Associative Document Retrieval Techniques Using Bibliographic Information*

GERARD SALTON

Harvard University,[†] *Cambridge, Massachusetts*

Abstract. Automatic documentation systems which use the words contained in the individual documents as a principal source of document identifications may not perform satisfactorily under all circumstances. Methods have therefore been devised within the last few years for computing association measures between words and between documents, and for using such associated words, or information contained in associated documents, to supplement and refine the original document identifications. It is suggested in this study that bibliographic citations may provide a simple means for obtaining associated documents to be incorporated in an automatic documentation system.

The standard associative retrieval techniques are first briefly reviewed. A computer experiment is then described which tends to confirm the hypothesis that documents exhibiting similar citation sets also deal with similar subject matter. Finally, a fully automatic document retrieval system is proposed which uses bibliographic information in addition to other standard criteria for the identification of document content, and for the detection of relevant information.

1. *Introduction*

In recent years considerable attention has been devoted to the design of automatic documentation systems. If the system is to operate fully automatically, the intervention of human experts for the analysis of document content and for the preparation of document identifications ought to be eliminated. Under these circumstances the retrieval system must of necessity be based primarily on the words occurring in the individual texts, and on the terms used to formulate the search requests.

It has been suggested [1] that an acceptable system can be generated by extracting from the texts and from the information requests those linguistic units which are believed to be representative of document content, and by defining a standard of comparison between words extracted from documents and words used in the requests for documents. To determine which words are particularly significant as an indication of document content a variety of criteria may be used, including the position of the words in the texts, the word types, the vocabulary size, and most importantly the frequency of occurrence of the individual words. The most significant words are then used as "index terms" to characterize the documents, and the most significant sentences, that is, those containing a large number of significant words, are used as abstracts for the documents.

A typical automatic indexing and abstracting system based on word frequency

* Received July, 1962; revised March, 1963. This study was supported in part by the Air Force Cambridge Research Laboratories and in part by Sylvania Electric Products, Inc.

† Computation Laboratory.

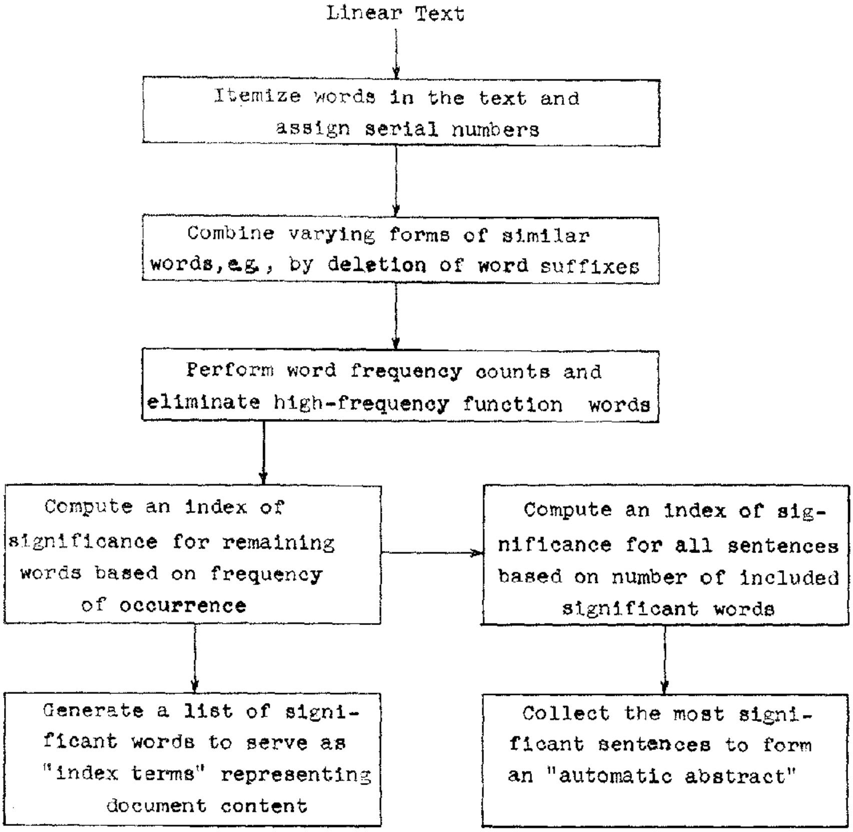
Linear Text
Itemize words in the text and assign serial numbers
Combine varying forms of similar words, e.g., by deletion of word suffixes
Perform word frequency counts and eliminate high-frequency function words
Compute an index of significance for remaining words based on frequency of occurrence
Compute an index of significance for all sentences based on number of included significant words
Generate a list of significant words to serve as "index terms" representing document content
Collect the most significant sentences to form an "automatic abstract"

FIG. 1. Typical automatic indexing and abstracting system based on word frequency counts.

counts is shown in Figure 1. The principal drawback of the system outlined in Figure 1 is the lack of any normalization procedure designed to take into account differences between individual authors or between individual document types. Thus, a given set of documents covering some homogeneous subject area may quite possibly give rise to many different index sets. Similarly, completely different document sets may be obtained in answer to only slightly differing search requests.

In order to reduce the importance attached to the individual words and to their frequencies of occurrence, the introduction of a synonym dictionary, or thesaurus, is often proposed. All words extracted from documents or search requests could then be replaced by standard thesaurus forms before being used. This solution, while attractive in theory, is difficult to implement because no definite criteria exist for the construction of good or useful thesauruses, and because the generation of any thesaurus is a complex and time-consuming undertaking. For this reason, several workers [2, 3, 4, 5] have been interested in automatic procedures designed to supplement the original terms extracted from the documents with new terms related to the old ones in various ways. Indexing techniques which make use of such "associated" terms have come to be known as "associative indexing," and the corresponding retrieval operations are known as "associative retrieval."

The present report suggests an extension of the usual associative retrieval techniques by taking into account bibliographic citations and other information peculiar to the author of a given document. It is suggested, specifically, that the set of identifying words extracted from the documents be supplemented by new words obtained in part from the bibliographic information provided with the documents; these new expanded sets of index terms may then give a more accurate representation of document content than the original ones and may thus provide a more effective retrieval mechanism.

The standard associative indexing techniques are first briefly reviewed. Thereafter, some properties of bibliographic citations are described, and the role of bibliographic information as an indication of document content is evaluated. A small computer experiment using citations is then summarized and the significance of the numeric results is discussed. Finally, a proposed fully automatic document retrieval system using bibliographic information in addition to other criteria is described.

