Toward a Dynamic Library

The library world has been forced to exist under troublesome conditions for many years. The difficulties are due to a variety of causes, including constantly increasing service demands, the great variety of library material that must be processed (tapes, cards and microforms, in addition to the normal printed materials) and, most important of all, the severity of the budget crisis. It is an unfortunate fact that library support levels have been shrinking at the very time when the cost of library services and materials is reaching a record high.

It was, perhaps, natural in these circumstances that library administrations should turn to the use of computing equipment as a means for coping with the increasing transactions and the cost explosion. Two main approaches were followed in the 1960s. The first, termed piecemeal mechanization, denotes the conversion of library operations to computer processing, one application at a time. Thus, one library would create an automated circulation system, while another concentrated on automating the acquisitions process.

It became clear very early that the piecemeal mechanization approach was fraught with difficulty. The files to be processed were often very large and subject to continual changes and updates, and a great many different processes had to be considered. An additional problem was the desire to maintain real-time control over all library items, that is, the status and whereabouts of each library item were to be ascertainable at any time. It is easy to understand, therefore, that a conversion to an automated processing system would not be simple and straightforward. This situation produced substantial disenchantment with the use of computers in the library, and some observers even claimed that computing equipment could not viably be incorporated into a library environment.1
TOWARD A DYNAMIC LIBRARY

The second main effort in library automation in the 1960s produced a number of prototypes for integrated library management systems which dealt with the complete library operation as a unit. Several complete management systems were designed by IBM, University of Chicago, Stanford University and others; some of these systems are most impressive in their conceptions and relatively easy to use. However, the recent trend in the direction of cooperative ventures among libraries has somewhat dampened the enthusiasm for the integrated stand-alone systems, and the feeling now seems to be that they are too costly to be supported by single library organizations without substantial outside aid.2

Presently, the library crisis remains undiminished; indeed, the budget situation may be more unfavorable now than it was ten or fifteen years ago. However, two fundamental changes have occurred in recent years. First, many librarians feel that the technical library processing tasks are becoming too big and too costly to be borne by an individual library. As a result, the opposition to the formation of compacts between libraries and library networks has substantially decreased, and some library administrators are reconciled to a small loss of autonomy in return for the benefits obtainable from cooperative endeavors.

In addition, a number of advances have taken place in the computer art that may be of substantial benefit in library processing. First, the increased storage capacity of the modern computing equipment has considerably simplified the processing of large library files. Recently, console terminals have also been developed which provide a friendly environment for user/system interaction, and these on-line systems have proved to be not only commercially successful, but essential for many types of applications, such as airline reservations, banking transactions, point-of-sale terminal processing, and so on. Later in this paper it will be shown that interactive processing methods can provide great benefits in library application. Finally, there has been progress in the design of computer networks and in the use of "distributive" computing, in which a process is separated into several pieces to be handled by different computers with interaction between processors to insure that the final product conforms to the initial specifications.

The combination of large memory capacities, intelligent front-end devices for user/system interaction, and distributive computing methodologies have changed the outlook for the mechanization of library processes. Accordingly, the current plans for the design of the library of the future differ from the earlier versions. The piecemeal mechanization efforts and the integration of library processes into a single management system are being replaced by the construction of cooperative library networks and by tentative plans toward a paperless library system which would operate at some future time in a totally new environment.
The main considerations in the design of library networks and paperless library systems are outlined in the next section. Some concepts are then introduced which may be utilized in the implementation of an alternative, so-called dynamic library. Finally, a number of specific processes are examined which may be incorporated into the proposed dynamic library system of the future.  

LIBRARY COOPERATION AND THE PAPERLESS FUTURE LIBRARY

Library Networks

It is generally recognized that a great deal of similarity exists among the technical processing cycles that form the basis for monograph and serials processing in libraries of comparable size and scope. Thus, the basic book-ordering, bill-paying and acquisitions operations are more or less similar; so is the cataloging process, and—possibly with somewhat greater variation—the circulation system. The basic tool for all these operations is a comprehensive library catalog which includes descriptive information for all library items. Such a catalog is consulted during the acquisitions process to determine whether an item on order may already exist in the collections. Furthermore, when a catalog description exists for a given item, it serves again for the creation of a more complete record during cataloging. Finally, the catalog entry is used repeatedly for charging and discharging materials during the circulation process.

