the first message. Similarly, one can easily find many examples demonstrating that postulate (3) is usually untenable in situations in which pragmatic information considerations, such as value or pleasure, prevail.

On the other hand, the postulates of information theory have empirical interpretations which hold very well for those aspects of information processes—in particular communication

processes—which involve objects and events as physical information carriers (specifically sign vehicles) and which are based on the laws of physics. It was indeed in this sense that Shannon developed his Information Theory as a model of communication. In the context of sign-based information processes, Information Theory can be more generally viewed as a theory focused on the syntactic dimension of these processes, in particular as a theory of sign media, sign shapes, and sign formation and transformation (coding). Since these sign vehicles are empirical prerequisites for information processes to occur, it is quite appropriate to interpret Shannon's Information Theory in the context of sign processes as a theory of *information potential* of sign vehicles.

THE PRINCIPLE OF LEAST EFFORT

Syntactic, semantic and pragmatic dimensions of sign processes are only convenient abstractions for the purposes of semiotic analysis; there are no actual sign processes which would be purely syntactic, or semantic, or pragmatic in nature. Rather, every sign process involves of necessity all three aspects, even though under certain conditions it may be expedient to concentrate the attention just on one of these three aspects. Furthermore, syntactic, semantic and pragmatic dimensions of a sign process are, in general, not independent. They interact and influence each other to a certain extent and the extent to which one dimension affects the other two varies, as a rule, with the type of signs, context, interpreter, purpose, etc. The investigation of these dependencies and their effects on information processes is one of the major challenges for information science as an empirical discipline.

The question, to what extent Shannon's Information Theory as a syntactic theory of information potential can be meaningfully made use of in semantic and pragmatic studies of information processes, is an open problem. More needs to be known about the phenomena which are characteristic of such interactions of semiotic dimensions, and more regularities and laws which hold for such phenomena need to be discovered before a definite answer can be provided on a firm empirical basis.

In the meantime, only tentative conjectures can be made in this respect from the rather limited store of knowledge about regularities governing phenomena related to information potential of sign vehicles for transmission of messages and phenomena on one hand and phenomana related to their semantic and pragmatic aspects on the other. Some of these regularities or laws might eventually be explained by a combination of Information Theory with the Principle of Least Effort (in one form or another), with the latter providing for the empirical import. The Principle of Least Effort, in applications to linguistic problems, is commonly associated with Zipf's name[13]. However, much more significant theoretical development, tying the principle to Information Theory, is due to Mandelbrot[6, 7]. Applied to words as sign vehicles (i.e. potential information carriers), the result can be briefly summarized as follows:

Assume "informationally" optimal system of word frequencies, i.e. word frequencies, for which discourse requires the smallest possible mean number of letters for a given entropy value H, or carries the largest amount of Shannon's information given the mean value of the number of letters. Then the shortest average word length, i.e.

$$\min \sum_{i=1}^{n} f(x_i)m(x_i)$$

subject to

$$H = -\sum_{i=1}^{n} f(x_i) \log f(x_i) = \text{const.}$$

is given by

$$f(x_i) = C[r(x_i) + V]^{-B} \qquad (*)$$

where $f(x_i)$ is the relative frequency of the word type x_i; $m(x_i)$ is the word length; $r(x_i)$ is the rank of the word x_i in a list of words ordered by decreasing frequency; and C, V and B are constants.

For the special case of $V = 0$, the rank-frequency relationship equation. (*) which was derived by Mandelbrot from the above stated theoretical assumptions on the basis of calculus of Information Theory and the Principle of Least Effort is known as hyperbolic distribution or Pareto law. It has been shown by Fairthorn[3] that a wide range of empirical regularities conform with the hyperbolic distribution law. Its cumulative distribution function, $F(x) = \sum_x f(x)$, approximates for $B = 1$ to

$$F(x) = a \log_e x + b$$

where a and b are constants.

Below are given selected examples of empirical regularities or laws of informational phenomena which admit a reasonable interpretation in terms of some least effort criterion and which have the hyperbolic distribution form or can be derived from hyperbolic distribution by straightforward mathematical transformations. Included in this set of examples is, for the sake of completeness, the original version of the Zipf's law as it was formulated by him based on extensive empirical observations. However, Mandelbrot's theoretically derived expression in the form of eqn (*) for word frequency distribution turned out to fit the data even better.

(1) *Zipf's Law.* Let $f(x_i)$ be the number of occurrences of the word type x_i in some given text. Let different word types x_i, $i = 1, 2, \ldots, n$ be arranged in order of decreasing frequency and let $r(x_i)$ be the order of the word type x_i in that list, called its rank [the most frequent word has rank 1]. Then

$$r(x_i) \cdot f(x_i) = C \quad i = 1, 2, \ldots, n$$

where C is a constant depending on a particular text[2].

(2) *Bradford's Law.* Let n be the total number of periodicals which publish articles on a given subject. If this set of periodicals is divided into k groups, each containing $n_1, n_2, \ldots, n_k$ periodicals and such that there is the same number of articles on that particular subject in each of the k groups, then

$$n_i = s_k n_{i-1} = s_k^{i-1} n_1$$

$i = 1, 2, \ldots, k$; $k = 1, 2, \ldots, m$; where n_1 is the so-called nucleus and $s_k > 1$ is the Bradford multiplier for k divisions of the n periodicals[3].

(3) *Lotka's Law.* The number of scientists who publish x papers in a given field is approximately $(1/x^2)$ the number of scientists who publish one paper only, i.e.

$$N(x) = \frac{N(1)}{x^2} = \frac{0.608}{x^2}$$

where $N(1)$ is the number of scientists who publish one paper, and $N(x)$ is the number of scientists who publish x papers[5].

