On Shannon, Fisher, and Algorithmic Entropy in Cognitive Systems

Marius Crisan

"Politehnica" University of Timisoara, Department of Computing and Software Engineering, Bd. V. Parvan 2, RO-300223 Timisoara, Romania
Phone: +40-256-403212, Fax: +40-256-403214. E-mail: crisan@cs.utt.ro

Abstract: *The paper's aim is to make a step forward in understanding and modelling cognitive systems in their interaction with the physical environment in the process of self-organization. In this regard, the information is conceived imbedded in physics reality and described by the conception of entropy. The concept of Shannon entropy, algorithmic randomness, and Fisher information is used as a tool in analysing the cognitive systems capacity of observation, measurement and acquiring knowledge upon the environmental universe.*

Keywords: cognitive modeling, self-organizing systems, Shannon entropy, algorithmic randomness, Fisher information

1 The Controversy

There is an opinion that the term entropy has different meanings in Shannon and Kolmogorov-Chaitin (K-C) information theories. This is related with the problem of increasing information in self-evolutionary system. In Shannon theory, entropy is randomness introduced by noise along the communication channel from source to destination. Noise can not increase the information content. The only source of information is provided by the transmitter, which tries to protect it against the influence of noise by coding techniques. In K-C information theory, the information is defined in terms of a nonstatistical definition of randomness, without considering the information origin. A series of numbers is random if it is patternless or incompressible, i.e. it cannot be found a shorter algorithm for specifying it than the length of the series itself. Therefore, a random series contains more information than a regular series, which can be specified by a much shorter algorithm.

Adding randomness to a system adds noise to the system, and by Shannon theory that means the information content of the system decreases. On the contrary, by K-C theory, randomness increases the information content of a system. If the term entropy means the same thing in both Shannon and K-C theory, then one can take a measure of information content of a system, say DNA, and then argue that on the basis of Shannon theory the information content of the system cannot increase, making evolution impossible [1].

A proposed solution considered that entropy in Shannon theory has a different meaning than in K-C theory. K-C information theory says nothing about how information content can change by a natural process, and Shannon theory applies to a very limited field where the transmitter is upper-bound on information content. The paper will discuss this issue.

2 Entropy Defined

Shannon defined information in terms of the probabilities of the discrete symbol types used in a string. The average quantity of information conveyed by one symbol in a string is $\Sigma_i p_i \log p_i$, where p_i is the probability of the ith symbol type in the set of available symbol types. In case that the probabilities of all members of the set of symbol types are equal for every location in a string, the information conveyed by the string is a theoretical maximum. The sequence of symbols is random in the sense that the recipient does not know beforehand which of the outcomes will be realized. According to Shannon's definition of quantity of information, the fewer the chances that a given message will be transmitted the higher one should express quantitatively the information obtained in the realization of that outcome. Therefore, the quantity of information contained in a message increases as its probability decreases. If the message consists of a series of binary digits, the quantity of information reaches a maximum when the sequence satisfies the statistical criterion that all possible subsequences should appear with roughly equal probability, in other words, the quantity of information is maximum when the sequence is random. If the probabilities of symbol types are not equal at any position in a string then the information conveyed by the string is less than a maximum. The difference between the theoretical maximum and the information contained in a given string is called redundancy. In other words, if a given pattern repeats in the reference string more often than the other possible patterns of the same size, then the string contains redundancy. The existence of redundancy in a string of symbols means that the string is not random. The ratio between the entropy H in such a case and the maximum entropy H_{max} is the relative entropy. Hence, the redundancy can be defined as $R = 1 - H / H_{max}$, showing the value with which the sequence is longer than the minimum length necessary to transmit the same information. If the redundancy is reduced to zero, the sequence becomes random and contains only pure information.

The previous approaches treated entropy as a probabilistic notion, in particular each individual micro state of a system having entropy equal 0. However, it is desirable to define a notion of entropy which assigns a non-zero entropy to each micro state of the system, as a measure of its individual disorder.

Firstly, we have to point out the relation between K-C complexity and Shannon's entropy. Briefly, classic information theory says a random variable X distributed according to $P(X=x)$ has entropy or complexity of a statistical form, where the interpretation is that $H(X)$ bits are on the average sufficient to describe an outcome x. Computational or algorithmic complexity says that an object x has complexity $C(x)$=the minimum length of a binary program for x. It is very interesting to

remark that these two notions turn out to be much the same. Thus, the intended interpretation of complexity $C(x)$ as a measure of the information content of an individual object x is supported by a tight quantitative relationship to Shannon's probabilistic notion. In particular, the entropy $H=-\Sigma_x P(x)\log P(x)$ of the distribution p is asymptotically equal to the expected complexity $\Sigma_x P(x)C(x)$.

3 An Information Content Approach

Both the statistical and algorithmic theory of information tries to define information in an objective way [2]. The statistical definition assumes that the recipient of information is able to extract the entire quantity of information transmitted from the source via a communication channel. The problem is to evaluate the capability of the receiver to extract a certain quantity of information from a given message mixed with random noise, being irrelevant the content of that information.

