DOCUMENT RESUME

ED 050 757 LI 002 817

AUTHOR	Upton, Charles C.
TITLE	Computerized Communications Citations Technology.
PUB DATE	Apr 71
NOTE	17p.; Paper prepared for a conference of the International Communication Association, Phoenix, Arizona, April 22 - 24, 1971
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	*Automatic Indexing, Communications, *Computer Programs, Conferences, *Information Processing, *Information Retrieval, Information Storage, Information Systems, *Man Machine Systems, Researchers, Technological Advancement
IDENTIFIERS	*Computer Software

ABSTRACT

Feasibility study findings indicate that a potential reduction in drudgery for the scholar may be obtained through a combination of heretofore divergent solutions to the increasingly critical problem of how researchers will interface with computers for an efficient and effective review of the literature. Presently existing computer software includes information retrieval programs with important traits: (1) they will combine so as to permit offsetting of otherwise major weaknesses in the several separate programs; (2) they will require minimal modification for adapting to a variety of machines due to their being written in higher level languages; (3) they are precision instruments produced at a combined cost of some half million dollars of research funds; (4) they do not require use fees, having been produced through federally financed research and development projects; and (5) they are amenable to operation with a unique means of preparing document abstracts to permit computerized indexing as well as filing and retrieval. At present there are both classification schemes and computer information retrieval programs of limited utility to a communications bibliography. Rather than abandon and ignore these efforts, their examination shows pitfalls to avoid by building on those schemes and programs presently available. (Author/NH)

COMPUTERIZED COMMUNICATIONS CITATIONS TECHNOLOGY

by
Charles C. Upton
School of Radio-Television
Ohio University
Athens, Ohio 45701

International Communication Association
Phoenix, Arizona
April 22 - 24, 1971

ABSTRACT

Feasibility study findings indicate a potential reduction in drudgery for the scholar may be obtained through a combination of heretofore divergent solutions to the increasingly critical problem of how researchers will interface with computers for an efficient and effective review of the literature whose explosion continues and accelerates. Presently existing computer software includes information retrieval programs with important traits: (1) they will combine so as to permit offsetting of otherwise major weaknesses in the several separate programs; (2) they will require minimal modification for adapting to a variety of machines due to their being written in higher level languages; (3) they are precision instruments produced at a combined cost of some half million dollars of research funds; (4) they do not require use fees, having been

produced through federally financed research and development projects; and (5) they are amenable to operation with a unique means of preparing document abstracts to permit computerized indexing as well as filing and retrieval.

At present there are both classification schemes and computer information retrieval programs of limited utility to communications bibliography. Rather than abandon and ignore these efforts, their examination shows pitfalls to avoid by building on those schemes and programs presently available.

A descriptor set is prepared according to the method of Calvin Mooers. Document abstracts are written in plain text, but with the descriptor set for a technical vocabulary. These abstracts are machine indexed by computer. Resultant codes and abstracts are packed and stored by a taxonomic information retrieval program. When users retrieve descriptons of the literature in response to a query, the English sentences previously used for indexing are made available to annotate the standard bibliographic citations. User queries are automatically analyzed by a computer program to determine changes in the current usage of descriptors and terminology in the real world of scholarly research. Whenever necessary, as indicated by the computer analysis, steps are taken to revise the descriptor set. Computer programs of interest include TAXIR by David Rogers, FAMULUS by Theodor B. Yerke, and SMART by Gerald Salton. Also of interest are techniques by Allen Kent and J. W. Perry, formerly of Case Western Reserve.

--//--

Introductin

Here is a proposal which might put computer assistance in your next review of the literature, or it might be another blind alley. Information storage and retrieval have become the perennial promise but after literally thousands of research dollars have been spent you are left with the same problem as before--how does one find the needed materials when the library systems are outmoded and yet the publications are accelerated into what is sometimes called an information explosion? The plan, simply enough, is to combine the best available features from some heretofore divergent solutions. A computer's speed and accuracy offer great potential in reducing the scholar's drudgery but a poorly programmed computer also offers great expense, even without delivering the speed and accuracy. The possible pitfall is that even those best available features are not good enough for production basis uses. Recognizing the risk as well as the possible reward, consider some feasibility study findings which suggest we have already solved the most difficult problems and could combine those results for a computer interface with communications researchers.[1]

Existing Software

Presently existing computer software includes information retrieval programs with five important traits. Combining programs will offset major weaknesses in the separate parts. Being written in higher level languages they require minimal modification to be adaptable on a variety of machines. They are precision instruments produced through well funded research. Federal financing for their research and development places them in the public domain. They will operate with a unique means of preparing document abstracts to permit computer indexing as well as storage and retrieval.

