joint session of ARL with its British counterpart, SCONUL (the Standing Conference of National and University Libraries). ARL is represented on Committees Z39 and Z85 of the United States of America Standards Institute. Other joint committees in which ARL participates deal with copyright issues, government publications, international education, national library and information systems, the National Union Catalog of Manuscript Collections, and the Union List of Serials. As a sponsoring organization, ARL is represented on the Board of the U.S. Book Exchange, Inc.

There have always been very close relations with the Library of Congress and other national libraries in Washington. As noted above, the Center for Research Libraries handles microreproduction projects for ARL and, after a special study of the Center, ARL voted, at its meeting of July 9, 1966, to recommend that its member libraries join the Center.

A grant from the National Science Foundation in 1962 enabled ARL to establish its full-time secretariat in Washington. Grants for a number of special projects have been received from the Council on Library Resources and other foundations.

Publications

The ARL *Minutes* were confidential documents until 1954. Since that date, however, the *Minutes* for meetings 1–42 (with an Index prepared by David C. Weber) have been available from ARL on microcards, and those for meeting 43 to date are available in the original format. The *Farmington Plan Handbook* by Edwin E. Williams (1961) is also published by the Association, as well as a *Farmington Plan Newsletter,* edited by Lloyd Griffin, which appears twice a year. This newsletter attempts to bring together all available information on the acquisition of library materials from foreign countries and their bibliographic control.

EDWIN E. WILLIAMS

ASSOCIATION TRAILS

Association trails, or the connections from idea to idea within the mind, have interested writers from Aristotle to present-day learning psychologists. However, Bush, in his now-famous article "As We May Think" (*1*), was the first to suggest that information retrieval systems should attempt to simulate "the intricate web of trails carried by the cells of the brain" and follow associations from topic to topic and from document to document. In the years since Bush's article, and particularly since the late 1950s, there have been many attempts to mechanize association mapping within a large corpus of documents. These have produced the so-called statistical association methods in information retrieval. While nothing has yet been

developed which will flit from topic to topic in the style of Bush's Memex, this research has stimulated a consideration both of the types of concept connections which are meaningful in an information retrieval context and the most effective and efficient methods for producing these connections.

This article will attempt to define the meaning of and appraise the usefulness of association trails in an information retrieval or library system, to review work in psycholinguistics and statistical methods which bears on this subject, to describe what is offered in nonmechanized systems in the way of document and term connections, and, finally, to review the recent work with statistical association techniques for mechanized indexing, classification, and retrieval.

Association, from a psychological point of view, is a relationship between two or more ideas, established in the course of individual experience and of such a nature that the presence of one tends to evoke the other (*2*). The operators and users of an information retrieval system, organized to provide documents relevant to informational requests, may be concerned with trails or sequences of associations from idea to idea for a number of reasons. Basically an information retrieval system is a file of documents which, for reasons of economy, must be searched by means of a key, the key generally being some form of subject index and/or classification scheme. Entry to the file will be based on the user's initial verbalization of his informational need, translated to the indexing language or classification grouping of the system. However, it is virtually impossible for an individual user to retain in his memory, or at least to recall initially, all the relevant vocabulary from his particular discipline. Thus relevant answers may be missed in the retrieval operation. If the user has access to the system, he may produce his own conceptual associations if he is dissatisfied by the initial output and follow his own trails through the index. If the user does not have access to the system, so that he cannot follow his association trails himself, experience has generally shown that the recall figure (proportion of relevant documents retrieved) will be unsatisfactory unless some procedure beyond a mere matching of request and index terms is employed. Even when the user has access to the file, he may not remember all the useful associations if the index, classification, or documents themselves do not somehow remind him.

An information retrieval system cannot associate ideas or concepts directly, since it deals with terms (which will here be used to include all such document attributes as subject headings, index terms, keywords, descriptors, and the like) and with documents (to include all such contexts as books, periodical articles, technical reports, abstracts, and arbitrarily specified units of text). Associations, therefore, must be made between terms or between documents; in the first case a thesaurus is produced, and in the second a classification hierarchy.

If an information retrieval system does incorporate some type of word or document connections, what is the relationship of these to relevance? The thesaurus can provide terms which are "relevant" to the request terminology, but documents indexed by these terms are not necessarily "relevant" to the request. The relevance or connectedness of terms (or documents) must be distinguished from the rele-

vance of a document to a request, which is a much more complex relationship. Whether term or document associations are employed, these must be considered simply as strategies or decision rules for searching a file, to be compared to other decision rules on the basis of expected risk or similar functions. The quality of the output will reflect the appropriateness of the assigned index terms or document classes and the amount of user-system rapport, in addition to the number and appropriateness of term and document connections. The association trails mapped by the system will produce relevant documents only if they correspond exactly to the association trails of ideas in the mind of the user.

