DOCUMENT RESUME

ED 053 779 LI 003 051

AUTHOR	Leimkuhler, Ferdinand F.
TITLE	On Information Storage Models.
INSTITUTION	Purdue Univ., Lafayette, Ind. School of Industrial Engineering.
SPONS AGENCY	National Science Foundation, Washington, D.C. Office of Science Information Services.
REPORT NO	RM-69-5
PUB DATE	69
NOTE	13p.; (14 References); Paper prepared for Seminar on Planning Library Services, Univ. of Lancaster, England, July 9-11, 1969
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Communication Problems, Costs, *Design, Electronic Data Processing, Evaluation, Evaluation Methods, *Information Retrieval, *Information Storage, Information Systems, *Mathematical Models, *Models, Relevance (Information Retrieval)
IDENTIFIERS	*Information Transfer

ABSTRACT

The transfer of information through space and time in communication systems is often accompanied by significant delays which give rise to meaningful storage problems. Mathematical models have been developed for the study of these kinds of problems which are applicable to the design of manual, library-type, or mechanized information storage and retrieval systems. This state-of-the-art review of such models divides the subject into three kinds of storage models: those concerned primarily with spatial efficiency, those concerned with usage and cost, and those concerned with retrieval accuracy. (Author)

ON INFORMATION STORAGE MODELS *

Ferdinand F. Leimkuhler

Research Memorandum No. 69-5

School of Industrial Engineering
Purdue University
Lafayette, Indiana 47907

This is a working paper. Comments are invited and should be directed to the author at the above address. Please do not reproduce in any way without the author's permission. A list of reports available is attached at the end of this report.

* Prepared for a Seminar on Planning Library Services, University of Lancaster, England, July 9-11, 1969. The first part of this paper was read at the Sixth Annual Information Retrieval Colloquium, Philadelphia, May 9, 1969.

Work supported in part by the National Science Foundation, Office of Science Information Services, Grant GN-759, while author was Visiting Professor, School of Librarianship, University of California, Berkeley.

ABSTRACT

The transfer of information through space and time in communication systems is often accompanied by significant delays which give rise to meaningful storage problems. Mathematical model have been developed for the study of these kinds of problems which are applicable to the design of manual, library-type, or mechanized information storage and retrieval systems. This state-of-the-art review of such models divides the subject into three kinds of storage models: those concerned primarily with spatial efficiency, those concerned with usage and cost, and those concerned with retrieval accuracy.

INTRODUCTION

Information storage theory is not a well-defined area of research in the formal sense and one is still free to make of it what he wants. My own viewpoint is that of an industrial engineer and operations researcher who has been seeking ways to develop mathematical models for the analysis and design of library-type information storage systems. The work "storage" can evoke some bad vibrations in library circles where it is associated with the least respectable aspects of librarianship, but I choose to use the word in a broad and inclusive sense. While it may be more meaningful to define libraries as communications systems which transfer information through time and space, such transfers are accompanied by significant delays which give rise to meaningful storage problems. The study of libraries from the storage viewpoint can identify some crucial aspects of information systems which are often overlooked or ignored when the focus is on communication.

Although I have been working at storage models for several years and have advocated its practical importance to libraries, I am quite aware that it is not an easy matter to translate theory into practice. It has been my experience that action follows from need, but that better action can result if some good theory is available to help diagnose the need and to guide the remedial efforts. Our present theory is quite rudimentary from a research viewpoint and has a good way to go before it reaches the sophisticated state of, say, modern inventory theory, but the beginnings are there and there is good promise of a rich harvest.

It seems a bit ironic to me that the first important exploitations of this new study of library systems and the best guarantee of its continued support are probably going to come from outside the library proper. It is in the design of automated special purpose information systems that one has the greatest freedom to make innovations and the greatest pressure to apply a uniform economic and technical yardstick to every facet of the system. Conventional libraries are already mature technological systems in their own right. They are predicated on an earlier piece of mechanical wizardry -- the printed book and a well-developed clerical work system to support its exploitation. Within these bounds most of the waste has been trimmed away in the long lean years of experience. As with a magnificent old clock, one doesn't tamper with it. If you want to keep it running, you will have to find parts made to the original specifications.