From these basic facts it follows that a comprehensive machine-readable library catalog accessible from a variety of geographic locations would be useful in controlling the technical processing performed by a number of different library organizations. Appropriate console entry devices must exist in each participating library to be used to access the common machine-readable catalog. This type of machine-readable union catalog forms the basis for the well-known catalog card ordering system managed by OCLC. An automated system in which a variety of different library organizations are hooked to one or more common mechanized catalogs is known as a library network.

It is clear that a great many different network configurations are possible in principle. Normally, a single centralized catalog is used to control the operations originating in various remote libraries; alternatively, each participating library in the network could manage its own mechanized catalog in such a way that the local user population is given access not only to the local catalog, but also to the remote catalogs of other libraries in the network. Such a system makes it possible to share library
resources and to reduce the burden of technical processing for any one organization. The following possibilities are immediately apparent:

1. Technical processing costs can be saved by sharing the burden of the operations; for example, a given library item might be cataloged once and other participating organizations would use the already-established cataloging record. This kind of argument was used in creating service organizations such as OCLC.

2. A shared mechanized catalog could support more sophisticated subject-accessing procedures than a conventional manual card catalog if additional types of content identifiers or a greater variety of conventional subject indicators were included.

3. The shared catalog constitutes a resource-sharing tool in the sense that the user population can be given access to the pooled resources of a number of different libraries; the network organization could then lead to a broad system of interlibrary loan procedures and cooperation.

4. The system could be used for shared collection development if each participating library were to orient its acquisitions policy toward particular subject areas; this would save resources by avoiding multiple acquisitions of rarely used materials.

While a resource-sharing library network could certainly provide a variety of actual and potential advantages to participating organizations, substantial difficulties still exist before such networks can actually fulfill the previously mentioned promises. There are problems of a nontechnical as well as of a technical nature. The nontechnical questions relate to the differing interpretations of aims and responsibilities of participating libraries: many libraries currently maintain different standards of growth and retirement policies; there may be user groups who deserve or expect specialized services of various kinds; in addition, administrative and other constraints imposed on the participating organizations may hamper the cooperative effort. The financial arrangements among the network participants would necessarily be difficult to manage because of the fundamentally uneven standing of the component libraries. Clearly, much of the service would be rendered by the wealthier units endowed with the best collections, and the weaker units may function mostly as recipients of the services. The question then arises of who pays how much to whom.

Whenever a number of user organizations share a common set of files, questions of privacy arise because it becomes necessary to preserve data confidentiality for items with restricted circulation characteristics. Finally, the effect of a library network system on outside organizations, such as the publishing industry, must be considered. Obviously, when fewer published items are bought by libraries, and most items circulate freely with-
out royalty payment, the publishing industry, and by extension authors, are liable to suffer. Many of these social and legal problems have not yet been considered in detail. One may hope that with goodwill on everyone's part, appropriate accommodations may, in time, be found.

Among the more technical problems of library and computer network organizations are those relating to the actual technical implementation. What, for example, should be the role of minicomputers in the network? What are the comparisons of communications costs, storage costs and data accessibility when each item is stored in a single, central location as opposed to data duplication at several points in the network? What are the software and hardware requirements implied by the latter, a distributed data base design, compared with the more normal centralized data base system? What problems are likely to arise when it becomes necessary to merge different technologies, such as computing equipment, communications lines and photographic technology?

The question of formal standardization and bibliographic control may also be expected to cause grief in a network situation. The current perception on the part of many librarians is that increasingly stringent controls are necessary as one moves from the level of a single-library item to the level of a complete institution, and from there to a larger system comprising several institutions, and finally to a comprehensive library network. Their idea is that a catalog item admitted into the network must conform to specific rules of description, formatting and control, and that these rules must be standardized. Such requirements would make it possible to use standardized query and search protocols to access any item, no matter where it is located:

There is growing realization that the authority file (which specifies established forms of headings and other bibliographic descriptions) is the foundation and basic building block of the automated library system.\footnote{6}

Protocols must be carefully formulated and followed; otherwise the network users will require a variety of manuals...and need to reformulate search requests each time a network component using a nonstandard form of indexing is accessed.\footnote{7}

If these precepts are to be followed, it is clear that substantial difficulties may arise in (1) deciding about appropriate standards, (2) converting nonstandard items to the common format, (3) deciding what items to admit into the network, and (4) exercising the quality control necessary for upgrading the items.

