ON FOUNDATIONS OF INFORMATION SCIENCE (SUR LE FOUNDATION DE LE SCIENCE DE INFORMATION)

Jack Belzer
University of Pittsburgh
Pittsburgh, PA 15260

ABSTRACT

Theories present systematic views of phenomena by specifying relations among variables. Relations among variables explain phenomena and their behavior can be predicted. Such theories form a *foundation* of a science. The paper presents several theories such as theory of organization, data structures, coding theory, representation and units of measure, and shows how they contribute to the understanding of phenomena within information science, and make predictions of their behavior possible. (Une theorie est une serie des construits qui sont en correlation avec des definitions et des propositions qui presentent une vue systematique des phenomenes en specifeant des relations entre les variables, avec l'intention d'experimer et de predire des phenomenes. La theorie nous permettre d'exprimer et de faire des predictions. Quand les theories qui sonte fonde sur la fonctionnement de phenomene d'une science particuliere peuvent etre explique et predicte avec de l'exactitude forment la base de cette science. Quand les theories peuvent resister les epreuves de consistance internale, c'est a dire, de ne pas creer des contradictions, et elles sont soumis de l'epreuves empiriques et sont accepte par ses egales, elles deviennent des lois. Dans la science physique, des lois de motion de Newton et des lois de Maxwell sur l'electromagnetique forment deux aspects de la foundation de cette science.)

INTRODUCTION

A department of systematized knowledge as an object of study is a *science*. Information science is such a department concerned with all aspects of information. The knowledge deals with the operation of general laws tested under rigorous conditions, the results of which are reproducible.

A *theory* is a set of interrelated constructs, definitions and propositions that presents a systematic view of phenomena by specifying relations among variables, with a purpose of explaining and predicting phenomena. Theory enables one to explain and to predict. When theories exist, based on which behavior of phenomena in a particular science can be explained and predicted with accuracy, these theories form *foundations* for that science. Theories which withstand the tests of internal consistency, i.e., create no contradictions, are subjected to empirical tests, and are accepted by its peers, become laws. In physics, Newton's laws of motion and Maxwell's laws of electromagnetics set foundations for the science of physics in two of its aspects.

THEORY OF ORGANIZATION

Locating or finding a specific item of recorded information, whether it be in books, journal articles, or records of transactions, on a random basis is a very difficult task because of the huge volume of recorded knowledge. The only way to locate the item is to examine each item until the one of interest is reached. Sometimes this will be reasonably soon, and sometimes it will be close toward the end of the file. On the average, if the file has N documents, (1 + N)/2 documents will be examined. Organizing these documents in some sequence supplies information about their relative location provided the search is made on the key for that sequence. Supposing we adopt a strategy to search by going to the middle of the file and examining the document there. If that document is not the one of interest, then it is either greater or smaller than the one being searched. This identifies which half of the file contains the document of interest. The other half is eliminated after the first search. We go again to the middle of the active half of the file, again eliminating half of what was left over. The process is repeated recursively until only one document is left. Anytime during the process it is possible to get a hit, i.e., the document examined may be the one which is being searched, at which time, the search is completed. However, at maximum, for N documents it would take $1+\log_2 N$ searches, a much samller number than that encountered previously in the completely random search, which was (1+N)/2. For example, if 100 documents were randomly stored in the file, (1+N)/2 or 50 searches on the average, would be required to locate a specific document. If they were organized sequentially and searched on a given key then a maximum of $(1+\log_2 N)$ searches would be required. This gives (1+6.644) or 8 searches. The reduction in effort of locating the document, i.e. (50-8) = 42 searches, is a measure of the amount of information provided by organizing the material. Later on we shall show Information and Communication Theory provides us with standard units for measuring information. The 42 searches will be replaced by units that make more sense.

DATA STRUCTURES

There are many different techniques for data structures. To determine which would be most effective and/or efficient for a given situation is of the essence. The file structure which would provide, on demand, the fastest access to any item in the file, is the one that reduces the uncertainty the most, with regard to the location and getting access to items in the file. The structure which produces the lowest entropy of the system is therefore most efficient.

Different demands are placed on systems by its users. Some require rapid retrieval like information retrieval systems, where updating can take place at a leisurely pace; some require rapid recording and storing of data, such as real-time satellite transmission of photographic data; and some require both rapid retrieval and rapid updating, like an airline reservation system.

Assuming random access memories, most file structures are modifications of the *sequential, random* and *list processing* file structures. Many combinations and varieties of file structures are possible. The question arises, how does one optimize on the most effective structure. Because files are accessed differently with each use, the optimum structure would depend on the frequency with which each use or access to the file is made. For example, how frequently is a personnel file used in obtaining information about an individual, say, Henry Stone rather than identifying individuals with specified talents such as, which physical chemists speak French. If the system was properly designed, i.e. if it maintains record of its own performance, then, the frequencies with which all the different ways the files have been searched most recently, are available. These are converted to probabilities, and the entropies of the system under various combinations of the file structure are determined. The combinations with the lowest entropy identifies the structure which provides the maximum amount of information with regard to the location and accessibility of each item in the file relative to the specified application. The probabilities used specify the application.

CODING THEORY

Information science like physics has many aspects. A language is used to communicate. Recorded knowledge requires representation of a language using an alphabet and a set of rules, a syntax. Coding is a form of representation and theory of coding provides a base for Foundations for Information Science.

Mapping of any arbitrary set into a set of mathematical or symbolic entities constitutes *coding*. A codijg set consists of strings of symbols. It is essential that mapping be one to one. A *redundant code* is one in which the mapping of the coded set does not cover the code; i.e., the code happens to have more words than necessary for the set to be coded, thus establishing unused invalid codes. The appearance of an invalid code during processing, signifies an error.