To use association trails of any sort in an information retrieval system implies some belief in common mental (or at least verbal) habits among the users of the system. If users work within the same discipline, this assumption may be justified. On the other hand, if users come from a variety of environments, the association trails within the system will likely represent a compromise among varying points of view. If the variability in verbal habits is great, the trails may be worse than useless. Paisley (*3*) pointed out that an associative cluster will serve best only the user whose associative patterns approximate the norms from which the index was constructed. Even for a single user, two or more terms may be related in one context and not related in another or differently related in another. For example, "mapping" and "function" would be strongly associated in a mathematical context, and associated weakly or not at all in a geographic one. If a collection is confined to a single discipline, the variation in word connotations and hence in word associations will not be so great from user to user, but even in this case some individual differences may be expected.

If all possible associations are included in a system, every term or every document will eventually be related to every other term or document through some association trail. As Locke put it (*4*):

> There is no one thing, whether simple idea, substance, mode, or relation, or name of either of them, which is not capable of almost an infinite number of considerations in reference to other things . . . one single man may be at once concerned in and sustain all these following relations and many more, viz, father, brother, son, grandfather, grandson, father-in-law, son-in-law, husband, friend, enemy, subject, general, judge, patron, client, professor, European, Englishman, islander, servant, master, possessor, captain, superior, inferior, bigger, less, older, younger, contemporary, like, unlike, etc., to an almost infinite number.

Term and document connections are not linear but multidimensional. A term or document is connected to a first term or document in one sense, and to a second in another sense. The term "volume" is connected to the term "area" in describing physical objects and to the term "loudness" in describing sound, but there is no strong relationship between "area" and "loudness," indicating the associations are along different dimensions or axes. The problem for the designer of an information retrieval system is how to control or limit the association trails so that only those relevant to the request will be followed.

Historical Background

In a survey of the historical foundations of research in statistical association techniques, Jones (*5*) listed three fields in which association formed a subject of inquiry long before the present interest by library and information scientists:

1. Philosophical studies of the principles of association
2. Psycholinguistic studies of the associative structure of thought on the basis of symbol (word) co-occurrences
3. Measures of statistical association used for data interpretation in a number of fields

Historically Jones finds two possible explanations for associations of ideas: the ideas represent entities which really do co-occur in real life, or the associations are made in the mind of an individual. Aristotle distinguished association by similarity, by contrast, and by contiguity. The phrase "association of ideas" was first introduced by Locke and extended by Berkeley, Hume, and other British empiricists. Locke attributed associations in verbal behavior to accidental contingencies in the real world. Hume found three principles of connection among ideas, namely, resemblance, contiguity in time or place, and cause and effect.

In the twentieth century the association of ideas was incorporated into psychological doctrine, but interest turned to experimental studies of, generally verbal, associations, with particular emphasis on learning. Word association tests, in which a subject is asked to give a response to a stimulus word under free or controlled conditions, became popular. Many of these tests were carried out for the purpose of studying various psychological attributes of the subjects, but a number of attempts of interest to information scientists were made to classify the semantic relationships between the stimulus word and its response. Woodworth and Schlosberg (*6*), for example, suggested the following classes, using, as an example, the word "needle" as the stimulus:

1. Definition, including synonyms and supraordination, e.g., instrument
2. Completion and predication, e.g., sharp, eye
3. Coordinate and contrasts, e.g., pin, thimble
4. Valuation and personal association, e.g., pricking, useful

Miller (*7*) produced a more extended list:

1. contrast
2. similar
3. subordinate
4. coordinate
5. superordinate
6. assonance
7. part-whole

8. completion
9. egocentrism
10. predication

Vickery (*8*), looking at classification from an information retrieval point of view, saw the following relationships:

1. generic
2. coordinate (terms subordinate to the same term)
3. analytic (all other types of relationships)

Perry and Kent (*9*) further classified the analytic relations as follows:

1. categorical
2. intrinsic
3. inclusive
4. comprehensive
5. productive
6. affective
7. instrumental
8. negative
9. attributive
10. simulative

Results from association tests have been little used in constructing either thesauruses or classifications. The reasons are clear. While it appears that the majority of responses depend on the meaning of the stimulus word, the personal experience of the subject and the context both of the subject and the stimulus word play an important enough role that it is difficult to derive from such tests any norm of language behavior. Also, in most word association tests, subjects are restricted to single responses, and what is desired, for retrieval purposes, are chains or clusters of associations. One study (*10*) of the use of word association tests to prepare a thesaurus has been reported, but there the tests were administered to determine which words the users would *not* associate by themselves so that these could be incorporated into the system.