In a mechanized library situation, it is likely, however, that storage space restrictions will be much less confining than has been customary for
a normal catalog using three-by-five-inch cards. It is not clear in these circumstances why a multiplicity of different indexing systems could not coexist quite peacefully. This possibility is examined in more detail later in this paper.

The Library of the Future

In addition to implementing plans for the construction of cooperative library systems and networking arrangements, a certain amount of attention is also given in some library circles to the role of the library in the society of the future. A number of blueprints are in existence which postulate the storage of all existing knowledge in machine-readable form. A huge, mechanized storage system would replace the normal library, and effective console-driven information retrieval protocols operating in an on-line mode would be used to locate stored items of interest to individual users. Conventional books and journals in the form of printed information on paper could be dispensed with in such a situation: "Any concept of a library that 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." 8

In fact, the replacement of the current labor-intensive, constant-productivity library setup with a remotely accessible, machine-readable data store, and the elimination of paper products exhibit substantial attractions:

1. a paperless, comprehensive machine-readable data store would eliminate the existing fragmentation of materials in a given subject over many different journals and books;
2. the large volume of material which necessarily will have to be processed and stored in the future would become much more manageable in a paperless system;
3. since the cost of standard (paper) publications is continually increasing, in part because of the labor-intensive nature of the publishing industry, substantial savings may be produced with respect to publishing in a paperless situation; and
4. the delays currently built into the standard publication system could largely be eliminated, and the dissemination of research results could be speeded by abandoning the conventional publishing chain producing paper products. 9

In a paperless society, many individuals would own personal computer terminal devices which could be used for a variety of purposes, such as maintenance of private files, composition of letters and articles, recording
of incoming messages and text and, incidentally, for library search and retrieval purposes. The role of the traditional library in such an environment is unclear; almost certainly the "library" would provide search services for users without personal on-line access. Printout services for bulky materials that could not economically be handled by the personal terminals might also be provided by a library center. Items of purely local interest might be collected and cataloged, and specialized search services could be provided for certain classes of customers.

It is not possible in the present context to go into a detailed examination of the merits and disadvantages of a completely automated, paperless communications system, or to assess the technical feasibility questions. Suffice it to say that a complete abandonment of books and journals as we know them would certainly produce inconveniences to large classes of the population: many people now use library materials in out-of-the-way places—on the beach, in bed, in buses and on airplanes—where computer access may not be immediately available. In any case, the use of computers to obtain on-line access to library materials (which may be expected in the foreseeable future) certainly does not imply the immediate elimination of printed materials. Furthermore, quite a few of the claims made in favor of the paperless library are almost certainly exaggerated: that a paperless on-line system is more "democratic" because everyone will have equal access to the vital information which is now confined to a few selected experts; that the library of the future could store items which under current conditions never make it through the publication process (as if it were advantageous to be able to access the bulk of materials that have never been subjected to quality control); that on-line communications systems would prevent the duplication of research and development efforts by making it easy to discover prior work; that a good deal of work now performed in offices and factories could be done at home using the computer console, thereby reviving the cottage industry and decreasing work alienation; and that the preparation of reports and articles from a computer console will help people improve their writing style and spelling ability.

In the next few paragraphs an alternate concept of the library of the future is outlined in which computer access plays a role in identifying pertinent materials, but where printed materials are maintained whenever possible. Whether this alternative future library has a greater chance of being implemented than the proposed paperless system remains to be seen.

