PROPOSALS FOR A DYNAMIC LIBRARY

G. Salton

72 - 138

August 1972

Department of Computer Science
Cornell University
Ithaca, New York 14850

Revised 9/13/72

Proposals for a Dynamic Library

G. Salton+

Abstract

The current library environment is first examined, and an attempt is made to explain why the standard approaches to the library problem have been less productive than had been anticipated. A new design is then introduced for modern library operations based on a two-fold strategy: on the input side, the widest possible utilization should be made of cooperative and shared operations, whereas dynamic, user-controlled procedures should be used for the subsequent internal processes. The dynamic environment applies in particular to the maintenance of the indexing vocabulary, the organization of the stored information files, the performance of search and retrieval operations, and the control of the library collection necessitated by document growth and retirement. Some experimental results are included as an illustration for the proposed dynamic operations.

1. Introduction

It is no longer controversial to affirm that all over the world there exists a library crisis and that some changes are in order if the traditional library services are to be maintained in the future.

+ Department of Computer Science, Cornell University, Ithaca, New York 14850

This study was supported in part by the National Science Foundation under grant GJ-344, and in part by the National Library of Medicine (NIH) under grant LM-00704.

Following a recent study of large university libraries, Mason put the situation in the following terms [1]:

> "the great university libraries are largely in a state of quiet crisis, ill-housed, ill-served, and ill-fed ...; within ten years, the full impact of overloading and under-supporting the research libraries will become painfully apparent ...."

The difficulties stem from a combination of factors: the increase in the amount of material to be processed and stored, the greater demands for service from a growing user population, the larger costs for goods and services, and the availability of new types of materials which cannot be treated by conventional methods. The obvious result is a breakdown of the established operations, an intellectual crisis among the people responsible for library management, and a deepening space and budget problem.

The latter, in particular, has led many observers to conjecture that the continually increasing requirements of the modern library and the growing budgetary needs cannot long be met by expanding the traditional library functions: "one cannot expect that an ever-increasing fraction of the available resources be allocated to some one function [the library], regardless of its importance." [2] Rather, it is argued the need arises for new organizational structures and new procedures which may better cope with the existing problems than conventional approaches.

In the remainder of this study, the current thinking concerning the library problem is briefly reviewed, and proposals are made for the implementation of a new environment for the future library.

2. Current Approaches to the Modern Library

Three main approaches to the library problem are apparent at the present time. The most obvious — although not necessarily the most painless — way of proceeding consists in retaining the current operations conceptually intact, while attempting to mechanize some of these procedures using data processing or computing equipment.

A second approach is based on the creation of compacts and cooperative arrangements whereby several libraries decide to share resources or experience to accomplish some common task; alternatively, a central agency may perform a given operation jointly for a number of different libraries. In either case, the hope is that the resources to be invested by individual participants become smaller, and the expenditure and upheaval inherent in the library processing turns out to be more bearable.

A final trend to be mentioned consists in postulating certain more drastic alterations of the present methodologies involving the creation of large library networks, the implementation of mechanized records for all library holdings, and the replacement of normal hard-copy storage and standard circulation procedures by a (long distance) transmission of the corresponding stored records, and the reproduction of the information as required by the various user populations.

A) The Mechanization of Standard Library Processes.

Most large libraries have introduced mechanized procedures to a greater or lesser extent. In decreasing order of popularity, the following operations are principally affected [3-6]:

i) the treatment of business operations, such as budget and accounting work, as well as the handling of book funds, bindery records, gifts and related matters;

ii) the processing of document acquisitions, including the ordering and control of serial and monograph items, and the handling of the accounts payable following the receipt of new documents;

iii) the control of circulation records, including loan procedures, check-out and check-in of material, automatic dissemination and transmission methods, and so on;

iv) the generation and maintenance of lists and records of many types, including the library catalog, acquisition lists, index and subject listings, shelf lists, technical abstract bulletins, dictionaries of various types, and so on;

v) the handling of reference and search work, often called "information retrieval" in the literature, including bibliographic (catalog) searches, demand searches by subject or content specification, and recurrent searches of the type used with selective dissemination systems where users are automatically notified of interesting materials on a continuing basis;

vi) the content processing, including the indexing, cataloging, abstracting and classification of documents and journals.

In the majority of libraries the mechanization efforts are restricted to business-type operations such as accounting and the production of printed lists of various kinds. Several hundred organizations have also mechanized their acquisition procedures and circulation records. Comparatively few use automatic search and retrieval operations — one major impediment being the creation of the voluminous mechanized files required for this purpose; finally, only very few are venturing into the areas involving the mechanization of intellectual tasks such as cataloging or classification.

For the most part, the mechanized housekeeping procedures are carried out off-line, in a batch processing mode; that is, any file changes that occur are accumulated (batched), and all records that require updating are simultaneously treated in a single sweep through the file. A complete new acquisitions file may then be created once per week, or a new circulation file once a day.

The main disadvantage of the batch-processing environment is the lack of real-time control. That is, the exact status of a given record may not always be ascertainable, since changes are not incorporated into the files as they occur. To obtain a continually updated set of records, the operations must be carried out on-line by using computer terminals to initiate individual transactions whenever they take place. An on-line implementation may in time lead to a complete mechanization of all library operations, using a multiterminal computer system, in which the same bibliographic record, in various stages of completeness, serves all library functions. [7]

A number of on-line library acquisitions systems are now in operation [8,9], as well as several on-line book circulation systems. Some of the latter may include provisions for queueing unfilled requests, followed by the automatic generation of notices to those customers whose requests had previously been held in abeyance. [10] Systems have also been implemented in which the user himself performs the necessary console operations for the charging and discharging of library items, [11] and in some cases, items may be charged out over the telephone. [12]

While the cataloging itself is normally carried out by trained experts, on-line methods, including, for example, the console display of bibliographic information and of vocabulary control data, may be used as an aid to the cataloger. [13] Furthermore, considerable thought has been devoted to the creation of mechanized procedures capable of using cataloging data from several different sources, such as locally generated data, and data originating at the Library of Congress. [14, 15, 16]

In summary, a considerable degree of sophistication has been achieved over the years in some areas of library mechanization, and while a great deal remains to be accomplished, most people concerned with library procedures feel that progress is being made in the right direction.

B) Centralization and Cooperative Efforts

The second notable trend in the library field is a tendency to consider seriously the advantages of joint efforts and cooperative ventures in the mechanization area. Many observers in fact, believe that a mechanization of library procedures should not be carried out separately by individual libraries, but that the investment is easier to justify in a somewhat different environment [2]:

> "one concludes that the problems met in attempting to serve a large, growing, diverse community are not to be solved locally with more buildings and more employees, and that if a solution is to be achieved, it must be done either by altering the problem or by going beyond the local level ..."

"Going beyond the local level" in this context implies the institution of centralized or cooperative ventures.

In principle, collaborative procedures are usable for all library procedures. In practice, the ends of cooperation are normally assumed to consist in the achievement of individual library goals rather than in the furtherance of a strong cooperative able to respond as a unit to its environment. As a result, most collaborative schemes are restricted in scope, and applied to simple, specific tasks such as card catalog production, or accounting, or printing of common catalogs and statistical data. [17, 18, 19]

A possible alternative to the introduction of collaborative arrangements is the delegation of certain tasks to a centralized agency. This is often easier to carry out, because such a transfer of responsibility does not raise the usual fears concerning loss of autonomy or freedom of action on the part of any individual library. By far the best known and most influential activity of this type is the utilization by a variety of libraries and information centers of the tape distribution services that now make available bibliographic information prerecorded on magnetic tape. In the library field, the principal services of this type make use of the MARC (Machine Readable Cataloging) records originating at the Library of Congress. [14-16, 20-22]

The MARC system makes available on tape in a standard format cataloging data for English language materials processed at the Library of Congress. Such data can be used by the recipients "passively" if the need for the particular citations has already been established within the library before the MARC records are received, or "actively" if the MARC record itself triggers an initial ordering or acquisitions operation. In either case, procedures must be available for searching MARC tapes by comparing the incoming information with individual current production needs, and for merging matched MARC records with internally generated data to produce a combined file for all items currently in process.

Centralized cataloging information can, of course, be used for a variety of purposes, including ordering, receiving, book accounting, cataloging, binding, labelling, and so on, and considerable savings may be produced, assuming that the tape formats are acceptable and that the service is timely and somewhat comprehensive. The same argument is put forward by a number of service bureaus that offer processing facilities to individual libraries, thereby eliminating, at least in theory, obvious overlapping.

A final approach toward cooperative endeavors in the library consists in the creation of library networks, where each component is connected to the others, and information can be exchanged by the several participants. The aim here is to institute common, or at least compatible, procedures, while at the same time making available to individual component members the resource of the whole network.

A number of such networks are now in existence, generally connecting organizations with common subject interests (Biomedical Information Network, Medline-Medlars On-Line, or with close regional contacts (Ohio College Library Center, New England Library Network, and so on). For the moment, the network organizations are mostly concerned with the institution of common housekeeping operations, and their influence over the total library scene is still somewhat limited. However, as communication costs go down, and the use of on-line processing becomes more prevalent, one can expect that the importance of the network concept will grow. Eventually, a meaningful pooling of network resources may be undertaken by eliminating some duplicate holdings, combining subscriptions to some specialized journals, and instituting common storage facilities for certain portions of the collections. [23,24] Such a development would dramatically change the library cost picture, but its realization requires a new understanding of the role of individual libraries, and some change in organization and attitude on the part of library managements.

C) The Library of the Future

A final development, illustrative of the thinking of some of the more influential observers on the library scene, is the conceptual design of the "library of the future". [25, 26, 27] In its most provocative form, such proposals postulate the complete elimination of books as we now know them [26]:

> "..any concept of the library which begins with books on shelves is sure to encounter trouble .... we should be prepared to reject the schema of the physical library — the arrangement of shelves, card indexes, check-out desks, reading rooms, and so forth..."

