Automation of Reference Work

CLAIRE K. SCHULTZ

THE EQUIPMENT NEEDED to automate reference work existed years before anyone tried to apply it. Perkins' patents, which led to the development of edge-notched punched cards, were issued in 1925 [1] and 1929; [2] Taylor, who is nationally credited with patenting the peek-a-boo principle, received his patent in 1915; [3] Hollerith developed internally punched cards and a sorter for them in preparation for the U.S. Census of 1890. The failure to apply this equipment indicated that librarians did not feel the need for automation until the 1940's, when experimentation with automation began. What, then, was the need which precipitated activity in the automation of reference work at that time?

Research information used to be published in books; however, because book publishing was too time-consuming for the articles to be of value, most research information came to be published in journals and reports. The increase of such articles necessitated up-to-date indexes to them. Until recently, however, indexes to these publications have been notoriously late in being issued. In addition, more flexible indexing approaches than those found in card catalogs or published indexes were needed. No matter how indexes or catalogs were arranged physically, questions always were asked of them that were difficult or impossible to answer in terms of the system used.

The reference librarian became progressively more sensitive to the inadequacy of his tools. The search for more adequate tools began in industrial libraries where librarians conducted the most specialized and intensive reference work for researchers.

The search led first to punched cards. In the mid-1940's several kinds of punched-card equipment were available. In addition, the

The author is Senior Research Associate, Institute for Advancement of Medical Communication, Bethesda, Maryland. The assistance of Mr. Bernard Epstein as technical editor is gratefully acknowledged.

Rapid Selector, based on the peek-a-boo principle, had been developed experimentally. This system, in which the codes assigned to indexing terms were recorded on rolls of film, provided a comparatively rapid method of searching for documents and printing out, from the film, copies of documents selected. Compared in cost and speed to edge-notched card systems, it was the "giant" equipment of the day. However, during the long development of the Rapid Selector, persons needing nonconventional indexing media became interested in manually operated card systems. These were either edge-notched systems such as those supplied by the McBee [4] or Zator [5] Companies, or the Uniterm System [6] introduced by Taube about 1950. Also beginning about 1950, IBM and Remington Rand punched cards were used by a few groups.[7] These were sorted by either a standard sorter or a program-controlled selective sorter.

An evolutionary pattern of development can be traced from punched cards to computers. The principles used in the application of edge-notched cards also were used in internally punched cards, except that with punched-card sorters more of the process could be automated. In addition, the internally punched cards made certain things feasible that were not feasible under less automated conditions, just as search methods not feasible with the card catalog are achieved simply with edge-notched cards. Introduction of the IBM 101 Electronic Statistical Machine around 1950 made selective sorting more powerful. With the 101, one could get not just all of the "nines" in the nine pocket and the "sevens" in the seven pocket; one could also direct cards into a particular pocket if all of several punches were present in a card. This ability was coupled with that of specifying certain patterns of cooccurrence of indexing terms, and dropping variations into different pockets. With computers, versatility and speed of selective sorting can be extended, and beyond that, a record can be maintained of what was done and what result the sorting produced. With each new generation of computers, sorting processes are speeded further, and thus more automation is made feasible.

Principles and Tools

INVERTED VS. NONINVERTED METHOD

The peek-a-boo principle of searching required storage-file arrangement different from that of other punched cards. The kind of filing used with peek-a-boo and Uniterm systems came to be known as the

inverted method. Inverted systems employ an indexing term as the unit record and display on the record a code for each document to which that indexing term applies. This method is like that used in the card catalog, where all pertinent document references are filed under the subject heading that applies to them. With both edge-notched and internally punched cards the document is the filing unit; the record for any document contains all of the indexing terms as well as any other pertinent data. The latter types of card systems produce one record per document, whereas peek-a-boo systems produce one record per term in the indexing vocabulary. Inverted systems are updated by adding new unit records. Inverted systems provide "random access" to stored information in that the user can choose an indexing term or group of indexing terms at random, locating them by means of the alphabetically filed unit records.

In contrast to inverted systems, noninverted systems usually are unordered and require total scan of the file, record by record, to locate desired information. Practical applications of inverted systems are limited by the number of document codes that can be stored in a single unit record because matching becomes involved if more than one card per term is to be matched. Noninverted systems do not require posting or file insertions; all additions are made by new records at the end of the file.