CONCEPTS OF THE DYNAMIC LIBRARY

The main idea behind the dynamic library is flexibility, and the use of customer/system interaction to control library operations. The descrip-
tion of a particular library item can be made to vary with the environment in which it is placed, and also with the judgments obtained from the user population about the usefulness and importance of the item. The classification system used to organize the library collection is similarly adjustable as the interests of the users change. Thus, when particular subject areas become of special interest, a more refined classification system can be used for them than for the remainder of the collection. This implies that the search system itself can also be adjusted with changes in classification. Finally, since the retrieval system operates in an on-line mode, information about previously retrieved items can be used to adjust the original query formulations in such a way that improved output may be retrieved from the collections.

The idea, then, is to avoid the imposition of outside rules in the form of authority lists, special indexing conventions and preestablished classification systems, and to treat the library like the dynamic environment it is, where the file contents as well as the user population are subject to continuous change. To give a brief example before turning to the technical details, suppose a document ostensibly dealing with the use of computers in medicine is to be indexed. How appropriate would it be to choose the term *computers* as a subject indicator? The answer is that no one knows. If this item is placed in an environment of medical documents to be accessed by medical people, the term *computer* may be precisely right, because it will help in distinguishing this particular document from other medical items in the collection. If, on the other hand, the document winds up in a computer science collection where it is accessed by computer experts, then the term would probably be inappropriate, because all the other documents deal with computers and the special nature of the item will not be recognized.

This example suggests that no one particular content analysis or cataloging system will be adequate for all purposes, but that the subject description must depend on the collection environment and the user population. In a computer environment where console access to the stored collection is available, there is no virtue in insisting on fully controlled, static indexing, classification and search procedures. Instead, each item can be described in many different ways, and each user can access the items in accordance with his own viewpoint.

The following main characteristics are important in the design of the dynamic library.

1. The operations are software procedures which facilitate access to the collections and retrieval of useful information; there is no intention to tamper with the storage and dissemination of printed materials.
2. Machine-readable information, consisting of at least abstract-length excerpts, is used to generate content identifications for each item, and the content analysis will depend on the general collection environment within which a given item is placed.

3. The files are interrogated remotely by the user population, and the relevance assessments obtained from the users about specific items are used to improve the available content descriptions. The same relevance assessments are also used for query reformulation purposes to enable the improved queries to retrieve more relevant and fewer nonrelevant items than the original formulations.

4. Dynamic classification systems, which consist of broad, general classes of related items for use by the casual client, as well as smaller, more refined classes to serve the experts in particular subject areas, are used to organize the stored collections.

5. The eventual value of a particular term for indexing purposes, or of a particular document in retrieval is dependent on the accrued experiences of the user population with that document, and on the total collection environment.

In order to understand the dynamic operations in detail, it is necessary to introduce the concept of similarity between items. It is obvious that similarities exist between library items; library personnel use this fact to arrange related items in adjacent places on library shelves. Unfortunately, the relationships between items are not recognized operationally in conventional library environments. In the dynamic library, the computation of similarities or distances between library items, or between terms, lies at the root of the operations.

Consider a collection of documents, and assume that each document is identified by a set of terms, or content identifiers. The identifiers might be words or phrases extracted from the documents, or entries found in a thesaurus. The collection may then be represented in matrix form:

\[
\begin{pmatrix}
D_1 & d_{11} & d_{12} & \ldots & d_{1t} \\
D_2 & d_{21} & d_{22} & \ldots & d_{2t} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
D_n & d_{n1} & d_{n2} & \ldots & d_{nt} \\
Q & q_1 & q_2 & \ldots & q_t
\end{pmatrix}
\]

As shown above, each row of the matrix represents a document (D), and a given entity \(d_{ij}\) represents the weight, or value, of the \(j^{th}\) identifier attached
TOWARD A DYNAMIC LIBRARY

69

to document i. A query \((Q)\) is similarly represented by a term vector; \(q_j\) represents the value of the \(j\)th identifier in query \(Q\).

Whereas the rows of the matrix are used to represent the documents, a particular column of the matrix identifies the assignment of a specific term to the items of the collection. This is indicated by the respective vector forms:

\[
\text{Document } D_i = (d_{i1}, d_{i2}, \ldots d_{it}) \\
\text{Term } T_k = (t_{1k}, t_{2k}, \ldots t_{nk}) \\
\text{Query } Q_j = (q_{j1}, q_{j2}, \ldots q_{jt})
\]

Where row \(D_i\) identifies the \(i\)th document, column \(T_k\) represents the \(k\)th term, and \(Q_j\) the \(j\)th query.