(4) *Skinner's Law of word association.* The rank-frequency of individual word responses to Jung's word association test is given by

$$F = C \cdot R^{-\theta}$$

where F is the frequency of a response word, R is its rank and C, θ are constants[11].

(5) *Law of vocabulary size.* The number of different words (i.e. word types), d, in a text N words long (i.e. containing N word tokens) is equal to

$$d = \frac{N}{k}(0.423 + k - \log_e N + \log_e k)$$

where k is a parameter depending on the type of text[4].

(6) *Response-time Law.* The reaction time t of a human subject making a sequence of choice-responses is proportional to the amount of information which needs to be processed (i.e. is proportional to the average uncertainty or entropy of the source of signals):

$$t = a - b \sum p_i \log p_i$$

where p_i is the probability of the ith signal being presented and a, b are constants [2].

(7) *Law of indexing exhaustivity.* The following relationship holds between the degree of exhaustivity of content representation when a document is indexed by a group of indexers and the number of indexers in the group:

$$h(n) = C_0 + k \log n$$

where n is the number of indexers, h is the degree of exhaustivity, C is a constant representing average exhaustivity of a single indexer and k is a parameter [14].

SUMMARY AND CONCLUSIONS

Shannon's Information Theory is often considered a branch of mathematics and rightfully so. It can be developed in its entirety from a set of formal axioms and neither its axioms nor the theorems need to be interpreted in any empirical sense. It is true that the development of an abstract mathematical theory might not have been Shannon's original intention. Rather, the theory may have been originally developed with reference to the very practical communication engineering problems of maximizing the efficiency of signal transmission in terms of cost and reliability. In Shannon's own words, the problem which motivated his research is described as follows:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is *one selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design [10].

The claims of enthusiastic proponents of Information Theory, who sought to apply it indiscriminantly to any kind of information processes have so far not been justified. In particular, Information Theory has made little impact on information science, which ought to be its main and natural domain of application. We attempted to show here that the cause of the failure is essentially this: concepts and postulates of Information Theory, which have adequate interpretations in the context of communication engineering problems or, more generally, with respect to syntactic phenomena of information processes, are not adequately interpretable on the semantic and pragmatic level. In other words, Information Theory derives its basic empirical import from phenomena associated with sign vehicles and media as physical objects and phenomena associated with transformation of sign shapes (forms).

On the other hand, the tendency to underestimate the significance of Information Theory for information science may prove to be premature and in the final result wrong. Combined with and augmented by appropriate empirical laws and general principles, Information Theory may still make important contributions to the study of semantic and pragmatic phenomena of information processes, in particular to the elucidation of effects of syntactic factors on such phenomena. A significant achievement in this respect was made by Mandelbrot who augmented Information Theory by the Criterion or Principle of Least Effort to derive the statistical structure of language which has an excellent match with empirical facts [6].

Efforts of extending the application of Information Theory by augmenting it with appropriate empirical principles of information science have so far been limited essentially to Mandelbrot's work. But even there, the "Principle of Least Effort" was considered by

Mandelbrot as fuzzy and neither empirically nor theoretically sufficiently well established, so that he took recourse to more specific optimality criteria, namely "cost" and "economical criterion of matching." Thus an open research problem is to study alternative operational definitions of the concept "least effort" and how they relate to particular informational phenomena. This should help us to sharpen the concept of the Principle of Least Effort and give us a better understanding of empirical foundations of information processes, which in turn may open the doors to new applications of Information Theory in information science.

REFERENCES

[1] R. ASH, *Information Theory.* Interscience, New York (1965).
[2] G. E. BRIGGS, Reaction time and uncertainty in human information processing. *Technical Rep. No.* 69-5. Computer and Information Science Research Center, Ohio State University, Columbus, Ohio (1969).
[3] R. A. FAIRTHORNE, Empirical hyperbolic distributions (Bradford–Zipf–Mandelbrot) for bibliometric description and prediction. *J. Documentation* 1969, **25**(4), 319–343.
[4] N. HOUSTON and E. WALL, The distribution of term usage in manipulative indexes. *Am. Documentation* 1964, **15**, 109–116.
[5] J. LOTKA, The frequency distribution of scientific productivity. *J. Washington Acad. Sci.* 1926, **16**(12), 317–323.
[6] B. MANDELBROT, An informational theory of the statistical structure of language. In *Communication Theory* (Edited by W. JACKSON), pp. 486–502. Butterworths, London (1953).
[7] B. MANDELBROT, Information theory and psycholinguistics: a theory of word frequencies. In *Readings in Mathematical Social Science* (Edited by P. F. LAZARSFELD and N. W. HENRY), pp. 350–368. Science Research Associates, Chicago.
[8] H. MARGENAU, *The Nature of Physical Reality.* McGraw-Hill, New York (1950).
[9] M. REZA, *An Introduction to Information Theory,* 496 pp. McGraw-Hill, New York (1961).
[10] C. E. SHANNON, A mathematical theory of communication. *Bell System Tech. J.* 1948, **27**(3), 379–423; 1948, **27**(4), 623–656.
[11] B. F. SKINNER, The distribution of associated words. *The Psychological Record.* 1937, **1**, 69–76.
[12] G. K. ZIPF, *Selected Studies of the Principle of Relative Frequencies of Language.* Cambridge, Mass. (1932).
[13] G. K. ZIPF, *Human Behavior and the Principle of Least Effort.* Cambridge, Mass. (1949).
[14] P. ZUNDE and M. E. DEXTER, Indexing consistency and quality. *Am. Docum.* 1969, **20**(3), 259–267.