2. *Associative Information Retrieval*

Most associative retrieval systems are based on the statistical word frequency counting procedures previously illustrated in Figure 1. Thus, given a document collection, it is possible to extract a set of n distinct high-frequency words $W_1, W_2, \cdots, W_n$, such that each document within the collection is initially identified by some subset of the set of n given words.

In practical retrieval systems, it becomes useful to provide for some additional flexibility. For example, given a search request expressed in terms of words in the natural language, it may be convenient to alter somewhat the original request, either by making it more specific and thus presumably reducing the size of the document set which fulfils the request or, alternatively, by making it more general. In the same way, given a set of terms identifying a specified document, it may be useful to alter somewhat the original set by deletion of old terms or addition of new ones in such a way that documents dealing with similar subject matter are identified by similar sets of index terms.

An analogous problem arises in connection with the document sets which are obtained in answer to certain search requests. It is often useful to alter these document sets by addition of further documents which may also have some relevance or, alternatively, by deletion of documents which are not directly relevant. Both questions can be treated by determining a *measure of association* between words or index terms on the one hand and between documents on the other, and by using this association measure for the alteration of the corresponding index term and document subsets.

Consider first the problem of word associations. Words may be related in

$$\begin{array}{c|cccccc}
\textit{Terms} & \multicolumn{6}{c}{\textit{Documents}} \\
 & D_1 & D_2 & & \cdots & & D_m \\ \hline
W_1 & \mathbf{C}_1^1 & \mathbf{C}_2^1 & \cdots & \mathbf{C}_j^1 & \cdots & \mathbf{C}_m^1 \\
 & & & \vdots & & & \\
W_2 & \mathbf{C}_1^i & \mathbf{C}_2^i & \cdots & \mathbf{C}_j^i & \cdots & \mathbf{C}_m^i \\
\vdots & & & \vdots & & & \\
W_n & \mathbf{C}_1^n & \mathbf{C}_2^n & \cdots & \mathbf{C}_j^n & \cdots & \mathbf{C}_m^n
\end{array} = \mathbf{C}$$

(a) Typical term-document incidence matrix $\mathbf{C}$ ($\mathbf{C}_j^i = n \leftrightarrow$ document D_j contains term W_i exactly n times)

$$\begin{array}{c|cccc}
\textit{Terms} & \multicolumn{4}{c}{\textit{Terms}} \\
 & W_1 & W_2 & \cdots & W_n \\ \hline
W_1 & \mathbf{R}_1^1 & \mathbf{R}_2^1 & \cdots & \mathbf{R}_n^1 \\
W_2 & \mathbf{R}_1^3 & \mathbf{R}_2^2 & \cdots & \mathbf{R}_n^2 \\
\vdots & & & \vdots & \\
W_n & \mathbf{R}_1^n & \mathbf{R}_2^n & \cdots & \mathbf{R}_n^n
\end{array} = \mathbf{R}$$

(b) Typical term-term similarity matrix $\mathbf{R}$

$$\left(\mathbf{R}_j^i = \mathbf{R}_i^j = \sum_{k=1}^{m} \mathbf{C}_k^i \mathbf{C}_k^j \Big/ \sqrt{\left(\sum_{k=1}^{m} (\mathbf{C}_k^i)^2 \sum_{k=1}^{m} (\mathbf{C}_k^j)^2\right)}\right)$$

FIG. 2. Matrices used for the generation of term associations

many different ways: for example, they may exhibit the same word stems, or they may have similar syntactic properties, or they may be usable in the same contexts, and so on. The criteria of association used in most automatic programs do not normally require a determination of syntactic or semantic properties. Rather, they are based on simple co-occurrence of words in the same texts or sentences, or on co-occurrence with individual or joint frequencies greater than some given threshold value.

Given a set of m documents and a set of n index terms, a typical procedure for the generation of term associations is as follows:

(a) a term-document *incidence matrix* $\mathbf{C}$ is constructed which lists index terms against documents; matrix element $\mathbf{C}_j^i$ is defined to be equal to k if and only if document j contains term i exactly k times;

(b) a coefficient of similarity between terms is then defined based on the frequency of co-occurrence of pairs of terms in the individual documents;

(c) a term-term *similarity matrix* $\mathbf{R}$ is then generated which exhibits all similarity coefficients between pairs of index terms;

(d) term associations are defined for those pairs whose associated similarity coefficient is greater than some stated threshold value.

A sample term-document incidence matrix $\mathbf{C}$ is shown in Figure 2(a). To obtain a coefficient of similarity between two terms based on the frequency of co-occurrence in the documents of a given collection, it is only necessary to perform a pairwise comparison of the corresponding rows of $\mathbf{C}$. Many different types of similarity coefficients have been suggested in the literature [2, 3, 4, 5]; a simple coefficient of similarity between rows of a numeric matrix, and one which may be as meaningful as any of the others, is the cosine of the angle between the corresponding m-dimensional vectors [6]. The similarity coefficients can be displayed in an $n \times n$ symmetric term-similarity matrix $\mathbf{R}$, where the coefficient of similarity $\mathbf{R}_j^i$ between term W_i and term W_j is

$$\mathbf{R}_j^i = \mathbf{R}_i^j = \frac{\sum_{k=1}^{m} \mathbf{C}_k^i \mathbf{C}_k^j}{\sqrt{\left(\sum_{k=1}^{m} (\mathbf{C}_k^i)^2 \sum_{k=1}^{m} (\mathbf{C}_k^j)^2\right)}} .$$

The term-similarity matrix $\mathbf{R}$ corresponding to the term-document matrix $\mathbf{C}$ of Figure 2 (a) is shown in Figure 2 (b). Since $\mathbf{R}$ is symmetric, only the right (or left) triangular part of $\mathbf{R}$ must be scanned in order to detect pairs of terms with large similarity coefficients.

To generate document associations instead of term associations the same procedures can be used, since the strength of association between documents may be conveniently assumed to be a function of the number and frequencies of the shared terms in their respective term lists. Document similarities are therefore obtained by comparing pairs of columns (instead of rows) of the term-document matrix $\mathbf{C}$, and a document-document similarity matrix is constructed and used in the same way as the previously described term-term matrix $\mathbf{R}$.