Instead a centralized repository is suggested based on a combination of computer storage, micro-reduction, and remote transmission of information, from which information could be extracted by a suitable interactive search process. In each case, a variety of linguistic, processing, and display tools would be used to establish appropriate connections between the customers and the corresponding relevant information items [26]:

> "we are assuming that the average man of that year (2000 AD) may make a capital investment in an "intermedium" or "console" — his intellectual Ford or Cadillac — comparable to the investment now made in an automobile .... the computer would be used as an interface in applying various sequences of procedures to named texts, graphs, and tables; observing the results; and intervening whenever a change or extension of plan is required."

Clearly, such a design is most appealing in that it eliminates the enormous duplication caused by the storage of multiple copies of the same material in many different places; it keeps the stored information continuously available for all comers because texts do not "circulate" in the normal sense of the word; it avoids delays in obtaining the desired information items and prevents losses, and, finally, it may improve not only the efficiency but also the effectiveness of the search. As Kemeny points out, [25]

> "I look forward with delight to being able to find in ten minutes everything relevant that has been written in a subject."

While no one has seriously proposed that a prototype of such an "ultimate" system be constructed at the present time, attempts have been made to implement more modest mechanized information transfer systems based on the simultaneous storage of full document texts as well as document surrogates and accessing information. [27] Such systems might be used to manipulate large data bases on-line, and to provide experience with a variety of information handling procedures, including the maintenance of stored library catalogs, the accessing of full document texts, the integration of different types of data bases, the utilization of text editing, reviewing, indexing and abstracting methods, the use of the on-line technology for the publication of new items, and the development of techniques for querying the data base.

Even this more modest on-line environment has not so far been realized for a variety of technical, financial, legal and political reasons. Some of the impediments which must be overcome before progress can be made in building the "new" library are examined in more detail, in the next few paragraphs.

3. Critique of the Current Approach

A) The Ultimate Library

Consider first the most advanced types of proposals which would eliminate books and libraries as we now know them, and replace these by man-computer interfaces for the dissemination of knowledge. While such systems may eventually come into being, it is unlikely that they will be seriously considered at the present time. The fact is that bound volumes of printed materials are likely to be in demand for some time, and the argument that books should be eliminated because they happen to be convenient for use only by human beings but not by machines appears insufficiently compelling.

Even if one were to grant that the users of information products might eventually get used to computer-produced copies of the stored information, the "procognitive" systems have little change of realization in the foreseeable future: on the one hand, the usual manual content analysis operations would not be adequate to analyze and classify a vast store of diverse knowledge; on the other hand, the free and unrestricted kind of user-system interaction that would be used to analyze and classify information, to find one's way in the information store, and to supply the necessary instructions for further processing cannot be utilized in any practical situation. Specifically, it is not possible to perform "a combined syntactic-semantic analysis of a text that reduces every sentance to a (linguistic) canonical form, or to a set of expressions in a (logical) predicate calculus;" [26] nor do we have capabilities that would easily mix question-answering and reference retrieval routines; [26] nor is it reasonable to assume that machine translation facilities could be used when needed; [25] nor can we implement in a meaningful way even simple systems to simulate the operations of a computerized handbook. [27]

In other words, it appears that the reality of the library crisis will not be affected for some time to come by proposals which are still several levels removed from current capabilities. For this reason, attention will be restricted to systems more directly applicable to the libraries as we know them.

B) Prototype Centralized Networks

A theoretically easy way of introducing some degree of standardization into library processing, and of testing the effectiveness of collaborative library arrangements consists in creating a prototype library network. Unfortunately, the actual realization of such projects is more difficult than expected, and a great many questions are immediately raised by any proposal aiming at network construction.

There are first problems of scale and of strategy. Should one aim immediately at the implementation of a national policy, such as a national depository, or the mechanization of a national union catalog, or should a start be made in a circumscribed subject area and geographical location. Should one implement only housekeeping operations that are well understood, or should the prototype network be used to experiment with new features — for example, a fact retrieval experiment, or an experiment to disseminate manuscripts prior to publication. Definite plans are difficult to generate in this area, since no experiences are available upon which the necessary decisions might be based.

The next question relates to costs and cost-benefits. At a time when library budgets are subjected to critical examination everywhere, cost studies are conducted with considerable seriousness. Unfortunately, while reliable current costs may sometimes be ascertainable, a quantification of benefit remains elusive. [28, 29] In particular, the quantitative contribution of information to the productivity or effectiveness of the user population is unknown and generally not determinable. This hampers the performance of cost-benefit analyses, and makes it difficult to adduce convincing arguments in favor of a mechanization of automatic information services.

The third question to be raised about the construction of prototype information networks covers legal and sociological problems, and more particularly the issues relating to copyright [30-32], and privacy. [33-37] The difficulty concerning the copyright is simply that the unlimited storage, reproduction, and transmission of copyrighted information tends to destroy the rights of copyright ownership and impairs the "fair use" concept under which copyrighted information can now be used for certain purposes. For this reason, the organizations that now depend on the copyright protection — including particularly the publishing industry — claim that the copyright restrictions should apply on input, that is, at the time the information is first entered into a computer network, rather than at the time of utilization when controls are more difficult to institute. Additionally, there is some concern among copyright owners that the "integrity" of a piece of information is difficult to preserve in the age of the computer utility, since the stored information can be manipulated in various ways — for example, by generating concordances, indexes, abstracts, translations, abridgements, and so on — again without controls over the utilization of these derivatives.

The privacy issue is equally tricky. While there is a realization that much information stored in mechanized data banks may be beneficial to the user population, there is also increasing concern about potential abuses. In particular, it may be desirable that information about private individuals or business enterprises be ethically or legally restricted to persons with a right to know. Unfortunately, privacy decisions which appropriately restrict or grant access are difficult to postulate and to enforce, since insight can often be gained by cross-correlating, or manipulating seemingly disparate and innocuous stored information items. Furthermore, even if the stored information were secured against unauthorized access, the accuracy of the stored information might not be guaranteed, so that harm might result even from authorized access and disclosure.

In summary, a variety of legal and ethical problems must be resolved before sophisticated information networks will prove acceptable for general use.

A last question pertaining to the implementation of information networks relates to the availability of adequate technologies. While the picture is still somewhat mixed in this area, improvements are expected soon where they are needed. [2] Large-scale computing equipment is already available, and the performance of the small computers that might be needed to connect individual user stations to a central computer is improving relative to the large ones. The internal core or disc storage capacity at acceptable cost and with practical access times is also already on hand for many purposes, and drastic improvements are foreseen in this area for the future. [2] At the moment, the less expensive microform technology is still mandatory as a primary medium for large-scale text and image storage.

The key problems relating to the computing and storage technologies remain lack of standardization and compatibility between various lines of equipment, and specifically the problems arising in the interplay between photographic microform and computer technologies.

The input-output operations are less well understood. In particular, there still exists no easy process for converting hard-copy text or images into machine-readable form. On the input side, optical character recognition equipment of sufficient sophistication and flexibility to process the needed variety of typefonts and character faces is still not available, and neither are output printers that can supply the character variations normally found on standard library cards. The long-term hope for the solution of the data entry problem rests on the continued evolution of the printing industry, and in particular on the use of computer-controlled typesetting routines able to generate data tapes that can later be read into a computer for storage and retrieval purposes. In the meantime, much of the source data information must be elaborately processed by keypunching techniques.

The terminal equipment necessary to carry out the normal user-computer interaction forms another problem area. A large variety of different consoles is currently used, ranging from simple keyboard devices and touch-tone telephones to elaborate graphics and audio terminals. However, users complain that the more sophisticated consoles are difficult to use and too expensive to acquire for many standard applications.

The data communications area presents what is likely to be the most serious challenge of all in the equipment area. Everyone agrees about the technological feasibility of transporting large volumes of data over long distances by electrical means. The problem is an economic one, in that channels of sufficient transmission capacity and speed are out of reach for normal library applications, whereas narrow-band transmission techniques — for example by facsimile methodology — are very slow.

In several hardware areas, including notably high capacity storage, input-output, and transmission equipment, a concurrent performance improvement and cost reduction will be required before information networks can be implemented comfortably. And overall, one must conclude that too many loose ends remain to be tied up before centralized networks may be expected to offer the kind of services now provided by the conventional environment.

In the immediate future, an improvement in the library situation must then come from a reevaluation of the standard library practices and procedures. This question is examined in the next subsection.

C) Library Software Assumptions

The strategies currently used for library mechanization appear to be based on a number of assumptions that are widely believed and accepted. The first of these states that <u>the clerical (housekeeping) operations constitute the most promising area for the institution of mechanized operations in the library</u>. Cost is believed to be the main obstacle to the implementation of mechanized housekeeping; as Locke says [38]: "We will bring in computers the minute they can do the same job for less money, or a better job at a price we can afford." The implication is that sooner or later the cost picture is bound to improve, and that the negative experiences that have accumulated in this respect in many places are a temporary phenomenon.

Unfortunately, an examination of library housekeeping operations from a data processing viewpoint reveals a considerable number of problem areas:

a) the library files to be processed are large, including sometimes many millions of items;

b) the file maintenance and updating operations required on a day-to-day basis are extensive;

c) a great variety of inputs are used and several dozen different output products are normally generated;

d) real-time control of the library collection is considered desirable in the sense that the whereabouts of each item should be ascertainable at any time.

Obviously, such a set of requirements is certain to impose great strains on the computing facilities, because no application involving large files subject to a great deal of updating is easy, or cheap, to mechanize; and the real-time control by itself requires for implementation a large complement of expensive, fast-access files. In these circumstances, it is not surprising that the application of computers in the library has not proven to be cost-effective for the most part. Nor is it likely, that the plea of many librarians for cheaper and more rational computational processes is likely to be heeded, because costs are not coming down by the orders of magnitude that are apparently needed to produce a viable mechanized process.