In this regard, Adami and Cerf claim that our intuition demands that the complexity of a random string ought to be zero, as it is not information [3]. According to this view, information is not a list of symbols or a description, yet it is given by the mutual entropy between two systems. Only that part of the description of an object X is information which is correlated or has a meaning in a given "universe" U that X is in contact with. That part of the entropy of X which is not shared by U ($H(X/U)$), does not convey any information about U, and is considered random. Consequently, a random string can be defined as the one that cannot be used to reduce the entropy of a closed system U under consideration, i.e. one that shares no entropy with U. One interesting consequence of this definition is that a string may be random in one system while extremely "meaningful" in another or, in other words, randomness is relative to the capacity of the observer to find a "meaning" in the series of events. However, this definition is not without its problems. The concept of 'meaning' has in turn to be precisely and unequivocally defined. If X can share only a part with U, that part being considered meaningful, and the rest valueless or random, then information will have only a relative and subjective relevance. If a string may be random in one system while meaningful in another, this implies that information is devoid of absolute value. If a subject has a lack of ability of recognizing the meaning of a certain piece of information, this doesn't mean that information is valueless. This definition applies to closed systems of physics and it proves to be inadequate for cognitive systems. Cognitive systems are not closed physical systems.

A closed system or even an open system connected only with a physical environment having an upper bound complexity cannot account for self-intelligence. Such a system would fail in front of challenging environments that are inaccessible, nondeterministic, nonepisodic, dynamic and continuous. If we are interested in an evolutionary model, we have to postulate the necessity for an open system of higher order, which allows for increasing information. In the rest of the paper this necessity will be demonstrated.

Crutchfield showed that due to measurement distortion the complexity of recognizing a pattern can be substantially higher than the complexity of generating that pattern, [4]. The indeterminacy caused by measurement distortion leads either to the appearance of effective randomness (and therefore, meaningless for the observer) or to the necessity to change the observer's model class. Thus, it results that the definition of randomness conditionally related with a closed system should be carefully reconsidered. Not always a random pattern of a measurement should be considered meaningless or valueless.

The capacity to extract information from a message X or perform an observation act depends on the recipient knowledge-base (KB). It may be useful de define the information state of a system as:

$$\Delta I = R + I(KB:X), \tag{1}$$

where ΔI accounts for the change of internal knowledge base, R is the randomness transferred to the system and $I(KB:X)$ is the useful or meaningful information. This equation is proposed based on the observation that randomness and meaningful information are two different forms of information.

Now, considering a closed informational system of n-bits of axioms, it is not possible to prove that a particular string has an algorithmic randomness greater than n [5]. In other words, we can only get out as much as we put in. The information is in the best situation conserved. This is the equivalent of the law of conservation of energy. Besides conservation of energy (mass, linear and angular momentum, electric charge), there must be a law governing conservation of information.

The efficiency of a cognitive system can be defined to be the net meaningful information or mutual entropy it can produce per unit of randomness taken in:

$$E = - I(KB:X)/R_i \tag{2}$$

The meaningful information is the difference between the input randomness R_i and the rejected randomness or noise, R_o. So,

$$E = (R_i - R_o)/R_i = 1 - R_o/R_i \tag{3}$$

This is the maximum possible efficiency for a cognitive system. Thus if all the randomness taken in is converted to meaningful information with no discharge of waste noise, then the cognitive system is 100% efficient. In other words, this is accomplished if $R_o = 0$ or R_i is infinite, neither of which is possible. However, the larger the randomness difference, the greater will be the efficiency of the system. In reality, the efficiency will always be less than 1.

The interaction between the observer and the observed system leads inherently to limitations on what the observer can infer from measurements. Similar to Gödel's incompleteness of formal systems, there is incompleteness in the observer's capacity to determine the states of the observed system, limiting the observer's ability to extract abstractions from the environment. Similar conclusions will be drawn further using Fisher information

Fisher information I has also the property to be a measure of the degree of disorder of a system. If a curve depicting the probability law $p(x)$ is broad and

smooth (unbiased probability density function) then I is small. This means high disorder or a lack of predictability of values x over its range. If $p(x)$ shows bias to particular x values then I becomes large, meaning also low disorder. This result is consistent with the values showed by the algorithmic information content applied on random and regular series. I acts also as entropy. Shannon entropy H is known to have the property of increasing monotonically with time, $dH(t)/dt \geq 0$. To that theorem (Boltzman H-theorem) there is a corresponding 'I-theorem':

$$dI(t)/dt \leq 0 \tag{4}$$

The 'I-theorem' can be seen as an equivalent of 'Boltzmann H-theorem,' which states that the level of disorder, the entropy H must increase with time. In other words, Fisher information of a physical system can only decrease (or remain constant) in time. Therefore, Fisher information can be also a measure of disorder.

According to Frieden, a similar result can be obtained starting from two premises of statistical mechanics: (i) the basic premise of statistical mechanics states that the probability density function (PDF) for a system that will occur is that one with the maximum probability; (ii) there is an ultimate uncertainty or resolution 'length' Δx in the actual value of the origin of a PDF $p(x)$ [6].