Consider now a typical system for document retrieval using term and document associations as shown in Figure 3. A list of high-frequency terms is first generated for each document by word frequency counting procedures. Normalization may or may not be effected by thesaurus lookup. A term-term similarity matrix is then constructed by using co-occurrence of terms within sentences, rather than within documents, as a criterion. It should be noted that as new term associations are defined, the original incidence matrix can be revised by inclusion in some of the matrix columns of new, associated terms which are not originally contained in the respective sentences or documents. The revised incidence matrix then gives rise to a new term-term similarity matrix, incorporating second-order associations, and so on. This feedback process is represented by an upward-pointing arrow in Figure 3.

To retrieve documents in answer to search requests, the programs already available can be used by adding to the term-document matrix $\mathbf{C}$ a new column $\mathbf{C}_{m+1}$, representing the request terms. Specifically, element $\mathbf{C}_{m+1}^k$ is set equal to w if term W_k is used in the search request with weight w; if word W_k is not used in the given search request $\mathbf{C}_{m+1}^k$ is set equal to 0. If no weights are specified by the requestor the values of the elements of column $\mathbf{C}_{m+1}$ are restricted to 0 and 1. An estimate of document relevance is then obtained by computing for each document the similarity coefficient between the request column $\mathbf{C}_{m+1}$ and the respective document column. The documents can be arranged in decreasing order of similarity coefficients, and all documents with a sufficiently large coefficient can be judged to be relevant to the given request. Clearly, the final relevance criterion depends not only on the terms assigned to the various documents or on the words used in the documents and search requests, but also on other terms associated with the original ones through co-occurrence in a given document collection.

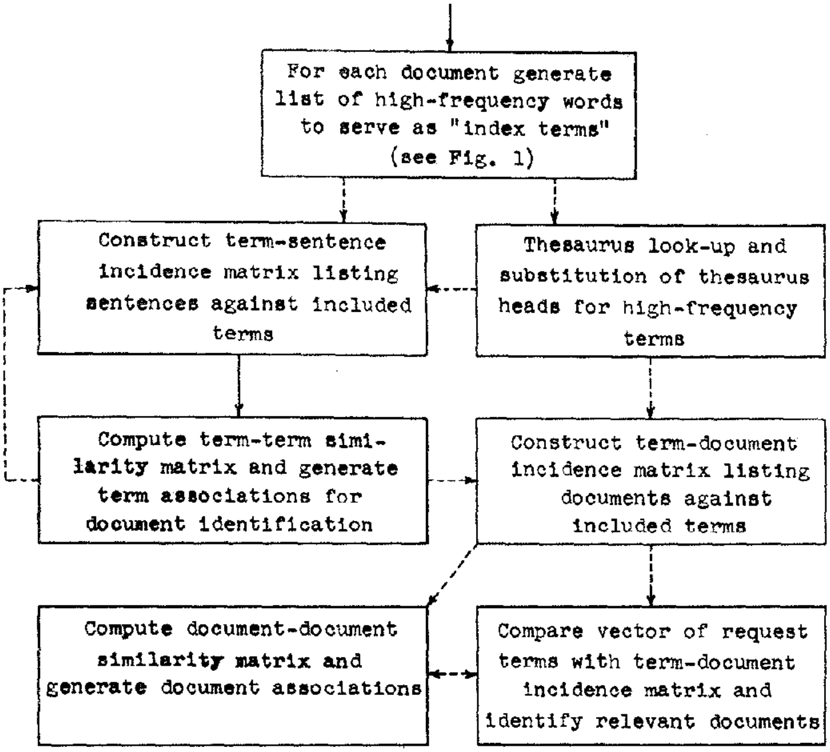
For each document generate list of high-frequency words to serve as "index terms" (see Fig. 1)
Construct term-sentence incidence matrix listing sentences against included terms
Thesaurus look-up and substitution of thesaurus heads for high-frequency terms
Compute term-term similarity matrix and generate term associations for document identification
Construct term-document incidence matrix listing documents against included terms
Compute document-document similarity matrix and generate document associations
Compare vector of request terms with term-document incidence matrix and identify relevant documents

FIG. 3. Typical automatic document retrieval system using term and document associations
⇢ optional paths → compulsory paths

3. *Bibliographic Information as a Factor in Content Analysis*

In the preceding section procedures were described for expanding the set of terms used as document identifications by inclusion of information derived from the texts of other documents in the same collection. Since the retrieval effectiveness depends to a large measure on the accuracy and completeness of the content identifications, it is of interest to inquire whether additional pertinent data available with many documents might not also be used to provide important content indications. In particular, it may be conjectured that information associated with the *author* of a given document, for example data contained in related publications of the same author, may furnish usable content indicators. The same considerations may also apply to information obtained from *publications cited* by a given author in his list of references, or from those citing the given document.

An attempt is therefore made in the next few sections to evaluate the utility of bibliographic citations as an aid to automatic content analysis. When this problem is first considered, the initial reaction must clearly be one of skepticism. Indeed, it is well known that many different practices are followed by individual authors of technical papers in the construction of bibliographies. Most important as a controlling element is the document type. Survey and tutorial articles carry more extensive bibliographies than specific research reports. Similarly, articles covering a wide variety of topics may be cited more frequently than others which are more specialized. As a result, two articles which cover identical topics from somewhat different points of view may include quite distinct bibliographies.

A second important criterion is the availability of the cited document. Thus, reports included in certain books or in important journals are likely to be cited more often than those not generally available to the public. By the same token, unclassified papers are cited more freely than classified ones. The date of publication is a related factor which also affects the probability of being cited. Very recent documents which have not had a chance to circulate, and very old ones which no longer circulate are, in general, cited more rarely than current articles which have been distributed within the recent past.

A final consideration pertains specifically to the author of a document. In many cases personal preferences are evident both as to number and type of papers cited; authors have varying backgrounds, and there may also exist a tendency toward self-citation regardless of relevancy.

Because of these and other variations, citation and reference lists[1] have not generally been used as an indication of document content. Rather, such lists are used to detect trends in the literature as a whole, and to serve as adjuncts to certain kinds of literature searches [7, 8]. Citation indexes have, for example, been used in attempts to identify significant research by equating frequency of citation with relative significance of the subject matter [9]; they have also served to trace the flow of information and to measure the relative importance of various journals to the scientific community [10].