Offsetting Weaknesses.--One information retrieval system operates economically with tape files of several thousand citations for personal collections, but the storage and retrieval operations themselves appear less efficient than in another system which lacks some of the latitude for individual preference in structuring data and in arranging items for printing. The more efficient program could make economical the handling of files much larger than one's personal collection, yet the system's indexing method lacks some flexibility needed with a sophistocated classification scheme. A third system handles indexing of a sophistocated classification scheme and permits individuality in the retrieval vocabulary, but operation is not economical and the retrieval is more elaborate than required outside of an experimental environment. A fourth system converts special abstracts to index codes which the machine can use without requiring additional error prone human processing, but inefficiency is introduced by generality beyond that necessary for a descriptor vocabulary of manageable size for communications. Combining the desirable aspects of these four systems could yield a compromise system with acceptable limitations on flexibility and generality but with superior storage, retrieval, and indexing. As The four systems presently exist there is a major weakness in each approach, however, offsetting these weaknesses appears viable.

Widely adaptable.--Being written in computer languages which are generally available on medium and large machines, the programs are readily adaptable to installations at most university and industry computing centers. Programming in a higher level language permits this relative universality by avoiding dialects which are not standardized among computer manufacturers and local implementations. Programs written in such languages are usually not as efficient as those written in the particular machine's own assembly language, although at least one of the systems discussed obtains an unusually high efficiency by algorithms which translate into assembly language instructions not ordinarily available with straightforward programming practices. This clever approach makes the large files economically possible through a program which takes minimal modification to operate on computers designed by various manufacturers. The system which offers individualized options in printing formats is similarly written in a higher level language so that these conveniences could also be made available even though their use is incidental to some researchers. Special sorting and merging operations would be as easily obtained for those needing such features because both systems are fairly readily adaptable to a wide variety of computers.

Precision Instruments.--The programs in question are not the hastily produced variety which are sometimes patched together just to get something limping along, but are well designed and carefully executed precision instruments which were produced from some half million dollars of research funds. Cost alone is not so important except as a measure of the resources available for research and development of some innovative approaches. Actual economical operation is another measure of instrument precision, and these programs seem to offer the most powerful operations for the least expenditure. Not all of the systems mentioned have been madè available in the same highly precise form, but taken as a group of programs, they offer a basic core of advanced programming. Even the less highly developed algorithms have been tested and show promise as experimental if not production solutions. None of the storage or retrieval routines per se is experimental. Some of the one time operations such as indexing from abstracts seem least satisfactory. This remains an area to beware of pitfalls.

Federally Financed.--While much federal support has now dwindeled there remains the indirect continuation of research by making such software available without use fees or royalties. New funding would be necessary to combine, modify, and expand present versions into the target program, but this expense is considerably reduced by taking advantage of existing programs already in the public domain. There are other developments under way which might permit using Library of Congress MARC-II tapes, whenever more of these federally financed research results are in. Presently there seems an advantage in available programs which can be modified for bargain investments, considering that the basic research costs have already been paid.

Unique Abstracts.--The programs of interest are amenable to operation with a unique means of preparing document abstracts.

Again combining divergent ideas seems to offer a workable solution. Experimental work with text processing indicates the state of the art is not yet sufficiently advanced for practical applications of indexing by computer except when specially prepared texts are used.[2] Two of the earlier mentioned systems permit conversion of special abstracts into indexing codes for information retrieval. Modifying and incorporating these procedures into an overall design would allow the production of an index directly from the document abstract, with little or no room for human error once the abstract was checked for content. Computer prepared index codes, together with the abstract and standard bibliographic citation would be compatible with other programs selected for efficient storage and retrieval. If present practice does not include direct handling of literary style abstracts, a combination of auto-abstracting and subsequent encoding is available. Full text machine translation is not available, although by exercising vocabulary and syntax control during the writing phase one could expect to have intelligible text for both man and machine. For more ambititious translations one may be required to wait a few years.