Thus, contrary to popular assumption, the library problem does not appear to be solvable by a superficial mechanization, one operation at a time, carried out separately in each individual library. Instead, a change in administrative practice is needed first, and a good deal of justification exists for waiting with the introduction of computers until common standardized procedures can be introduced cooperatively into many different environments. When each item will be cataloged once only, instead of hundreds of times, and a given mechanized card catalog will prove acceptable for use by many organizations, instead of only by a single one, the existing cost projections may no longer apply, and mechanized library housekeeping operations will have a much better chance of proving cost effective than under present operational conditions.*

The second assumption relating to the introduction of computers in the library field states that computers are not suitable for carrying out the content analysis operations required for document indexing and cataloging. The arguments brought forward are based largely on sentiment rather than facts. Specifically, one demonstrates (correctly) that a complete semantic analysis of document texts cannot be carried out reliably by automatic means, and one concludes (falsely) that the resulting automatic indexing products are necessarily inferior to manual ones produced by trained catalogers or subject experts. In the process one forgets that manual indexing and cataloging products are also unreliable and that perfect performance, either manual or automatic, is not attainable in the library field. As Swanson points out: "though machines may never enjoy more than a partial success in library indexing, ... people are even less promising." [39]

* This statement is likely to prove correct, even though the local processing costs in each library are liable to be higher for a mechanized card catalog than for a conventional manual one.

The available evidence, collected over the last ten to fifteen years, indicates that practically useful indexing products are obtainable fully automatically with comparatively little effort. [39-43] In the most recent comparison between a conventional, intellectual indexing process performed under controlled conditions by trained subject experts (Medlars) and an automatic term extraction process using document abstracts as the principal machine input (SMART), it was found that a deficiency in performance of 30 to 40 percent in recall and precision* produced by a raw, automatic term extraction process can be turned into an advantage of 15 to 30 percent in recall and precision over the conventional, manual process by using a variety of mechanized processing features such as automatically generated dictionaries and thesauruses for language normalization, ranked document output in decreasing query-document similarity order, and interactive searching based on feedback information supplied by the user [43].

In any case, the mechanization of content analysis and indexing operations should not automatically be excluded from the design of the future library.

---

* <u>Recall</u> is the proportion of relevant material actually retrieved, while <u>precision</u> is the proportion of retrieved material actually relevant. Ideally, everything relevant is retrieved while at the same time everything extraneous is rejected, producing recall and precision values equal to 1.

The third assumption in the library mechanization area is that a combination of standard library classification methods and inverted indexes of subject indicators and keywords are adequate for storage, searching and retrieval. Concerning first the standard library classifications, every available test result indicates that classificatory indexing languages — that is, languages in which hierarchical relations can be specified between terms — are unrewarding in a retrieval environment. In this connection, Keen says that "the theory is sound enough, but no way has been found to realize the benefits of classificatory languages in practice". [44] The evaluation results seem to indicate that the use of the hierarchical linkages for indexing purposes leads to an over-specification of document content, and that it is easier to supply any missing controls at the search stage in the form of a thesaurus or other language normalizing tools, or through interactive search methods, than to have them built into the indexing language.

Misgivings have also been voiced about the existing library classification systems in their role as links between catalogue and the classified arrays of volumes, and as an aid in promoting browsing at the shelves. The main complaints relate to the static quality of most classification systems, the difficulties of updating and maintaining the schedules, and the need to classify a given item uniquely into the tree structure provided for that purpose.

While it is, of course, necessary to choose a sensible shelving arrangement — for example, by placing items dealing with a common subject area in adjacent positions in the library stacks — and to keep this reasonably fixed, there is no need to tie the shelving order to a hierarchically classified catalog. Instead, a flexible cataloging method can be used in which the classes are easily adjusted, and items may be simultaneously entered into several overlapping classes. Moreover, the same class identifiers should be used for query formulation and for search and retrieval, as well as for the normal classification purposes.

It should be noted in this connection that the inverted file organization method now considered standard for search and retrieval purposes exhibits flaws somewhat similar to those inherent in the conventional library classification methods. An inverted file search is normally conducted by consulting an inverted directory which supplies for each index term acceptable to the system a list of references, or accession numbers, for all documents identified by the given term. The accession lists corresponding to the various query terms are then merged to reflect the correct query formulation, and the document citations for all items exhibiting the appropriate combination of terms are retrieved from a citation file. As usual for keyword retrieval systems, the retrieval process separates the collection into two parts, containing respectively the retrieved items — those identified by the appropriate term combination — and the rest which are not so identified. No ranking, or order of preference is normally provided within either set, and each retrieved item is inherently considered to be as important as any other. Thus to obtain a desired level of performance in terms of recall and precision, it is necessary to find just the right kind of query formulation. For if the formulation is too broad, the retrieved set is likely to be very large, thereby imposing a great burden on the user and producing low precision even when the recall is fairly high; on the other hand, a narrow formulation is likely to restrict severely the level of recall which may be attained.

The main virtue of an inverted file search is the search speed, since it is normally necessary only to examine those file portions corresponding to index terms actually used in the query statement.

The disadvantages, however, of the standard Boolean searches used with inverted files are serious and varied and a reexamination of the normal search and retrieval methodology may well be needed:

a) an inverted file organization makes it difficult to vary the search depth so as to retrieve a greater or lesser amount of material, depending on user requirements; specifically, to obtain adequate performance variations, a user must be intimately familiar with the properties of the vocabulary, and the effects to be obtained by adding and/or subtracting terms from the query statements, or by changing the Boolean formulation;

b) it is awkward to carry out "partial matches" between queries and documents, so as to retrieve items that exhibit varying degrees of closeness with the Boolean search statement; while the construction of a set of nested subsearches can ideally produce the desired performance, the required multiple query formulations are difficult to generate for the average user;

c) the set of output documents cannot be presented to the user in ranked order, for example in decreasing order of query-document similarity; instead a partial ranking must be simulated by using weighted query terms, and by listing the output in decreasing order of the matching term weights;

d) finally, the indexing information attached to queries and documents cannot easily be changed in an inverted file organization, because the set of keywords belonging to any one item is scattered throughout the file, and in general as many file accesses are needed to collect the full term set belonging to a given item, as there are index terms attached to each document.

The output of Table 1 exhibits partial results obtained in the previously referred to comparison between the conventional Boolean search environment used at Medlars, and the automatic SMART system. [43] It is seen that the "partial match" policy used by SMART (which utilizes a correlation coefficient ranging from 0 to 1 to reflect query-document similarity), followed by a ranking of the documents in strictly decreasing order of the similarity coefficient, improves the SMART output by about twenty percent in recall and precision.

This assessment is confirmed by Lancaster's evaluation studies of currently existing retrieval services, all of which operate with the conventional inverted file organization: "the use of Boolean algebra for querying computer-based retrieval systems may have been a mistake ... arising through the similarity of early computer-based systems to systems employing edge-notched cards or the optical coincidence principle .. ... There was no possibility for partial match and the retrieval of lists of references ranked according to degree of match with the search strategy. Computer-based systems, on the other hand have no reason to be so restricted ...." [45]

Indeed they haven't, and the conclusion is that the standard file organizations currently used in library and information centers for both classification and search and retrieval might usefully be replaced by more flexible methodologies affording the same search speeds without the accompanying disadvantages.

A last assumption currently popular in the library environment relates to the role of the library user in influencing systems and procedures. Specifically, it is taken for granted that <u>the library is user responsive in its service aspects</u> (for example by acquiring items of interest to the user population, and by improving circulation and related services), <u>but users do not influence internal processes such as classification and indexing methods, storage organizations, or collection control methods.</u> That is, although the library lives in a dynamic environment characterized by a constant interaction with a changing user population, the feedback information which could potentially be secured from the users concerning various aspects of the collection processing is wholly ignored, and the users are insulated from the library environment which moves with a momentum of its own, regardless of what happens around it.

The picture that emerges from the current assumptions regarding library software is one based on unproven assertions, and on habits stemming back to the days of punched-card technologies: that the clerical operations are the ones that can easily be mechanized; that intellectual operations must be carried out by humans; that current library file organizations have proven themselves; and finally, that there is no place for user control over the internal library processing.

A new dynamic library environment is outlined in the next section in which clerical procedures are carried out cooperatively, while the intellectual search and retrieval and collection control operations are influenced by the user population. It is believed that the proposed system can alleviate the current library crisis without requiring an abandonment of books and journals, and without leading to dream situations that cannot be implemented under current conditions.

4. The Dynamic Library

A) Basic Principles

The library processing should ideally be based on three main principles:

a) a "total" library system in which one basic input serves to initiate a chain of successive processing steps — the basic inputs undergoing successive modifications as a result of the various steps in the processing chain; [46]

b) The widest possible use of cooperative and shared operations, including collaborative or centralized acquisitions policies, shared cataloging, and standardized library housekeeping operations. [17-24];

and c) an adaptive environment in which the user population influences the main intellectual processes such as the indexing vocabulary and practices, the storage organization, the search and retrieval operations, and finally the collection control necessitated by document growth and retirement. [47]

Specifically, the mechanization of the basic input and housekeeping operations cannot be justified if performed separately in each individual organization. It should therefore be undertaken cooperatively and/or centrally, using the greatest possible degree of procedure standardization and work sharing. Ideally, the work might be performed on-line using real-time procedures; however, the type of centralized off-line processing now inherent in the MARC tape distribution service is also most encouraging. It is already the case that fewer and fewer organizations now claim convincingly that their requirements are so specialized, and their user population so unusual, that a standardized process will not fit. One may hope that the developments in the direction of greater standardization and cooperation in the housekeeping area will continue to be pushed rapidly, because they constitute the best hope for an eventual improvement in the cost-benefit picture.

