Essays of an Information Scientist, Vol:1, p.124-125, 1962-73
Current Contents, #28, September 9, 1970

"current comments"

Information Theory and the Evaluation of Information Retrieval Systems.

September 9, 1970

Recently a rather interesting study was reported by A.R. Meetham[1] of the National Physical Laboratory in England on the use of information theory in evaluation of information retrieval systems. Meetham shows that Shannon's basic theorems provide a valid instrument for comparative ranking of the effectiveness of a variety of indexing systems, including citation indexing; and he concludes that "retrieval by citation indexing and bibliographic coupling must rank among the best valued of the conventional indexing languages."

I find it at once gratifying and surprising that citation indexing should so soon be described as one of the "conventional" indexing languages of science. Such a description may indeed seem premature, for even as Meetham was drawing his conclusion, a survey being conducted by N.B. Hannay of the Bell Telephone Laboratories was to reveal that only one in five chemists was aware of the availability of the *Science Citation Index*® [2]. Considering in addition how recently many bibliographic experts were propounding that citation indexing simply could not work in information retrieval, much less in sociometric research, it is difficult to restrain astonishment at finding our *Science Citation Index* so suddenly and so respectably labeled "conventional". (I should add that the Hannay report does draw a conclusion that does not surprise me and one with which I heartily agree, namely that "many people not using *Science Citation Index* would find it profitable to learn to use it.")

For readers who may be unfamiliar with the distinction between citation indexing and bibliographic coupling[3] as retrieval techniques, let me say only that bibliographic coupling is an extension of citation indexing wherein one attempts to measure the substantive similarity of documents, and thus the relevance of their content to search queries, by determining their "coupling strength", that is, the degree to which both cite the same sources. But it has been my experience that bibliographic coupling is by no means an exclusive measure of one's possible interest in a new paper - - a paper, for example, that may cite only one of the source references that is part of my *ASCA*® profile. It is not really difficult to understand why this should be so. To the extent that two or more members of an invisible college are familiar with the literature, their papers are likely to exhibit high coupling strength. But if a nonmember cites just one pertinent reference, it may put me in touch with work that is sufficiently novel to warrant further investigation.

In any case, Meetham has performed

a very useful service in investigating methods of evaluating the performance of retrieval systems. I might add that he was not the first to suggest the use of information theory in this field[4].

1. Meetham, A.R. Communication theory and the evaluation of information retrieval systems. *Information Storage and Retrieval* 5(3): 129-134 (1969).
2. Anonymous. ACS report rates information system efficiency. *Chem. Eng. News* 47(31): 45-46 (July 28, 1969).
3. Kessler, M.M. Bibliographic coupling extended in time; ten case histories. *Information Storage and Retrieval* 1: 169 (1963).
4. Garfield, E. Information theory and other quantitative factors in code design for document card systems. *Journal of Chemical Documentation* 1: 70 (1961).

Information Theory and Other Quantitative Factors in Code Design for Document Card Systems*

Eugene Garfield

Institute for Scientific Information
325 Chestnut Street
Philadelphia, Pennsylvania 19106

In the past ten years, the field of information retrieval has witnessed the development of many new systems, devices, and theories. In particular, two opposing "schools" of thought on card indexing systems have developed. One claims that the term card (unit term) or "collating" system is the most desirable. The other advocates the document card (unit record) or "scanning" system. Dr. Whaley has noted many of the advantages and disadvantages of collating and scanning systems, and I am glad to adopt his terminology and agree with most of his comments.[1] For the record, however, I wish to remind the proponents of term card systems that theirs was no new finding. Costello[2] says Batten[3] anticipated Taube [4] by 15 years. Batten was anticipated by at least another 35 years.

One term card system began at the turn of the century at Johns Hopkins Hospital. Subsequently, it went through all the evolutionary stages which clearly demonstrate the inherent similarities between term card and document card systems. This does not mean that the rediscovery of the term card system was an insignificant development. After all, many useful ideas and inventions are rediscovered and we are grateful for these discoveries. However, when appropriate, our precursors ought to be given credit. Even the ten column posting card was anticipated by Paul Otlet, founder of the modern documentation movement.[5] Indeed, long ago, the term card system was used in several medical institutions, including Johns Hopkins Hospital and the Mayo Clinic.

Texts on medical records management demonstrate such systems.[6] These consist of one 3 x 5 card for each disease (term). Each card then lists the case history document numbers for all patients so diagnosed. Ultimately, the number of case history numbers grew larger

Reprinted with permission from the *Journal of Chemical Documentation* 1:70, 1961.

* Presented at the American Documentation Institute Annual Meeting, 22 October 1959, Lehigh University, Bethlehem, Pa.

and the time required to make any correlations between two diagnostic term cards increased to ridiculous, exponential proportions. Somewhere along the line it was decided that the document card system should be employed. At Johns Hopkins and Mayo, Hollerith cards were in use as early as the 1920's. The School of Public Health at Johns Hopkins was one of the earliest users of punched-card machines. Their equipment is still of early vintage. At Johns Hopkins, even the IBM card finally became a problem as the volume of patients grew into the hundreds of thousands. The "vicious circle" was continued when it was decided to use duplicate sets of cards—*i.e.*, rotated files, not unlike the system used at the Chemical-Biological Coordination Center (CBCC) several years ago.[7] Finally, this semi-collating,semi-scanning system was abandoned because of the high cost of storing millions of cards. The entire file was tabulated on printed sheets and the punched-cards thrown out. This printed index arrangement is very similar to the original term card arrangement. However, in a separate section, the equivalent of the document card is also printed. Thus, one is able to do a search by both methods. Depending upon the individual search either one or both may be used. Pre-coordinations were made where appropriate before printing the index.

The Mayo Clinic long ago attacked the space problem in another fashion. The storage density of the IBM card was increased by a system of binary coding.[8] These IBM methods, I believe, are still used there. The binary coding utilizes all of the 4024 combinations possible in a 12 position punched-card column. It is understandable that a group of statisticians would discover this method. After all, statisticians work with probability data constantly. However, it is interesting that many people, including the statisticians, have been clever in finding ways of increasing the number of codes that can be crammed on a card (Wise,[9] Mooers,[10] *et al.*). However, the problem of how many times each was used was not considered as important.

