Information Processing & Management, Vol. 69–77. Pergamon Press 1977. Printed in Great Britain

ON THE NOTIONS OF AMBIGUITY AND INFORMATION LOSS

JAMES L. DOLBY

San Jose' State University, San Jose', CA 95112, U.S.A.

Abstract—One of the fundamental problems in information science is to distinguish various objects (such as books or journal articles) on the basis of associated values (such as authors and titles). Where the values fail to distinguish two distinct objects we say that the objects are ambigious under the given value assignment. To obtain a measure of ambiguity, it is only necessary to count the number of ways that the objects can be arranged for each set of ambiguous objects, multiply these counts and take logarithms. It is shown that such an approach leads to a measure in the formal sense and that the measure depends only on the definition of equality of values so that it can be simply extended to sets of values and ordered sets of values. It is also shown that it is possible to construct a function of ambiguity that one can call "information" and that the information loss that occurs when distinct values are grouped into equivalence classes (as in the use of search and sort keys) is also a measure. Finally, it is shown that ambiguity and information as here defined are directly related to Shannon's definition of "information" thus tieing this approach to that portion of information theory associated with the derivation of optimal distributions frequently used in information science models.

> "... *if one is concerned with messages to be transmitted, and if there is reasonable freedom in coding the message for transmission, entropy is clearly a quantity of interest and importance. But this fact makes it no easier to describe. One may find (its) description ... to be simple, completely perspicuous, and immediately intuitive. But neither the writer nor any of his friends did*".
>
> *John W. Tukey* (1963)

INTRODUCTION

Despite a long history and a growing body of significant literature in what is now generally called "bibliometrics", the field of information science has long suffered in comparison with the field of communication theory for lack of a sound mathematical base. A number of authors have shown that most of the important distributions found in the information field can be derived as optimal solutions to various functions of Shannon's notion of entropy (or information) but it has never been quite clear why a quantity derived in response to a basic problem in coding theory should be optimized in response to an entirely different problem involving no reference to codes whatsoever.†

In this paper we suggest that the fundamental problems of information science revolve around the need to distinguish various objects on the basis of the value or values connected to those objects (as an author and title is connected to a book or article) and that it is possible to set forth a measure, in the formal mathematical sense, of the ambiguity present in such a situation (as when two books have the same author). We also show that this measure can be related to Shannon's notion of information and hence provide a linkage to the various optimality problems and the use of maximal information as reasonable criterion for "goodness".

AMBIGUITY IN INFORMATION SETS

We begin with the assumption that one of the fundamental problems in information processing can be stated in the following terms: we are given a set of order pairs, $\{(0_i, X_i)\}$, where the first element of the pair is an object (such as a book or article) and the second element is an associated value such as the surname of the senior author. As the value presumably provides some information about the object we call this set of ordered pairs, the information set.

†Shannon's work can be found in, A mathematical theory of communication. *Bell Syst. Tech. J.* 1948, 27, 379–423; 623–656. One of the first papers showing that linguistic distributions known to occur in practice could be derived from functions of information is, B. MANDELBROT, An information theory of the statistical structure of language. *Proc. of the Symp. on Applications of Commun. Theory*, London, Sept. 1952.

We make two basic assumptions about the nature of these variables:

(1) The objects are all distinct.

(2) The values are subject to a linear order.

The first assumption insures that the set of ordered pairs is a function. (Any set of ordered pairs $\{(A_i, B_i)\}$, is a function if for every case where $A_i = A_j$ it must be true that $B_i = B_j$. Clearly, if all the A_i are distinct, this condition is met.)

The second condition provides the underlying mathematical structure for the process we usually call "sorting", that is, the process of putting the values and/or objects into an ordering. More specifically, a linear order requires two further conditions:

(3) If X_i precedes (is less than) X_j and X_j in turn precedes X_k we must have that X_i precedes X_k.

(4) For any two values, X_i and X_j one and only one of the following situation obtains: $X_i = X_j$, X_i precedes X_j or X_j precedes X_i.

The first condition is certainly a natural one in the information context and need not concern us further. The second condition gets us immediately to the heart of the problem.