Still, I find it fruitful and hopeful to pursue the study of information storage systems in the context of conventual libraries. They offer a rich source of experimental data, and a wealth of ingenuity in an operational setting. As a "going" system, it is a good place to test the validity of one's models. Furthermore, there are plenty of indications that some radical changes in library operations are going to have to occur in the not too distant future.

Space Models

A good first example of the nature and implications of storage theory is the book shelving model which was developed at Purdue several years ago (1, 2). The model assigns a given collection of books to a set of shelves with those lengths and heights which will minimize the shelf area required. The direct application of this model to some representative library collections has indicated that relatively efficient storage can be achieved by using only three or four different heights; and, in fact, by shelving large books on their fore edge, one can do remarkably well with only two shelf heights (3). This result poses an interesting question to those large libraries which presently employ eight or more size classifications in their depository-type storage areas. It also calls into question the wisdom of spending extra money on variable height shelving and the practice of adjusting shelves up and down as new books are added. But these are rather minor benefits; and, perhaps, the greatest immediate importance of the model is its ability to show rigorously that one is not going to achieve dramatic reductions in space utilization through shelf arrangement alone. If all books could be stored by size on their fore edge, the best one could do is to double shelving capacity (4). This is not by any means a long run solution to library storage problems.

Of considerable interest are two recent applications which are peripheral to the shelving problem. In one instance, the MARC catalog tapes produced by the Library of Congress were examined for the distribution of lengths of the records they contain (5). Some 65 different record lengths were found in a random sequence. When ordered by size, they formed a bell-shaped distribution. The book shelving model was applied to find the optimal record lengths to use for blocking the tape so as to produce a fixed record length tape with one, two, three, etc., different block sizes. The records would be in conventional sequence within each block size, and shorter records would cause some loss of storage space. As with books, it was found that the use of only a few block sizes could make fixed length processing relatively efficient at the expense of more storage capacity. Again, as with flipping large books on their fore edge, the model could be used in conjunction with a program for selective code compaction of longer records so as to achieve an optimal balance of processing and storage costs.

An even more esoteric application of the basic shelving model is the possibility of using it in the production of microform records. For example, in producing microforms of conventional book material, one has to compensate for the variable sizes of book pages. What set of fixed frame sizes would achieve an optimal balance between the cost of handling variable frame sizes and the cost of reproducing blank spaces? If a single frame is used, it must accommodate the largest page size at the expense of much excess capacity for smaller pages. If only two sizes are used, what smaller size is optimal? The use of more sizes decreases lost area but increase the complexity of the system. Again, as with the fore edge storage of books and the compression coding of MARC tapes, variable magnification might be used in conjunction with the selection of optimal frame sizes. This complicates the analysis considerably but allows for many more options and the possibility of a much better solution.

Another different sort of application of the book shelving model was made recently to the design of industrial warehouses, where the problem was that of determining optimal bay configurations and the assignment of variable size lots of palletized materials (6). This might have useful implications for the design of library building and the assignment of subject groupings of varying size to different areas so as to minimize the sum of paging and space costs.

Usage and Cost Models

Space models of the kind considered above have the analytic advantage of dealing with the physical measurement of inanimate objects and avoiding the more difficult problems of measuring human behavior and judgment. It is the absence of the human element which makes them most amenable to mechanical applications and which evokes the strongest suspicions of practical librarians. There are two basic ways of approaching the role of human intervention in man-machine systems. One way is to take the direct approach and concentrate on people, their perceptions and reactions to the system. This is the approach of the behavioral scientist.

A second approach is an indirect one of focusing on the physical components and attributing to them attributes which are really the net effect of some prior human action. For example, we speak of a book circulating, of it containing certain information, or of it having so much worth and relevance. This approach permits the reduction of much of the human element to measurable quantities which can be related directly to other aspects of a system. This is the approach of the economist who can infer a value measurement from the limited availability of certain resources and the desire to have more of everything rather than less.

A good example of this approach is in the work of Philip Morse and his recent book on Library Effectiveness (7). Most of his models depend heavily on the notion of "randomness" in the behavior of library patrons. Tossing coins to retrieve information is an idea which seems patently absurd, if applied to some individual researcher, but is remarkably useful in measuring the collective effects of many individual choices and actions on the performance of a service system. Once we accept these measures as good approximations, we are in a position to make meaningful comparisons and recommendations for system improvement.