Whether or not the file is inverted is a mechanical consideration, important to the efficiency of the system, but of little importance to the intellectual aspects of reference work. In contrast, the freedom of rearrangement provided by all punched cards is highly important to the quality of reference work. Coordinate indexes are more amenable to rearrangement than are indexes with indented subject headings; thus, with appropriate insight, system designers have combined the advantages of coordinate indexing with those of punched cards.

THE THESAURUS

Quality of automated reference service is directly dependent on the system's authority list or thesaurus. Librarians understand that a card catalog cannot function effectively unless an authority list is maintained for catalogers and indexers and unless the cross references established are made available to the catalog users. The authority list in automated systems customarily is called a thesaurus. It may justifiably be called by this different name because a thesaurus contains some features not found in the usual authority list.

A thesaurus should include the terminology that represents the subject matter of interest to the users of the system in which it functions. It should not contain terminology chosen systematically for "all of knowledge" as the Dewey Relative Index does, but it should include terminology established empirically, that is, according to its use in documents and questions. Terms should be included according to the degree of specificity that will make the indexing most useful. One can determine the degree of specificity empirically by studying how questions are asked, for example.

In the thesaurus, similar and related terminology are cross referenced so that the documents indexed are most accessible. Cross referencing is like that in a conventional authority list, except that the most desirable cross references to use can be determined by statistical analysis. Entries in the thesaurus should be arranged so that they will be used consistently and can be accessed from the storage file efficiently. For example, entries in both the ASTIA Thesaurus [8] and the Medical Subject Headings of the National Library of Medicine [9] are arranged both alphabetically and categorically. Because such thesauri contain no overall structure of arrangement as in the Dewey system, the thesaurus user cannot take for granted that general-specific relationships have been incorporated. For that reason, the general-specific relationships among terms are made much more explicitly in a thesaurus than in an authority list.

SEARCH STRATEGY

Search strategy is involved every time a search is performed, whether it is done manually, with punched cards, or with a computer. Knowing how to start, what to do next, and how to separate the relevant from the irrelevant, is part of the built-in equipment of the reference librarian. The ambition to transfer these abilities to automated equipment has made system analysts aware that the decision process of the human being had to be objective, deliberate, and machine-like.

A librarian begins to develop a search strategy when she receives a request for information. Suppose, for example, a little boy asks a children's reference librarian, "What do you have on pets?" She would immediately suspect that the boy was not interested in all possible pets, and she would probably ask questions to discover his more exact interests. In the system designer's language, when she does this she is performing feedback; she is also establishing some of the parameters of the search. Let us assume that at the end of the feedback process

she knows that the little boy is interested only in how to keep a kitten. She might supply the answer to this question by producing a book she has right at her fingertips; for the sake of the example, however, let us handle the question as if it were a research request.

The librarian might divide the question into "care and feeding of kittens" and "having fun with kittens." This is part of the process of translating the question into the terminology of the system; the reference librarian is a very important part of the system. Consulting the card catalog (which also functions as her authority list) she finds that the subject heading *Kittens* says, "see Cats." After proceeding to *Cats* she finds it subdivided; she chooses the pertinent subheadings and finds about twenty apparently suitable references. From these references, she intuitively selects one or two books and sends the boy on his way.

If this had been a true research question, the librarian might have reviewed the method by which she selected the two books. Did she locate *all* of the pertinent material from which to choose? For example, perhaps a book on pets in general would have a chapter on kittens more useful than the material she actually gave the boy. If she found little under *cats,* she probably would have looked under *pets,* but she risked missing some references and did not look under the general heading. The cataloger could have obviated the question of whether or not all pertinent references had been obtained by having indexed the book on pets under *cats.*

In the human procedure, then, are many uncertainties, many steps in answering even a simple reference question, and many decisions to make. These all become important when the procedure is mechanized.

Some of the processes the reference librarian just performed have been automated. Equipment capability for a potential system plays a large role in the amount of the procedure that can be automated. The power of the system also is dependent on the search strategy developed for it. A weak strategy can be applied to a powerful computer, for example. In general, the more capable the machine, the more sophisticated the search strategy can be. In the following paragraphs each part of the search process is explored, and the degree of automation, to date, for each of the parts is described. It should be pointed out that this paper cannot discuss military systems that are classified. Also, certain systems are singled out for discussion because they represent either very large collections or because they seem to be leading the state of the art.