If, for example, a three digit code was assigned to identify uniquely 100 items. A three digit code would have a range of 000-999 a thousand unique codes. If the codes assigned to the 100 items were scattered across the entire range of the code separated approximately by 10 of each other, then an error in a code would, in all probability produce an invalid code. This would constitute an error detection code. If the error was less than 5, it would be closer to the correct code than any other valid code and it could be replaced by the correct code, thus making it an error correction code. Based on redundant coding utilizing minimum differences* among codes, a theory for error detection and error correction exists.

REPRESENTATION AND UNITS OF MEASURE

A record of events requires representation. An event is a point in sample space. Basic unit is a symbol. All available symbols is the alphabet. Each event if translated into a single symbol, would require a very large alphabet making representation inflexible and cumbersome. Using a small alphabet, by a combination of symbols a wide range of possibilities can be represented. For developing a general theory, we use the simplest of all alphabets, the binary alphabet, where "0" and "1" represent two possible and alternative events. One set of binary symbols can partition a set of real events into two classes, 0 and 1. Two symbols make up four different words. In general there are 2^n possible code groups consisting of n binary symbols. Conversely, if there are n possible choices, or n distinct representations, then $log_2{}^n$ binary digits are required to represent each uniquely.

In a given situation if there are n possible outcomes, then the uncertainty of an outcome is only as to which of the n occurs at a given time. When this is revealed, there is no longer any uncertainty. The amount of information is measured by the amount of uncertainty it

---

* The theory of error detection and error correction codes was developed by Richard Hamming of Bell Labs making use of geometric representations of binary codes where the differences were represented as geometric distances, known as Hamming distances. Distances between two codes are measured by the number of binary positions by which the two codes differ.

removes. Representation of n possible outcomes can be made by $log_2 n$ binary digits. If each outcome can be represented uniquely, without any ambiguity, by $log_2 n$ bits of information, then none of the outcomes can possess any more information than that. In the simplest case, where only two outcomes are possible, i.e., where $n = 2$, like tossing a coin, $log_2 n = 1$. One bit, a "1" for a head and a "0" for a tail can represent each possible outcome of the toss of the coin.

The behavior of any system is determined by the laws governing the system. These laws may be too complex to make their representation by differential equations impractical or perhaps impossible. In such a case, it is sometimes possible to consider the behavior of the system as a *random process* and apply the theory of probability for its representation and analysis. We would thus be dealing with *functions of a random variable* which can assume one of many different values and anyone can occur at any given time. If these values form a continuous function, then the random variable is continuous. A discrete random variable can assume a finite number of values only. When one of these values occur it is called a *random event*. Although we cannot predict the occurrence of an individual random event, probability theory enables us to determine the probability of occurrence of a random event.

In the toss of a coin, the probability of a head coming up is p_1 and that of a tail is p_0, where $p_1 + p_0 = 1$. On the average, each comes up with a frequency of its probability of occurrence, i.e., on the average, each comes up with a frequency of its probability of occurrence, i.e., on the average p_1 of the time the toss comes up head and p_0 of the time tail. The information supplied when the toss comes up head is $\log_2 p_1$, but it comes up p_1 of the time only, therefore its information contribution is $p_1 \log_2 p_1$ only. Similarly the information contribution of a tail coming up is $p_0 \log_2 p_0$. The total amount of information supplied by a toss of a coin is $I = p_0 \log_2 p_0 + p_1 \log_2 p_1$. The uncertainty of the toss is the negative of that and it is called *entropy*, represented by H. In the toss of a coin it is:

$$H = -(p_0 \log_2 p_0 + p_1 \log_2 p_1)$$

In an unbiased coin $p_0 = 1/2$ and $p_1 = 1/2$ and

$$H = -(1/2 \log_2 1/2 + 1/2 \log_2 1/2)$$

$$H = -[1/2(-1) + 1/2(-1)] = 1 \text{ bit per toss.}$$

This defines a unit of measure of information. It measures the amount of information required to resolve the uncertainty of two possible equiprobable outcomes. When the outcomes are not equiprobable, then we know something about the schema. This reduces the uncertainty about the outcome, and its entropy would be reduced to less than one

bit per event. In a biased coin such that $p_0 = 1/4$ and $p_1 = 3/4$, the outcome of the toss is constrained and we know that the chances are higher for the coin to fall head. We therefore have less uncertainty about the outcome of the toss and its entropy should be lower.

$$H = -(1/4 \log_2 1/4 + 3/4 \log_2 3/4)$$

$$H = -[1/4(-2) + 3/4 (-.415)] = .811 \text{ bit per toss}$$

Not to confuse these units with a "bit" in binary digits the International Standards Committee is recommending that this unit be called a *shannon*, named after the man who is responsible for Information and Communication Theory.

The *shannon* thus is an excellent unit of measure of information, measuring the amount of uncertainty it removes. This also provides a more rigorous meaning or definition of *information*. In this context, information is the resolution of uncertainty. If we should consider a source and a recipient, then entropy is the uncertainty of the recipient about the message being transmitted by the source. The uncertainty is resolved on receipt of the message.

IN CONCLUSION

In essence, theory of organization, coding theory, representation of information and units of measure, and data structures provide foundations for information science in some of its aspects. Fuzzy sets, some aspects of behavioral science and linguistics are other aspects for which foundations for information science exist. They can be brought to bear on explaining and predicting phenomena in information science with reasonable accuracy. These are the foundations on which information science is being built.