On the other hand, while the clerical input operations should be standardized, the collection management itself can certainly be handled by each library in accordance with its own environment. Not everyone must necessarily maintain the same file arrangement, nor must the same indexing vocabulary be used everywhere regardless of the orientation of the user population; and certainly, when the time comes to retire the less-used materials to an auxiliary store, each organization should use its own experience to decide what portion of the collection can best be demoted.

The dynamic system outlined in the next few paragraphs makes it possible for each library to be directly responsive to its user population, and overcomes the common failures which now arise from the use of static indexing and classification systems, the need to learn specialized, controlled indexing languages, and the lack of user-system interaction during information search and retrieval.

The main idea is to maintain the library cataloging system in a continuous state of flux. Documents are first tentatively indexed — hopefully automatically — by assigning to each item a set of weighted terms representing document content. These term vectors are then used automatically to classify the items into affinity groups, in such a way that items with similar vectors wind up in the same classes. During the processing, both query and document vectors are subjected to small changes — for example, by increasing or decreasing the weights of certain terms, or by adding a new term, or subtracting an old one. As the individual term vectors change, so will the document classification. In the end, a file is obtained in which "interesting" items wind up in the center of the file, whereas unwanted items wind up on the periphery.

The file organization and the main processing steps are described in the remainder of this study.*

B) Clustered file Organization

A direct file organization in which documents are not stored in subject order but in some random document order — for example, in chronological order of the acquisition dates, would obviously

* It should be understood that the dynamic operations described in the remainder of this study are applicable to the organization and processing of the mechanized catalog files, and to the software aids used in obtaining access to the stored items. The physical arrangement of books on shelves need not be affected, except of course for the addition of new items, and the removal to remote storage of items designated for retirement.

not be suitable for library processing, since subject or author searches would then require a complete match of every document with a search request, thereby producing excessively long search times. An inverted file, on the other hand, organized in order by permissible keywords leads to a static indexing system and inadequate search strategies. A clustered file organization is therefore recommended in which documents carrying somewhat similar content descriptions are automatically grouped into clusters. [48] Each cluster is identified by a representative cluster profile, or centroid, somewhat akin to the center of gravity of a set of mass points. A cluster centroid is simply a weighted set of terms, derived from the document vectors included in the corresponding cluster.

A clustered file is then similar in concept to a normal classified library file, except that the document classes are automatically generated, and that some overlap may exist between classes, that is, certain documents may be included in more than one class. A search in such a clustered file is carried out in several steps: each query is first compared with the index file of centroid vectors; for those centroids which exhibit a sufficiently high similarity with the query, the individual document vectors in the corresponding clusters are examined next, and the document citations are ranked for output purposes in decreasing query-document order, as previously explained.

It is clear that the "depth" of the search, as measured by the number of query-document comparisons, can be controlled in a clustered file, since it is possible to search only the "best" cluster — the one exhibiting the highest query-centroid similarity — or the top two clusters, or the top ten, as may be required. Moreover, since all document vectors and citations belonging to a given cluster are stored adjacently in the same storage area — for example, on the same track or cylinder of a given disc assembly — only one access operation is needed for each document cluster, as opposed to one access for each document citation in an inverted file.

A detailed comparison of inverted and clustered file organizations comes to the conclusion that the clustered file is more economical of storage, leads to faster retrieval operations, and permits more flexible, feedback-type searches. [49] The one disadvantage of the clustered organization, namely the requirement to generate the clusters in the first place, is not as serious as might appear, because fairly inexpensive clustering methods are known which require of the order of $n \log n$ vector comparisons to cluster n document vectors (as opposed to n^2 comparisons for the conventional methods that are based on a comparison of each document vector with all others). [48, 50]

Since the feedback searches and file reorganization methods are considered essential in future library processing systems, a clustered file organization is assumed in the remainder of this study.

C) Automatic Analysis with Dynamic Vocabulary

An automatic generation of document or query vectors must necessarily rely on the analysis of query and document texts. Normally, a term extraction process is used as a basis for all subsequent processing. [40-43] The following principal steps may be involved in a typical automatic word stem analysis and retrieval process:

a) the individual words included in document abstracts and in query texts are isolated;

b) certain words, such as function words listed in a negative dictionary are removed from the document and query word lists;

c) one of two types of suffix cut-off procedures is used to reduce the word lists to word stem lists: the suffix 's', or word form method removes only final 's' endings, and the regular word stem process cuts off all normal suffixes to produce word stems;

d) weights are assigned to the word stems based on the frequency of occurrence of the stems in the document abstracts or query formulations;

e) the resulting weighted word stem vectors representing documents and queries respectively are compared, and a correlation coefficient is computed for each query-document pair reflecting the similarity between the corresponding vectors;

f) document citations are presented to the user in decreasing order of the correlation coefficients.

It was seen in Table 1 that the retrieval effectiveness produced by such a system comes to within ten or fifteen percent of a conventional manual indexing system, assuming that the cluster search process and output ranking systems are properly used.

In order to obtain output that is completely comparable in effectiveness to systems using intellectual indexing, it is necessary to introduce some word control in the form of dictionaries or control lists. Two of the more useful tools in this connection are first the word discrimination list which identifies any term determined to be a common identifier, or nondiscriminator, and second, the thesaurus which specifies a grouping of thesaurus entries into affinity groups to provide synonym recognition. [43]

Both dictionaries can be automatically constructed by using the standard word stem vectors, representing the documents of a collection. A typical set of sample vectors is shown in matrix form in Fig. 1. Using the matrix representation of Fig. 1, it is possible to compute a similarity coefficient between any pair of documents by comparing the corresponding rows of the matrix. This process is, in fact, used as a basis for the document clustering process previously mentioned.

The same process is also used to generate the automatic term discrimination list. Specifically, a nondiscriminator is defined as a term which increases the interdocument similarity (that is, which increases the average similarity coefficient between pairs of rows in the matrix) when the corresponding term is incorporated in the normal document vectors; the contrary is true for discriminators, which reduce the interdocument similarity when used for indexing purposes. By computing for each term a document space density function, or equivalently, some measure of average document-pair similarity, it is easy to obtain term discrimination values that can be used as criteria for judging the goodness of each term for purposes of content representation. [43, 51, 52]

To construct an automatic thesaurus, it is necessary to group those terms judged to be valuable as content indicators into affinity classes. Clearly, the same clustering methods that generate document classes from the document-pair similarity coefficients can construct term classes by using term-pair similarities. Instead of comparing the rows of the matrix of Fig. 1 to judge the similarity among documents, one uses pairs of columns of the matrix to obtain the term similarities leading to the thesaurus classes.

In the dynamic retrieval environment to be described later, information obtained from the user population during the processing will eventually be reflected in changes in certain document vectors. As the rows of the matrix of Fig. 1 are changed, so will the columns be altered. Thus, a change in document indexing leads to a similar change in the thesaurus organization.

A number of experiments were recently conducted to evaluate the operations of such a <u>dynamic thesaurus</u> and to determine how the thesaurus classes would change with time. [53,54] In particular, regular feedback searches to be described later were conducted using about 120 original user queries. The feedback operations were then used to modify the document-term matrix, and this in turn was utilized to produce new thesaurus classes. Thirty new user queries, not previously utilized in the feedback operations, were then processed both with the original and the final thesauruses, and a performance improvement of about four or five percent in recall and precision resulted for the new updated thesaurus. [53]

No fundamental difficulty appears to exist about the utilization of changes in vocabulary usage and in user interest to effect automatic alterations in the corresponding vocabulary control tools.

D) Feedback Searches in a Dynamic Library

The main idea behind the dynamic library file is the utilization of customer experience in improving service and updating the file organization. A principal contribution to such a dynamic system is the incorporation of user inputs in an effort to improve the formulation of search requests submitted by the user population. The preferred method of query alteration used with the SMART system is known as relevance feedback, because the queries are automatically updated based on relevance information furnished by the user about previously retrieved documents. [48,55] Specifically, the relevance feedback process assumes that an initial search is first performed for each query entering the system. A small amount of output, consisting of some of the highest scoring documents is then presented to the user, and the user is asked to identify some of these documents as being either relevant to his information need (R), or nonrelevant (S). These relevance judgments are then returned to the system, and used automatically to adjust

a) the document vector for an item identified as relevant during the feedback process is altered by adding query terms, or incrementing the weights of terms jointly present in the document and query vectors; on the other hand, document terms absent from the query are decreased in importance by being assigned a lower weight;

b) similarly, for documents identified as nonrelevant, the document terms jointly included in the document and query vectors are reduced in weight, while document terms absent from the query are increased in weight.

In both cases, the assumption is that the query vectors represent active subject areas. Thus documents moved closer to the active queries are "promoted" by being placed closer to the centers of interest. The reverse is true for documents removed from the main query areas.

The procedure was tested by first using a set of 125 user queries to modify the given document space. A new set of 30 queries was then processed against the original document collection (prior to document vector modification), and also against the final, modified space, following the processing of the earlier 125 queries. The evaluation output shows that the thirty new users profited from the earlier user interaction, since the retrieval results with the new, modified space are improved by 3 percent in normalized recall, and 8 percent in normalized precision* over the results obtained with the original space. [57]

---

*Normalized recall and precision are global measures derived from the standard recall and precision parameters. [55]

In a practical library environment, the document space modification process may be considered to be a permanent feature of the file system: if the user population is fairly homogeneous, an equilibrium condition will be attained rapidly with most relevant items clustering around the corresponding query areas. As user interests shift, or new subjects become popular, the document organization will slowly change to take care of the new conditions.

Obviously, when a large number of document vectors change as a function of incoming user queries, the corresponding document class (centroid) vectors will in time become ineffective as a representation of the altered documents. Several possible strategies can be used to update document centroids, including the addition of new and the deletion of old terms from the centroids, or weight alterations in the existing terms. Eventually, as the number and magnitude of the document space alterations increases, it will become necessary to allow for a shifting of documents from cluster to cluster. Such a reclustering operation is envisaged in the next section.