This aspect first troubled me while working with the IBM 101 at the Welch Medical Library Indexing Project.[11] Some readers may recall the experimental 101 system we demonstrated in 1953 using five digit decimal codes, randomly strung along the first sixty columns of an IBM card.[12] For each subject heading or descriptor there was one five digit decimal number. Each card contained 12 such numbers. The details are described in the final report of the project. To use the same code length for all descriptors regardless of their frequency was rather inefficient in terms of space utilization, input time and searching cost. Obviously, others have arrived at similar conclusions because their coding systems *intuitively* employ a statistical approach. It is surprising, however, how many extant systems still do not make provisions for "normal distribution." A good example is the CBCC system, and the same is true of Uniterm,[13] Zatocoding[14] and others. To reiterate: they all use the same amount of coding space for each descriptor, regardless of its

frequency of use.

Working with the CBCC system, and utilizing Heumann's statistical data[15] on about 25,000 chemical compounds coded with this system, it was possible to design a code which reduced significantly card space and the time and cost of searching. For the moment it is sufficient to state briefly that the statistical information available on the CBCC file was used to construct a normal distribution curve giving the frequency of use of each alphanumerical code. One then arbitrarily breaks into the frequency curves in various sections to determine the space allocations for the descriptors. If a descriptor, such as benzene, occurs in half the chemicals and the code for uranium occurs rarely, why devote the same amount of space to both. Obviously, as Wiswesser,[16] Steidle[17] and many others have found, it is quite sufficient to assign permanent card locations to frequently occurring codes. On the other hand, descriptors which occur infrequently can be assigned some coding configuration which requires, relatively, a great deal of card space. This will be of little consequence since it will crop up so rarely. These "rare" birds are treated as a class and codes are used that permit many combinations in a larger space. The Mayo system is one example; another is the Zator system, as applied by Schultz.[18] Indeed, one of the primary shortcomings of Mooers' Zator system is the indiscriminate, *i.e.*, random assignment of an equal number of code symbols regardless of actual occurrence in the file.[19] This results in excess noise, *i.e.*, false drops. Incidentally, I wish to point out that I am well aware of Mooers' early attempt in *American Documentation* to set Wise straight on the folly of a superimposed coding scheme for the now defunct Rapid Selector.[9,10] However, to use probability theory is one thing—to use information theory is something else. We all readily can visualize methods of utilizing card space that will *grossly* take advantage of the facts revealed by a statistical analysis of the use made of a particular descriptor dictionary or subject heading list. The theoretician, however, wants precise quantitative criteria for allocating code space to individual descriptors or groups of descriptors. Here is where Information Theory comes to the rescue. The design of the most efficient coding system does not depend upon the meaning of terms. The terms, by themselves, have no informational value. Rather, it is the frequency of use of a particular descriptor which determines its informational content. One can only measure the amount of information in the word benzene when transmitting it in English text. As a code or term in a document collection dictionary, the word has no value. It is only significant in so far as it occurs with a particular frequency. If half of the chemicals coded contain benzene then the knowledge that a particular chemical contains benzene reduces the remaining choices to one half.

Having cleared the cobwebs on what the real "coding" problem is in documentation systems it is then relatively simple to apply Shannon's basic formula for measuring informational content.[20] I might mention that it is difficult, at first, to

think of the card searching problem as a transmission problem. However, if you think in terms of magnetic tape systems (Univac) or paper tape systems such as the Western Reserve Scanner, it is easier to see an analogy between "transmission" and searching.

The information content of a document file is neither the number of descriptors used, nor the number of documents which the various combinations of descriptors constitute. The information content of a document collection is a function of the *probabilities* of the descriptors in the dictionary. *H*, the familiar thermodynamic entropy function, and Shannon's measure of information, is equal to the sum of the individual probabilities multiplied by the logarithm of the individual probabilities, *ie.*, $H = -(P_1 \log P_1 + P_2 \log P_2 + \ldots + P_n \log P_n)$.

From this we are able to draw many interesting conclusions. For example, a document collection of 1,000 documents may contain no more information than a document collection of one million documents. This fact accounts for the intuitive decision of the Patent Office to use a "composited" card, which in certain cases is quite justifiable.[21] It also can be shown that the informational equality in two such files can be changed readily if the depth of indexing is altered. Indeed, if the informational content remains constant during such a growth one must either conclude that unnecessary cards remain in the file, new sub-dividing terms are required, or noise is present during a search. This situation is illustrated perfectly by our experience in coding steroid chemicals using the Patent Office code. In many instances a dozen different steroids were coded exactly alike. If the code dictionary is not changed, it is properly concluded that it is more economical to "composite" the 12 cards into one. However, one could increase the specificity of the coding. From the point of view of the Patent Office, with emphasis on the generic approach, the former conclusion, compositing, may appear simplest. From the point of view of the research chemist the latter approach, more specificity in coding, is more desirable. Taube's paper at the ICSI Conference implies that a term card system for the same steroid file could be used as readily as the Patent Office document card system.[22] This has a theoretical validity in view of the fact that in *both* systems no attention whatsoever is devoted to the frequency of occurrence of the various codes. (The Patent Office uses one punched hole position for each descriptor and the Uniterm system uses a 4 digit document number for each descriptor.) Indeed, from a tabulation of the coding done by the Patent Office of over 2500 U.S. patents, involving about 35,000 codes, it is no coincidence to find that seven descriptors account for over 9,200 codes, 16 additional account for another 9,100, the next 52 another 9,400 and all the remaining 359 descriptors 6,800.[23] Deciding the relative merits of working with a term card involving 1,500 document numbers (the highest frequency code) or the time to run 2,500 cards through a machine with a speed varying (according to price) from 500 to 2,000 cards per minute is meaningless. This becomes particularly ludicrous if one

then considers the time required to find those chemicals containing both a 3-Hydroxy Steroid code and a 17-Hydroxy steroid which occurs with almost equal frequency (1,200 occurrences). Instead of matching numbers on Uniterm cards by eye, one can speed this up by "collating" on an IBM machine at speeds comparable to the sorting operation. Using a Ramac system or a high speed computer this can be speeded further.[24] The point is that each system, according to the circumstances, has advantages and for this reason, in certain cases, I have used a combination of both—even going so far as to maintain two independent systems. This is commonly done, but not admitted, in many installations.