Locating Sources

Surprisingly the major problems with computerized searches have little to do with the computer. Classification schemes may be the weakest link in the whole process. Beginning graduate students and seasoned researchers share the hazards of finding pertinent documents too late for use, even when the material is known to exist. From such frustrations grew a series of informal investigations of problems associated with computerized assistance for library searches. Attempting to devise some classification scheme more in keeping with recent communications developments, one investigation terminated by concluding there is no real need for another scheme tailored to one man's perceptions and speculations. Dewey's was successful for a time, but his scheme lost favor for research uses as being innappropriate for the needs of many researchers.

Representing an opposite extreme, the Library of Congress Classification (LCC) is not always suited to research needs of specific individuals. While newer than Dewey Decimal Classification (DDC), the LCC barely postdates a horse and buggy era so that both DDC and LCC literally have no place for live telecasts from the moon's surface (now part of the history of communications) except by improvising rather freely.[3] Both DDC and LCC were developed to solve problems of shelving books, the former by Dewey's idea of "all human knowledge in print" and the latter by a library committee's view of which books are used together in a specific collection.[4] Despite the disadvantages of LCC for communications literature one may expect continued use of that scheme as a shelving system, especially with machine readable cataloging information readily available through MARC-II computer tapes, supplementing cards.

Faceting, highly developed by Ranganathan in the Colon Classification (CC), permits breaking down a subject classification into relational aspects much as English permits "typewriter ribbon," "typewriter carriage," and when appropriate, "horseless carriage," "moving pictures," "wireless telephony," or "mass communications."[5]

Without benefit of a computer CC was difficult because faceting involves complexities. CC also suffered notational problems and the expense of converting an already cataloged collection, for DDC and LCC were already widely accepted when CC was introduced. If notational problems were solved a faceted classification might easily be adapted to a computer without disturbing the use of LCC for a shelving device. Both faceting and translating of the call number are more readily done by computers than by researchers. Faceting permits developing new relationships by combining symbols so that live telecasts from the moon's surface should present no great difficulty to the cataloger, even when considered as part of the history of communications.

Another scheme deserving mention despite being little used is the Expansive Classification (EC) of Cutter.[6] EC was actually seven separate schemes for as many different sizes of collections, from a budding village library to the national collection. Many of Cutter's ideas were adopted by the Library of Congress but his EC fell into disuse. Interestingly the EC remains highly regarded by librarians even though the scheme lacked backers and so remained unfinished with Cutter's death in 1903.

Dewey's scheme was something of a breakthrough in 1873, and perhaps for that reason the classification based on one man's ideas proved tolerable. He certainly was plagued with demands to alter the DDC but he generally resisted modification. One attempt at a facted version of the DDC became popular in Europe as the Universal Decimal Classification (UDC).[7] Note that DDC met resistance for failing to serve users as conveniently as the designer imagined. One suspects a pitfall in one man schemes when adding CC and EC to the example of DDC. At the risk of gross oversimplification the LCC might be thought successful in part because several men pooled their ideas into a compromise scheme. This compromise feature is important to consider if the one man scheme is an inherent pitfall just because one man's perceptions differ from those of other men.

Mooers describes what appears to be a useful method for constructing a classification scheme, whereby he meets with a select group and essentially arbitrates a compromise scheme to their needs.[8] His select group might reasonably consist of the leaders who either generate research or supervise those who do generate research (including the graduate faculty and the industry researchers alike). In practice Mooers forces these "high level users" to make decisions regarding the contents of a collection as well as bringing them to agreement on a descriptor set (or a common vocabulary for subject classification). He then arranges the descriptors for convenience such as grouping brand names, relational concepts, items of equipment, and so forth. Words and phrases representing concepts are used rather than terms directly appearing in the text and sometimes called keywords. In this way vocabulary control is maintained even to the extent that faceting is manageable because some three hundred or so descriptors serve instead of several thousand terms when synonyms are listed as they occur in text.

Special Abstracts

Given a descriptor set of the sort just described, an abstract of the document in question could be prepared by using plain English text with the concepts being expressed by combining words and phrases from those descriptors nearest in meaning to the ideas expressed by an author. An annotator working in his native English language and restricting himself to descriptors should produce rather acceptable descriptions of documents which could be read by man and by machine alike, without quite the room for error that is found with arbitrary codes. Proofreaders of these abstracts would again have an advantage in using their own native tongue rather than some code designed for pleasing the innards of a machine. Similarly a researcher would be able to read and understand an annotated bibliography without translating a machine code.