If all of the X-values are distinct the objects can be unambiguously ordered according to their corresponding X-values. In such nice circumstances we shall say that the information set is an unambiguous function. (Mathematically, such functions are variously called "strict", "one-to-one", or "injective".) However, it takes but a moment's reflection to see that not all functions are unambiguous. Different authors have the same surname and, for that matter, the same author can write more than one book. It therefore becomes important to measure the degree of ambiguity for a particular function.

To derive such a measure we proceed as follows. First, by way of example, let us suppose that in a given situation we have the following order imposed on the objects by the X-values:

$$P_X(0) = (\{0_1, 0_2\}, \{0_3\}, \{0_4, 0_5, 0_6\}, \{0_7, 0_8\}).$$

Here we have grouped the 8 objects into four subsets in each of which the corresponding X-values are the same and then ordered the four subsets according to those X-values. In other words, both 0_1 and 0_2 precede 0_3 but we are unable to establish the order for 0_1 and 0_2 because each has the same X value in this example, so that we could either have 0_1 *precede* 0_2 or 0_2 precede 0_1. 0_3 has a unique position in the order, but there are six ways to order the three objects in the third set and two more ways to order the objects in the fourth and final set. More generally, if there are n objects in such a subset, there are $n!$ (n factorial) ways to order those values and as each of these sets of orderings is independent of the other the total number of ways to order all of the objects is

$$A(X) = n_1! \cdot n_2! \cdot \ldots \cdot n_r! = \prod_{i=1}^{r} n_i!$$

for the r subsets of the partition, $P_X(0)$, where n_i is the number of objects in the ith subset of the partition. Call $A(X)$ the *ambiguity* of x.

It is easy to show that $A(X)$ has three fundamental properties:

(1) $A(X) \geq 1$.

(2) $A(\phi) = 0! = 1$.

(3) if X and Y are disjoint sets, $A(X \cup Y) = A(X) \cdot A(Y)$.

The first property follows from the fact that n factorial is a positive integer and a product of positive integers is, in turn, a positive integer. The second property is a bit of triviality but important to the structure: if the information set is empty (usually represented by the Greek symbol , ϕ) the ambiguity is zero-factorial which is unity. The third property is only a bit more difficult: if we wish to combine two information sets (and sets are combined by the operation of union, denoted here by $\cup$) wherein the values of one set are not to be found in the set of values of the other set, we need only multiply the ambiguities to obtain the ambiguity of the union.

Now suppose we replace $A(X)$ by $a(X) = \log A(X)$. Then the three properties of $A(X)$ become:

(1′) $a(X) \geq 0$.

(2′) $a(\phi) = 0$.

(3′) if X and Y are disjoint sets, $a(X \cup Y) = a(X) + a(Y)$.

The log transformation provides several useful by-products. Factorials increase very, very rapidly with n, ($10! = 3{,}628{,}800$), while the log of n-factorial increases, roughly, as $n \log n$. It seems reasonable to have the minimum value of the measure to be zero and it is certainly nicer to add (property 3′) than to multiply.

From the mathematical point of view, however, the difference between $A(X)$ and $a(X)$ is crucial: $a(X)$ is a *measure* precisely because it satisfies the three properties 1′, 2′ and 3′ while $A(X)$ is not. Thus we can immediately appeal to a large literature of known results about measures rather than have to derive special results as we proceed. Furthermore, the log transformation is the unique transformation to carry the properties given for $A(X)$ into the desired properties found for $a(X)$. Proving *that* statement is not trivial, but it is useful to know that we need not search for alternative transformations to see if they have nicer properties.

(To one not familiar with the literature of measure theory an appeal to its existence may be less than convincing. Hence, it might be well to point out that the three properties 1′, 2′ and 3′ are shared by the most fundamental measure we have, namely counting. If, for instance, one wished to count the number of people in a building, it is clear that the count could not be negative—property 1′—the count would be zero if the building were empty—property 2′—and the total count could be obtained by counting the people in each room and adding if we insure that no one can change rooms while the counting process is going on—property 3′.)