Returning to the discussion of the now measurable quantity H of an information file, to explain how this measure of information is determined and used, I must resort to basic Information Theory. For that I have paraphrased Shannon's own words, to which I refer those who are not yet familiar with Information Theory.[20]

Information Theory is concerned with the discovery of mathematical laws governing systems designed to communicate or manipulate information. It sets up quantitative measures of information and the capacity to transmit, store and process information. Information is interpreted to include the messages occurring in standard communication media, computers, and even the nerve networks of animals. The signals or messages need not be meaningful in any ordinary sense. Information Theory is quite different from classical communication engineering theory, which deals with the devices used—not with that which is communicated.

I submit that most of the polemics concerning devices, *i.e.*, term card *vs.* document card systems have kept us in the dark ages of conventional engineering theory. Relatively speaking, we have paid little attention to the nature of the information itself. This led to the failure to design really efficient searching devices; anyone who rents an IBM machine knows this. The measure of information, H, is important because it determines the saving in transmission time that is possible, by proper encoding, due to the statistics of the message source. Consider a model language in which there are only four letters—A, B, C, and D. These letters have probabilities 1/2, 1/4, 1/8 and 1/8. In a long text, A will occur 1/2 the time, B one quarter, and C and D each 1/8. Suppose this language is to be encoded into binary digits, 0 or 1 as in a pulse system with two types of pulse. The most direct code is: A equal 00, B equal 01, C equal 10, and D equal 11. This code requires 2 binary digits per letter. However, a better code can be constructed, with A equal 0, B equal 10, C equal 110 and D equal 111. The number of binary digits used in this code is smaller ***on the average.*** It will equal $1/2\ (1) + 1/4\ (2) + 1/8\ (3) + 1/8\ (3) = 1\ 3/4$, where the first term derives from letter A, second B, *etc.* This is just the value of H found if the probability functions are calculated.

The result verified for this special case holds generally—if the information rate of the message is H bits per letter, it is possible to encode it

into binary digits using, on the average, only H binary digits per letter of text. There is no method of encoding which uses less than this amount if the original message is to be recovered without noise. An average of 1 1/4 bits is possible if the message is allowed to be noisy, *i.e.*, not a completely faithful rendition of the original message.

Before we can consider how information is to be measured it is necessary to clarify the precise meaning of "Information" to the communication engineer. In general, messages to be transmitted have "meaning," but have no bearing on the problem of transmitting the information. It is as difficult to transmit nonsense words or syllables as meaningful text (more so in fact). The significant point is that one particular message is chosen from a set of possible messages. What must be transmitted is a specification of the particular message chosen by the information source. The original message can be reconstructed at the receiving point only if such an unambiguous specification is transmitted. Thus "information" is associated with the notion of a choice of a set of possibilities. Furthermore, these choices occur with certain probabilities; some messages are more frequent than others.

The simplest type of choice is from two possibilities, each with probability 1/2, as when a coin is tossed. It is convenient, but not necessary, to use as the basic unit the binary digit or bit. *If* there are N possibilities, all equally likely, the amount of information is given by $\log_2 N$. If the probabilities are not equal, the formula is more complicated. When the choices have probabilities $P_1, P_2, \ldots, P_n$, the amount of information H is given by the equation above. An information source produces a message which consists not of a single choice but of a sequence of choices, for example, the letters of a printed text or the elementary words or sounds of speech. In these cases, by an application of a generalized formula for H, the rate of production of information can be calculated. This "information" rate for English text is roughly one bit per letter, when statistical structure out to sentence length is considered (see *Bell System Tech. J.*, October 1949[25] or "Encyclopedia Britannica" article on Information Theory[26]).

The problem of applying information theory to documentation, I believe, is to be solved in properly defining the information source, which is the totality of descriptors assigned in any file. The next problem is defining the language units, *i.e.*, the descriptors and/or their components. A classification number, *e.g.*, has built into it much more information than a Uniterm. Each facet of the class number must be taken into consideration when measuring the information content of a classification system. It is then necessary to determine the probabilities of the units involved.

I will further hazard the statement that in the design of a document card of the IBM type *the most efficient space utilization will be obtained when the informational content of all card fields approach equality.* For example, in the case of the steroid file mentioned above, a card of four basic fields could be designed in which about 25% of the

information was contained in each. The first "field" would consist of one column of 12 punches. The twelve most frequently occurring codes would be assigned to each of the twelve locations. The next eighteen codes would be accommodated in another column divided into six sections, each of which could accommodate three different mutually exclusive codes. You cannot have a steroid which is both an 11-keto and an 11-hydroxy compound. In actual punched-card application I suspect that one would continue to use the first five columns, at least, for direct codes covering the first 60 most frequently occurring descriptors. If not, another field could be used to accommodate the next 28 codes dividing one or more columns into 4 sections, each containing 3 punches. To accommodate the remaining 359 codes in one field would be quite simple by using all the 495 combinations (binary) of four hole punching patterns possible. The number of columns in the field would depend upon the average number of such codes possible in a single compound. Specific characteristics of existing equipment may modify this decision.

The preceding example of applying measures of information content to the design of an IBM card has been very brief and may not be entirely clear to those not familiar with IBM machines. It is important, at this point, to make clear the similarity between this simple code for an IBM card and a similar code that can be used for a variety of document card or scanning card systems. Let us take up a brief discussion of the qualitative aspects of document cards systems, particularly as they relate to coding.