There exists considerable evidence, however, that in addition to being useful for the above-mentioned standard applications, bibliographic information might also help in content analysis. First, it is clear that for the large majority of authors, at least some of the items included in a reference list will be highly pertinent documents whose subject matter overlaps drastically with that of the citing document. The same is true of other documents published by the same author whether they appear as part of the references or not. Second, it has been stated that many experts who are reasonably familiar with research and development in their field of activity use a citation index in preference to a subject index by following up the references to the standard, well-known works which are judged important. In such cases bibliographic references are then in fact used as content indicators. Third, citations are relatively simple to incorporate in an automatic system, since they are directly available with most technical documents, and since they can be manipulated automatically nearly as easily as ordinary running text.

If it could be shown that citations were usable as content indicators, then the associative techniques described in Section 2 could be further refined by adding

[1] A *citation index* consists of a set of bibliographic references (the set of cited documents), each followed by a list of all those documents (the citing documents) which include the given cited document as a reference. A reference index, on the other hand, lists all cited documents under each citing document.

$$\begin{array}{c|cccc}
\textit{Cited} & \multicolumn{4}{c}{\textit{Citing documents}} \\
\textit{documents} & D_1 & D_2 & \cdots & D_m \\
\hline
D_1 & \mathbf{X}_1^1 & \mathbf{X}_2^1 & \cdots & \mathbf{X}_m^1 \\
D_2 & \mathbf{X}_1^2 & \mathbf{X}_2^2 & \cdots & \mathbf{X}_m^2 \\
\vdots & \vdots & & & \vdots \\
D_m & \mathbf{X}_1^m & \mathbf{X}_2^m & \cdots & \mathbf{X}_m^m
\end{array} \quad = \quad \mathbf{X}$$

($\mathbf{X}_j^i = 1 \leftrightarrow$ document D_i is cited by document D_j)

FIG. 4. Matrix $\mathbf{X}$ exhibiting direct citations

to the term-document matrix illustrated in Figure 2(a) further document columns representing cited documents, citing documents, or documents written by the same author. These new documents would then provide new associated terms which might be equally as important as the term associations derived from other documents in the same collection.

To test the significance of bibliographic citations, a comparison was made between citation similarities and index term similarities for an indexed document collection. Specifically, a measure of similarity was computed between each pair of documents in the collection, based on the number of overlapping index terms a similar measure was then computed for the same pairs of documents, based on the number of overlapping citations; finally, the similarity measures obtained from index terms and citations respectively were compared by calculating a similarity index between citation similarities and index term similarities. An overall measure was also computed for the complete document collection by taking into account the similarity measures between all document pairs.

If use were to be made of bibliographic information for purposes of content analysis, one would hope that large similarity measures between specific document pairs due to overlapping index terms would also be reflected in large measures between the same document pairs due to overlapping citations, or vice-versa To test this hypothesis, the data obtained for the actual document collection were compared with data obtained from a randomly constructed, fictitious collection for which no correlation should exist between index terms and citations. The actual procedures used in the experiment are summarized in the next section.[2]

4. *Comparison of Citation Similarities with Index Term Similarities*

Consider a collection of m documents each of which is characterized by the property of being cited by one or more of the other documents in the same collection. Each document can then be represented by an m-dimensional logical vecto $\mathbf{X}^i$, where $\mathbf{X}_j^i = 1$ if and only if document i is cited by document j, and $\mathbf{X}_j^i =$ (otherwise. If these m vectors are arranged in rows one below the other a squar logical incidence matrix is formed similar to the matrix exhibited in Figure 4 The number of ones in the row vectors of $\mathbf{X}$ represents the degree of "citedness' for the documents listed at the head of the rows. Similarly, the number of one

[2] A more complete exposition of the experiment is given in [6].

in the column vectors of **X** measures the amount of "citing" for the documents listed at the head of the columns. For a closed document collection, the set of row identifiers is the same as the set of column identifiers as shown in Figure 4.

A measure of similarity between row (column) vectors can be obtained by calculating the cosine factor, previously exhibited in Section 2, for each pair of rows (columns). The result of such a computation can again be represented by a similarity matrix **R**, similar to that shown in Figure 2(b), where $\mathbf{R}_j^i$ is the value of the similarity coefficient between the *i*th and *j*th rows (columns) of **X**.

The coefficients of **R** now represent a measure of similarity between documents, based on the number of overlapping *direct* citations. This concept may be extended by using as a basis for the calculation of similarity coefficients not the existence of direct links between documents (links of length one), but links of length two, three, four, or more. Consider, as an example, a document collection in which document A cites document B, or B cites A. The corresponding documents are then said to be linked directly. On the other hand, if A does not cite B, but A cites (or is cited by) C which in turn cites (or is cited by) B, no direct link exists between A and B. Instead, A and B are then linked by a path of length two, since an extraneous document C exists between documents A and B. Similarly, if the path between two documents includes two extraneous documents, they are linked by a path of length three, and so on.

Given a square citation matrix **X** it is possible by matrix multiplication to obtain matrices $\mathbf{X}'$, $\mathbf{X}''$, etc., exhibiting respectively the existence of paths of length two, three, and so on [11]. Specifically,

$$[\mathbf{X}']_j^i = \bigvee_{k=1}^{m} (\mathbf{X}_k^i \wedge \mathbf{X}_j^k),$$

$$[\mathbf{X}'']_j^i = \bigvee_{k=1}^{m} (\mathbf{X}_k^i \wedge (\mathbf{X}')_j^k), \text{ and so on.}$$

Boolean multiplication is used, since the new connection matrices $\mathbf{X}'$, $\mathbf{X}''$, etc., are again defined as logical matrices. $(\mathbf{X}')_j^i$ is then equal to 1 if and only if at least one path of length two exists between documents D_i and D_j; otherwise, $(\mathbf{X}')_j^i$ is equal to 0. It may be noted that $\mathbf{X}'$, unlike **X**, can have nonzero diagonal elements, corresponding to the case where two documents mutually cite each other.

As before, the cosine measure can again be used to obtain a row or column similarity matrix $\mathbf{R}'$ from citation matrix $\mathbf{X}'$, or $\mathbf{R}''$ from $\mathbf{X}''$, and so on. These new correlation matrices measure document similarity based on common citation links of length two, three, and so on. While it is theoretically possible to work with correlation matrices $\mathbf{R}^n$ based on overlapping citation links of length *n*, it is likely that the similarity between subject matter and citations will diminish rapidly as the length of the citation links increases. For present purposes, the investigation of overlapping citations is therefore restricted to the consideration of direct links and links of length two, three, and four.