THE MEASUREMENT OF INFORMATION

It is nice to know that log-ambiguity is a measure, particularly if one is a mathematician. However, the application of such a measure requires that we obtain some knowledge of how the measure changes in changing circumstances. For instance, it is not uncommon to replace the author's surname by a "search key" built, say, by choosing the first three letters of the author's surname. We will restrict our study of such changes to changes which are functions. That is, we now consider sets of ordered pairs $\{(X_i, Y_i)\}$ where the X_i are the original values in the information set and the Y_i are some new set of values derived from or based on the X-values. The restriction to functions in this case implies that the Y-values cannot break any of the ties that existed in the set of X-values. Thus JONES is always replaced by JON in a three letter search key. This restriction in turn implies that ambiguity cannot be decreased since the only way to decrease ambiguity would be to break up the subsets in the partition $P_X(0)$ and we are deliberately inhibited from doing this by requiring that Y be a function of X. Thus we have the first elementary result that

If Y is a function of X, then $a(Y) \geq a(X)$.

(It should be acknowledged that non-functional values are, at times used in information processing. Librarians do, in fact, deliberately assign copy numbers to books to make otherwise identical things distinct and computer programs do provide distinct sorts of data even when there are ties in the data. Further, these arbitrary tie-breaking operations are useful and, at times necessary. However, they are also arbitrary and hence contribute little to the discussion that follows. At the proper time, however, we shall return to this aspect of the problem and make good use of it.)

There are, of course, circumstances wherein the Y-assignments do not increase the ambiguity. That is,

If Y is an unambiguous function on X, then $a(Y) = a(X)$.

In such a case, nothing is lost by substituting Y for X and if Y requires less storage than X, as in a search-key, there may be positive advantage to doing so. In the more general situation something is lost and this can be measured by the increase in ambiguity:

$$a(Y) - a(X) = \log A(Y) - \log A(X) = \log [A(Y)/A(X)]$$

In general, the difference of two measures is not a measure because differences can, of course, be negative and property 1′ would be violated. However, in this case we already know that $a(Y)$ is never less than $a(X)$ and the difference is in fact a measure. (This result in turn follows directly from the fact that Y is a function of X and is thus a good reason for restricting our investigations to functions.)

However, the form of the difference in log-ambiguities ends up as a logarithm of a ratio and this suggests that there might be some profit in studying such forms more generally. We thus return for the moment to the study of X alone and note that the maximum possible value of $A(X)$ is $n!$ and that this situation obtains when all of the X-values are the same, that is, when the X-values provide absolutely no information as to how the objects are to be ordered. Consider then, the *relative ambiguity of X*, $R(X) = A(X)/A_{max}(X) = A(X)/n!$ and its logarithm $\log R(X)$. As $n!$ is the maximum value for $A(X)$, $R(X)$ is always less than or equal to unity and hence its logarithm is always negative. We remove the "always negative" condition by simply changing the sign so that we have:

$$I(X) = -\log R(X) = \log [A_{max}(X)/A(X)].$$

The change of sign implies that as $A(X)$ increases, $I(X)$ decreases and conversely. Thus $I(X)$ should have a name whose connation is diametrically opposed to that of ambiguity and we choose to call this quantity the *information* of X. In this view, information improves as the subsets of the partition $P_X(0)$ get smaller and is maximal when every subset contains precisely one object, namely when X is an unambiguous function.

Returning to the comparative situation we now consider the *information loss* in replacing X by Y, namely

$$\begin{aligned}\text{Information Loss} &= I(X) - I(Y) = -\log A(X)/n! + \log A(Y)/n! \\ &= \log [A(Y)/A(X)] = a(Y) - a(X)\end{aligned}$$

In other words, information loss is precisely equal to increase in *log-ambiguity*. Now however we are in a position to break up the right hand side of the equation is such a way as to pin point precisely where the information loss occurs.

It is, perhaps, easier to see how this works in an example. Suppose we have a set of documents (labeled A through H) together with the surname and the three-letter author key formed from the first three letters of the surname.

Table 1.