INPUT

Input starts with receipt of the question. The only automation of this has been by means of intercommunication such as mail delivery, telephone, personal secretary, and the like. In some instances indirect communication hinders rather than helps the procedure; if so, it is undesirable automation. At some point, a human being must receive the question for further handling. If there is a feedback process to establish additional parameters for the question, this, too, must be done by a human being. Translation of the question into the language of the system, that is, establishing what is wanted and the terminology to be used for finding it, is done by a human being in most cases. In one system, MEDLARS, (National Library of Medicine, Medical Literature and Retrieval System) the computer helps determine the terminology used to refine a search as it progresses.

After the terminology has been established, the logical connectives that are to be used among the terms must be determined. Logical connectives were used by the reference librarian who found books on kittens for the little boy, but their use did not have to be made explicit. If that same question were asked of a librarian using any mechanized system, even if it were as simple as a peek-a-boo system, the formal logic of searching would become more apparent. For example, to find an answer to the little boy's question, the librarian cannot search at the same time for both *cats* AND *pets,* because (assuming a peek-a-boo system) if two cards were held to a light source to discover what they had in common no matches might result; at best, only those references on *pets* that were also about *cats* would be indicated. All references on *cats* that were not also about *pets* in general would be missed. If the librarian wants to know which books on either *cats* OR *pets* discuss playing with cats, he can match the card for *cats* with the card for *play.* In this system the *or* relationship was established by searching for two *ands*: *cats* AND *play; pets* AND *play.*

To demonstrate how *not* might enter into a search, assume that *kittens* would be indexed in the peek-a-boo system separately from *cats.* In the thesaurus, in this case, the entry for *cats* would read, "See also kittens." In the example, information is wanted that is specifically about feeding kittens and *not* cats. (The librarian is looking for diets for young rather than adult cats.) With the peek-a-boo system this information is searched for by matching the cards for *kittens* AND *feeding* and recording the document numbers common to both cards.

Then *cats* AND *feeding* are matched and the matching document numbers recorded. The document numbers common to the two matches presumably should be excluded, but this presumption is fallacious. The same book may have a chapter on feeding cats and one on feeding kittens; if so, and the reference is categorically rejected because it is about feeding cats, pertinent information about kittens is being lost. For that reason the logical connective *not* is seldom used in machine searching. In some cases, however, this kind of undesirable effect would not occur and *not* could be used to advantage.

In the mechanics of search, in the preceding example, *or* and *not* were derived in terms of *and*. These mechanics are followed in all mechanized systems, even computers, although such derivation may be obscured when one is unaware of the procedural steps. The logic of search is very simple: no matter how a question is expressed it will eventually be answered by systematically inserting *and* among all of the search terms involved.

Suppose the little boy in the example wanted to know whether feeding potato chips to kittens would harm them. Assume that both *potatoes* and *kittens* are accepted terms in our peek-a-boo system. Assume, too, that when *potatoes* AND *diet* AND *kittens* are searched for, no pertinent references are found. Reference librarians would not stop searching at this point; they would broaden the search with the hope of finding pertinent material. One way to broaden the search is to delete a search term; perhaps searching for *potatoes* AND *kittens* would be productive, or, more likely, *diet* AND *kittens*, since material on diets for kittens could be scanned for information about feeding kittens starches and fats. Another method of broadening the search, instead of dropping a term from *potatoes* AND *diet* AND *kittens*, is to substitute a general term for one of the specific terms: *starches* for *potato chips*, for example.

Edge-notched and peek-a-boo systems always need the human operator to do this kind of broadening. Internally punched card systems, if they use equipment as capable as the IBM 101 Electronic Statistical Machine, can do a little of such broadening automatically, through a plugboard wired for alternative searches to be done during one pass of the cards. For example, references indexed by *kittens* AND *diet* AND *potatoes* are programmed to be sorted into a particular pocket; references indexed by just *kittens* AND *diet* in another; *starches* AND *diet* AND *kittens* in another; and so on. Most computer systems for automating reference work have been designed to accomp-

lish approximately the same thing. Their greatest difference from punched-card systems is that they can search faster and can be programmed for more alternatives at the same time. Larger computers, however, may be programmed to have access to a thesaurus that defines for the computer terms that will broaden or narrow the search. Of the present systems, MEDLARS is the only one to incorporate this feature.

