Reprinted in Essays of an Information Scientist, Vol:6, p.454-458, 1983

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.[2] 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 tecnhiques 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[8] 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

[8] 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.