Object (Z)	Author (X)	Author Key (Y)
A	Johns	JOH
B	Johnson	JOH
C	Johnson	JOH
D	Johnson	JOH
E	Johnston	JOH
F	Jones	JON
G	Jones	JON
H	Jonson	JON

In the first column we have given each object a unique identification, the one-letter codes A through H. (Social Security numbers and International Standard Book Numbers are both attempts at unique identification of objects for information handling exercises.) Under such a value assignment, the information is maximal (ambiguity is minimal) and is equal to $\log 8!$. To make this concrete we must choose a base for the system of logarithms. Such a choice is arbitrary (and will change the results only by a constant multiplier) and so we will follow tradition in the information field and use base two. If Z is the name of the identification number we than have:

$$I(Z) = \log_2 [8!/1] = 15.299$$

as the numeric value for the information provided by the unique identification code, Z.

Now let X be the name of the author surname assignment. Then

$$I(X) = \log_2 [8!/1! \cdot 3! \cdot 1! \cdot 2! \cdot 1!] = 11.714$$

The information loss, $I(Z) - I(X) = 15.299 - 11.714 = 3.585$. This information loss stems from the fact that Z identified each item specifically while X does not. In particular, the surname Johnson is assigned to three different objects and the surname Jones is assigned to two different objects. As the other author surnames (Johns, Johnston and Jonson) each identified objects uniquely, the information loss must stem solely from the ambiguity in the Johnson and Jones assignments. Initially, the three objects now assigned to Johnson had three distinct names (B, C and D) and the information available with that assignment (and lost when the three distinct codes are replaced by the single code Johnson) is:

$$I_{\text{Johnson}}(Z) = \log_2 [3!/1] = 2.585$$

Similarly, the information originally given by the codes F and G now lost if these codes are replaced with the surname Jones is

$$I_{\text{Jones}}(Z) = \log_2 [2!/1] = 1.000$$

These two losses, 2.585 and 1.000, sum to 3.585, precisely the total information loss found by subtracting the information in X from the information in Z. This is not a coincidence, of course. Simple rearrangement of terms is sufficient to show that

$$I(Z) - I(X) = I_{\text{Johnson}}(Z) + I_{\text{Jones}}(Z).$$

Suppose further that Y is the name of the three letter author key of Table 1. Here we have even more information loss because Johns, Johnson and Johnston are all combined into JOH and Jones and Jonson are combined into JON. Substitution into the information formula gives

$$I(Y) = \log_2 [8!/5! \cdot 3!] = 5.807$$

a further information loss (when compared to $I(X)$) equal to 5.907. This information loss can, in turn, be decomposed into two components: one representing the loss due to grouping three names in JOH and the other due to the loss in grouping the two names into JON:

$$I(X) - I(Y) = I_{\text{JOH}}(X) + I_{\text{JON}}(X) = 4.322 + 1.585 = 5.907.$$

The additive property of information loss makes it relatively easy to evaluate various strategies in value assignments. For instance, one could reduce the information loss in the assignment of surnames by adding the first initial of the first given name, assuming that the Johnsons and Jones have differing first initials. One could reduce the information loss found in three-letter author keys by going to four-letter authors keys. (This would help JON by splitting Jones and Jonson but would not help the JOH class.) One could then ask whether it were better to use a four-letter author key formed by the first four letters of the surname or by the first three letters of the surname followed by the initial of the given name. The latter procedure is almost always better, but it is useful to have a specific measurement of how much better particularly as there are other competing criteria of goodness in any information system.

It is useful to add a remark that may not have been obvious in the above calculations. Suppose, as in the example above, that Z is an unambiguous function defined on the objects. Then consider the information loss of X relative to Z:

$$\begin{aligned} I(Z) - I(X) &= \log [n!/A(Z)] - \log [n!/A(X)] \\ &= \log n! - \log 1 - \log n! + \log A(X) \\ &= \log A(X) = a(X). \end{aligned}$$

In other words, for any X, the information loss of X relative to the unambiguous function (or sequence number assignment) is simply the log-ambiguity of X. Thus these two measures are closely related and it is reasonable to say that "information loss" is a natural generalization of "log-ambiguity" that enables us to measure what is going on when we apply functions (such as search keys) to the given values of X.