A similarity measure \((s)\) may be computed between pairs of items as a function of the global similarity between items. The use of a global similarity measure makes it unnecessary to insist on the presence or absence of a particular identifier, because the eventual similarity value depends on the values of the complete collection of identifiers. Typical similarity measures between the two vectors

\[
X_i = (x_{i1}, x_{i2}, \ldots x_{in}) \text{ and } X_j = (x_{j1}, x_{j2}, \ldots x_{jn})
\]

might be expressed by the following formulas.

\[
s_1 (X_i, X_j) = \sum_{k=1}^{n} x_{ik} x_{jk}
\]

\[
s_2 (X_i, X_j) = \frac{\sum_{k=1}^{n} x_{ik} x_{jk}}{\sum_{k=1}^{n} (x_{ik})^2 + \sum_{k=1}^{n} (x_{jk})^2 - \sum_{k=1}^{n} x_{ik} x_{jk}}
\]

The similarity computations between pairs of document vectors and pairs of term vectors, and between a document and a particular query, are illustrated as follows:

\[
s (D_i, D_j) \quad \text{Similarity computation between documents } D_i \text{ and } D_j
\]

\[
s (T_k, T_l) \quad \text{Similarity computation between terms } T_k \text{ and } T_l
\]

\[
s (Q_j, D_i) \quad \text{Similarity computation between query } Q_j \text{ and document } D_i
\]
It is possible to represent a collection of items on a two-dimensional map in such a way that the similarity between items is inversely related to the distance between them in the space. This is shown in Figure 1 for three items, where the similarity between items A and B is clearly much greater than between A and C.

These preliminaries are used in the remainder of this study for the description of the dynamic library operations.

\[ x = \text{individual items} \]

![Figure 1](image)

**FIGURE 1. SPACE REPRESENTATION OF SIMILARITY MEASUREMENTS**

Similarity inversely related to distance

\[ s(A, B) >> s(A, C) \]

**Dynamic Classification**

In a conventional library environment, a classification system serves to group into common classes items exhibiting certain similarities. When documents are represented by sets of content identifiers, it is possible in principle to compute a similarity coefficient between all pairs of documents, and to group into a common class all items with sufficient similarity (i.e., sufficiently small distance between them). This produces a clustered arrangement of documents such as that shown in Figure 2a, where each “x” represents a document, the large circular structures are the classes, and the small squares at the center of each circle are dummy documents, called centroids, which represent the given classes.

It should be noted that the classification process outlined above represents a global vector processing operation involving all document vectors. In an automatic retrieval environment, it is desirable to replace global operations by local ones involving only small subsets of items. For
classification purposes, global operations are required; however, comparatively inexpensive automatic classification systems can be used in practice.¹¹

The classification system of Figure 2a is similar to conventional library classifications, except that the classes are automatically constructed, and that overlap may exist because some items appear in several classes. To search a classified or clustered file, it is convenient first to compare a given query with the class centroids—by computing the similarity coefficients $s(A_i,C_j)$—and next to consider for individual comparison with the query all those documents located in classes with sufficiently high query-centroid similarity.

As indicated earlier, it is possible to tailor the classification system to the interests of the user population by altering the threshold in the similarity measure needed to enter a given item into a given class. Thus, a small number of large classes, obtained by using fairly low threshold values in the required similarity computations between items, may be adequate for casual, nonexpert users. When it becomes necessary to perform more directed searches in a particular subject area, each large class may be broken into a number of smaller classes by raising the magnitude of the similarities needed to group items into common classes. The effect of this operation is illustrated in Figure 2b in which two large initial classes are broken down into five smaller ones.

In principle, the cluster refinement operations can, of course, be repeated by constructing still smaller and more homogeneous classes in subsequent iterations. Just as it is possible to use a variety of content description or indexing systems which allow the user to choose query formulations tailored to his own background and experience, so can several different classification systems be stored simultaneously in an automatic environment, thereby accommodating many different user interests. A standard search would use the broadest or least-refined classes. As the user became more interested in a given subject area, refined classes could be used in subsequent searches. This makes it possible successively to reject more and more extraneous items, and to concentrate the search in the few specific areas that are actually of interest.