A measure of similarity based on the number of overlapping index terms between documents can be obtained as explained previously, by starting with a term-document incidence matrix **C** similar to that shown in Figure 2(a), and using the cosine measure to compute the elements $\mathbf{S}_j^i$ of an m by m symmetric similarity matrix **S**. If the documents appear as column headings as in Figure 2(a), the similarity coefficients are obtained by matching pairs of columns of **C**.

Since the term-document matrix **C** is not in general a square matrix, matrix multiplication cannot be used to obtain second order effects, similar to the citation links of length two or more. Instead, it is first necessary to compare the index terms by performing a row comparison of the rows of **C**. This produces a new n by n symmetric term matrix $\mathbf{C}^*$ which displays similarity between index terms. This matrix can be used to eliminate from the set of index terms those terms which exhibit a large number of joint occurrences with other terms. A reduced set of index terms can then be formed and a new term-document matrix $\mathbf{C}'$ constructed, from which a new correlation matrix $\mathbf{S}'$ is formed. Higher order term similarity can of course be obtained if desired by squaring or cubing $\mathbf{C}^*$.

The last step needed in the testing procedure is a comparison between the document similarity coefficients obtained from the citations and the coefficients obtained from the index terms. Specifically, it is necessary to compare the element of matrices **R**, $\mathbf{R}'$, $\mathbf{R}''$, and so on, with the equivalent ones in **S** or $\mathbf{S}'$. Since the comparison of individual pairs of equivalent coefficients may not be very meaningful, one coefficient will be computed for each document by comparing *equivalent document rows* in **R** and **S**. The cosine measure can again be used for that purpose in the form

$$\mathbf{x}_i = \cos\left(\mathbf{R}^i, \mathbf{S}^i\right) = \frac{\sum_{k=1}^{m} \mathbf{R}_k^i \mathbf{S}_k^i}{\sqrt{\left[\left(\sum_{k=1}^{m} (\mathbf{R}_k^i)^2\right)\left(\sum_{k=1}^{m} (\mathbf{S}_k^i)^2\right)\right]}}$$

to obtain a "cross-correlation vector" **x**. Each element of a cross-correlation vector is thus a measure of similarity for a *given document*, derived by comparing similarity coefficients obtained from citations with the corresponding similarity coefficients obtained from index terms for that given document. Large values of the vector element $\mathbf{x}_i$ will indicate a close similarity between the measures obtained from citations and those obtained from index terms, since then nonzero terms in $\mathbf{R}^i$ will correspond to nonzero terms in $\mathbf{S}^i$. Small values of the element $\mathbf{x}_i$, on the other hand, will indicate no similarity in the two types of measures.

A single "overall" cross-correlation coefficient can also be obtained for the complete document set, in addition to the cross-correlation vector, by comparing the complete matrix **R** with the complete matrix **S**. This is done by considering all elements within a given matrix as belonging to a single vector of dimension m^2, and comparing the two resulting vectors by means of the cosine measure. The "overall" cross-correlation coefficient will be large if many of the cross-correlation vector elements are large, that is, if for many documents large similarity coefficients obtained from citations correspond to large coefficients obtained from index terms.

The complete procedure is summarized in the flow-chart of Figure 5. For the

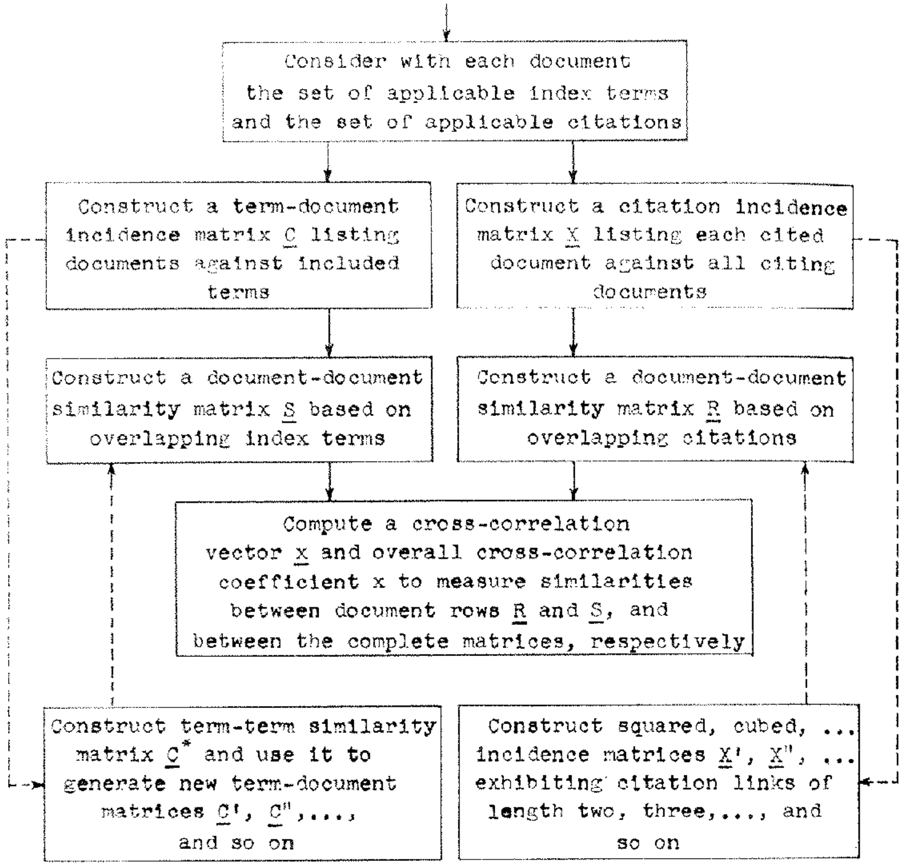
Consider with each document the set of applicable index terms and the set of applicable citations
Construct a term-document incidence matrix C listing documents against included terms
Construct a citation incidence matrix X listing each cited document against all citing documents
Construct a document-document similarity matrix S based on overlapping index terms
Construct a document-document similarity matrix R based on overlapping citations
Compute a cross-correlation vector x and overall cross-correlation coefficient x to measure similarities between document rows R and S, and between the complete matrices, respectively
Construct term-term similarity matrix C* and use it to generate new term-document matrices C', C",..., and so on
Construct squared, cubed, ... incidence matrices X', X", ... exhibiting citation links of length two, three,..., and so on