By document card systems, as contrasted to term card systems, we mean systems wherein all descriptors, or codes for descriptors, are retained together in the particular storage medium involved. Thus, in a punched-card document card system, *i.e.*, McBee, E-Z Sort, IBM, Remington Rand, Underwood-Samas, *etc.*, the holes or perforations are used to encode descriptors assigned to individual documents.[27] In a limited sense, the card *is* the document. Indeed, if the coding were sufficiently elaborate and detailed the card could be the document. The original Luhn Scanner employed an IBM card in which semantically factored words were stretched across the card to form an encoded telegraphic style message.[28] The IBM card employed was the standard 80 column card with a total of 960 punching positions.

Punched-card document card systems have their counterparts in film (Filmore[29] and Minicard[30]) where again all the descriptor codes are assembled together on a single piece of unitized film. The coding patterns may or may not be exactly of the type found on punched-cards. However, black or white spots correspond to perforations or the lack of perforations. The film-card (microfiche) may also contain a micro image of the original document. Similarly, an IBM card could contain the same micro image in a microfilm insert (Filmsort).[31] Similarly, the Magnacard[32] is the magnetic analog of a punched card. In this case information is coded as magnetized spots on magnetic tape.

The unit-card characteristic com-

mon to punched-cards, film cards, and magnetic cards is not only found in document-card systems. The same information found on Magna-cards can be stored on continuous magnetic tape. This is done on Univac and the IBM 700 series computers. The mechanisms employed to scan the "card" (sections of tape) are naturally somewhat different. Similarly, the defunct Rapid Selector was a continuous series of Filmorex cards strung out on one reel of film.[33] In the Benson-Lehner Flip system, the Rapid Selector system is partially revived.[34] A compromise between Filmorex and the Rapid Selector was suggested in the AMFIS system by Avakian.[35] The serial counterpart of perforated cards can be found in the Flexowriter tape used at Western Reserve where each document is represented by a series of codes exactly as in the fasion of the Luhn scanner.[36] This is no different from teletype tape except for the number of channels involved and the selector circuitry.

The Zator card is another version of the punched card.[37] The coding method employed has no basic dependence upon the card. It can be used with any type of document card system. Superimposition of codes is employed to make more efficient use of space. I mentioned earlier some of the limitations of Zator coding theory.

There are, obviously, many factors to consider in evaluating document card systems. Cost is one factor, but I believe its relative importance has been overly stressed by Taube and others.[38] Document card systems are not inherently expensive, nor small collections of manual punched-cards. Dr. Whaley has covered more than adequately many other factors which may favor the document-card or scanning card system.[1] He particularly stressed the need, sometimes, to retain relationships between various descriptors. He did not stress adequately the advantages in terms of input convenience and cost, where it is equally advantageous to keep codes together. Preparing a single IBM card is simpler than posting a dozen or more document numbers to individual term cards. It is also simpler than duplicating the same card a dozen times, each to be filed in twelve different file locations.

At the present time, punching a really efficient IBM card is difficult because the IBM machines are not designed for retrieval purposes exclusively. However, in my own experience, preparing elaborately punched cards is not an insurmountable obstacle. Key-punching costs are not considered major problems when a file is used repeatedly. Another factor to consider is searching time for large files. This can be cut down by converting to speedier machines—if time is a problem.

The major criticism of existing document-card systems is the need to operate in a "scanning" sense, *i.e.*, each card or each unit of tape or file must physically pass by a scanning unit. When there are large volumes of records involved very high speeds may be required. This is not only costly, but it will be obvious that there is a limit to the speeds we can reach in *mechanically* transporting cards, film, *etc.* It is phenomenal how fast some sorting and scanning devices do work, and possibly these speeds will satisfy

most requirements for a long time. However, these speeds are generally available only at a relatively high price. IBM machine rentals are higher in proportion to the speed at which they work, presumably because of greater maintenance and engineering cost. IBM tabulator rentals also vary according to the speed at which they are operated.

An ideal document card system would be one in which the basic advantages are retained—unit record input and storage, logical capabilities, *etc.* However, one would like to eliminate the need to scan the entire document file, in a *physical* sense, *i.e.*, by passing cards through a sorter, or magnetic tape past a reading lead, running film by a photoelectric cell. I believe such a system is possible and required particularly if we are to achieve the ultimate in access time. Such a system would be a truly random-access system and not a term card system using so-called random access. Systems such as RAMAC or AMFIS do not appear to be as energy consuming as high speed tape readers or sorters on punched cards, but their mechanical characteristics would seem to be limiting. It is comparable to solving the problem of sorting at high speeds by using a dozen sorters all at once. Similarly to use the equivalent of a dozen magnetic tape readers is no fundamental solution. In the ideal, the file will remain completely stationary and the scanning mechanism will be able to identify the existence of desired codes by scanning in a non-mechanical fashion. An approach in this direction is seen in the Bell Telephone system of routing long distance calls by use of special punched cards. Verner W. Clapp once asked me why you couldn't wave a flashlight at a file and have it throw out the answers. This is not impossible. I have been exploring a similar principle utilizing electromagnetic phenomena which I have called Radio Retrieval.

In conclusion, I have tried to show the fundamental similarities between so-called term card and document card systems by tracing the cyclical evolution of a term card system into a document card system, then into a semi-document card system employing collating methods, and finally back to a term card printed index arrangement. I maintain that the differences between term and document card systems are basically illusory. You will find vigorous proponents for each system depending upon the circumstances. If one had no indexing system at all in the first place, any system is an improvement. Once a system is adopted, thereby improving access to documents, a proposal to merely change the mechanics will not usually excite people.

An area of research which requires more fundamental work is in coding. No matter what system is used, the same amount of information is produced if one uses the same code dictionary and code frequencies.

The Patent Office Steroid Code would be, theoretically, equally efficient with a term card system as in its present document card system. From a practical point of view, it would not. Using Information Theory the coding space required in a document card system can be reduced considerably. It is possible that similar efficiencies are possible

in designing term card systems, but these are not yet apparent and may be difficult to find. In other words, term card systems are inherently inefficient because they seemingly cannot take advantage of the variations in code frequencies which are inherent to *all* information systems. According to Keckley, "there is a central tendency for 90% of the activity to be concentrated within 25% of the classifications."[39] This appears to be well substantiated in the coding of 2,500 steroid compounds from the literature. Furthermore, term card space requirements may increase exponentially as the size of the collection grows. A collection of 1,000 documents requires less than 7 bits per descriptor assignment, a collection of 10,000 about 12 bits per descriptor assignment, 100,000 16 bits, and 1,000,000 20 bits.