E) Library Growth and Retirement*

The document vector modifications discussed up to now are performed in response to the normal query processing, the intention being to keep the collection updated with changes in user interests. There are, however, even more important reasons for insisting on a

* Some of the experiments examined in this section are described in more detail in reference [47].

dynamic catalog environment, and these have to do with collection growth and retirement. For the most part, library and information center personnel are aware of the severity of the problems created by normal collection growth, and by the lack of viable retirement policies for journal articles and monographs. Unfortunately, the state-of-the-art is such that the inevitable response of library administrations faced with accelerating document growth rates is an insistence on additional buildings, funds and personnel.

In the document environment stipulated in the present study, it is possible to institute document retirement policies based on actual experiences with the local user populations. Furthermore, new documents introduced into existing collections can be accommodated within the framework of the existing operations. Both the collection growth and retirement operations affect initially only the library files, by keeping the catalog current, and removing to a secondary file obsolete items that clutter up the catalog and impede the search operations. Subsequently, the addition and/or deletion of catalog items may improve also the physical arrangement of the library shelves.

Consider first the question of collection growth. In the clustered environment, the principal question to be resolved concerns the manner of updating the clusters when new documents are introduced. The illustration of Fig. 2 shows two typical clusters on the left side of the figure. If new documents are simply added to the cluster identified by the closest centroid vectors, a situation similar to that illustrated in the center of Fig. 2 results after some time.

Eventually, it becomes necessary to generate completely new centroid vectors if an accurate representation of the collection is wanted. The results of the required reclustering operation are shown for the sample document space on the right-hand side of Fig. 2.

In principle, it is desirable to reorganize the file as often as possible. In practice, the work involved in reclustering a sizeable file is considerable and alternative methods must be used whenever possible. Murray [49] finds that the normalized recall and precision figures decrease by about 4 percent for a fifty percent updating range (that is, when about fifty percent of the document vectors in the collection are modified once or are newly introduced), and that the loss increases to 8 percent for a seventy-five percent updating rate. Murray concludes that "enough decay in the file organization occurs with 25 to 50 percent updating to warrant a complete reclustering operation; the break-even point is probably on the low side of this range." [49]

Between reclustering operations, the collection can be updated by suitable changes in the centroid vectors as new documents are introduced into the clusters. Three modes of centroid alteration appear possible:

a) new document vectors are associated with the best existing cluster (the cluster for which the centroid-document similarity is highest), but all centroid vectors remain unchanged;

b) new document vectors are associated with the existing clusters, and the centroids are changed by updating only existing centroid terms, that is, the weights of existing centroid terms may be altered, but no new terms added, thereby insuring a constant centroid dimension;

c) new documents are associated with the existing clusters, and the centroids are changed by updating the existing terms and introducing new terms taken from the documents added to the clusters.

To determine which of the centroid updating methods operates most satisfactorily, an experiment was performed by Kerchner [58] in which a collection of 700 documents in aerodynamics was augmented by a new set of 700 additional documents in the same subject area, thus accounting for a total updating rate of fifty percent for the complete collection. For each added document all three centroid updating methods were utilized, that is, no updating, updating of existing profile terms only, and updating of existing as well as addition of new terms. Furthermore, the organization of the collection was also regenerated by a complete reclustering procedure.

The output of Table 3 represents the average performance of 50 queries using each of the four document organizations. As expected for a fifty percent updating rate, the completely reclustered collection provides the best performance, and the unaltered profiles are least effective. Of the two centroid alteration methods, the one keeping a constant profile length is somewhat preferable, particularly for high recall values; it is of course also the preferred method from the viewpoint of storage space utilization.

One other collection growth policy deserves investigation, and that involves the maintenance of user query-clusters, in addition to the normal document clusters. Under this policy, incoming queries are clustered in a manner similar to the documents, each query cluster thus representing a composite user profile. New documents entering into the collection can then be matched against the existing query profiles (rather than the document profiles) before being assigned to the appropriate document clusters. Such a strategy remains to be fully investigated. [59, 60]

Possibly the most important current problem in library management — one which has been attacked with remarkably little success — is the question of document retirement. By retirement is meant not a complete loss of an item, but merely its removal from the central file system — the one searched every time — to an auxiliary storage area which may be accessed only in special circumstances.

The classical approach to collection retirement consists in introducing concepts such as the half-life of a collection, that is, the time required for obsolescence of one-half of the currently published literature [61,62,63], or the utility of a document in terms of the number of references that it can be expected to attract in its context during the remainder of its existence [64,65]. In each case, a short half-life, or a low utility factor implies rapid obsolescence, and therefore potential retirement of an item. In order to translate concepts such as these into an active retirement policy, measurements must be made to be translated into an obsolescence threshold, and these measurements often cover one or more of the following forms:

a count of the number of citations to a given item in the literature which post-date the item in question (e.g. the number of bibliographic references in later journal papers); the usage of a given item as measured by the number of times it is removed from the library shelf, or the length of the borrowing period when the item circulates; or the age of the item, that is, the number of years since it was first published. In each case, a low citation count, a small rate of usage, and high age lead to a high obsolescence figure.

Unfortunately, the approach based on measurements is difficult to implement, for practical purposes, because accurate values of these parameters (except for the age factor) are basically unobtainable. First of all, each measure depends on a specific library or user environment, and the calculated values cannot be translated into other environments. Second, it is not clear what periods of observation, and sample sizes are required in order to obtain reliable measurements. [66] Finally, for the technical literature, at least, it is important to distinguish the general scientific usage of a document from its historical usage (for survey and other retrospective purposes). Obviously, the rates of obsolescence are different for the two cases. In summary, a retirement policy based on measures which cannot accurately be determined is not likely to be successful.

A policy is therefore proposed which directly uses the dynamic document environment. Specifically, a generalized document vector modification policy is suggested based on the following three factors:

a) the closeness of a given document to the set of query centroids, measured by the size of the similarity coefficients between query centroids and the documents;

b) the rank of a given document in the list of retrieved items, whenever it is retrieved in response to a given user query;

c) the user response whenever a given document is retrieved, that is, the judgment which the user renders concerning its potential utility to his information needs.*

The idea is that documents located close to the centers of user interest (close to existing query sets), or retrieved with a low rank in response to a query (say in the top 50 documents), or known to be relevant to the users' needs, should be promoted by being shifted closer to the respective queries where user interest is concentrated; at the same time, documents which are far removed from the query sets, or which are retrieved with very high rank (in the bottom 50 for example), or which are known to be nonrelevant to the users' needs, are demoted by being shifted away from the current query positions. If such a policy is implemented correctly, it is clear that items which are never wanted, or which are always near the bottom of the retrieved list, will be shifted toward the periphery, away from the active part of the file, and will eventually become irretrievable. Simultaneously, documents which are promoted will become more easily retrievable in the future if new queries, similar to the currently active ones should be received.

A number of problems will be met in such a dynamic retirement environment. First, it is important to pick adjustment parameters for

---

* Such a policy judges not only document obsolescence but also document quality, because negative user responses may imply that the documents are useless rather than obsolete.

the term weights that are large enough to make themselves felt eventually, but not so high as to cause violent perturbations in the document space. In particular, it is likely that different adjustment rates must be used for items about which something specific is known, for example in the form of a user relevance judgment, and for those which merely happen to be retrieved with high or low document ranks. This leads to the fast and slow increment functions which have been used successfully to update user profiles in certain information dissemination systems. [67] Second, when documents are demoted, that is, when the shift is away from a given item (rather than toward a specified area), it is necessary to take special precautions to prevent the "disappearance" of a vector, that is, the reduction of all term weights to zero. [68].

An experiment was recently performed in which the candidates for retrieval were chosen from among the following sets [69]:

a) documents that are retrievable — that is, that exhibit a sufficiently high query-document correlation — but that are repeatedly judged to be nonrelevant;

b) documents that are not retrievable because the query-document correlation is too low;

c) document <u>clusters</u> that are never expanded, the corresponding documents being never matched against any query because the query-centroid correlations are too low.

The retirement policy actually implemented consisted of an upweighting of relevant document vectors retrieved above a given rank r , up to a maximum value of the average term weight beyond which

a) the closeness of a given document to the set of query centroids, measured by the size of the similarity coefficients between query centroids and the documents;

b) the rank of a given document in the list of retrieved items, whenever it is retrieved in response to a given user query;

c) the user response whenever a given document is retrieved, that is, the judgment which the user renders concerning its potential utility to his information needs.*

The idea is that documents located close to the centers of user interest (close to existing query sets), or retrieved with a low rank in response to a query (say in the top 50 documents), or known to be relevant to the users' needs, should be promoted by being shifted closer to the respective queries where user interest is concentrated; at the same time, documents which are far removed from the query sets, or which are retrieved with very high rank (in the bottom 50 for example), or which are known to be nonrelevant to the users' needs, are demoted by being shifted away from the current query positions. If such a policy is implemented correctly, it is clear that items which are never wanted, or which are always near the bottom of the retrieved list, will be shifted toward the periphery, away from the active part of the file, and will eventually become irretrievable. Simultaneously, documents which are promoted will become more easily retrievable in the future if new queries, similar to the currently active ones should be received.

A number of problems will be met in such a dynamic retirement environment. First, it is important to pick adjustment parameters for

---

* Such a policy judges not only document obsolescence but also document quality, because negative user responses may imply that the documents are useless rather than obsolete.

the term weights that are large enough to make themselves felt eventually, but not so high as to cause violent perturbations in the document space. In particular, it is likely that different adjustment rates must be used for items about which something specific is known, for example in the form of a user relevance judgment, and for those which merely happen to be retrieved with high or low document ranks. This leads to the fast and slow increment functions which have been used successfully to update user profiles in certain information dissemination systems. [67] Second, when documents are demoted, that is, when the shift is away from a given item (rather than toward a specified area), it is necessary to take special precautions to prevent the "disappearance" of a vector, that is, the reduction of all term weights to zero. [68].