EXTENSION TO HIGHER DIMENSIONS

In the preceding discussion, we assumed throughout that each object, 0_i, had a single associated value, X_i, and that the essence of the problem was to order the objects through the linear ordering provided by the values; failure to obtain a unique ordering led to the definition of ambiguity. Now we extend the problem to include a second value, Y_i, where this second value is not necessarily a function of the first value. We might have, for instance, the first value as the author of a document and the second value as the title of the document. Formally, we denote this situation by the set of ordered pairs: $\{(0_i(X_i, Y_i))\}$ wherein the value of the ordered pair is itself an ordered pair. As before, we assume that the objects are all distinct so that the set of ordered pairs is a function. We also assume that both X and Y are subject to a linear order. Note that this is *not* equivalent to assuming that the pairs, (X_i, Y_i), are subject to a linear ordering.

Indeed, the first problem is to define what we are to mean by equality of two ordered pairs as the definition of linear order for the two values does not in and of itself imply such a definition. We adopt the following, quite standard, definition:

Two ordered pairs, (X_i, Y_i) and (X_j, Y_j) will be said to be equal if and only if: $X_i = X_j$ and $Y_i = Y_j$.

With such a definition we are in a position to determine for every pair of pairs whether they are equal or not and hence to determine for each distinct type of pair the number of pairs of that type. Let n_i be the number of pairs of the ith type. Then, the ambiguity of the set of ordered pairs is, as before, $A(X, Y) = \prod_{i=1}^{r} n_i!$ and all other definitions (log-ambiguity, relative ambiguity, information, information loss) follow immediately and have precisely the same properties! Further the generalization to three or more values is the obvious one: two ordered sets are equal if and only if the values are equal position by position, and once equality is defined the n_i can be determined and the rest of the definitions follow immediately.

It only remains to return to the question of ordering the objects. In particular, we must ask whether there is a function of X and Y (satisfying the same linear order as X and Y) that will order the objects and provide the same information (or ambiguity) as the set of pairs does. That is, does there exist a function, $f(X, Y)$, such that $A(f(X, Y)) = A(X, Y)$?

In general there will be several such functions, but it will suffice to exhibit two of them. As we are interested in ordering the objects, it seems natural to consider the standard sorting procedures. Suppose, for instance, we were to sort the objects by author and then to break ties with titles (to the extent that titles could break such ties). Failure to break the ties would imply that not only were the authors the same but also the titles (as would occur if there were two editions of the same book with no change in author). But this notion of equality is precisely the notion expressed previously: two pairs will be equal if and only if the respective elements of the pairs are equal.

It is a bit easier to see that this procedure is equivalent to a function on X and Y if we accomplish the ordering by way of a "sort key". To make it simple, let us suppose that the author's names are all stored in a fixed length store and that p characters are sufficient for this purpose. Suppose further that the titles are stored in another fixed length store and that q characters are sufficient for this purpose. The sort key is then a string of characters of length $p + q$ with the author's name in the first p characters of the key (padded out with blanks as necessary) and the title in the next q characters (again padded with blanks as necessary). Let us call this process of laying the two values into a single (larger) store, "concatenation" and represent it by the symbol, $\oplus$. Then,

$$f(X, Y) = X \oplus Y = X \cdot k^p + Y$$

where k is the number of bits necessary per character. It is easy to see that $X \oplus Y$ provides a unique result for any given pairs of values (X, Y) so that $X \oplus Y$ is a function. It is also easy to see that $X_i \oplus Y_i = X_j \oplus Y_j$ if and only if $X_i = X_j$ and $Y_i = Y_j$ so that we have:

$$A(X \oplus Y) = A(X, Y) = A(Y \oplus X)$$

where the second equality reminds us that it does not matter whether we choose to have X dominate (i.e. be in the left portion of) the search key or have Y dominate. More generally, if we have V values associated with each object (author, titles, subject, pub date, etc., etc.) there will be $V!$ different search keys each of which has the same ambiguity as the ordered set of values itself.