Thus, I is proportional to the cross-entropy between the PDF $p(x)$ and its displaced version $p(x + \Delta x)$.

$$I = - (2/\Delta x^2) \int dx p(x) \ln[p(x + \Delta x)/p(x)], \tag{5}$$

So, I appears as an approximation on the scale of Δx of the empirical PDF $p(x)$, that is different from the ideal PDF $p(x + \Delta x)$. Now, by maximizing the logarithm of the probability of PDF $p(x)$ we obtain a condition of minimum Fisher information,

$$I[p(x)] \equiv I = Min. \tag{6}$$

Now, if I is only an approximation on the scale of Δx, then the derived equations will loose their validity at scales finer than Δx. It appears an endless subdivision process of Δx to which a corresponding finer theory has to be found. At present, this issue is controversial between two alternatives: (i) we have to determine an ultimate finest resolution length, or (ii) the physics equation are valid down to all scales.

We'll show further using a version of Richard's paradox that the first alternative is true. First, we have to make a distinction between what is observed and the intrinsic information of the observed phenomenon. According to the efficiency of the cognitive system defined above, this implies the distinction between how much one knows about a phenomenon by observation, i.e. the mutual entropy, and how much it is possible to know about it, i.e. the input randomness. The source of information is the phenomenon itself. At this level it has the value R_i, which is the information carrying the phenomenon. The observed information is $I(KB:X)$. By observation a cognitive system (CS) makes a transformation of data into parameter estimation. In general, for a given CS and a finite value of $I(KB:X)$, we can define $OBS(I)$ to be the estimation of R_i given by the following definition:

$OBS(I)$ = The real information R_i, if any, whose I-name is coded up by I(KB:X),

$OBS(I) = 0$, otherwise.

We say that the finite data $I(KB:X)$ is an 'I-name' for the real phenomenon R_i exactly when CS is able to give R_i on the basis of the observed data $I(KB:X)$. For a natural number n, we may have a sequence of estimations $OBS(I_1) = k\ R_{i,1}$, $OBS(I_2) = k\ R_{i,2}$, … $OBS(I_n) = k\ R_{i,n}$, where $0 \leq k \leq 1$. Only under ideal conditions of knowledge acquisition, $I(KB:X) = R_i$. The field of R_i values of intrinsic information is composed by a square array of real numbers of the form r_{ij}, which comprises the whole intrinsic information of the universe. Now, consider a closed CS. This one cannot estimate the diagonal number composed by the values r_{kk}, since this number was artificially constructed to differ from every real number with an I-name. Therefore, CS does not have an I-name for the whole R_i. If CS would have a name for the whole R_i then since CS is closed, it will have an I-name for the diagonal number, but this is impossible. So, the conclusion is that any closed CS cannot have an I-name for the whole intrinsic R_i. No CS can give a finite description of how it connects the real and the ideal, the physical and the mental, or the language and thought. Consequently, in order to be self-intelligent, the system must be connected to an external source of information.

Finally, based on observations upon informational systems, we may suggest two formulations of the corresponding version of the second law applied to information: (i) no process is possible whose sole result is the producing of information equivalent to the amount of algorithmic randomness received from a generator of random events; (ii) no process is possible whose sole result is the transfer of random (or compressed) information from a less complex system to a more complex one. The entropy is not conserved, but is increasing in time. A cognitive system must keep lowering the entropy that is within itself, in order to be autonomous cognitive. This implies to be connected to a source of low entropy. The form of algorithmic entropy would suggest that the system must be supplied with low-entropy energetic combination, according to Boltzman component, and with low-entropy complexity forms, for satisfying the complexity component. This means a connection with the external world both physical and informational. Any other combination, would fail to keep the entropy of the system low, because of the entropy balance.

Conclusions

The capacity of cognitive system to extract new information from environment has been discussed. A learning machine would manifest intelligence in the proper sense of term when it will be capable of finding new solutions and rules of higher complexity. Therefore, it was drawn, from different entropic viewpoints, the conclusion that the interaction between the observer and the observed system leads inherently to limitations on what the observer can infer from measurements. Information is in the best situation conserved, and it seems to follow a similar principle as energy conservation, but at a higher level. Also, based on observations upon informational systems, we suggested a corresponding version of the second law.

References

[1] Yockey, H.: “Information Theory and Molecular Biology,” Cambridge University Press, 1992.

[2] Machta, J.: "Entropy, information, and computation." Am. J. Phys., Vol. 67, No. 12, December 1999.

[3] Adami, C. and Cerf, N.J.: “Complexity, computation, and measurement,” in T. Toffoli, M. Biafore, and J. Leao (ed.), PhysComp96, New England Complex Systems Institute (1996), 7-11.

[4] Crutchfield, J.P.: “Observing Complexity and The Complexity of Observation,” Proc. of the International Workshop on Endo/Exo-Problems in Physics, Ringberg Castle, Tegernsee, 1993.

[5] Chaitin G. J.: “Algorithmic Information Theory.” Cambridge: Cambridge University Press, 1987.

[6] Frieden, B.R.: “Physics from Fisher Information-A Unification,” Cambridge University Press, 1998.