An experiment was recently performed in which the candidates for retrieval were chosen from among the following sets [69]:

a) documents that are retrievable — that is, that exhibit a sufficiently high query-document correlation — but that are repeatedly judged to be nonrelevant;

b) documents that are not retrievable because the query-document correlation is too low;

c) document <u>clusters</u> that are never expanded, the corresponding documents being never matched against any query because the query-centroid correlations are too low.

The retirement policy actually implemented consisted of an upweighting of relevant document vectors retrieved above a given rank r, up to a maximum value of the average term weight beyond which no further upweighting would occur; a downweighting of nonrelevant documents retrieved above rank n , down to a minimum value of the average document term weight, below which the document would be retrieved; a downweighting of the b bottom documents in the list ranked according to query-document correlation, down to a minimum value of the average document-term weight below which the document would be retired; and finally, a downweighting of all document centroids with a query-centroid correlation less than a threshold value, down to a minimum value of the average centroid term weight below which the centroid (and all included documents) would be retired. A simplified flowchart of the retirement policy is shown in Fig. 3.

A set of 155 queries was used together with 424 clustered documents in aerodynamics. The particular strategy chosen for the test led to the retirement after 155 user responses of 24 documents — about 6 percent of the collection. The recall-precision output of Fig. 4 shows that the reduced collection gives somewhat better results because of the downweighting of unwanted items, and the retirement of the useless ones.

Obviously more extensive experimentation is needed in the area of document growth and retirement. The rate of retirement can be altered by parameter variations, and the updating method of document and centroid vectors can be modified — for example by shifting close to, or away from, the queries instead of multiplying the term weights by a constant parameter.

Still, a retirement policy appears feasible which is self-correcting in the sense that user input forms the basis for decisions concerning document promotions or demotions. As user opinions change, items that are suddenly needed, but have been pushed out of the way can be rescued, or contrariwise, items found near the top can eventually be discarded. Moreover, the policy is implemented without computing or maintaining statistical data concerning age or utility of each document, so that the entire storage space gained through retirement can be utilized to accommodate collection growth. Finally, the retirement strategy produces gains in retrieval performance.

Eventually, appropriate cost studies will have to be made for the dynamic environment. Any dynamic file process will obviously require some apparatus (such as document vector modifications) not needed for static file management. On the other hand, the costs presently inherent in the maintenance of ever-growing files are large, and the trained personnel, storage space, and new buildings are becoming scarcer. A self-monitoring, dynamic environment such as the one proposed may then be easier to justify than would appear, both economically and technically.

5. Conclusions

It may be time to discard the old assumptions and priorities in the library mechanization area. The static environment in which an item is cataloged for eternity and the user population contributes little to the efficiency of the procedures is not conducive to effective library management. Neither are the one-shot operations in which every organization goes off on its own, each one in a somewhat different direction. No amount of additional land, buildings, space, people, or money will solve the library problem, but a combination of new ideas and the rejection of old prejudices may.

In the present study, a new dynamic library file environment is proposed. It emphasizes procedures oriented toward the customer population, and generates for each organization a catalog organization optimized for its own user group. It does not require new technologies not now available, nor is it based on a replacement of the current information carriers (books and journals) by alternative, unproven products.

The dynamic environment should appeal to library users because services will be far more effective; it should also help librarians overcome the current operational crisis conditions and reduce the growing budget problems. It may constitute the best hope for a viable library system for the future.

References

[1] E. Mason, Along the Academic Way, Library Journal, May 15, 1971, p. 1671-76.

[2] Libraries and Information Technology — A National System Challenge, National Academy of Sciences, Washington, 1972.

[3] B.E. Markuson, An Overview of Library Systems and Automation, Datamation, February 1970.

[4] R.H. Parker, Library Automation, in Annual Review of Information Science and Technology, Vol. 5, C. Cuadra, editor, Encyclopedia Britannica, Chicago, 1970.

[5] G. Salton, On the Development of Libraries and Information Centers, Library Journal, October 15, 1970, p. 3433-3442.

[6] R.A. Wall, Computer Applications in Libraries, Int. Journal of Math. Educ. in Science and Technology, Vol. 2, 1971, p. 51-74.

[7] R.W. Alexander and R.W. Harvey, An Overview of the Experimental Library Management System (ELMS), Laboratory Report 16.197, IBM-ASDD Division, Los Gatos, California, May 1970.

[8] T. Burgess and L. Ames, Lola — Library On-Line Acquisitions Subsystem, Washington State University Library, Pullman, Washington, July 1968.

[9] F.G. Spigai, and T. Mahan, On-Line Acquisitions by Lolita, Journal of Library Automation, Vol. 3, No. 4, December 1970, p. 276-294.

[10] R.A. Kennedy, Bell Laboratories Library Real-Time Loan System (Bellrel), Journal of Library Automation, Vol. 4, No. 2, June 1968, p. 128-146.

[11] J.S. Aagaard, An Interactive Computer-based Circulation System: Design and Development, Journal of Library Automation, Vol. 5, No. 1, March 1972, p. 3-11.

[12] G.D. Guthrie, An On-Line Remote Catalog Access and Circulation Control System, Proc. ASIS National Meeting, Washington 1971, p. 305-309.

[13] J.L. Bennett, On Line Access to Information: NSF as an Aid to the Indexer/Cataloger, American Documentation, Vol. 20, No. 3, July 1969, p. 213-220.

[14] F.G. Kilgour, Concept of an On-Line Computerized Library Catalog, Journal of Library Automation, Vol. 3, No. 1, March 1970, p. 1-11.

[15] C.T. Payne and R.S. McGee, The University of Chicago Bibliographic Data Processing System, Reports PB 195836 and PB 204205, University of Chicago, October 1970 and April 1971.

[16] A.H. Epstein and A.B. Veaner, A User's View of Ballots, Stanford University, June 1972.

[17] P.J. Fasana and A. Veaner, editors, Collaborative Library Systems Development, MIT Press, 1971.

[18] D.C. Weber and F.C. Lyndon, Survey of Interlibrary Cooperation, in Interlibrary Communications and Information Networks, J. Becker, editor, Am. Library Association, Chicago 1971, p. 69-81.

[19] A.E. Jeffreys, Centralized Services for University Librarians, in Interface, C.K. Balinforth and N.S.M. Cox, editors, Oriel Press, 1971, p. 115-125.

[20] R. McGee and R.C. Miller, Marc Utilization in the University of Chicago Library Bibliographic Data Processing System, Eighth Annual Clinic on Library Applications of Data Processing, University of Illinois, April 1970.

[21] R. Bregzis, The University of Toronto Marc Pilot Project, in Organization and Handling of Bibliographic Records by Computer, N.S.M. Cox and M.W. Grose, editors, Oriel Press Ltd., Newcastle, 1968.

[22] F.G. Kilgour, Computerized Cooperation, presented at Annual Meeting of the Association of Research Libraries, Columbus, May 1972.

[23] F.B. Ludington, Hampshire Inter-Library Center, ALA Bulletin, Vol. 46, January 1952, p. 10-11.

[24] H.J. Harrar, Cooperative Storage Warehouses, Ph.D. Thesis, Rutgers University, June 1962.

[25] J.G. Kemeny, A Library for 2000 AD, in Management and the Computer of the Future, M. Greenberger, editor, MIT Press and John Wiley and Sons, 1962, p. 134-179.

[26] J.C.R. Licklider, Libraries of the Future, MIT Press, 1965.

[27] C.F.J. Overhage, Project Intrex Planning Conference, Summary of Report, Woods Hole, August 1965.

[28] F.W. Lancaster, The Cost Effectiveness Analysis of Information Retrieval and Dissemination Systems, JASIS, Jan. Feb. 1971, p. 22-27.

[29] F.W. Lancaster, Information Retrieval Systems — Characteristics, Testing and Evaluation, John Wiley and Sons, 1968, Chapter 13.

[30] S. Allen, S. Green, J. Friedman, B.E. Harrington and L.R. Johnson, Information Storage and Retrieval by Computer — An Emerging Problem for the Law of Copyright, in Reader in Library Services and the Computer, L. Kaplan, editor, Microcard Editions, 1971, p. 149-168.

[31] 1967 ABA Copyright Symposium, Computers and Copyright: The New Technology and Revision of the Old Law, Bulletin of the Copyright Society of America, Vol. 15, No. 1, Oct. 1967.

[32] R.P. Bigelow, The Computer as a Copyright Infringer, Law and Computer Technology, Vol. 1, No. 7, May 1968, p. 2-5.

[33] R.W. Conway, W.L. Maxwell, and H.L. Morgan, On the Implementation of Security Measures in Information Systems, Communications of the ACM, Vol. 15, No. 4, April 1972, p. 211-220.

[34] R.M. Graham, Protection in an Information Processing Utility, Communications of the ACM, Vol. 11, No. 5, May 1968.

[35] L.J. Hoffman, Computers and Privacy — A Survey, Computing Surveys, Vol. 1, No. 2, June 1969, p. 85-103.

[36] D.N. Michael, Speculations on the Relation of the Computer to Individual Freedom and the Right to Privacy, George Washington Law Review, Vol. 33, 1964, p. 270-286.

[37] A. Westin, Privacy and Freedom, Atheneum, New York, 1967.

[38] W.N. Locke, Computer Costs for Large Libraries, Datamation, February 1970.

[39] D.R. Swanson, Searching Natural Language Text by Computer, Science, Vol. 132, 21 October 1960, p. 1099-1104.

[40] C.W. Cleverdon and E.M. Keen, Factors Determining the Performance of Indexing Systems, Aslib-Cranfield Research Project Reports, Vol. 1 and 2, Cranfield (England), 1966.