Mooers deserves credit for recognizing the value of Information Theory for retrieval theory.[40] However, it is just as inefficient to use five punched holes for every descriptor on a document card as it is to use a five digit document number on a term card. By proper application of descriptor probabilities Information Theory can make Zato coding even more powerful.

It has been shown that one can quantitatively measure the amount of information in a document collection by the Shannon formula $H = -(P_1 \log P_1 + P_2 \log P_2 + \ldots P_n \log P_n)$

As a result of this expression, it is concluded that the size of a document collection is no realistic measure of its "information content." Indeed, *two collections of entirely different size contain the "same" information if they use exactly the same code or dictionary with the same percentage distribution of descriptors.* Thus, in this sense the Library of Congress Subject Catalog contains no more information than the local Public Library Catalog. This may sound startling or ridiculous to librarians. However, as long as the local Library uses the LC Subject Heading Authority List, it may even contain *more* information because it may add further refinements to the existing LC dictionary or use it with varying frequency assignments. A special library is of more use to its clientele than is the Library of Congress. *To alter the information content of a collection one must index in greater depth—not index more documents.* This point is most important in industry.

Analysis of the Patent Office steroid code frequencies illustrates in a simple case how Information Theory may be put to use.[41] A brief summary and review of Shannon's Information Theory has been presented to show that the past preoccupation of documentalists with devices is comparable to the earlier preoccupation of communication engineers with machines rather than the information they were transmitting. The main problem in applying information theory in documentation is in defining the "information source" and the "channel." A completely successful retrieval system must combine the advantages of both term and document card systems in such a way that all inertial characteristics of existing systems are removed.

REFERENCES

1. **Whaley F R.** The manipulation of nonconventional indexing systems. *American Documentation* 12:101-107, 1961. Presented at the American Documentation Institute Annual Meeting, 22 October 1959, Lehigh University, Bethlehem, Pa.
2. **Costello J C.** Uniterm indexing principles, problems and solutions. *American Documentation* 12:20, 1961.
3. **Batten W E.** Specialized files for patent searching. *Punched cards: their application to science and industry.* (Casey R S & Perry J W, eds.) New York: Reinhold, 1951, p. 169-181.
4. **Taube M. & Associates.** *Studies in coordinate indexing.* Vol. 1. Washington, D.C.: Documentation, Inc., 1953.
5. **Otlet P.** *Traité de documentation: le livre sur le livre: théorie et pratique.* Brussels: Editiones Mundaneum, Palais Mondial, 1934, p. 388.
6. **Cristina S X.** History of writing and records. *Hospital Management.* Part 1, 72:111, 1958. Part 2, 86:82, 1958.
7. **Beard R L & Heumann K F.** The chemical-biological coordination center: an experiment in documentation. *Science* 116:553, 1952.
8. **Berkson J.** A system of codification of medical diagnoses for application to punched cards with a plan of operation. *American Journal of Public Health* 26:606, 1936.
9. **Wise C S & Perry J W.** Multiple coding and the rapid selector. *American Documentation* 3:223, 1952.
10. **Mooers C N.** Coding, information retrieval and the rapid selector. *American Documentation* 1:225, 1950.
11. **Himwich W A, Garfield E, Field H G, Whittock J M & Larkey S V.** *Final report on machine methods for information searching: Welch Medical Library Indexing Project.* Baltimore: Johns Hopkins University, 1955.
12. **Garfield E.** Preliminary report on the mechanical analysis of information by use of the 101 statistical punch card machines. *American Documentation* 5:7, 1954.
13. *The uniterm system of indexing: operating manual.* Washington, D.C.: Documentation, Inc., 1955.
14. **Mooers C N.** Zatocoding applied to mechanical organization of knowledge. *American Documentation* 2:20, 1951.
15. **Heumann K F & Dale E.** *Statistical survey of chemical structure.* Presented at the American Chemical Society Meeting, 12 September 1955, Minneapolis, Minn.
16. **Wiswesser W J.** *A line-formula chemical notation.* New York: Thomas Crowell Co., 1954.
17. **Steidle W.** Possibilities of mechanical documentation in organic chemistry. *Pharmazeutische Industrie* 19:88, 1957.
18. **Schultz C K.** An application of random codes for literature searching. *Punched cards: their application to science and industry.* (Casey R S, Perry J W, Berry M M & Kent R A, eds.) New York: Reinhold, 1958, p. 232-247.
19. **Mooers C N.** Zatocoding and developments in information retrieval. *ASLIB Proceedings* 8:3, 1956.
20. **Shannon C E & Weaver W.** *The mathematical theory of communication.* Urbana: University of Illinois, 1949.
21. **Baily M E, Lanham B E & Leibowitz J.** Mechanized searching in the US Patent Office. *Journal of the Patent Office Society* 35:566, 1953.