Dynamic Query Reformulation

One of the major advantages of an on-line information search-and-retrieval environment is the ability to assess the usefulness of the retrieved items as soon as a given search operation is terminated. This enables immediate query reformulations when the initial search output is unsatisfactory, followed by reassessment and reformulation processes until a satisfactory search output is obtained. All on-line retrieval systems make
provisions for query reformulation; normally, vocabulary displays are provided, consisting of terms similar to those used in the original query formulation, and the user is required to choose the new query terms that may help improve the search results.

Instead of giving a detailed description of the existing query reformulation procedures, it may be more useful to provide a model for query improvement based on the vector space processing previously discussed. Consider a typical document space, and assume that a number of items

FIGURE 2a. DYNAMIC CLASSIFICATION: CLUSTERED DOCUMENT COLLECTION

FIGURE 2b. DYNAMIC CLASSIFICATION: REFINEMENT
located in a given region of the space have been retrieved in response to a search request. The user may then be asked to provide assessments of relevance for some of the retrieved items. This normally identifies some relevant and some nonrelevant items, and makes it possible to construct an improved query which is closer to the relevant and further from the nonrelevant than the original query. In other words, the query/document similarity coefficient for the new query should be large for the relevant items and small for the nonrelevant ones.

The well-known relevance feedback process is an automatic query reformulation process based on relevance assessments supplied by the user population. The query transformation is executed in two steps:

1. the new query is moved close to the items identified as relevant by the addition of terms taken from the relevant items, or alternatively, by increasing the weight of those original query terms that are present in the relevant items; and
2. at the same time, the new query is moved away from the nonrelevant items by subtracting from the original query those terms also present in the nonrelevant items, or alternatively, by decreasing the corresponding query term weights.\(^\text{12}\)

A typical document space with added relevance assessments is shown in Figure 3a, and the relevance feedback operation is illustrated in Figure 3b. It is clear that the query transformation process can be iterated several times by constructing new query vectors based in each case on relevance assessments for items retrieved by a previous query formulation. Strong experimental evidence indicates that such a feedback operation can provide substantial improvements in retrieval effectiveness.\(^\text{13}\)

Relevance judgments can also be used as a basis for query reformulation in conventional environments where Boolean query statements are used to retrieve documents manually indexed by keywords. The user feedback process devised for the retrieval service of the Commission on the European Communities consists of the following main steps.

1. Relevance assessments are obtained for some of the documents retrieved in response to an initial search request.
2. The set of terms used to index some of the items known as relevant is examined; for example, \((A \text{ and } B \text{ and } C \text{ and } D)\) and also \((E \text{ and } F \text{ and } G)\).
3. Some terms from the query statements chosen in (2) are removed in order to broaden the resulting search statements; for example, statements \((A \text{ and } B \text{ and } D)\) and also \((E \text{ and } F)\) are constructed.
4. These shortened queries are used as new search statements to retrieve additional documents; the relevance of some of these newly retrieved documents is then assessed.
5. For each of the new query statements a "query quality factor" is computed as the ratio of the new relevant items retrieved to the new non-relevant items retrieved.

6. Those partial queries with sufficiently high query quality factors are chosen and a final feedback query is constructed by inserting or connectives between the corresponding partial query formulations; for example, the new statement used could be [(A and B and D) or (E and F)].

7. The newly constructed query is used for search purposes and the process is repeated if desired.\textsuperscript{14}

Additional feedback techniques incorporating slight variations of such a process can easily be devised.

One of the virtues of the relevance feedback and related query reformulation methods is the local nature of the operations; normally, only the previously retrieved documents are used, rather than the whole document set. Such consideration lies at the root of a number of local clustering systems designed to improve the final search output. It is thus possible to
use the automatic classification procedure previously mentioned to cluster
the (local) set of documents retrieved in response to a particular search. The
appropriate document classes can then be used to determine a specific
ranking order in which the output items can be brought to the user's
attention. By displaying whole groups of related documents, and bringing
them to the user's attention simultaneously, the choice of new terms to be
incorporated in a feedback query may be simplified.