[41] G. Salton and M.E. Lesk, Computer Evaluation of Indexing and Text Processing, Journal of the Association for Computing Machinery, Vol. 15, No. 1, January 1968, p. 8-36.

[42] G. Salton, Automatic Text Analysis, Science, Vol. 168, 17 April 1970, p. 335-343.

[43] G. Salton, A New Comparison Between Conventional Indexing (Medlars) and Text Processing (SMART), Journal of the ASIS, Vol. 23, No. 2, March-April 1972, p. 75-84.

[44] E.M. Keen, Prospects for Classification Suggested by Evaluation Tests Carried Out 1957-1970, in Classification in the 1970's, A. Maltby, editor, Linnet Books and Clive Bingley, 1972, p. 193-210.

[45] F.W. Lancaster, Evaluation of On-line Searching in Medlars (AIM-TWX) by Biomedical Practitioners, Report No. 101, University of Illinois Graduate School of Library Science, February 1972.

[46] R.M. Hayes, The Concept of an On-Line, Total Library System, Library Technology Reports, American Library Association, May 1965.

[47] G. Salton, Dynamic Document Processing, Communications of the ACM, Vol. 15, No. 7, July 1972.

[48] G. Salton, The SMART System — Experiments in Automatic Document Processing, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1971, Chapters 10-13.

[49] D.M. Murray, Document Retrieval Based on Clustered Files, Doctoral Thesis, Cornell University, Scientific Report No. ISR-20, Dept. of Computer Science, June 1972.

[50] D.B. Johnson and J.M. Lafuente, A Controlled Single-Pass Classification Algorithm with Application to Multilevel Clustering, Scientific Report No. ISR-18, Section XII, Department of Computer Science, Cornell University, October 1970.

[51] K. Bonwit and J. Aste Tonsman, Negative Dictionaries, Scientific Report No. ISR-18, Section 6, Cornell University, Dept. of Computer Science, October 1970.

[52] G. Salton, Experiments in Automatic Thesaurus Construction for Information Retrieval, Proc. IFIP Congress-71, North Holland Publishing Company, Amsterdam 1972.

[53] K. Hamilton and J. Leslie, Thesaurus Class Modification Based on User Feedback, Scientific Report No. ISR-21, Dept. of Computer Science, Cornell University, 1972.

[54] A. Wong, R. Peck and A. van der Meulen, An Adaptive Dictionary in a Feedback Environment, Scientific Report No. ISR-21, Dept. of Computer Science, Cornell University, 1972.

[55] G. Salton, Automatic Information Organization and Retrieval, McGraw Hill Book Co., New York, 1968.

[56] G. Salton, The Performance of Interactive Information Retrieval, Information Processing Letters, Vol. 1, 1971, p. 35-41.

[57] T.L. Brauen, Document Vector Modification, in The SMART Retrieval System — Experiments in Automatic Document Processing, G. Salton, editor, Chapter 24, Prentice Hall Inc., Englewood Cliffs, N.J. 1971, p. 456-484.

[58] M.D. Kerchner, Dynamic Document Processing in Clustered Collections, Doctoral Thesis, Scientific Report No. ISR-19, Dept. of Computer Science, Cornell University, October 1971.

[59] V.R. Lesser, A Modified Two-Level Search Algorithm Using Request Clustering, Scientific Report No. ISR-11, Section VII, Dept. of Computer Science, Cornell University, June 1966.

[60] S. Worona, Query Clustering in a Large Document Space, in the SMART Retrieval System — Experiments in Automatic Document Processing, G. Salton, editor, Chapter 13, Prentice Hall Inc., Englewood Cliffs, N.J., p. 298-310.

[61] R.E. Burton and R.W. Kebler, The Half-Life of Some Scientific and Technical Literatures, American Documentation, Vol. 11, 1960, p. 18-22.

[62] A. Sandison, The Use of Older Literature and its Obsolescence, Journal of Documentation, Vol. 27, No. 3, Sept. 1971, p. 167-183.

[63] R.S. Grant, Predicting the Need for Multiple Copies of Books, Journal of Library Automation, Vol. 4, No. 2, June 1971.

[64] B.C. Brookes, Obsolescence of Special Library Periodicals: Sampling Errors and Utility Contours, Journal of the ASIS, Sept.-Oct. 1970, p. 320-329.

[65] B.C. Brookes, The Growth, Utility and Obsolescence of Scientific Periodical Literature, Journal of Documentation, Vol. 26, No. 4, December 1970, p. 283-294.

[66] A.K. Jain, Sampling In-Library Book Use, Journal of the ASIS, Vol. 2, No. 3, May-June 1972, p. 150-155.

[67] C.R. Sage, R.R. Anderson, and D.R. Fitzwater, Adaptive Information Dissemination, American Documentation, Vol. 16, No. 3, July 1965, p. 185-200.

[68] J. Kelly, Negative Response Relevance Feedback, in The SMART Retrieval System — Experiments in Automatic Document Processing, G. Salton, editor, Chapter 20, Prentice Hall Inc., Englewood Cliffs, N.J. 1971, p. 403-411.

[69] K.C. Tai and C.S. Yang, Document Retirement in Automatic Retrieval Systems, Scientific Report No. ISR-21, Dept. of Computer Science, Cornell University, 1972.

Analysis Method	Recall	Precision
Medlars (controlled terms assigned by trained indexers)	0.3117	0.6110
SMART (word stems extracted from abstracts; simulation of Boolean search)	0.1814 (-42%)	0.4141 (-32%)
SMART (word stems using ranking feature, and displaying documents in decreasing query-document order)	0.2622 (-16%)	0.4901 (-19%)
SMART (thesaurus consisting of groups of related terms, and using ranking feature)	0.3232 (+ 4%)	0.6106 (0%)

SMART-Medlars Comparison Showing
Importance of Document Ranking

(450 documents, 29 queries; adapted from [43])

Table 1

Analysis Method	Recall	Precision
Medlars (controlled terms)	0.3117	0.6110
SMART word stem		
0 - initial search	0.2622 (-16%)	0.4901 (-19%)
1 - iteration feedback	0.3235 (+ 4%)	0.6385 (+ 5%)
2 - iteration feedback	0.3433 (+10%)	0.6892 (+13%)
SMART thesaurus		
0 - initial search	0.3232 (+ 4%)	0.6106 (0%)
1 - iteration feedback	0.3915 (+25%)	0.7427 (+18%)
2 - iteration feedback	0.4029 (+29%)	0.7438 (+22%)

SMART - Medlars Comparison Using Feedback Feature

(450 documents, 29 queries; from reference [43])

Table 2

Precision / Recall	Original Space	Centroid Mod. 1 (updating of existing terms)	Centroid Mod. 2 (updating and addition of new terms)	Reclustered Space (addition of 700 documents
0.1	0.3981	0.4234 (+ 6%)	0.4296 (+ 8%)	0.4313 (+ 8%
0.3	0.2190	0.2432 (+11%)	0.2460 (+12%)	0.2631 (+20%
0.5	0.1390	0.1547 (+11%)	0.1526 (+10%)	0.1408 (+ 1%
0.7	0.0422	0.0458 (+ 9%)	0.0444 (+ 5%)	0.0717 (+70%
0.9	0.0282	0.0287 (+ 2%)	0.0285 (+ 1%)	0.0388 (+38%

Document Profile Updating following Addition of New Items

(1400 documents, 50 queries; adapted from [58])

Table 3

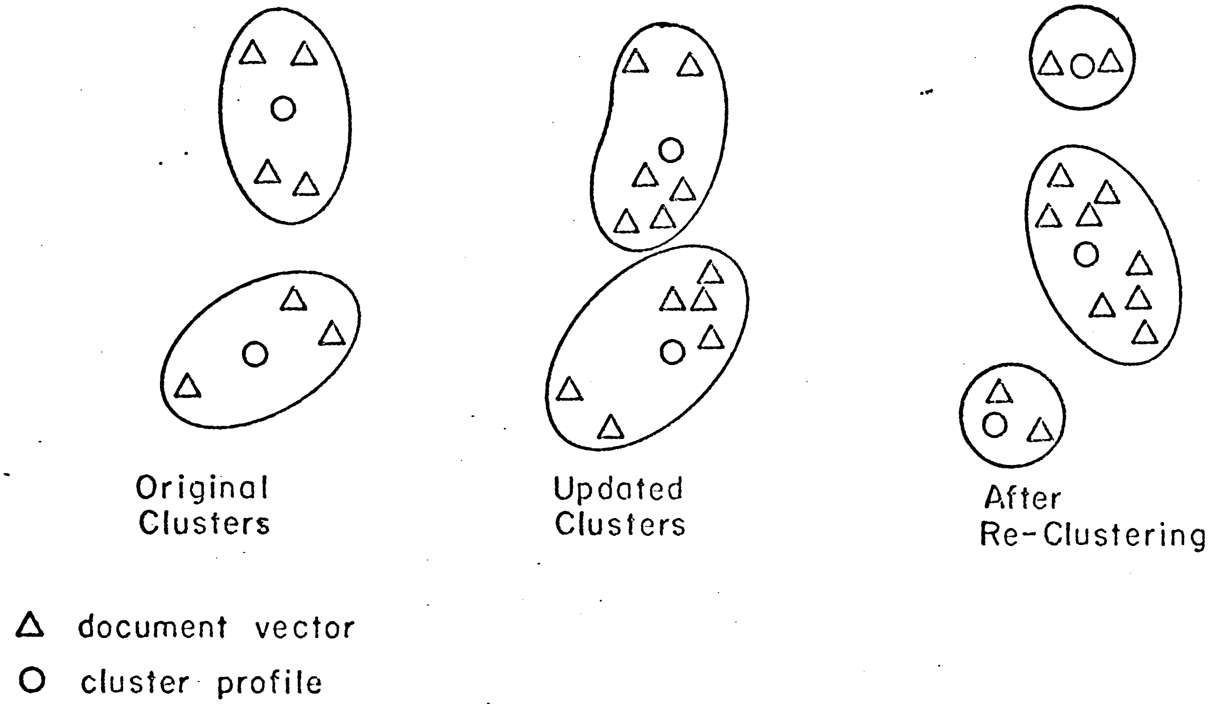
$$
\begin{array}{l|c}
 & \text{Term 1} \quad \text{Term 2} \qquad\qquad \text{Term k} \\
\hline
\begin{array}{l} \text{Doc. 1} \\ \text{Doc. 2} \\ \\ \text{Doc. n} \end{array} &
\left\{\begin{array}{llll}
t_1^1, w_1^1; & t_2^1, w_2^1; & \ldots; & t_k^1, w_k^1 \\
t_1^2, w_1^2; & t_2^2, w_2^2; & \ldots; & t_k^2, w_k^2 \\
 & & & \\
t_1^n, w_1^n; & t_2^n, w_2^n; & \ldots; & t_k^n, w_k^n
\end{array}\right\}
\end{array}
$$