22. **Taube M.** The Comac: an efficient punched card collating system for the storage and retrieval of information. *Proceedings of the International Conference on Scientific Information.* Vol. 2. Washington, D.C.: National Academy of Sciences/National Research Council, 1959, p. 1245-1254.
23. **Ball N T.** Searching patents by machine. *American Documentation* 6:88, 1955.
24. **Nolan J J.** Information storage and retrieval using a large scale random access memory. *American Documentation* 10:27, 1959.
25. **Shannon C E.** Communication theory of secrecy systems. *Bell System Technical Journal* 28:656, 1949.
26. ———————. Information theory. *Encyclopedia Britannica* 14th edition. Vol. 12. Chicago: Encyclopedia Britannica, Inc., 1958, p. 350.
27. **Casey R S, Perry J W, Berry M M & Kent A,** eds. *Punched cards: their application to science and industry.* New York: Reinhold, 1958.
28. **Luhn H P.** The IBM universal card scanner for punched cards information searching systems. *Emerging solutions for mechanizing the storage & retrieval of information: studies in coordinate indexing.* Vol. 5. (Taube M, ed.) Washington, D.C.: Documentation, Inc., 1959, p. 112-140.
29. **Samain J.** *Filmorex. Une nouvelle technique de classement et de sélection des documents et des informations.* Paris 1952.
30. **Taube M.** The Minicard system: a case study in the application of storage and retrieval theory. *The mechanization of data retrieval: studies in coordinate indexing.* Vol. 5. (Taube M, ed.) Washington, D.C.: Documentation, Inc., 1957, p. 55-100.
31. **Rees T H.** Commercially available equipment and supplies. *Punched cards: their application to science and industry.* (Casey R S, Perry J W, Berry M M & Kent A, eds.) New York: Reinhold, 1958, p. 80-90.
32. **Hayes R M.** *The Magnacard system.* Presented at the International Conference on Information Processing, June 1959, Paris, France.
33. **Shaw R R.** The rapid selector. *Journal of Documentation* 5:164, 1949.
34. Introduction to the FLIP (film library instantaneous presentation). *American Documentation* 8:330, 1957.
35. **Avakian E & Garfield E.** AMFIS—the automatic microfilm information system. *Special Libraries* 48:145, 1957.
36. **Perry J W.** The Western Reserve University searching selector. *Tools for machine literature searching.* (Perry J W & Kent A, eds.) New York: Interscience, 1958, p. 489-579.
37. **Mooers C N.** Scientific information retrieval systems for machine operation: case studies in design. *Zator Technical Bulletin #66.* Boston: Zator Company, 1951.
38. **Taube M.** *Studies in coordinate indexing.* 5 volumes. Washington, D.C.: Documentation, Inc., 1953-1959.
39. Unidentified reference. [I would appreciate hearing from any reader who is able to supply this reference.]
40. **Mooers C N.** Information retrieval viewed as temporal signalling. *International Congress of Mathematicians, Harvard University, August 30-September 6, 1950. Proceedings.* Providence, R.I.: American Mathematical Society, 1952, p. 572-573.
41. **Andrews D D, Frome J, Koller H R, Leibowitz J & Pfeffer H.** Recent advances in patent office searching: steroid compounds and ILAS. *Advances in documentation and library science. Vol. 2. Information systems in documentation.* (Shera J H, ed.) New York: Interscience, 1957, p. 447-477.

Amer. Documentation 5:7-12(1954)

PRELIMINARY REPORT ON THE MECHANICAL ANALYSIS OF INFORMATION BY USE OF THE 101 STATISTICAL PUNCHED CARD MACHINE

EUGENE GARFIELD*

The "analysis of information" has a variety of meanings. One may have a file of documents and wish to select in various ways types of information with any number of criteria as the basis of selection. Thus one may be interested in certain statistics, even if only the number of documents stored. More often one is interested in the number of units meeting certain requirements as in census counts. One may wish to select and ultimately remove from the file units which meet certain of these criteria. This is often referred to as selection or in the case of scientific documents or references – literature searching. These techniques may also be employed to establish correlations between previously unrelated data. The tremendous increase in the size of information files has made these problems most difficult to solve by conventional methods. It is felt that there may exist some remedy to this quantitative problem if we can in some way mechanize the procedures involved. This paper discusses one approach to the mechanization of information analysis, as embodied in the use of the IBM 101 Electronic Statistical Machine.

Before discussing the use of the 101 it is important that we consider why we use machines at all. With the rapid intermingling and overlapping of subject disciplines, especially in science, one might say that if the time were available some of us might, of necessity, read and digest all of the recorded information available. If this were possible, as it once was, we might not be so concerned with this problem of selection of information. But we do not have unlimited time. Indeed, the time factor is probably the essense of the problem. It is, therefore, necessary to speed up the process of selection by mechanization. Like most other applications of machines, we use the machine to do a task we could do ourselves, but we have neither the time nor the energy to do it. We use machines to facilitate operations we now do manually. It is not pertinent to discuss at this time whether machines "think" or not, but it should be mentioned that because machines are more efficient than man in repetitive operations, we find that we are today performing numerous tasks, especially in information analysis, that we would never have contemplated before. True, the

*Grolier Society Fellow, School of Library Service, Columbia University.

information must be fed into the machine – it must have been there in some form in the first place. But before using the machine the information was useless and without the machine would have remained dormant.

Having established why we use machines at all, we must consider some of the techniques required for using machines. This subject is sufficiently broad as to require separate treatment elsewhere. Deutsch[1] has analyzed the fundamentals of this problem admirably. However, we shall briefly discuss the concept of coding.

In order to employ machines efficiently it is necessary to translate information into a form more amenable to the mechanical operations one wishes to perform. This requires that somewhere along the line an encoding process take place. It may be possible to use a typewriter similar to the one preparing this page to record a name on a magnetic tape. The keyboard of the machine looks exactly the same as any other keyboard. However, the keys cause patterns of magnetic spots to appear on the tape. The typist is not aware that a coding process is taking place. The resultant tape can be fed into another machine which causes typewriter keys or type bars to be activated, typing the same name or item of information on a piece of paper. Externally one is not aware that a coding operation has taken place. The coding was done mechanically, but nevertheless coding took place. In other, less sophisticated machines it is necessary to perform coding operations that are quite apparent to the observer. For example, one may represent a name by a number. Adams may be coded as 1125, Jones as 3456 and Smith as 8698. This would enable certain machines to manipulate the information more easily than in the "original" form. This is true of punched-card machines which handle alphabetic information through numerical coding or by coding the letters of the alphabet into two-hole patterns. Thus, in the case of the names above it would be possible to arrange the names alphabetically in two ways. One could prepare a file of cards where the individual names are punched in letter codes on a card and then arrange the cards by machine in alphabetical order. Or one could merely punch the numerical code number on a card and arrange the cards in numerical order. It can be seen that such a numerical arrangement of the cards simultaneously alphabetizes the cards because the code numbers were assigned in increasing numerical value starting at A on through the alphabet. An added degree of machine efficiency is obtained if one has to deal with a four digit number rather than an eleven letter name. If one repeatedly alphabetizes the same file, the saving in time can be quite large. This might also apply in hand sorting such a file. Once the coding operation has been performed one has established the basis for mechanization. Consequently, these techniques may apply to the use of humans as well as machines.