FIG. 5. Comparison of citation similarities with index term similarities

actual experiment, a collection of sixty-two documents dealing with linguistics and machine translation was chosen. A set of fifty-six index terms was used for manual indexing of the documents. The two basic inputs used for the computer experiments were thus logical matrices of dimension 62 by 62 and 62 by 56, listing, respectively, cited versus citing documents, and documents versus terms. These two input matrices correspond in format to the examples of Figures 4 and 2(a). From the two logical input matrices a set of three principal and six auxiliary similarity matrices was obtained by performing comparisons between pairs of rows or columns. The similarity matrices all correspond in format to the example of Figure 2(b). The CITED and CITNG similarity matrices of dimension 62 by 62 were obtained from the original citation matrix by row and column comparisons, respectively. The TDCMP similarity matrix, also of dimension 62 by 62, was similarly obtained by column comparisons from the original term-document matrix. Additional citation similarity matrices, designated CTD2, CTD3, CTD4, and CNG2, CNG3, CNG4 were obtained from the squared, cubed, and fourth power logical citation matrices, as previously explained.

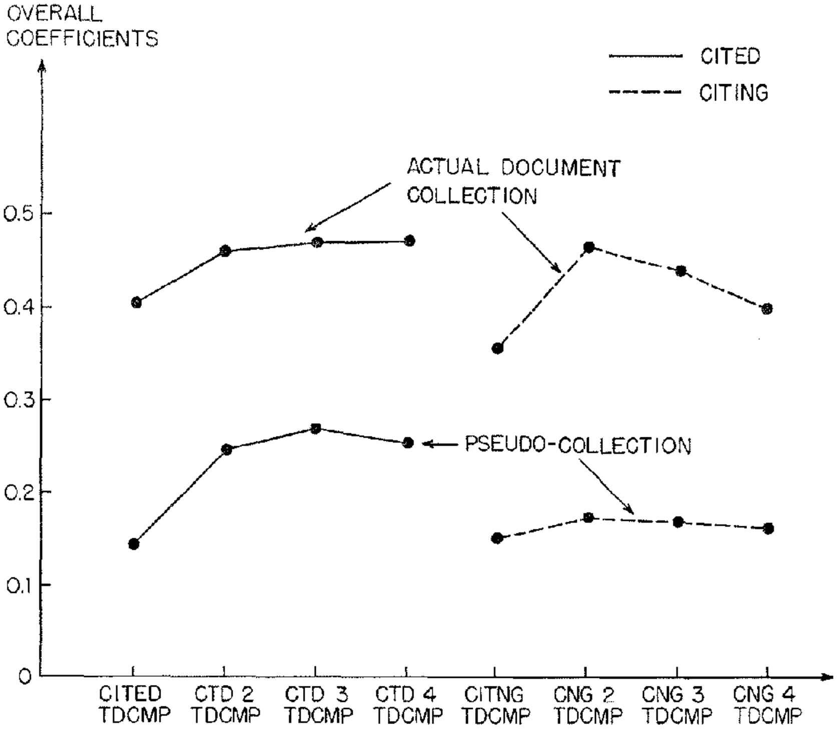

FIG. 6. Comparison of overall similarity coefficients

Eight cross-correlation operations were performed by correlating each of the eight citation similarity matrices with the term-document similarity matrix TDCMP. Each cross-correlation operation produced a cross-correlation vector of dimension sixty-two. An overall cross-correlation coefficient was also computed in each case. The internal processing was performed by means of four principal operations: matrix transposition, Boolean matrix multiplication, row correlation for logical matrix, and cross correlation between pairs of numeric matrices. An evaluation of the numeric results follows.

5. *Evaluation of Associative Techniques and Similarity Measurements*

The eight overall cross-correlation coefficients obtained for the sample document collection are shown at the top of Figure 6. The value of the overall similarity coefficient first rises as the length of the citation links increases, and then drops again as the length of the links becomes still greater [6]. This is due to the fact that as the length of the links increases, the total number of links of any length increases also; an increased number of links results in a larger number of ones in the original logical citation matrix, and thus in a higher probability of overlapping ones and a larger overall similarity coefficient. At the same time, as the length of the links increases, two factors also tend to decrease the magnitude of the overall similarity coefficient. First, the number of documents which exhibit citation links of length n but which do not exhibit links of length greater than n increases as n becomes larger. Thus more and more documents will exhibit individual similarity coefficients of zero value, thus tending to decrease the value of the overall coefficient. Second, as the length of the links increases and the citations thus become increasingly less accurate indications of document content, the magnitude of the cross-correlation coefficients obtained from the citation matrix and the term-document matrix would be expected to decrease, even for those documents for which a large number of citation links can still be found.

To test the significance of the values obtained for the overall cross-correlation coefficients, the experiment was repeated with a pseudo-document collection for which the initial term-document and citation-incidence matrices were constructed by a random assignment procedure. The random input simulates closely the conditions which obtain for the actual document collection in that the total number of links increases, as before, with the length of the links, as does the number of "documents" with a zero similarity coefficient. Nevertheless, the values of the overall cross-correlation coefficients exhibited in the lower section of Figure 6 are less than half as large for the random input as for the actual document collection. The differences can only be due to the fact that the comparison of citation similarities with index term similarities is meaningless for the random input because of the random assignment of citations and index terms, whereas there do, in fact, exist similarities for the actual document collection.

The situation which obtains for the overall document collection holds also for the cross-correlation coefficients of the individual documents. Each of four different cross-correlations is exhibited in the upper part of Figure 7 for the eight documents with the largest correlation coefficients. The same correlations are shown for the pseudo-collection in the lower part of Figure 7. A comparison of the magnitudes reveals that the individual cross-correlation elements are again much larger for the coefficients involving the document collection than they are for the random input.