Studies in psycholinguistics of measures of meaning similarity preceded similar work in information retrieval, but it does not appear that the influence of the earlier studies has been great. Osgood's "semantic differential" (*11*) measures the meaning of a word for a particular subject as a point or vector in multidimensional space. The operation of measurement is likened to a game of "twenty questions." The subject judges the concept against a series of scales defined by polar adjectives; as an example, for the word "father" some of the scales are as follows:

happy - - - - - - - - sad
hard - - - - - - - - soft
slow - - - - - - - - fast

The n major factors or dimensions along which meaning varies are then selected by the multivariate technique of factor analysis. Similarity between two terms is them measured by the distance formula

$$\sqrt{\sum_{k=1}^{n} (f_{1k} - f_{2k})^2}$$

where f_{1k} is the kth factor loading or weight for the first term and f_{2k} is the $k$tlh factor loading or weight for the second term. Osgood used this formula to determine the similarity of concepts as judged by different individuals or groups. Fronn an information retrieval point of view, the semantic differential is more a measure of emotional attitude toward a term than semantic meaning, and Osgood himself noted that concepts which are distinct semantically may have the same semantic profile. Nevertheless, Osgood's work is significant as the first attempt to quantify some aspect of meaning and word relationships.

Deese (*12*) defined a particular type of meaning which he calls "associative." This meaning is distinct from, if not independent of, the other senses of meaning, that is, the dictionary and grammatical relationships listed above. The associative meaning of a word is given by the distribution of responses to the word by a group of subjects in a word association test. Two stimuli or words have the same meaning when the distribution of responses is identical. Two stimuli have similar meanings to the extent they have the same distribution of associations. Garskof and Houston (*13*) extend Deese's associative meaning to associative hierarchies rather than the single response per subject which Deese employed. Their relatedness measure is a function both of the number of common associates and the congruence of their rank orders. Studies by other investigators in this general area are listed in the references for the above two papers. Associative meaning predicts the words which will co-occur in the verbal environment of a word and thus is similar to statistical association measures in information retrieval, which measure the tendency of terms to co-occur in a written environment.

Nonmechanized Systems

Most conventional library and retrieval systems employ some type of associative apparatus. Index term association trails are provided by cross referencing ("see" and "see also"), subject-authority files, and specially compiled thesauruses. Document trails are provided by the hierarchical arrangement of document groups or classes on the shelf or by a classified catalog. In addition, it must be recognized that the need for association trails is dependent on the amount of control in the indexing vocabulary. Vickery points out that control of index terms is essentially a matter of establishing associations among words. Where a controlled vocabulary is used, as in descriptor systems or such subject-heading lists as that of the Library of Congress, synonyms and near synonyms will be assigned the same heading so

that little associative apparatus is needed. When little control is exercised over the natural language of the author in the selection of index terms, as in a pure uniterm system, the need for some kind of associational mechanism soon becomes evident. That statistical associative retrieval techniques first arose in connection with coordinate indexing systems is not surprising, since vocabulary control in such systems is minimal.

If conventional indexing and classification schemes are judged against the Woodworth classification of word relationships given above, it will be seen that most of the relationships are definitional (synonymic, partially synonymic, or hierarchical) or, to a smaller extent, coordinate. Using Vickery's categorization, they are generic or coordinate rather than analytic. As an example, consider the treatment of the term "needle," the example used by Woodworth, in the Library of Congress subject headings list, the Library of Congress classification, and the Dewey decimal classification (seventeenth edition).

L.C subject-headings list
- Needles. See Pins and Needles.
- Needles, Phonograph. See Phonograph Needles.
- Pins and Needles. (Classify under HD9999.P6–63 or GT2280 or TS2301.P5)

L.C classification schedules
- H – Economics
- HD – Economic History
- HD9999 – Miscellaneous Trades and Industries
- HD9999.P6 – Pins, Needles, etc. (other miscellaneous industries include Pens, Phonograph Records, etc.)
- GT – Manners and Customs
- GT0500–2370 – Dress, Costume
- GT2280 – Pins
- T – Technology
- TS – Manufactures
- TS2301 – Miscellaneous
- TS2301.P5 – Pins and Needles

Dewey decimal classification (seventeenth edition)
- Needles. See Notions. (in index)
- Notions
  - garmentmaking
    - commercial, 687.8
    - domestic, 646.1
      - 680 Handcrafted, assembled final products
      - 687 Clothing
      - 687.8 Items auxiliary to clothing construction (notions): buttons, hooks, eyes, needles, pins, thread, snaps, slide fasteners, padding, shields
      - 646 Clothing and Care of Body
      - 646.1 Textiles, Fibers, Other Materials Used

LC relates "needles" to the coordinate term "pins" and Dewey to the generic term "notions." Both classifications relate needles to coordinate (pins, phonograph records, buttons) and generic (notions, clothing, manufacturing) subjects from several points of view. That other types of associations, such as "sharp," "prick-

ing," and "eye," do not appear is probably not so much an oversight as an indication of their relative uselessness in an information retrieval context. As Vickerry has said, relationships other than generic are not used because their value has not yet been established, while generic relations have proved valuable from the earliest days of library work. Problems arise with many classification schemes not because associations are only on the generic-specific level, but because documents may be associated from a number of points of view, as can be seen in the example. The faceted classification approach is an attempt to solve this problem by grouping items not in a single hierarchy but with respect to a number of fundamental facets. Similar problems arise in many indexing systems, not because cross references are solely on the coordinate or generic-specific level, but because the appropriateness of the cross references does not remain invariant.