While not quite as easy for the programmer perhaps, the machine treats properly prepared textual material more slowly but just as "easily" as coded data. Procedures developed by Kent, Perry, and Shera included a computer technique which takes special abstracts and reduces them to the coded data for faster processing and more efficient storage.[9] This experimental work used special symbols to indicate relationships, with the symbols being added to clarify usage of words and phrases in abstracts that could then be read by man and machine alike. Translation by machine was accomplished on the abstracts directly and coded information was then available for machine searching. While this process used a vocabulary taken from the text rather than from a descriptor set as described earlier the automatic indexing of an abstract could be accomplished with equal or greater efficiency when using a more closely controlled vocabulary with fewer entries and with words rather than special symbols to represent relationships. Both the steps taken to produce the special abstracts actually used and the steps necessary to write somewhat different special abstracts require some editorial judgment but the abstracts are essentially the same when considered as an input for machine translation. Faceting appears easily handled by either means, but when relationships are expressed in natural English text there is no need for writer or reader to use code techniques. Of course a skilled writer might wish to use abreviations which the machine would replace with corresponding English equivalents.

Information Storage and Retrieval

Three distinctly different kinds of information are needed for machine storage and retrieval: indexing codes, bibliographic information, and annotations in the form of special abstracts previously discussed. Indexing codes are produced by machine translation of these abstracts, and the two are then associated with the bibliographic information for the document of interest. Rogers devised a system which converts descriptors into compactly coded form and permits search operations to be performed directly on the encoded descriptors in a binary coded rather than a decimal coded form.[10] This solution offers an advantage of both time and main storage utilization when compared with more commonly used procedures in program construction. Time requirements are brought to a minimum through the use of Boolean operations directly on the binary coded descriptors, which is relatively fast, and through the elimination of much wasted transfering from auxiliary storage to main storage by virtue of the more efficient binary coded indexing information. The compact binary codes also reduce the necessity for much of the repetitious moving and comparing of decimal codes.

To provide ample storage of descriptor codes this information is stored separately from the bibliographic information and the two files are linked by simply numbering both lists of information. For each coded item with the desired descriptor configuration the corresponding item number is stored during the descriptor search phase. These numbers are then used to retrieve bibliographic information for listing in some convenient format according to the researcher's preference. Abstracts could be referenced in the same fashion and listed along with the bibliographic information, or the standard bibliography format could be printed to guide one in finding the appropriate abstracts in book form if that seemed preferable. The choice is purely a matter of economics versus convenience, when considering the access method for annotations—the previously indicated savings of time and storage to process searches on the descriptor index file is a more basic matter of placing operations within reasonable economic range when they have not been there before except for experimental purposes. This system is operational at everyday production budget levels.[11]

Maintaining Current Descriptors

Vickery points out the tendency of books to lag in reporting developments which appear first in the article literature.[12] While terms taken directly from the document's text would reflect that phenomenon, specially prepared abstracts using descriptors based on concepts would be able to use the same descriptor to cover both articles and books so that a difference in terminology would not mask actual differences in what the materials discuss as content. Eventually there would be changes needed, for terminology changes in both articles and books until there is a datedness about "wireless telephony" or even "moving pictures." Just as "wireless telephony" is a different concept from "radio" there is need of descriptors to reflect that difference, provided of course that "wireless telephony" is a concept of interest to the researcher. "Film" might be used interchangeably with "moving pictures" under some circumstances, but only another high level user's conference would tell us what we mean separately and collectively by those terms, and whether a new concept demands a new descriptor.

To some extent the work of that group of leaders might also be taken over by the computer, especially in helping to determine the twilight conditions of changing usage. Salton has described an experimental system which permits the computer to collect and count which descriptors are used in an information retrieval search.[13] Indeed, his system does many other information retrieval operations but the experimental nature of his work permits features which are presently beyond the everyday production budget levels. This one feature is attractive because the computer may reasonably be used to translate a researcher's plain English text into descriptors rather like a reference librarian suggests subject headings the patron has not thought to try. While this may seem experimental and therefore risking too much to luck in the vicinity of a known pitfall, there is every reason to believe this experimental phase is drawing to a close as remote terminal installations continue to offer conversational interaction with researchers on everyday production budget levels. Spreading the terminals throughout an area which extends literally hundreds of miles from the computer has brought about a favorable ratio of users to facilities and the cost sharing has put these formerly experimental installations in the reach of widely separated researchers.