Turning now to an evaluation of the experiment, it is clear that a single comparison between one small document collection and one randomized pseudo-collection cannot possibly be of significance, no matter how great the difference between the respective coefficients, or how large their size. To obtain more significant measurements a large number of document collections would have to be examined and compared with an equally large number of randomly generated pseudo-collections. The distribution of the cross-correlations obtained for the actual collections could then be compared with that obtained for the pseudo-collections. Moreover, if a sufficiently large number of experiments were run, the distributions of the cross-correlations might reasonably be approximated by normal distributions, in which case it would be possible to determine whether certain sample values derived from the document collections belong to the same distribution as those for the random cases. A negative answer would then confirm the hypothesis that citations are useful for content analysis, since the distribution of correlations derived from the pseudo-collections would be expected

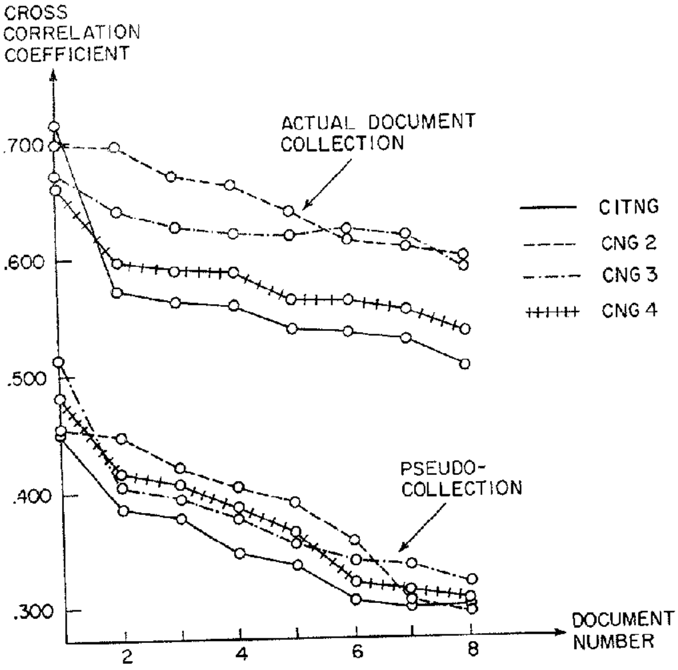

Fig. 7. Comparison of cross correlation coefficients for individual documents

to exhibit a different, possibly larger, variance and a smaller mean than that derived from the actual collections.

It might also be possible to derive the mean value and variance of the distribution of correlations for the randomly generated collections by analytic means. In that case, useful information applicable to the actual collections might be obtainable by assuming that the same distribution of correlations is valid except for a translation due to the difference in mean values.

Consider now the actual correlation measures used and the corresponding values obtained for the cross-correlation coefficients in the present experiment. For reasons already stated, it is difficult to draw very definite conclusions. In particular, it is not possible to say anything about the "goodness" of the specific measure used, since no comparison was made with other correlation measures. On the other hand, in view of the large differences in correlation coefficients exhibited in Figures 6 and 7, and the relatively large size of the coefficients obtained for the actual documents, ranging in value from 0.45 to 0.70, a more thorough examination of the procedures used may be warranted. In fact, an evaluation method applicable to the present experiment might also be used more generally for other associative processes which are based on term or document comparisons.

Let **x** and **y** be two arbitrary logical document vectors appearing in a typical term-document or citation incidence matrix. Let m_1 be the number of ones in **x**, and $n-m_1$ the number of zeros. Similarly, let m_2 and $n-m_2$ be respectively the number of ones and zeros in **y**. The problem of finding the probability of exactly p matches of ones among elements of **x** and **y** is equivalent to the sampling problem of choosing a random set of m_2 elements out of the possible n, and finding the probability of obtaining exactly p ones out of the possible m_1 (or alternatively, of finding exactly m_2-p zeros out of the possible $n-m_1$). This probability is given by

$$P_{(p)} = \frac{\binom{m_1}{p}\binom{n-m_1}{m_2-p}}{\binom{n}{m_2}}.$$

The probability of finding p or more matches of ones is then

$$C_{(p)} = \sum_{k=p}^{\mathrm{Min}\,(m_1,\,m_2)} \frac{\binom{m_1}{k}\binom{n-m_1}{m_2-k}}{\binom{n}{m_2}}.$$

The larger p, the smaller is $C_{(p)}$, that is, the smaller is the likelihood of obtaining p or more matches purely by chance. If p is so large, for example, that $C_{(p)}$ is less than .05, then a number of matches as large as p between the corresponding vector elements would be expected in less than five percent of the cases. The number of matches p may then be said to be significant at the 95 percent level.

Given two arbitrary n-dimensional vectors **x** and **y** and the corresponding number of ones m_1 and m_2, it is then possible to perform a *test of significance* for term-term and document-document associations by comparing the actual number of matches found with the theoretical number required at a given level

								m_1								
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	1	1	1													
	2	1	1	2	2	2	2	2	2							
	3		2	2	2	2	2	3	3	3	3	3	3	3	3	
	4		2	2	2	3	3	3	3	3	3	4	4	4	4	4
	5		2	2	3	3	3	3	3	4	4	4	4	4	4	5
	6		2	2	3	3	3	3	4	4	4	4	4	5	5	5
	7		2	3	3	3	3	4	4	4	4	4	5	5	5	5
m_2	8		2	3	3	3	4	4	4	4	5	5	5	5	5	6
	9			3	3	4	4	4	4	5	5	5	5	6	6	6
	10			3	3	4	4	4	5	5	5	5	6	6	6	6
	11			3	4	4	4	4	5	5	5	6	6	6	7	7
	12			3	4	4	4	5	5	5	6	6	6	7	7	7
	13			3	4	4	5	5	5	6	6	6	7	7	7	7
	14			3	4	4	5	5	5	6	6	7	7	7	7	8
	15				4	5	5	5	6	6	6	7	7	7	8	8

FIG. 8. (a) Significant number of matches p (n = 62; significance level 98%)

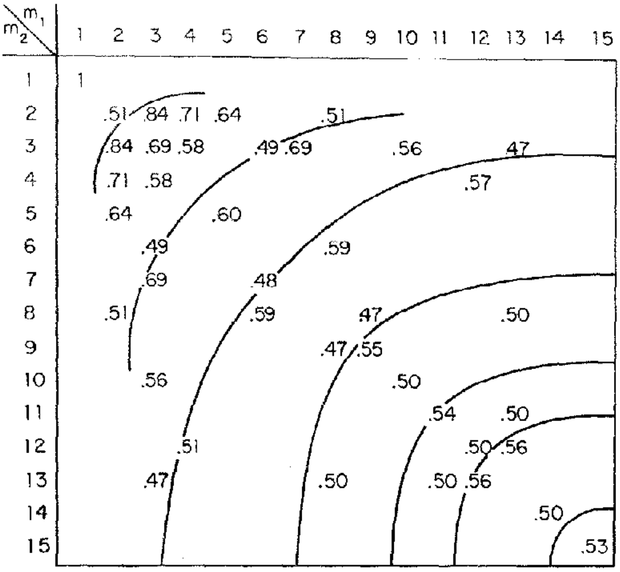

FIG. 8. (b) Significant similarity coefficients R ($n = 62$; significance level 98%)

of significance; alternatively, it is possible to compare the actual similarity coefficient computed by using the cosine measure R, with the theoretical similarity coefficient obtained by using the p value at a given level of significance. A valid significance level is thus effectively obtained for each similarity coefficient.