Does this brief survey of the approach to association trails in a conventional library justify Bush's reproach that "our ineptitude at getting at the record is caused by ineptitude in indexing"? Twenty years of new, improved mechanized systems have not brought forth a Memex which is universally accepted as superior to the norm of Bush's day: the conventional library system with a card catalog and open, classified shelves of books. In fact, it is now clear that the older systems have advantages with respect to association mapping which are yet to be achieved by mechanized systems. The first of these is multiple access. A user in a conventional library may locate relevant material through author, title, or subject headings in the catalog, through books on the shelf adjacent to those which were listed in the catalog, through references in these books, through cross references in the catalog, and through subject headings in the catalog adjacent to his initial entry (i.e., with similar spelling). In addition, the conventional library provides browsability. The user may scan many possible trails because of the multiaccess capability, but need follow them only so long as they provide relevant material. The user controls the direction and the length of each association trail. By contrast, in most mechanized systems access is linear because following such a multiplicity of trails would result in large outputs with low relevance ratios. The user, since he does not control the mapping, cannot immediately eliminate nonproductive trails before they are followed through the system. A third advantage of the conventional system is its opportunities for librarian-user interaction. In well-administered libraries, an individual experienced with the system and with user verbal habits is able to interpret informational needs and steer the user along useful association trails.

In several important respects, then, mechanization of library procedures, by removing the user from close contact with the system, has tended to reduce rather than increase the possibilities for following trails of fruitful associations. Of course, one problem with conventional libraries is that they do not provide the same degree of access to periodical articles, government documents, technical reports, and the like, and this disadvantage has been an important impetus toward mechanization, since indexers and classifiers are apparently not available in sufficient numbers, or perhaps at a low enough cost, to cope with such a quantity.

Mechanized Systems

Mechanized association mapping implies methods which, at least partially, employ a repetitive procedure or algorithm capable of being automated to extract term and document associations. The incorporation into a mechanized system of associative features, such as cross references, from a conventionally constructed index or classification scheme will not be considered here. Association trails in mechanized systems may be in the nature of term or document connections. Term connections may be incorporated into the index or calculated as part of the searching algorithm. Document classification, on the other hand, is generally applied to the entire file, independent of any particular request. In either case the techniques used may be statistical, i.e., based on word counts, or nonstatistical, employing other methods. In the development of mechanized association methods since the late 1950s, interest was first concentrated on statistical associative retrieval, later on associative indexing, and most recently on automatic classification. Nonstatistical methods, with the notable exception of citation indexing, have attracted less attention. In the discussion of these methods, this approximate chronology will be followed.

Statistical Associative Retrieval

Statistical associative retrieval includes all techniques for extending the direct translation of the user's informational request to the language of the system by additional, associated terms, where degree of term association is based on term co-occurrences within the file. The term "association" has a somewhat different connotation in statistics than in psychology. The association of two variables or two properties is a numerical measure of their tendency to vary together among members of some set or population. Many different measures have been proposed; surveys of these have been made by Yule (*14*) and Goodman and Kruskal (*15*). To apply a measure of association, it must be possible to measure the two variables or properties on at least a classificatory scale. The particular measure chosen will depend on whether the scales are simply classificatory (e.g., yes-no, red-white-blue), ordered (e.g., low, medium, high), numeric and discrete (e.g., 1, 2, 3, 4, . . .), or numeric and continuous (e.g., may assume any real value between 0 and 100). Thus measures of term or document association are possible only if the terms or documents can be classified on some scale. The notion that terms may be classified by their relationship to a document—either on a yes-no (1-0) scale (assigned or not assigned to the document) or on a discrete numeric scale (the number, for example, indicating the frequency of the word in the document)—is basic to the development of statistical associative retrieval. Similarly statistical document classification is based on the assumption that documents may be classified *vis-a-vis* a set of terms on a yes-no or numeric scale based on frequencies or other weightings.

The technique of associative retrieval assumes that words which are statistically associated in a document collection (i.e., occur together in a large number of documents or are strongly weighted in the same documents) are also semantically associated. Once a set of m terms has been classified with respect to a set of n documents, an nxm document-term connection matrix C may be constructed, with the n rows representing documents and the m columns representing terms (see Figure 1). The ijth element c_{ij} indicates the relationship of the ith document to