Thus the simple set of measures (log-ambiguity and information loss) derived from the one-variable case is simply extendable to any number of variables, thus opening the door to operations on one or more of the variables—including the operation of removing a variable from the system—with an associated measure to determine how much the underlying situation is changed by these operations if at all. For example, the Library of Congress provides magnetic tape records of *LC* cataloging information for use by other libraries in their automated library operations. These records (called MARC records) contain more fields (and subfields) of information than are necessary for many applications so that the systems analyst must determine which fields and subfields are to be included in the records to be used, for instance, in the circulation system. Decisions must also be made as to whether to use fixed length or variable length fields and, in the former case, one must decide how many characters to include in the field and whether to use a straight truncation rule when input records are too long or a more complex abbreviation procedure. Knowledge of the impact of these decisions on the measures of ambiguity and information loss will frequently be useful in such decision making situations.

As an indication of the potential we offer the following simple theorem, the proof of which is left to the reader:

$I(X, Y)$ is greater than the maximum of $I(X)$ and $I(Y)$ if, and only if, X is not a function of Y and Y is not a function of X.

The significance of this modest theorem is that if Y, say, is a function of X: then Y contributes no information to the situation in the sense that $I(X) = I(X, Y)$ so that $I(X, Y) - I(X) = 0$. This is not to say that Y is not useful as it may provide a useful ordering that is different from X or require fewer bits to store than X. However, in such circumstances, Y does not improve our ability to distinguish one object from another given that we already had X and it is that situation that this measure of information speaks to.

EXTENSION TO SETS OF VALUES

The third basic way that we describe objects is to associate an unordered set (properly, just "set") of values to each object. Thus, it is common to associate a set of subject headings, or descriptors, or index terms to a book or article. In such circumstances, the size of the set is not fixed; that is one book may have one subject heading while another has three or four. The extension of the definitions of ambiguity and information again depends only on the choice of the definition of equality. The standard definition for equality of sets is:

Two sets are equal *if and only if they contain precisely the same elements.*

Two objects would then be equivalent if and only if they had the same set of descriptors (subject headings, ets.) and this definition would provide the basis for determining the set of counts, $\{n_i\}$, which in turn determine the ambiguity and information for the information set.

One of the key questions in the study of information sets of this type is the question of the control of the vocabulary as new items are added to the store. Hence it is useful to note how log-ambiguity and information change with additions of new ordered pairs. The addition of a new ordered pair (say, a new book with its associated set of subject headings) will either increase the log-ambiguity (if the set of subject headings has been used for a book already in the set) or leave

it unchanged (if the set of subject headings is new to the system). In this sense, vocabulary control is at least partially concerned with the control in the growth in log-ambiguity. The addition of new subject headings to the vocabulary tends to reduce the growth in log-ambiguity. Increasing the number of subject headings per book also tends to reduce the growth in log-ambiguity as it permits a larger number of subsets for a vocabulary of a given size.

Except for the trivial case where the information is zero, the addition of another ordered pair also increases the information. At first glance this appears paradoxical as information was defined in such a manner as to be inversely proportional to ambiguity; however, the definition was made in terms of a particular value of n and we are now considering what happens as n increases. Specifically, if we denote the information present with n items in the information set by $I_n(X)$ and the information when new item is added to that set by $I_{n+1}(X)$, the increase in information is given by:

$$I_{n+1}(X) - I_n(X) = \frac{n+1}{n_i+1}$$

where $n_i = 0$ if the set of terms associated with the new item has not previously appeared in the information set and is otherwise the number of previous uses of that set of terms. On average, the ratio of $(n+1)$ to (n_i+1) will be approximately r, the number of distinct sets of terms used in the information set, so that the typical gain in information for a new item will be of the order, $\log r$. Thus the incremental gain in information when a single item is added to the store is roughly proportional to the logarithm of the number of distinct sets of terms in the store at the time. This seems intuitively reasonable and provides a starting point for the formal arguments to justify various types of information distributions.

The same procedures can, of course, be used in the analysis of the "inverted file", the information set containing the index term (subject heading, etc.) as its first element and the set of books (articles, etc.) attached to that index term as the second element. In this sense, we would be measuring the information that the books provide about the terms rather than vice-versa and the incremental gain in information through the addition of a new index term would depend—as it should—on which set of items were to be associated with that term.