A somewhat more formal process of this kind has been used experi-
mentally with apparently good results:

1. documents obtained in response to an initial search request are retrieved
   as before, and relevance assessments are obtained;
2. the similarity coefficients between pairs of terms extracted from the
   relevant retrieved documents are computed by comparing pairs of
columns of the reduced term assignment array (see matrix, p. 68);
3. clusters are constructed of similar terms by using as cluster centroids the
   original query terms, and grouping around them the sets of related terms
   identified in step (2); and
4. the clusters of related terms are then used for query reformulation
   purposes.\textsuperscript{15}

Again, related methods are easy to construct. In each case, the dynamic
nature of the process is evident, because all methodologies involve user
relevance assessments obtained by user/system interaction, and all pro-
cesses are based on local rather than global vector operations.

Dynamic Generation of Term Values

The document vector model discussed earlier is based on a knowledge
of the value (or weight) of each term incorporated in a given document
vector. In the absence of information about the appropriateness of a
particular term for content identification purposes, it is always possible to
assign initially a neutral weight of 1 to all terms present in a given vector,
and a weight of 0 to terms absent from the vector. In general, however, it is
preferable to discriminate further by distinguishing terms that are particu-
larly important in describing a given item from terms that are less impor-
tant; this can be done by assigning higher weights to the former than to the
latter. As mentioned earlier, it is preferable to use the collection environ-
ment to determine appropriate values of terms than to proceed by hunch or
fiat, as is now often done in conventional retrieval situations.

Two main considerations must be made: first, the environment in
which a given document is placed exerts an important influence on term
value (the term \textit{computer} may be fine in a collection of medical items, but
not in a collection in computer science); second, user assessments of document relevance should also be taken into account when available, because terms that congregate in documents judged relevant in a given subject area may be expected to be more important than terms found mostly in the nonrelevant items.

When user relevance assessments are not available, the value of the individual terms or content identifiers may be determined by considering only the context of the given collection. Consider the situation in which a document collection has already been indexed, that is, in which term vectors of the kind shown previously are already available, and assume that a new term k is to be assigned to the document collection. It is interesting to examine the expected effect of assignment of a new term on the complete document space configuration. Under normal conditions, a dual effect will be noticed:

1. the items to which term k will have been assigned may be expected to resemble each other more than before, because all these items will now exhibit an additional term in common; hence the similarity coefficient between any pair of such items will increase, and the distance between them will correspondingly diminish; and
2. at the same time, the average distance between the set of items with term k and those without term k will become larger, since the corresponding similarities between pairs will become smaller after the new term assignment than before.

This effect is illustrated in Figure 4, where the documents inside the dotted area are those to which term k is to be assigned. In Figure 4, items changing position as a result of the term assignment are transferred from the original “x” position to a new “o” position. It should be noted that the dual operation of compressing certain items (by reducing the distance between them) and increasing their distance from the remainder of the collection is precisely what is needed to enhance retrieval effectiveness, under the assumption that the compressed items can be identified with the relevant document set. Indeed, when the relevant set of items appears clustered tightly in the space, the corresponding documents can easily be retrieved together; hence the recall will be high. When these same items are simultaneously removed from the remaining items, the search precision will also be high, because extraneous, nonrelevant items are then easily rejected. (Recall is the proportion of relevant items retrieved, and precision is the proportion of retrieved items that are relevant. One normally postulates that the average user desires high recall as well as high precision.)

A question now arises about the frequency characteristics of terms capable of effecting the type of space transformation illustrated in Figure 4.
TOWARD A DYNAMIC LIBRARY

FIGURE 4. BASIC DOCUMENT SPACE ALTERATION

Three cases may be considered: first, a high-frequency term \( k \) assigned to nearly all documents in the collection; next, a low-frequency term assigned to almost no documents in the collection; and finally, a medium-frequency term assigned to a few documents but not to all. The corresponding space transformations are illustrated in Figures 5a, 5b and 5c, respectively.

1. The high-frequency term assignment will pull all the documents closer together (Figure 5a). A compressed space in which all items appear close together is unfavorable to retrieval, because it then becomes difficult to distinguish the relevant from the nonrelevant items.