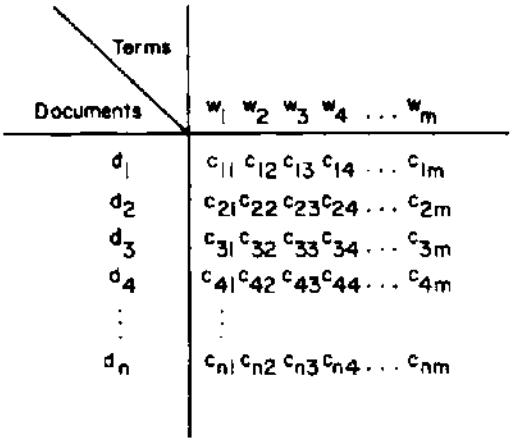

FIGURE 1. *Document-term association matrix* C.

the jth term. In the simplest model c_{ij} will be 1 if the jth term has been assigned to the ith document; otherwise, it will be 0. In other models c_{ij} may indicate weighting of the jth term for the ith document and will thus assume a range of values. The scale, in this case, will generally be discrete, although continuous scales are possible. Given this matrix, the association measure for two terms or two documents is defined as a function of the column or row vectors.

Goodman and Kruskal suggest that it is desriable to choose a measure which has contextual meaning, i.e., which may be given a probabilistic interpretation, instead of, as a matter of course, one of the traditional measures such as the product-moment correlation coefficient. A number of other properties are generally considered desirable, or at least conventional, for any measure of association r:

1. r assumes values between -1 and 1 (for signed variables) or between 0 and 1 (for unsigned variables).
2. If two properties are statistically independent, $r = 0$. Specifically, if the assignment of index term A to a document has no effect on the assignment of index term B to a document, A and B are considered to be independent and $r_{AB} = 0$.
3. If the two properties are equivalent, $r = 1$. Specifically, if A and B are always assigned together to a document, $r_{AB} = 1$.
4. r increases monotonically with increasing association. This means that as the proportion of documents described by both A and B increases, so should r_{AB}.
5. r is symmetric with respect to the two variables. In other words, $r_{AB} = r_{BA}$.

It will be seen that some of the proposed association measures do have contextual meaning and satisfy the other desirable properties, whereas others satisfy only some of them. The development of these measures will now be described in more or less chronological sequence. The following symbols will be used throughout:

C	document-term connection matrix
R	term-term or document-document association matrix, in which the element r_{AB} measures the association between term (or document) A and term (or document) B
$N(A)$	number of documents assigned term A (or, dually, number of terms assigned to document A)
$N(A+B)$	number of documents assigned term A or term B or both (or number of terms assigned to document A or document B or both)
$N(AB)$	number of documents assigned both term A and term B (or number of terms assigned to both document A or document B)
n	total number of documents in the file
m	total number of distinct terms in the file

Where dual (term or document) interpretations are possible, the context will make clear which one is required.

Statistical associative retrieval had its roots in the coordinate indexing systems (uniterm systems) which became popular in the 1950s. It soon became apparent that, while any combination of terms could be used in searching such a file, the user had no guide as to which combinations would be useful.

In 1955 Taube (*16*) suggested that a requestor should in some manner be provided with a display of all the uniterms (index terms) which co-occurred as document indexing terms with the terms of his request. Although he did not use a quantitative measure involving the frequency of co-occurrence, this article might be considered the forerunner of later numeric coefficients.

The use of quantitative measures of document and term similarity appears to have been first suggested by Maron and Kuhns (*17*) in a 1960 paper on the technique of "probabilistic indexing." Index terms assigned to a document are weighted in proportion to the probability that if a user uses the term he will be interested in that document. When the computer is given a request, represented as a Boolean polynomial of index term (i.e., employing set union and intersection), it derives for each document a relevance number which is proportional to the probability that the document will satisfy the request. Following the initial selection based on relevance number, further documents are selected by means of measures of term or document association (closeness). Three term association measures were discussed:

$$r_{AB} = \frac{N(AB)}{N(A)} \qquad r_{AB} = \frac{N(AB)}{N(B)} \qquad r_{AB} = \frac{N(X)N(Y) - N(U)N(V)}{N(X)N(Y) + N(U)N(V)}$$

where A is a request term, B an index term assigned to a document, and $N(X)$, $N(Y)$, $N(U)$, and $N(V)$ are the numbers of documents in the respective classes

pictured in the Venn diagram of Figure 2. The first two may be interpreted probabilistically as conditional probabilities of a document being indexed by both terms, given that it is indexed by A, in the first formula, or B, in the second. The distance of any document from another was defined by the Pythagorean distance formula used by Osgood.