If this translation of a researcher's request is available then terms for which there was no anticipated demand should be recorded for bringing to the attention of the high level users at their next conference--perhaps on an annual basis so that updating the descriptor set keeps pace with changing terminology. Of course the computer would be expected to inform the researcher when a term was not being translated into a descriptor, so that an additional request might be tried, however, the important point to be made is that the computer should also provide feedback on current usage to the leaders responsible for vocabulary control. Failure to do so seems to walk into the known pitfall that has given us too many classification schemes which fail because they do not meet the researchers' needs, especially with passing time.[14]

Summary and Conclusions

Several computer programs and techniques suitable for information retrieval already exist and should be used to reduce time-consuming literature searches for scholarly research. These programs and techniques are not presently combined into a single working system. Their successful combination appears possible. There is an even greater than ordinary need to proceed carefully for the computer offers not only speed and accuracy, but expense and even failure to obtain the speed and accuracy. The programs and techniques appear especially well suited to the communications researchers' needs.

The programs should combine to offset weaknesses. Minimal modification of the programs would permit adapting them to many other computers using higher level languages. These programs are precision instruments produced by well supported research. Federal funding has provided these programs without need of use fees. They are amenable to operation with a unique technique for preparing document abstracts to permit computerized indexing as well as filing and retrieval.

The plan appears quite attractive and serviceable as an overall design but may contain pitfalls which escaped disclosure. Some researchers, desiring an expedient solution to problems of library research, lend support to the proposal. Knowing they may not have thought of everything, those supporting the proposal solicit constructive criticism.

NOTES

[1] U. S. National Science Foundation. *Current Research and Development in Scientific Documentation, No. 15.* Washington, D. C.: Government Printing Office, 1969. (NS2.10:15). Pp. 343-344, 354, 506-507. Projects 7.73, 7.86, and 10.15 describe FAMULUS, TAXIR, and SMART, respectively. See also: Shera, Jesse H. *Documentation and the Organization of Knowledge.* Hamden, Connecticut: Archon Books, 1966.

[2] U. S. Department of Commerce. *Automatic Indexing: A State-of-the-Art Report.* National Bureau of Standards Monograph No. 91. Washington, D. C.: Government Printing Office, 1970. (C13.44:91).

[3] Dunkin, Paul S. *Cataloging U.S.A.* Chicago: American Library Association, 1969. Pp. 105, 144-147.

[4] *Ibid.*, pp. 98-100.

[5] *Ibid.*, pp. 131-135.

[6] *Ibid.*, pp. 100-101.

[7] *Ibid.*, pp. 135-137.

[8] Mooers, Calvin N. "The Indexing Language of an Information Retrieval System." *Information Retrieval Today: Papers Presented at the Institute Conducted by the Library School and the Center for Continuation Study, University of Minnesota, September 19-22, 1962.* Edited by Wesley Simonton. Minneapolis: Center for Continuation Study, University of Minnesota, 1963. Pp. 21-36.

[9] Perry, James W. "Exploitation of Abstracts by Applying Machine Translation Techniques." *Advances in Documentation and Library Science. Vol. III. Information Retrieval and Machine Translation. Part II.* Edited by Allen Kent. New York: Interscience Publications, Inc., 1961. Pp. 787-810 (Chapter 29).

[10] Rogers, David J., Henry S. Fleming, and George Estabrook. "Use of Computers in Studies of Taxonomy and Evolution." *Evolutionary Biology. Vol. I.* Edited by Th. Dobzhansky, M. K. Hecht, and Wm. C. Steere. New York: Appleton-Century-Crofts, 1967. Pp. 169-196 (Chapter 6).

[11] See also: Meetham, Roger. *Information Retrieval: The Essential Technology.* Garden City, New York: Doubleday, 1970.

[12] Vickery, Brian C. *Classification and Indexing in Science.* New York: Academic Press, Inc., 1959 (Second Edition). P. 176.

[13] Salton, Gerard. *Automatic Information Organization and Retrieval.* New York: McGraw-Hill, 1968.

[14] Taylor, Archer. *General Subject-Indexes Since 1548.* Philadelphia: University of Pennsylvania Press, 1966.

--//--