A search can be narrowed by either substituting more specific terminology, as was just discussed, or by adding more terms to the search. To date, the latter procedure must be done by humans; if it is anticipated that the search may need to be narrowed, the reference librarian specifies alternative searches containing the additional terms.

If it is anticipated that the system will produce too many references in answer to a question, the number can be decreased by narrowing the search as just described or by other methods. In the example about the boy, the reference librarian found twenty books on how to keep a kitten. By some intuitive process, which the state of the art has not yet defined, she was able to choose two which she presumably thought most suitable to give to the boy.

Most automated search strategies have not attempted to deal with the problem of limiting the output, because it is an unchartered process. MEDLARS has made a first step in solving the problem arbitrarily. Before the search is begun, the MEDLARS client is asked to state whether he desires a few (1 to 10), a moderate number (11 to 100), or many (101 or more) references. The system is programmed to comply with his wishes by every other means of search strategy previously discussed; if more references result than is desired, the computer is instructed to print out only enough of the most recent references to fulfill the search requirement.

Still another way to reduce the output is to design the system so that indexing terms point out whether a document is a review, a textbook, a report, or is in some other bibliographic form. When the system is so indexed, the client can ask, for example, only for reviews on the subject in which he is interested. The MEDLARS system makes use of this technique. Because only a few of the indexing terms assigned to a document are used in preparing their published indexes, the DDC (Document Defense Center, formerly called ASTIA, Arlington, Va.) and MEDLARS staff weight their indexing terms. The terms chosen for publication are supposedly the most important ones to have been assigned to the document; the additional (nonpublication) in-

dexing terms are used only for searching within the computer system. The two kinds of indexing terms are distinguished by a label. To limit the number of references retrieved, terms labelled one way or the other can be specified when the search is formulated.

Preparing a question for processing involves still more considerations. How soon is the answer needed? Reference librarians assign priority ratings to search requests and process them sequentially. Computers usually can process batches of questions; even with computers, however, it is important to know whether routine or nonroutine scheduling is needed. Are foreign-language references wanted? If each article has been indexed by language, a sorting device of any kind can be used to include or exclude particular languages. Can a date range for the search be set? For example, does the client want only recent material, or is his subject one of recent origin? If references are added to the file in serial order, or if the date of the document can be accessed by the system, then date ranges can be imposed for the machine search. What form should the output have? For example, does the client want a bibliography, a group of abstract cards, or photostatic copies of the documents for which references are retrieved? If he wants a bibliography, does he want it arranged by date, subject, author, or language? In a system where one has choices in these matters, the choices must be made explicit before the computer search is begun.

PROCESSING

One of the first decisions a reference librarian must make when processing a search is where to look first; it is probably more efficient to start with one source than with some other. An analogous situation is sometimes found within a file or set of files available to a computer system. If the system uses noninverted filing, that is, the indexing for each document is stored as a unit within that file, the search always starts at the beginning and proceeds to the end of it. If inverted filing is used, that is, storage is arranged according to the entries in the thesaurus, then only the term records pertinent to the batch of questions must be searched (but in a stepwise fashion that is sometimes deceptively long). In an inverted system, it is most economic to find the least heavily posted of the required search terms for a particular question and to match the next most heavily posted term against it. This procedure ensures a minimal number of comparisons by the computer and thus makes the search faster and less expensive. The DDC

system has an inverted file but does not make use of the latter feature at present. The MEDLARS system uses a noninverted file for storing reference citations; however, MEDLARS makes use of features of both approaches to file organization by maintaining a running index to the file; the index is maintained according to the arrangement of the thesaurus. As additions are made to the file, each term is checked against the thesaurus and the index is updated to maintain a tally of how frequently each term has been used. As a first processing step for a search, then, the magnetic-tape index to the file is consulted to rank the specificity of each of the terms for the search. This statistical information is used to develop the processing formula for the noninverted file; that is, only a document record that contains the least-posted search term (for each question in the batch) will be examined to determine whether other required terms are present; for some searches use of the index reduces search time to a small fraction of what it would be without the index.