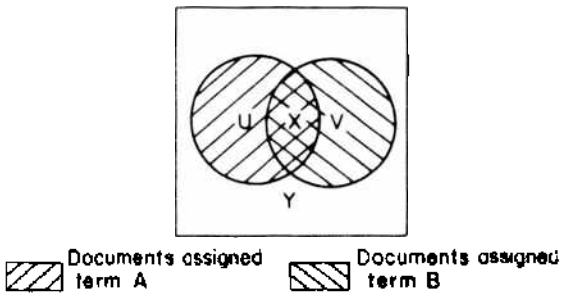

FIGURE 2. *Venn diagram showing classes of documents used in computation of association measures.*

A second landmark in the development of associative retrieval techniques was Stiles' 1961 paper on association factors (*18*). Stiles proposed that requests be expanded by statistically associated terms and that document relevance numbers then be computed by summing the association factors over all terms in the request and document. Stiles suggested a formula which is the logarithm of the Chi-square value obtained in a two-by-two contingency table test of independence in the assignment of the two terms. In such a test, high Chi-square values indicate a lack of independence. Both Maron and Kuhns carried out small tests of their proposed procedures. Stiles was also the first to distinguish two types of statistical association and to suggest that these were related to two corresponding types of semantic association. Words which co-occur are frequently called first-generation terms and result not from synonymity or closeness in meaning but from the nature of the real-word facts described by the document. Second-generation terms are terms both of which co-occur frequently with a third term or set of terms, i.e., terms which have the same first-generation associates. Stiles believed that these would tend to be synonymous or partially synonymous.

In his earlier papers Doyle (*19*) also advocated the use of statistical measures of word co-occurrence for document retrieval, suggesting that because authors in a particular field are obliged to use a particular professional vocabulary, there should be a "consensus of co-occurrence" among authors which should be measurable. A computer-based indexing system could be augmented with a program which measured associations among terms and then printed them out in order of associative strength or displayed them, via a television-like device, in a "semantic road map" (see Figure 3), in which connecting lines would show words which co-occur frequently. The user, presented with the primary and secondary associations to the terms of his request, would select those word clusters which were of interest. The system would then give the user more information about the documents which caused these clusters, or, if the number of co-occurrences were large,

a subassociation map, based on this subset of the library. As a measure of association, Doyle used

$$r_{AB} = \frac{N(AB)}{N(A + B)}$$

which is the conditional probability a document will be indexed by both A and B, given that it is indexed by either. In terms of the Venn diagram of Figure 2,

$$r_{AB} = \frac{N(X)}{N(X) + N(U) + N(V)}$$

Doyle also suggested that hierarchical maps could be generated by eliminating connections between words of approximately equal frequency, and equating words of high frequency to more general categories and words of low frequency to more specific categories.

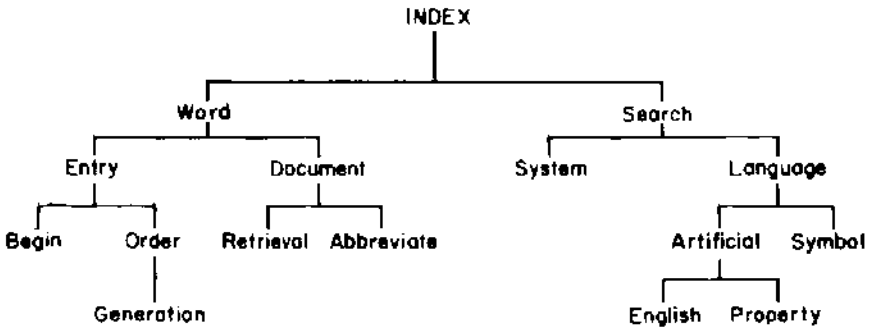

FIGURE 3. *Example of hierarchical word association map.* [*Reproduced from Ref.* (34), *courtesy of L. B. Doyle.*]

Becker and Hayes (*20*) spoke of the relevancy of file items (documents, terms, requests) in much the same way as other authors speak of term and document associations, with the difference that file items are considered to be relevant "not in any absolute sense but to the extent that they are considered as relevant from some point of view." File items are represented as points in space, and the relevancy of any two points is measured by the cosine of the angle subtended by them at a third point P (see Figure 4). It is suggested that P might in some way represent the position of the user, for example, by some type of user profile.

Guiliano and Jones (*21*) presented a procedure for document retrieval based on linear transformations, both for establishing associations between terms and for ranking. In their model, the request is represented by an m-dimensional column vector $\mathbf{Q}$, in which ith component represents the weight assigned to the jth index term by the user; and the answer or output of a retrieval system by an n-dimensional column vector A, in which the jth component represents the value assigned to the jth document in response to $\mathbf{Q}$. The retrieval process was considered to be the product of three linear transformations or matrices:

$$\mathbf{A} = SCR\mathbf{Q}$$

where C is the document-term association matrix, S a document-document association matrix, and R a term-term association matrix. If term associations through several generations are to be included, the matrix R will be replaced by a sum of the form

$$I + YR + (YR)^2 + (YR)^3 + \cdots$$

where Y is a diagonal matrix whose diagonal elements represent weights assigned to each term, based on the cost of associating through that term. An element of R^2, for example, gives a measure of the interconnection between two terms via all pairs of documents such that one contains the first, the other the second, and both share one or more index terms. Normalizations are introduced so that association strengths for longer and longer paths approach zero.