The analysis of depository schemes for libraries is a good example of what I call usage models. In general, depository models have argued that a considerable portion of a library's collection is so rarely used that these items could be stored elsewhere at less cost or to make room for new material. On the basis of out-of-pocket library costs alone, Lister (8) argued that several science libraries at Purdue could justify the storage of up to 60% of their holdings and achieve a small reduction in total costs. However, if some significant user delay cost is added to the charges, the advantages of depository storage are reduced drastically. The net effect of his study is to show that depositories do not provide an easy solution to library storage problems. Where space is limited and storage is the only answer, however, Lister's models do show how a rational, suboptimal policy can be developed.

A variation on this theme is seen in the recent study by the Center for Research Libraries (9) of the feasibility and potential benefits of a cooperative storage and lending facility for periodicals. This study is notable for its analysis of pertinent cost data from several libraries. A similar preliminary study was made in England at the University of Lancaster which showed that university libraries might utilize a national lending service for 10 to 30% of their demands depending on the user delay cost.

Perhaps a better prototype storage model of the usage variety is the one proposed by Cole (10) and refined and extended by Buckland (11). Cole showed that a 2000 volume petroleum library could expect to satisfy the greatest number of user requests by subscribing to approximately 190 journals and holding them for about 11 years. He assumed exponential obsolescence of older volumes and a Ziph-type pattern for the marginal productivity of additional journal titles. Buckland introduced some considerable mathematical refinement to these basic relationships and was able to go beyond Cole's results and show how to include such additional features as variable retention periods for different journals and use of interlibrary loan options. He also looked at how to meet a given level of service and minimizing a cost function that gives explicit recognition to different storage policies. An excellent review of the history of library use studies and models was made by A. K. Jain (12).

Somehow the usage models, like the space models, seem to fall short of the mark in an attempt to come to grips with the critical problems of libraries. Libraries do have something in common with warehouses and bookstores but there is still a residual difference which cannot be ignored. The further development of economic models of library-type systems must focus on investment as well as operational costs. Because of their patterns of long-term storage and exponential growth, investment models may provide the better approach to the understanding of library economics. This approach would seem to be better suited to the development of system planning models and the justification of technical innovations.

Retrieval Models

One cannot pursue information storage models very far without confronting difficult problems of information retrieval. It is interesting to observe how operations researchers and industrial engineers have tended to focus on the storage side of library systems while library scientists focus on the retrieval side. M.E. Maron has defined "the library problem" as the problem of retrieval and not of storage. He points out that the space problem is largely a problem of technique and economics -- a matter of miniaturization, for which the necessary physical theory is already available, but that any use of miniaturization or mechanical storage is going to necessitate the development of a sophisticated remote access capability. The theoretical work on information retrieval which is necessary for such a development is not available now.

The separation of storage and retrieval is the critical factor in the automation of information systems and libraries. Conventional libraries must depend on direct user access to keep costs within bounds and to make card catalog systems work, i.e., we really have catalog-aided manual retrieval. Interlibrary loans, for example, are one of the most expensive kinds of services a library offers; and yet even this is cheap when compared to the cost of providing remote reference service, as for example, in the specialized information centers which the government has funded.

A thorough review of retrieval models is too large a subject to cover here and is beyond my competence to review. There appears to be a wide variety of approaches and classes of models, among which are those based on behavioristic studies of how man uses language in the transfer of information; and then, there are the computer-oriented approaches which concentrate on the algebraic and physical capabilities of electronic devices.

The approach I have taken is an operational one in that it attempts to model the patterns observed in existing working systems, and to draw inferences for local optimization and evolutionary development. This is the method of operations research as opposed to basic research and is not offered as a substitute for the latter but as a complimentary approach.

It is characteristic of OR work to look for analogies from other fields and to draw heavily on the selective experiences of past observers so as to attempt a sort restatement of what is known about a system in the language of applied mathematics. An example is the model which appeared recently in American Documentation (13) where ideas were taken from the theory of military search and reconnaissance and from the earlier empirical work and wisdom of the English documentalist, S. C. Bradford. These ideas were used to formulate an analytic model that also incorporates a mathematical approach which is similar to the math used in the book shelving model, and which, by the way, is developing its own separate history within OR circles under such names as the assignment problem, the cutting stock problem, the packaging problem, and other such titles.

