THE INDEXING PROBLEM*

BY CHARLES L. BERNIER
Chemical Abstracts Service, Ohio State University, Columbus, Ohio

Electronic equipment that very rapidly performs programmed mathematical and other operations has stimulated much thinking as to how it can be used for the storage and retrieval of verbal (not oral) information. During the last twenty years subject indexing has been "rediscovered" by many whose training has been largely in the fields other than documentation and librarianship. Documentalists have been re-examining their methods. During this period of awakened interest in dealing with information, ways of keying information have been studied again. An example is the correlative index, which facilitates a selection of documents by the correlation of two or more terms and is said to date back to the time of cuneiform writing. Chemical Abstracts pioneered in the indexing of chemical information and has grown in this direction by continually re-examining its methods and procedures.

The study of electronic and other equipment sometimes has proved to be a distraction from investigation of the fundamentals of the indexing problem. These fundamentals are important no matter how machines may enter into the picture. Sometimes it helps in studying a problem to start with a very simple model.

Let us start our study of the indexing problem by imagining a rudimentary library of unclassified periodicals. The Library has no subject catalog or similar retrieval device. We can assume, to make the picture more rational, that there is no librarian.

Let us say that we are assigned to the task of finding all information about a given subject in the stacks of periodicals piled in order of acquisition on the shelves. Let us further assume that we have no knowledge of library techniques and no way of knowing at what date the desired information may have been published.

Our naivete saves us from being disheartened or resentful.

So we roll up our mental and physical sleeves, provide ourselves with paper, ballpoint pen, table, and chair. Resolutely we start.

As we examine the first issue of the first journal from the first pile of periodicals on the first shelf of the first row of shelves, we wonder by exactly what clues the needed information will be detected. We turn through the issue page by page, reading every word.

Several days later we find, after careful experiment, that it is possible to scan pages and yet discover significant information without missing any. We have already found a few bits of information and have jotted them and their references onto a pad. We have also taken precautions to keep track of the issues, journals, shelves, and rows that we have already examined.

As the first week comes to a close, we wonder if it would not be possible at times to pass by entire journals or whole issues after merely examining the journal titles or tables of contents. We discover that none of the information we select comes from pages of advertising and that, from the nature of our search, these pages will likely always prove to be barren. During the second week, before scanning the pages of papers we try to predict, from journal titles and tables of contents, those papers that will turn out to be fruitful. We carefully keep a record of how well our predictions have turned out. From the success of our predictions during this and following weeks, we gradually gain confidence in our ability to exclude whole issues and even some journals. This discovery makes the work go much more rapidly since we can lay aside issues with only a glance or merely a rapid reading of the tables of contents.

Being of a cautious nature, we check further from time to time, using a random sample of those issues that we have predicted to be barren, just to make sure that we are not missing information. The discovery that we can predict some barren issues and titles and can omit pages of advertising reduces our scanning time to perhaps ten per cent. of what it was before.

As weeks pass, and our mind wanders momentarily from the task, we hypothesize that since our response to material rejected and selected is triggered solely by symbols (usually words) on paper, it might have been possible to have preselected these symbols, and thus to have saved ourselves much of the time taken to examine all periodicals, one by one.

---

*Gordon Conference, New Hampton, N. H., July, 1961.

In order to test this hypothesis, we experiment by predicting terms in titles and tables of contents that led to rejection of barren periodicals and papers. As our list of these rejection terms grows longer day after day, we slowly come to see that our list will eventually include an unabridged dictionary of words, and nearly all new words that have come into the language. We decide to abandon the use of a comprehensive rejection list; its use will be impractical because of (1) the time required to compile it, (2) the time needed to consult it, and (3) the lack of completeness at the moment we need it; there will always be new words coming in so frequently that the list will need revision every day or oftener.

The use of a list of terms for selection rather than rejection of documents comes to seem more promising as we work along and thoughtfully mull over the problem. As the next experiment, we try predicting and recording terms in titles and tables of contents that are actually found to lead us to select fruitful periodicals and papers. Our success is immediate and heartening. We find that predicted title terms of papers lead to selection of about 21% of the papers that actually contain information related to our search. The reason why more than 21% of relevant papers cannot be predicted by terms in titles of papers and periodicals is mainly that the titles do not contain terms that we are using for selection. About 66% of the papers we select are chosen by terms (not necessarily those on our list) that we find in the bodies of papers rather than in their titles.

An analysis of the terms in the experimental list we use for selection shows that about 64% of them are identical with the terms actually used in the papers. About 17% are synonymous, and about 18% are more generic, more specific, or merely suggestive, in ill-defined and vague ways, of terms actually in the papers.

As progress is made in preparing a list of terms for use in selection of periodicals and papers, our statistics show that we are selecting about 18% of documents, as just mentioned, not on the basis of our list of terms, nor sometimes on the basis of anything that we have been able to predict, but on the basis of vague and often poorly defined associations in our minds between the problem that started the search and the contents of these documents. As we select a document that falls into this 18% category, we may say to ourselves, "The terms I am using for selection are not in the paper, but it is about the subject," or "I have a hunch that this paper will be useful," or "This paper seems to be related to the problem in its larger aspects, but I can't tell exactly how," or "If the solution to the problem takes a certain turn, then this document must be used," or "This paper deals with the solution to an analogous problem (and don't ask me to define 'analogous')."

In spite of the failures of our listed terms in about 36% of the cases, the success is so obvious that we are encouraged to pursue this line of thought further.