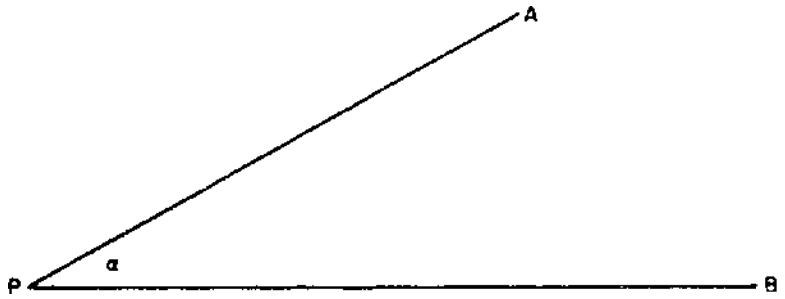

FIGURE 4. *Association between terms* A *and* B *from point of view* P *is measured by cosine* α.

Like Stiles, Guiliano and Jones distinguished two types of semantic association: contiguous and synonymous. In a later paper Guiliano (*22*) introduced statistical measures of contiguity and synonymity, which, although applicable specifically to natural language text in which adjacent word pairs are used as contexts or "documents," are capable of extension to the term-document matrix. Two words are contiguous if the objects denoted by them have to do with one another in the real world. Contiguous associations thus are similar to the analytic relations mentioned earlier. The proposed contiguity coefficient measured "the degree of surprise connected with finding $N(AB)$ pairs, when chance and statistical independence dictate finding $N(A)N(B)/n$ pairs." The formula is

$$r_{AB} = \frac{nN(AB)}{N(A)N(B)}$$

Synonymity is defined in terms of interchangeability of usage. Two words are perfect synonyms if and only if one can always be used in place of the other and; partial synonyms if one can sometimes be used in place of the other. The synonymity measure is

$$r_{AB} = \frac{\Sigma n_i N(Ai)N(Bi)/N(i)}{N(A)N(B)}$$

Salton has described in a number of papers (*23*) the evolution of the SMART retrieval system, which makes possible the extension of document specifications and/or search request using one or more of three optional operations:

1. expansion of "semantic labels" through statistical term and document correlations
2. substitution by more general or more specific concepts in a concept hierarchy
3. syntactic analysis of document sentences and search request and a matching of term clusters only if syntactic relationships are identical.

The system allows search requests to be altered successively by any of these three operations until a satisfactory response is achieved. Term similarity is measured by a modified correlation coefficient which is identical to Becker and Hayes' measure if their point P is taken as the origin. Salton's system for document retrieval using term and document associations involves constructing a term-term association matrix and then revising the original document-term matrix so that the rows (documents) include new terms having a high association value with those already assigned to the document. Then a request is represented as an additional row in the document-term matrix and document-request association measures computed. In another paper (*24*) he suggested an extension of the usual associative retrieval techniques to the bibliographic citations to and from the document. Further document rows could be added to the matrix representing cited documents, citing documents, and documents written by the same author.

Spiegel and Bennett (*25*) pointed out that the techniques used by Maron and Kuhns, Stiles, and Doyle do not take into consideration the fact that the more lengthy the string of terms representing a document, the more likely the co-occurrences involving terms in the string are due to chance. This omission is minor if the number of terms used to index all documents is more or less constant, but becomes important when "document" means natural language contexts. They then derived the following measure of association which is independent of the individual frequencies of the terms, sentence (i.e., document) length, number of different terms, and total number of pairs within the corpus or file:

$$r_{AB} = \frac{m(m+1)}{2}\left[\frac{N(AB)}{m_0} - \frac{N_0(A)N_0(B)}{2m_0^2}\right]$$

where m_0 is the total frequency of co-occurrence, $N_0(A)$ the number of co-occurrences which contain term A, and $N_0(B)$ the number of co-occurrences which contain term B. A more complicated formula is given which takes into consideration the direction of co-occurrence.

Hillman's topological models for term and document relations (*26*) could be mentioned here. Although theoretical in nature, they lead to retrieval algorithms which may be compared with those already proposed. A set of pairwise term similarity coefficients (based on similarity of contiguity profiles) defines a graph G (see Figure 5) in which vertices representing terms are joined by edges if the

similarity coefficient r of the terms is greater than some threshold value (here 0). Connected subgraphs represent genera (sets of conceptually related terms) of the system. The length of an edge is defined as $1/r - 1$ and the distance between two vertices as the length of the shortest path between them. The affiliation c of a term for a document is a function of the number of different quantifiers for the term in the document. This affiliation relation defines a directed graph G' in which arcs join vertices representing terms to vertices representing documents and distance is again defined as $1/c - 1$. The two distance measures are compatible, so that a single graph may be formed by the union of the arcs of G and G' and distance between documents may be calculated in terms of the length of the chain joining the two documents. Documents will not be adjacent to documents in a chain, so that

Term \ Term	A	B	C	D	E	F
A	1	0	0.2	0.5	0	0
B	0	1	0.2	0	0	0
C	0.2	0.2	1	0.4	0	0
D	0.5	0	0.4	1	0	0
E	0	0	0	0	1	0.3
F	0	0	0	0	0.3	1

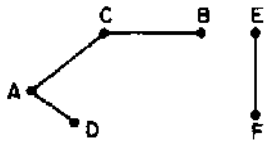

FIGURE 5. *Graph defined by matrix of term association measures.*

every chain will be via at least one term vertex. Neighborhoods for documents (or terms) are defined as all those documents (terms) which fall within a certain distance of the document (term).