This approach seems promising, expecially in the connection it made with Bradford's "scattering" studies, which can be related to Ziph's law and which, in turn, opens the door to some promising extensions into information theory, linguistics, and economic theory. Furthermore, the first model has called into question the proper measurement of search effectiveness and its relation to user preferences, perceptions, and behavior (14), and the relation of the latter to expect judgment of the relevance and content specification of a chunk of printed matter we call information. Thus, the models have a double payoff: they can lead to practical applications and can also open doors to new theory.

References

(1) Leimkuhler, F. F., and Cox, J. G., "Compact Book Storage in Libraries", Operations Research, 12:3, pp. 419-427, May-June, 1964.

(2) Cox, J. G., "Optimum Storage of Library Material", Ph.D. Dissertation, Purdue University, June, 1964. See Reviews in Library Quarterly, 35:3, College and Research Libraries, 26:3, and Library Review, 20:2.

(3) Popovich, J. D., "Compact Book Storage", M.S.I.E. Project, Purdue University, 1966.

(4) Raffel, L. J., "Compact Book Storage Models", M.S.I.E. Thesis, Purdue University, June, 1963.

(5) Stirling, K., Cost Exchange Analysis of Variable Length versus Fixed Length Marc II Bibliographic Records, Course 244-C, School of Librarianship, University of California, Berkeley, Fall, 1968.

(6) Roberts, S. D., "Warehouse Size and Design", Ph.D., 1968, Purdue University, Major Prof. Ruddell Reed, Jr.

(7) Morse, P. M., Library Effectiveness, The M.I.T. Press, 1968.

(8) Lister, W. C., "Least Cost Decision Rules for the Selection of Library Materials for Compact Storage", Ph.D. Thesis, Purdue University, January, 1967. U.S. Clearinghouse Report PB 174441.

(9) Williams, G., "Library Cost Models: Owning Versus Borrowing Serial Publications", Center for Research Libraries, Chicago, 1968.

(10) Cole, P. F., "A New Look at Reference Scattering", J. of Documentation, Vol. 18, No. 2, June 1962.

(11) Buckland, M. K., and Woodburn, I., "Some Implications for Library Management of Scattering and Obsolescence," University of Lancaster Library, Occasional Papers No. 1, 1968.

(12) Jain, A. K., "A Statistical Study of Book Use," Ph.D. Thesis, Purdue University, January, 1968, U.S. Clearinghouse Report PB 176525.

(13) Leimkuhler, F. F., "A Literature Search and File Organization Model", American Documentation, 19:2, pp.131-136, April, 1968.

(14) Baker, N.R., "Optimal User Search Sequences and Implications for Information Systems Operation", School of Industrial Engineering and The University Libraries, Purdue University, 1968.

RESEARCH MEMORANDUM SERIES

No.

69-1 "Computer Planning and Optimization of Automated Manufacturing Processes", P. B. Berra and M. Barash, May 1969.

69-2 "The Nature of the Distribution of the Life of HSS Tools and Its Significance in Manufacturing", J. G. Wager and M. Barash, May 1969.

69-3 "Some Studies of the Effect of Service Time Variability on the Design of In-Process Storage Areas and the Dynamic Operation of Production Lines", D. R. Anderson and C. L. Moodie, May 1969.

69-4 "Optimal Harvest Policies for Natural Animal Populations in Which Members are Indistinguishable as to Sex at the Time of Harvest", S. H. Mann, May 1969.

69-5 "On Information Storage Models", F. F. Leimkuhler, June 1969.

69-6 "A Basis for Time and Cost Evaluation of Information Systems", R. R. Korfhage and T. G. DeLutis, June 1969.

69-7 "Comparative Design Considerations of the Engineering Plastics and Die Casting Alloys", R. F. Adams, June 1969.

69-8 "Storage Policies for Information Systems", F. F. Leimkuhler, June 1969.

69-9 "A Mathematical Theory for the Harvest of Natural Animal Population in the Case of Male and Female Dependent Birth Rates", S. H. Mann, July 1969.

69-10 "A Mathematical Theory for the Harvest of Natural Animal Populations when Birth Rates are Dependent on Total Population Size", S. H. Mann, July 1969.

69-11 "A Mathematical Theory for the Control of Pest Populations", S. H. Mann, August 1969.