Figure 8(a) exhibits a table of significant p values at a ninety-eight percent level of significance for logical vectors of dimension 62. Figure 8(b) contains the corresponding similarity coefficients R at the same significance level. A cursory comparison of actual similarity coefficients obtained in the present experiment with the theoretical values shown in Figure 8(b) reveals that very few (less than five percent) of the values are significant at the ninety-eight percent level. However, the indicated procedure is nevertheless important, since it permits the separation of those document pairs which exhibit significant similarity coefficients at some given level from those which do not. The former can then be used in an extended associative retrieval system.

6. *Extended Associative Retrieval*

Clearly no proof has been presented in this study that citations do in fact play a significant role in automatic document retrieval systems. The results of the present experiment are, however, sufficiently encouraging to warrant further work with other document collections, as suggested in the preceding section. Additional evidence might then be collected to show that a given document collection could usefully be supplemented by the addition of new classes of documents or, alternatively, that a given set of index terms could be extended by addition of further index terms derived from such new document classes. The following document classes must be considered in particular:

$$
\begin{array}{ll|cccc|ccc}
 & & \multicolumn{4}{c|}{\overbrace{\hspace{10em}}^{\textit{Original documents}}} & \multicolumn{2}{c}{\overbrace{\hspace{5em}}^{\textit{Related documents through citations}}} & \overbrace{\hspace{4em}}^{\textit{Related documents through authorship}} \\
 & & D_1 & D_2 & \ldots\ldots & D_m & D_{m+1} \ldots & D_p & D_{p+1} \ldots D_q \\
\hline
 & W_1 & \mathbf{C}_1^1 & \mathbf{C}_2^1 & \cdots & \mathbf{C}_m^1 & \mathbf{C}_{m+1}^1 & \cdots & \mathbf{C}_q^1 \\
\textit{Original terms} & W_2 & \mathbf{C}_1^2 & \mathbf{C}_2^2 & \cdots & \mathbf{C}_m^2 & \vdots & & \vdots \\
 & \vdots & \vdots & & & \vdots & & & \\
 & W_n & \mathbf{C}_1^n & \mathbf{C}_2^n & \cdots & \mathbf{C}_m^n & & & \\
\hline
\textit{New terms provided by related documents} & W_{n+1} & & & & & \mathbf{C}_{m+1}^{n+1} & \cdots & \mathbf{C}_q^{n+1} \\
 & \vdots & & & 0 & & & & \vdots \\
 & W_r & & & & & \mathbf{C}_{m+1}^r & \cdots & \mathbf{C}_q^r
\end{array} = \mathbf{C}
$$

FIG. 9. Basic term-document incidence matrix usable for extended associative retrieval

(a) extraneous documents whose similarity coefficients with the original documents based on overlapping index terms are significant at some specified level of significance;

(b) extraneous documents whose similarity coefficients with the original documents based on overlapping citations are significant at some specified level of significance;

(c) extraneous documents with specified authorship, whose similarity coefficients with the original documents are significant at some specified level of significance.[3]

A typical extended term-document matrix is shown in Figure 9. This extended matrix can be used directly for associative retrieval in accordance with the procedures of Figure 3. It should be noted that, in general, no manual analysis of the extraneous documents is necessary, since the associations are directly determined by the citation or author similarities.

Citations might also be used directly for the assignment of index terms to unknown documents, by computing citation similarity coefficients between these unknown documents and the known documents of an indexed collection. Index terms assigned to "similar" documents in the known collection could then be used directly for the unknown documents. Experimentation would of course be required to determine whether the resulting index sets would in fact provide a satisfactory retrieval mechanism.

The following tentative conclusions can be drawn from the foregoing experiment: the similarity coefficients obtained by comparing overlapping citations for a sample document collection with overlapping, manually generated index terms are much larger than those obtained by assuming a random assignment of citations and terms to the documents; relatively large similarity coefficients are generated for nearly all documents which exhibit at least a minimum number of citations. If the foregoing results were confirmed by experiments with other document collections, citations could provide a large number of relevant index

[3] A proposal for generating "citation" and "author images" of specified documents is given in [12].

terms not originally available with a given document collection, and thereby create a much more flexible retrieval process. Presently available programs for associative retrieval could be used unchanged in an extended system.

REFERENCES

1. LUHN, H. P. Auto-encoding of documents for information retrieval systems. In *Modern Trends in Documentation*, Pergamon Press, 1959.
2. DOYLE, L. B. Indexing and abstracting by association. *Am. Doc. 13* (Oct. 1962).
3. GIULIANO, V. E. Automatic message retrieval by associative techniques. First Congress on the Information System Sciences, Hot Springs, Va., Nov. 1962.
4. SPIEGEL, J., BENNETT, E., HAINES, E., VICKSELL, R., AND BAKER, J. Statistical association procedures for message content analysis. Report No. SR-79, Mitre Corp., Bedford, Mass., Oct. 1962.
5. STILES, H. E. The association factor in information retrieval. *J. ACM 8* (1961).
6. SALTON, G. Some experiments in the generation of word and document associations. Proc. AFIPS Fall Joint Comput. Conf., Spartan Books, Philadelphia, 1962.
7. DUTTA, S., AND RAJAGOPALAN, T. S. Literature citations in scientific and technical periodicals—a survey. *J. Sci. Ind. Res. 17A* (1958), 259–261.
8. GARFIELD, E. Citation indexes for science. *Science 122* (July 1955), 109–111.
9. WESTBROOK, J. H. Identifying significant research. *Science 132* (Oct. 1960).
10. KESSLER, M. M. Technical information flow patterns. Proc. Western Joint Comput. Conf., Los Angeles, 1961, 247–257.
11. HOHN, F. E., SESHU, S., AND AUFENKAMP, D. D. The theory of nets. *IRE Trans. EC-6* (1954), 154–161.
12. SWITZER, P. Vector images in document retrieval. Term paper (1962); available from Widener Library, Harvard University, Cambridge, Mass.