Typical Term-Document Matrix

Fig. 1

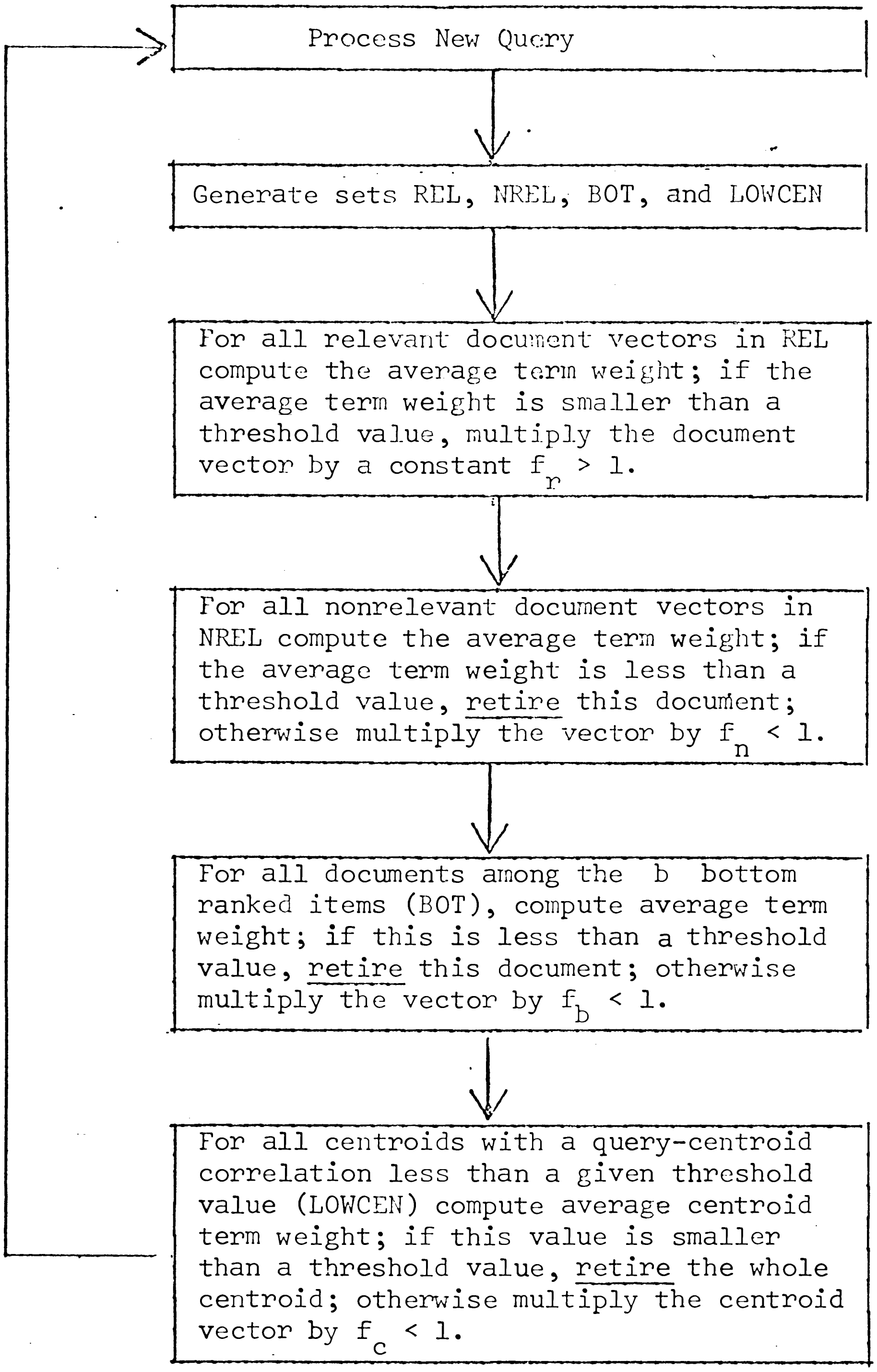
t_j^i = term j in document i

w_j^i = weight of term j in document i



Cluster Updating with Profile Alteration
(new document addition)

Figure 2



Simplified Document Retirement Strategy

Fig. 3

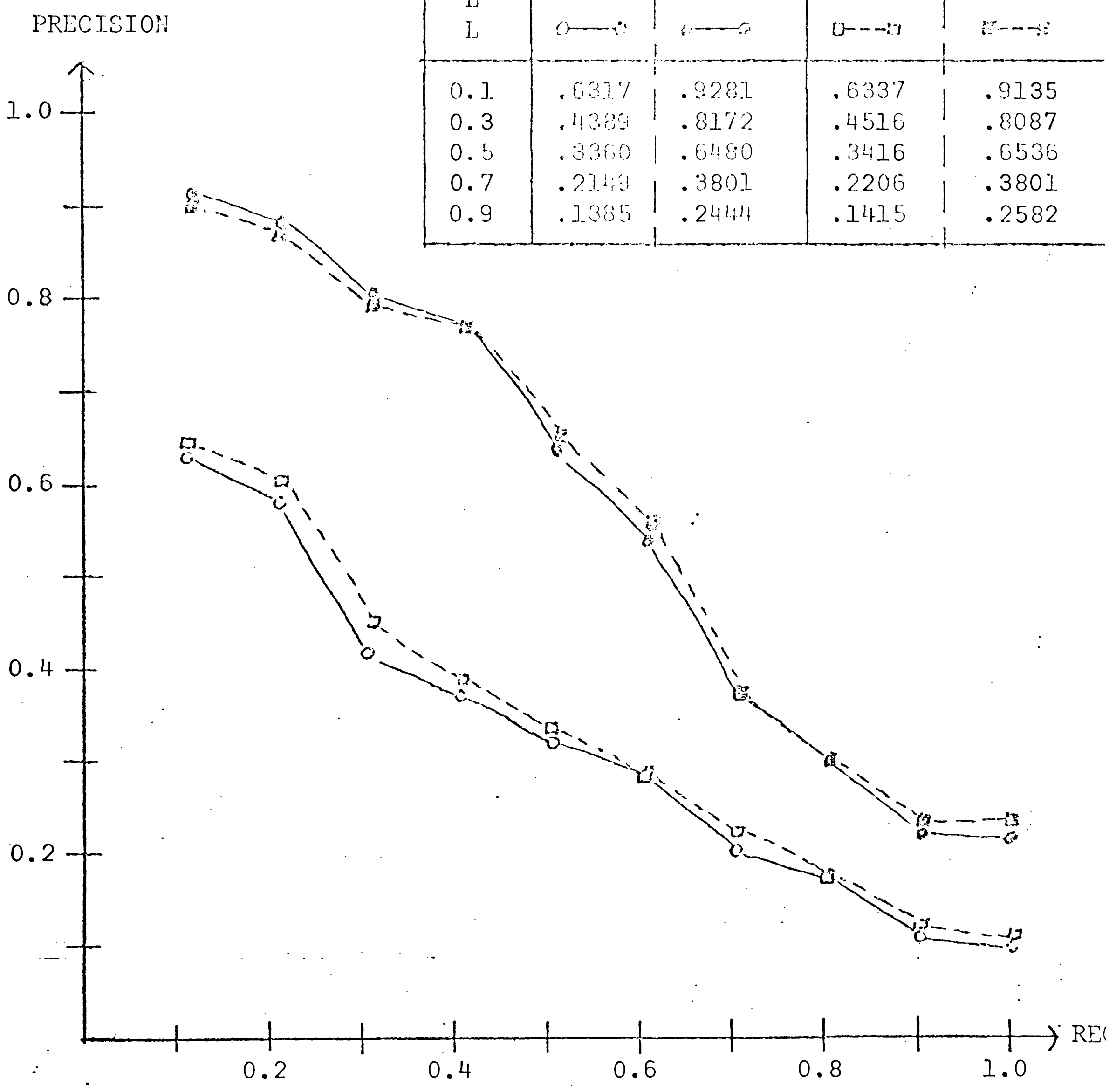
LOWCEN: subset of document centroids whose correlation with query is smaller than a threshold value

REL : subset of documents judged relevant retrieved above rank r

NREL : subset of documents judged not relevant retrieved above rank n

BOT : subset of documents including the b bottom ranked items

RECALL	Original Items (424 Docs) Precision Run 0	Original Items (424 Docs) Precision Run 1	Modified After Retirement (400) Precision Run 0	Modified After Retirement (400) Precision Run 1
0.1	.6317	.9281	.6337	.9135
0.3	.4389	.8172	.4516	.8087
0.5	.3360	.6480	.3416	.6536
0.7	.2149	.3801	.2206	.3801
0.9	.1385	.2444	.1415	.2582



Recall-Precision Comparison of Original Collection (424 documents) and Modified Items following Retirement (400 documents)

Fig. 4

Run 0: Original Search

Run 1: First Feedback Search