In this paper the problem of literature searching shall be emphasized. The principles apply to information analysis of all kinds. The present work was initiated, however, with the specific problem of searching scientific literature in mind. In literature searching problems there is, prior to the coding operation mentioned above, a most important step necessary to implementing searches, mechanical and otherwise. We usually refer to this as indexing or cataloguing. In this operation we attempt to decide what avenues may lead to the particular document involved.**

Indexing decisions are usually based solely on the contents of the document. In certain specialized indexing operations the indexing

[1]Karl W. Deutsch, "Communication Theory and Social Science," The American Journal of Orthopsychiatry, vol. XXII, No. 3, July 1952, p. 469-83.

**This step is quite inefficient because we index every item even though a large percentage may never be desired or called for. However, it has as yet been impossible to decide in advance which items will be desired or what criteria will be required in making a search. Therefore, we must index everything – in advance. Perhaps we may someday find new methods of handling information that will obviate this very costly step. Until that time, however, indexing is fundamental to all searching systems. The indexing dilemma has its analogues in communication problems of all sorts. If the telephone company knew, in advance, those telephone numbers to be searched for in directories, it would be possible to prepare much smaller directories. It would be interesting to learn the number of names that are never consulted in the directories. A preliminary statistic would be the number of unlisted telephones.

will also include considerations of the users interests, e. g., a document may concern the budget of a certain industrial corporation. A medical indexing staff may decide that this document may be of interest to members of the medical profession, even though there is not the slightest mention of medicine. However, it is impossible to anticipate all of the possible avenues of approach to a particular document. To facilitate indexing, indexers select a number of descriptors which most adequately cover the subject material of the document. These are referred to as subject headings, terms, rubrics, etc. It will be seen that these descriptors taken together often constitute the basic subject matter of a document. Thus, a study on the use of DDT in agriculture may be adequately described by the subject headings DDT and AGRICULTURE.

In preparing documents for coding, the selection of these subject headings is therefore a most important step. Once this has been done coding can proceed or perhaps indexing and coding can be combined. Coding obviously cannot precede indexing. In the present study an indexer selects a subject heading and a coder assigns a code number to that heading once it is selected. Indexing would produce a data sheet or marks on original copy. The code numbers would usually be added to these data sheets or original copy by the coder. Once this is done it is possible to prepare a punched card. (If some other device than punched card equipment were used then the appropriate medium would be prepared as e.g., a strip of magnetized tape.) Once the punched card has been prepared we have established the "machine index." Efficient use of the index depends on the intervention of a machine.

The use of punched-cards in literature searching is not new. Punched-card installations of various kinds have been in existence for some time. However, the range of information problems handled by punched-card machines has been severely limited until recently because of limited flexibility. (This does not mean that in certain specific applications such as accountancy these machines are not capable of amazing flexibility.) It shall be shown that with the use of the 101 punched-card machine even greater versatility is possible if combined with well planned operations.

It will be useful to review the problem further and consider what have been the major difficulties in using standard punched-card equipment for the purposes of information analysis. One difficulty in using the punched-card is the physical limitation of the card itself. A 3 x 5 file card has an amazing storage capacity. The difficulty there is that printed matter is as yet impossible to search mechanically. The standard punched-card is larger than the 3 x 5 card but actually one is limited to the amount of information that can be placed in 80 columns or 960 different punching positions, i. e., 12 to a column. One must add to this great physical limitation the limitations imposed by the various punched-card machines in their ability to manipulate these cards. Thus, the standard sorting machine can only operate on one column at a time. This is the equivalent of reading one letter on a printed page. With certain attachments one can increase the number of columns that can be searched simultaneously. In other machines like the collator there is increased searching ability. Suffice it to say that these limitations of card capacity and machine flexibility have necessitated many laborious techniques in preparing punched-card files. One of these is the technique of placing in a designated area of the card a specific category of information. This results in what is called the fixed field card. Thus, if one has specified that all chemical information is to be punched in the first ten columns of the card it is only necessary to search one eighth of the card to locate certain items of chemical interest. One difficulty that immediately arises here is that there is considerable waste of space. In a medical file perhaps only ten to fifteen percent of the information is of a chemical nature. On the other hand, those documents that do deal with chemical concepts may require several chemical descriptors. If the card has room for only one chemical descriptor it is necessary to prepare a card for each such descriptor. However, one may ask why use the fixed field card? This is reasonable. If this is not done one loses efficiency in employing the machines since it would be necessary to search the entire card if punching were random. On a standard sorter this might mean a fantastic increase in sorting time. In the present study we asked the same question. Would it be possible to search a card which was not of the fixed field type? This is basically the same approach used in IBM's photoelectric

scanning punched card machine.[3] Briefly then, it is intended to show primarily that it is possible to prepare a rather efficient punched card file, which can be searched with the 101 with extreme versatility. This machine, if these new techniques are employed, can be useful in extremely complicated information selection problems as well as various other standard searching problems.

The Welch Medical Indexing Project has specified certain criteria in approaching the use of machines for the searching of scientific literature.[3] It was felt that simplicity was paramount to our operations. This applies to punching as well as coding. This further applies to searching. Mauchly[4] has stated that since coding and punching is done only once this may be too harsh a requirement. In principle he is correct. But in terms of the immediately practical problems of indexing medical literature it was felt that this requirement could not be overlooked. The various avenues that brought us to the techniques employed will not be discussed. It merely remains to describe the capabilities of the system, as well as some of its operating features.