During processing of input information, the reference librarian frequently finds that the search parameters supplied by a client are not precise enough to obtain a satisfactory answer; or, an interesting tangent can develop during a manual search that is judged to be worth pursuing further. In either case the reference librarian can be more effective if the client is informed of these developments. Can a computer be programmed to react to such situations? Theoretically, yes; but in practice little of such programming is done, except to set up alternative searches to pursue tangents that can be envisioned in advance or to formulate new questions for the next computer run on the basis of references located.

OUTPUT

The output from peek-a-boo systems is a set of document numbers. From edge-notched cards, one usually obtains a full citation which has been written on the face of the card and either must be read directly or copied manually or photographically. Almost all internally punched card systems yield a number, such as a class or serial number, which must be looked up to get more complete information.

Computer systems can produce almost any kind of output that a client could desire, if the system is so built. In most systems, however, only one or two output formats are feasible, such as a list of document numbers or a list of alphabetically arranged references. To design a system so that only document numbers are given to the client seems

a misuse of both the system and its clients. A few systems have tried giving abstracts that had been stored on magnetic tape, but have found this uneconomic. As a substitute, DDC and other systems retrieve document numbers from the computer and then manually extract the correspondingly numbered abstracts from a printed-card file. When photostatic copies of full documents are supplied, the same process as for abstracts is usually followed; that is, the references are retrieved by computer, and the documents or their film versions taken from the shelf and copied by equipment independent of the computer.

MEDLARS plans to supply bibliographies complete with author(s), title, source, date, and language directly from the computer. This is a two-step process. First, document numbers are retrieved through a tape file of indexing terminology arranged according to document; after the pertinent document numbers have been located and recorded, the file containing the complete citations is searched for the remainder of the data. MEDLARS offers a number of choices for arrangement of the bibliography. It can be grouped by author, title, source, date, or language. In addition, the bibliography can be printed in a wide variety of formats.

State of the Art

The MEDLARS system design [10] incorporates the most advanced search strategy for automating reference work to date. Also, from an engineering standpoint, it has advanced the art by sponsoring the development of computer-driven equipment that can compose rapidly (440 characters per second) by optical means; the product of this equipment is at least as good as that usually provided by typesetting. Thus reference librarians will find indexes provided through this system more legible than those provided by a system using a conventional high-speed printer.

Most nonconventional systems used to automate reference work employ punched cards rather than computers for information storage and retrieval.[11] Such systems usually contain fewer than 50,000 documents. This number is not a large reference store or reference potential; however, if all such retrieval systems were compatible and covered different material systematically, they could then be linked to form a network. As one might expect, however, because they are experimental systems, they are highly disjointed and incompatible. Experience gained with these systems will prepare their designers, operators and, to some extent, their clients, for more sophisticated systems.

If the individual systems continue to grow, the volume of documents accumulated will require more powerful searching methods and equipment. If the need for so many more or less duplicating systems (as are found in libraries of competing companies in the same industry) disappears, the more powerful methods and equipment of centralized information centers will replace them. The few computer applications thus far designed for automating reference work are leading the way to future development.

References

1. Perkins, Alfred. "Sorting and Classifying of Flat Sheets, Tallies, Cards, or the Like," U.S. Patent 1,544,172. 1925.

2. Perkins, Alfred. "Card and the Like for Classificatory Systems," U.S. Patent 1,739,087. 1929.

3. Taylor, Horace. "Selective Device," U.S. Patent 1,165,465. 1915.

4. The Royal McBee Co., 850 Third Avenue, N.Y. 22, N.Y.

5. Zator Company, Calvin Mooers proprietor, Cambridge, Mass.

6. Marketed through Documentation, Inc., Washington, D.C.

7. *Non-Conventional Technical Information Systems in Current Use, No. 1.* Washington, D.C., National Science Foundation, 1958.

8. *Thesaurus of ASTIA Descriptors.* 2nd ed. Arlington, Va., U.S. Armed Services Technical Information Agency, Dec. 1962.

9. "Medical Subject Headings, 2nd ed.," *Index Medicus,* Part 2, Vol. 4, No. 1. National Library of Medicine, Jan. 1963.

10. *The MEDLARS Story at the National Library of Medicine.* Washington, D.C., U.S. Dept. of Health, Education and Welfare, Public Health Service, 1963.

11. *Non-Conventional Technical Information Systems in Current Use, No. 3.* Washington, D.C., National Science Foundation, Oct. 1962.