It should be noted that this method for determining information for this case is not the only one possible. For instance, one could let n_i be the number of terms associated with the ith book or, in the inverted file case, the number of books associated with the ith term. The latter situation can be derived from first principles by the following artifact: For each ordered pair in the information set defined earlier in the section of the form, $(0_i, \{X_{i1}, \ldots, X_{in_i}\}$, we create n_i new ordered pairs, $(0_{i1}, X_{i1}), \ldots, (0_{in_i}, X_{in_i})$ where the n_i "copies" of the original object, 0_i, are tagged with "copy number" subscripts. The tagging assures that each set of ordered pairs so created and the whole set will be functions. This new expanded set is then of the same form as that with which we began: the value of each ordered pair is a singleton, and the resulting values of n_i will be precisely the number of books with the ith tag. Nor is the artifact inherently unreasonable. It is easy to argue that the subject heading file in a library is a surrogate for a subject heading ordered shelving of the books—with multiple copies of the books used to allow multiple placement of these books with multiple subject headings.

In this expanded mode, the information gain of adding a new subject heading with n_{r+1} books assigned to it to a system that previously had r subject headings and n expanded pairs would be, with a slight abuse of notation,

$$I_{r+1}(X) - I_r(X) = \log \frac{(n+n_{r+1})!}{n!(n_{r+1})!}$$

where the right hand side increases (very roughly) linearly with the size of n_{r+1}. This is, of course, consistent with the earlier observation that increased assignment of subject headings per object allowed for greater distinction among the sets of terms assigned to each book. However, it should be noted that log-ambiguity also increases with increases in n_{r+1} and ambiguity in this context corresponds to the inherent difficulty in using a system that returns "too many" items per lookup.

INFORMATION AND SHANNON'S ENTROPY

Some thirty-odd years ago, Shannon considered certain questions in coding (or communication) theory that are closely related to, and helped stimulate, the ideas considered in this paper. In Shannon's problem, the X-values are characters in some alphabet and their multiplicities allow the message sender to construct a variety of messages of a given length. The number of such messages is easily seen to be:

$$N = \frac{n!}{n_1!n_2!\ldots n_r!}$$

where n_1 is the number of characters of the first kind, n_2 the number of characters of the second kind, and so forth, and n the total number of characters (including multiplicities) in the message. Shannon then considered the quantity

$$H(X) = \frac{1}{n}\log N \simeq -\sum_{i=1}^{r} p_i \log p_i$$

where $p_i = n_i/n$, the proportion of the total number of characters that are of the ith type. The approximate equality in the above expression is found by replacing each factorial term in the preceding expression by the Sterling approximation, $n! \simeq n \log n$. The quantity, $H(X)$ is variously called, "information", "entropy", or "expected entropy" in accordance with terminology long used in statistical mechanics.

Shannon's work has in turn led to a rich development of results that are not only mathematically pretty but quite useful in constructing practical systems of communication.

If one compares Shannon's $H(X)$ with the $I(X)$ defined earlier in this paper, it is immediately clear that

$$H(X) = I(X)/n$$

that is, entropy is "information per object" (in our terms) or "information per character" (in Shannon's terms). It should be noted that although both definitions are made in terms of "information" problems and lead to the same mathematical form, they are, in fact, different. Shannon was interested in the number of ordered sets of X-values (messages) and had no need to consider the "objects" behind the X-values whereas we have been totally concerned with the number of ways the objects could be ordered given the X-values. Nonetheless, it is well to note the close relation between the two definitions, particularly since there is a substantial body of results that can be borrowed from the communication problem and re-applied to the "object" problem.

CONCLUSION

We have shown that it is possible to make an elementary definition of ambiguity that leads to reasonable measures of log-ambiguity and information loss; reasonable in the sense that these measures perform in intuitively reasonable ways in the face of several of the basic operations in information science: sorting, creation of search keys and sort keys, and the use of taxonomic numbers. We have also shown that these measures bear a close resemblance to the work of Shannon and hence opened the door to further application of the tresults of measure theory and communication theory to information problems.

Acknowledgements—MARTIN BILLIK, DEREK DE SOLLA PRICE and HOWARD L. RESNIKOFF were all kind enough to read early versions of this paper and provide helpful comments.