The punched-card is divided into areas of a specified number of columns. Thus, in figure 1 sixteen five column areas are shown. Five digit code numbers are punched in each of these areas. These numbers are punched without any reference to category as is necessary on the fixed field card. As many as sixteen code numbers could be punched on the card. Indeed, all sixteen could be from the same discipline such as sixteen symptoms in a medical case history. The code numbers presently used are numerical. However, they could be alphabetical or a combination of the two. If a document requires more than sixteen five digit descriptors it is possible to use as many additional cards as required. This might be the case in purchasing and supply files where items are described according to dozens of criteria as in steamship parts. Chemical documents may contain information on hundreds of compounds. The code numbers which are employed by the Indexing Project are the same as the serial numbers used in connection with punched card operations intended for the preparation of printed indexes[5] as contrasted with the present operation involving machine searching.

The details of the actual 101 machine functions will be explained elsewhere, as well as certain mathematical considerations pertinent to our use of the machine. The important point now is – what is the 101 capable of?

It is possible to search the punched-card file for any code number desired on a single pass of the cards. Since the card does not use fixed fields it is not necessary to specify that the code number will be found in a certain location. This has been obviated by special wiring of the control panels of the 101. The ability to search for any particular code number is important. However, what does this mean in practical terms? In conducting a literature search one establishes certain criteria for making that search. Thus, in searching for all documents on antibiotics one must assume that in the indexing procedure all pertinent documents were indexed under antibiotics and that the code number for antibiotics appears in any card that will be selected by the machine.* In the language of symbolic logic the ability to search for a single code number may be stated as meeting the requirements of a first order search. What about the higher order searches which may involve what are called logical sums, products and differences? One may specify in a search that all desired documents should have been coded for antibiotics (code number A). One may further specify that any document coded for antihistaminics (code number B) will also be desired.

[2] "Mechanized System Launches New Era for Literature Searching," Chemical and Engineering News, vol. 30, No. 27, July 7, 1952, p. 2806-10.

[3] Sanford V. Larkey, Williamina A. Himwich, and Helen G. Field, "Categorization as a Basis for Machine Coding," unpublished report.

[4] John W. Mauchly, Personal Communication.

[5] Eugene Garfield, "The Preparation of the CURRENT LIST OF MEDICAL LITERATURE by Punched-Card Methods," unpublished report.

*It is not irrelevant to mention at this point that using a machine of this type should probably not be considered if one is searching for an article written by John Jones in 1952. One should not confuse the problems involved in printed indexes and "machine indexes." You do not need a Cadillac to cross the street. The failure of the punched card equipment at Harwell (6) was not surprising, since one should only contemplate using machines for tasks which are too difficult if not impossible to perform by existing techniques.

This is a logical sum, i.e., A + B. One may specify that documents coded for A are desired but only if they do not contain B. This is a logical difference, i. e., A - B. One may finally specify that selected documents be coded for both A and B. This is a logical product, i.e., AB. These examples are second order searches, i.e., they involve two descriptors. Using our 101 techniques it is possible to make all of the above searches. Furthermore, it is possible to make searches theoretically of the 50th order.** A fifth order search might be A + BC - (D + E). The requirements of this search are that if either D or E appear, the document is not desired; if B and C occur in the same document or if A appears then the document is desired providing D or E do not appear. In the language of the 101 one would first "test" for D or E. If either were present the card would not be selected. If neither D nor E were present the 101 would then "test" for A. If it were present the card would be selected. If A were not present the 101 would then "test" for the presence of B and C and only if both were present would the card be selected. Of course, all of the "tests" would be performed simultaneously.

This type of versatility is not available in most punched-card selection systems. However, this is not the limit of one's abilities to make searches with the 101. Careful consideration was given to the fact that in making searches by machine it is unfortunately necessary to scan every card in the file, unless special prefiling is done. Without specifying prefiling this (searching the entire file) is a most inefficient feature of mechanized searching. This is the case in the Rapid Selector[7] where thousands of frames of microfilm may be scanned in order to find one or a few desired documents. It was felt that this shortcoming could in part be minimized if it were possible to perform several searches simultaneously. In the case of the Harwell[6] experiment the complaint was that several searches could not be made simultaneously. Notwithstanding the fact that they were attempting to make searches that are more properly made with printed indexes or files of 3 x 5 cards, they erroneously concluded that simultaneous searches are not possible with punched-card equipment. Using the 101 it is possible to make simultaneous searches. Indeed it is possible to make as many as nine or ten fifth order searches at one time. The significance of this feature should not be overlooked, since it increases the effective speed of the machine as much as ten fold. Thus a search of one million cards that requires about 40 hours work is made considerably more practical when the same time is required to do ten searches simultaneously.

If we now take into consideration the possibilities of prefiling the punched card file it may be possible to speed up searches considerably. Several possibilities exist here. However, we shall at present only consider approaches which do not require duplication of cards, because this is one of the defects we are trying to remove by introducing more versatile equipment. (It is common practice in many centers to prepare a card for each descriptor used in indexing documents and by suitable prefiling it is possible to reduce the number of cards required for searching to a small number.) However, ultimately one runs into a space problem. If one has a million case histories with an average of ten symptoms per case one has to deal with ten million cards. Nevertheless, if one has extremely large files it is possible to visualize that even such duplication of cards would not obviate the need for the searching systems described here, since one may still search for combinations of criteria that appear many thousands of times in the file. Such is the case, e.g., in searching for all material on antibiotics in respiratory infections, or any other combination of generic terms. Possibly the right combination of prefiling and judicious programming will provide the most economical solution.

In dealing with a single card per document it is still possible to prefile cards in such a way as to make searching more efficient. One approach is to take into consideration the number

[6] H. D. Ashthorpe, "The Punched Card Indexing Experiment at the Library of the Atomic Energy Research Establishment, Harwell," ASLIB Proceedings, vol. 4, May 1952, p. 101-104.

[7] Ralph R. Shaw, "Machines and the Bibliographical Problems of the Twentieth Century," Bibliography in an Age of Science, U. of Illinois Press, Urbana, 1951, p. 58-62.

**The average document rarely requires more than a dozen descriptors. It is therefore unnecessary to make a search of higher order than the maximum number of descriptors assigned to any one document.