2. The low-frequency term assignment will leave the document space more or less unchanged (Figure 5b), because such a term is assigned to very few items; again, the relevant items (assuming there exist more than one or two) are not separated from the nonrelevant.

3. The only favorable situation is produced by the medium-frequency terms assigned to some items (presumably the relevant set) and not to the others (Figure 5c).

Thus, when no information is available about the relevance characteristics of the terms, the medium-frequency terms are the only ones exhibiting favorable space transformation characteristics. If, for example, the space density is measured as the average similarity between pairs of items (or as the sum of the pairwise similarities), it may be seen from Figure 5 that for the high-frequency terms, the overall space density increases, the low-frequency terms leave the density more or less unchanged, while the medium-frequency term assignment may be expected to decrease the space density. The discrimination value theory, described elsewhere in the literature, assigns the highest weight to those terms capable of producing the
greatest decrease in space density upon assignment to a collection of documents. If the discrimination value of term k ($DV_k$) is defined as the space density before assignment of term k minus the space density after its assignment, a typical indicator of term value is given by $d_{ik} = f_{ik}.DV_k$
where $d_{ik}$ represents the weight of term $k$ in document $i$, $f_{ik}$ is the frequency of the term in the document (the number of times the term occurs in the document), and $DV_k$ is the discrimination value of the term $k$.

Consider the case where user relevance assessments are available for certain documents. In these circumstances it may be possible to compute the values of the probability parameters $p_k$ and $q_k$, where $p_k$ is the probability that a relevant document contains term $k$, and $q_k$ is the corresponding probability that a nonrelevant document contains term $k$. It may be shown that an excellent indicator of term value is given by the ratio of the relevant items containing term $k$, divided by the ratio of the nonrelevant containing term $k$, or

$$w_k = \log \left( \frac{p_k}{1-p_k} \div \frac{q_k}{1-q_k} \right)$$  \hspace{1cm} (1)

When binary vectors are used to identify the documents (i.e., term weights are restricted to 0 and 1 only) and the terms are assigned to the documents independently of each other, the term weight assignment $w_k$ can be shown to be optimal.\(^{17}\)

This development is of no practical use unless actual values can be substituted for the parameters $p_k$ and $q_k$. Once again, the interactive retrieval environment comes to the rescue. Indeed, after a number of feedback iterations, it is possible to construct for each term $k$ a table of frequency values, as shown in Figure 6. In the figure, $r_k$ represents the number of all documents (N) which are identified as both relevant and containing term $k$ (out of a total of $R$ relevant documents); similarly, $(n_k-r_k)$ represents the number of nonrelevant documents containing term $k$ out of all nonrelevant $(N-R)$ documents. Using the relevance information obtained from the user population in the course of the search operations, approximations can be generated to the term assignment values of Figure 6. This makes it possible to substitute the following actual frequency values for the probability parameters of Expression 1:

$$w_k = \left[ \frac{(r_k/R-r_k)}{(n_k-r_k)/(N-R-n_k+r_k)} \right]$$  \hspace{1cm} (2)

The more accurate the relevance information obtained from the user population, the closer the values indicated in Expression 2 will be to the theoretically optimal values of Expression 1.

Summary

In light of the foregoing discussion, currently favored plans for a national library network using a completely secure network kernel appear to be the reverse of what is actually needed. When the user becomes part of the system, as he or she necessarily does in an on-line search environment, there is no need to impose strict controls on the input; there is no need for a
unique, controlled, standardized indexing system; there is no need for a unique, agreed-upon classification system; and there is no need for a static search environment.

Instead, it becomes reasonable to relax the input controls by eliminating to the greatest extent possible authority lists and cataloging rules. Each item can be described by merging the content indicators assigned by a large variety of different procedures, and access can be obtained by the user with respect to a variety of different viewpoints. A range of different classification systems can be used, including broad classes for casual users and narrow ones for experts with specialized needs. Finally, the user population itself can help in selecting useful information search strategies and in adjusting the term weights based on the collection context and on prior experience with the search environment.

REFERENCES

3. This paper is drawn from a study supported in part by the National Science Foundation (DSI 77-04843).