Titles of journals are often seen to have terms more generic than those used in the selection of papers. This gives us a clue as to a way of organizing our list of terms. We reason that there should be fewer of these generic terms than specific searching terms. An examination of our list shows this to be true. Thus, one generic term can nearly always be used to stand for several, or many, of the more specific search terms. Here is the way that we can organize our thinking and our list. The more specific terms can be arranged under the more general. Thus, we rediscover hierarchical classification. Also we find that the more general terms are more useful in selecting by title.

If the terms on our selection list of terms had been applied to titles and bodies of papers before we made our search, if the journals had been arranged by title or subject on the shelves, and if all documents had been related in a list of references to each term by means of abbreviations for the journal name, volume number, serial number, issue number, and pages covering the paper, then our present task would have been speeded beyond imagination for about 60% of the significant documents that we actually located. The months of work that we expended in examining every periodical would have been, in fact, reduced to using the list of references related to our problem in selecting periodicals and papers from the stacks and to copying the information found there. This would have given us about 60% of the references needed. The only references that would not have been on the list of references would have been those without selection terms, those that seemed "analogous," those on which we "had a hunch," and the like.

With our original search finally completed and, with this background knowledge, we start to devise a system that would have saved us these months of careful, patient, page-by-page, issue-by-issue, and journal-by-journal searching.

It is clear that the prepared list of references about which we dream would have solved our present search problem to the extent of about 60%. It is equally apparent that it would not have been effective in an entirely different search. In order to take care of all searches, it will be necessary for us to develop a universal searching tool that will save page-by-page scanning of documents in looking for the answers to all questions.

We have become convinced that our searching is based solely on intelligible symbols (usually words) on paper and that these symbols are to be used principally for selection rather than rejection. We know that the terms used for selection or rejection can be grouped into hierarchies under generic terms. Also, and most

important, we became convinced during our tedious task that, "something must be done about this literature problem." With these facts and this motivation, we set out to plan and conquer in the following way.

Words representing new ideas, new facts, new data, in fact all news, will be associated in our planned system with references to documents. Novelty or newness is an important criterion since we do not want to waste time of ourselves or other searchers on what is old -- on what is already well known. Another criterion for selecting terms will be their relation to the subject of the document, i.e., based on new concepts that the author intended to communicate, not on what words he used. Thus, we must "subject index" and not "word index." As we think about it, this criterion will be a very subtle one and difficult to explain. Another criterion will be that the terms must be commonly used or popular ones that the searcher will most probably seek first. Still another criterion will be selection of the most specific term justified by the document and avoidance of generalization unless the author has sanctioned our doing so, i.e., we will be truthful in our selection and not push the author beyond limits he has already set. It is clear to us that if we index one document with general terms, thus pushing the author beyond the limits that he has set, then we must index all documents in the same way and with a pre-selected set of terms in order to ensure consistency. This will, we see, create a generic or classified index.

With our collection of selected terms and their associated references, our next problem is how to organize them. We find that the terms can be arranged either into straight alphabetical order or into groups with like meanings to form a classification. The former will give an alphabetical index (of sorts) and the latter a classified index (of sorts). If all of the terms for indexing had been chosen from a prepared list of general terms, then we would have a generic index (of sorts).

We decide in favor of the alphabetical arrangement of specific terms since the searcher can approach it directly with the words that he knows rather than first having to turn to a classification schedule and from there to the classified index. We could have chosen a classified arrangement in order to help with generic questions and to suggest analogous answers in the event that the exact information sought is unavailable. During our page-by-page search, we made the discovery that analogous information can often be used.

Use of this simple alphabetical index of terms and references will, we believe, be much better than going through the literature page by page. However, as we visualize this simple index, it will present difficulties in actual use. We can see that some index terms will be very popular and have many references associated with them. We can picture how frustrating it will be to look up all undifferentiated references under a popular term and find that, perhaps, no reference is suitable. Because of this difficulty we decide to use modifying phrases, or expressions associated with each of the original indexing terms to help in differentiating among references under the terms. We suspect that the optimum grammar and diction of these modifying phrases or auxiliary terms will need careful study and development.

In an alphabetical index, we realize that terms related by meaning often will be separated by the arbitrary order of the alphabet. We plan to solve this problem by indicating semantic relationships among the terms locked into alphabetical order. We can choose indicators external to the index, e.g., a thesaurus or a list of cross references, or internal indicators, e.g., cross references and notes alphabetized along with the index entries. We chose internal indicators for greater convenience to the index user.

Alternatively, we could have chosen a classified index in which like concepts were brought together so far as possible. Also, we could have chosen to produce both kinds of indexes.

We believe that it will be best to use alphabetical order for terms in both indexes because no other order is almost universally remembered.

In our search through chronological, unclassified stacks of periodicals, we also came to understand that shelf arrangement and a record of this arrangement would have saved us weeks of work in handling piles of journals, most of which actually proved to be barren.

CONCLUSION

Through our imaginary, model search we have re-invented library classification, shelf arrangement, subject indexes, cross references, classified indexes, generic indexes, etc. We have decided that, no matter how inadequate classifications and indexes may be, they are far superior to none at all, and we suspect that these retrieval tools can be vastly different in effectiveness and cost, depending on how they are built.

We have seen that documents are selected by some, but not all, of the intelligible terms in or on them, that listing terms for rejection of documents is not so effective as listing them for selection, that 34% of the documents are selected by terms in titles, that 66% are selected by terms in the bodies of documents, that 18 to 36% are selected by unpredicted terms, some of which seem always to be unpredictable. It should be pointed out that true subject indexing as practiced by the indexing staff of Chemical Abstracts uses index entries generated by all of the 18 to 36% of unpredicted and unpredictable terms. We have come, perhaps, to an appreciation of the fact that both classified and alphabetical indexes serve useful, and somewhat different, purposes.