The Symposium on Statistical Association Methods for Mechanized Documentation, held in Washington in 1964, brought together most of the investigators whose work has been described above, and others, equally interesting, whose work is more or less built on these foundations. Individual points of view and techniques were predominant, but one paper of a more general nature, by Kuhns (27), classified the various measures of association between properties and from this derived a generalized association measure which is the product of a parameter and the deviation δ of the observed data from expectation if the two properties were unassociated or independent. δ is given by

$$\delta_{AB} = N(AB) - N(A)N(B)/n$$

Most of the suggested coefficients are comprised in the general form

$$r_{AB} = \frac{\delta_{AB}}{\alpha}$$

where the parameter α specifies the coefficient. Coefficients may then be ordered as to α, i.e.,

Minimum $[N(A),N(B)]$

.
.
.

Intermediate values, e.g., $N(A) + N(B)$, $\sqrt{N(A)N(B)}$

.
.
.

Maximum $[N(A),N(B)]$

In the postscript to the symposium *Proceedings,* Guiliano distinguished three areas for further investigation—studies which, for the most part, have yet to be undertaken, or, at least, published. One is an appraisal, from a statistical point of view, of the various procedures and formulas, indicating the conditions under which each is applicable. The second involves carefully controlled experiments, using large collections of documents, to evaluate the usefulness of the proposed techniques. The third is to question that the techniques will be most useful in completely automated rather than merely machine-aided document retrieval involving some form of man-machine interaction.

Automatic Classification

Although many of the pioneers in associative retrieval suggested that document as well as term associations be used in retrieval operations, the notion that statistical (i.e., word counting) techniques could provide the basis for a document classification schedule is, at least from an experimental point of view, of more recent interest. Associative retrieval, as Doyle and others have pointed out, is simply an extension of coordinate indexing. In a request the original request terms and their associated terms are combined by the logical connectives "and," "or," and possibly "but not." If the file is in inverted (term) order, then the appropriate lists of documents assigned to each term in the request are isolated in the file and matched or combined as required, and the relevant document numbers extracted or ordered if relevance numbers are computed. This is a comparatively rapid procedure, since only a part of the total file need be scanned. Unfortunately computer retrieval systems use magnetic tape storage and in this case inverted order is not economical, particularly if the file is a growing rather than a static one. For this reason magnetic tape files are generally in document order, and, in a search, the entire file must be matched against the request to determine all relevant answers. In small-scale experiments, the problem of scanning time is insignificant; however, in considering the mechanization of a document collection of several hundred thousand items, it becomes an overriding consideration. Thus workers in the area of statistical association methods became concerned, as librarians had before them, with methods for classifying the documents of a collection so that the entire file would not have to be scanned in response to a request. Document

classification still involves some rewriting of tapes for a growing collection, but not so much as for an inverted file. In addition, a document classification determines a shelf arrangement for the file and hence can, as with conventional systems, provide multiple access to an indexed file. The representation of the classification groupings is of aid to a searcher in defining his request and exploring appropriate association trails.

An early theoretical paper in this area is Tanimoto's discussion of mathematical procedures for classifying qualitative data (*28*). The similarity of a pair of objects with respect to a set of attributes is defined by the formula used later by Doyle. All objects in the set are then considered as points in a semimetric space, with distance d between points defined as

$$d_{AB} = -\log_2 r_{AB}$$

If two points are connected when $d < \infty$, the points form a graph. The clusters of points in the graph determine the classification; the radius of clusters is determined by subjective judgment. Thus his approach is a forerunner of the model developed by Hillman.

Clump analysis (*29*), the creation of workers at the Cambridge Language Research Unit in England, is primarily a method for classifying documents. Starting with the document association matrix R, which is computed from the original document-term connection matrix C, clump analysis then finds subsets of objects (documents) with the property that there tends to be a stronger connection between two members of a group than between a member and a nonmember. Various types of clumps are defined, the most useful being R clumps, which allow multiple classification. A subset is an R clump if for any member of the clump the sum of the associations to the remaining members of the clump exceeds the sum of connections to documents outside the clump.

Borko's approach to document classification (*30*) is through factor analysis, a statistical technique which extracts from a correlation matrix such as the matrix R a smaller number of vectors which account for most of the variance. He begins with sets of high-frequency words assigned to a set of documents, determines the correlation coefficient for each term pair, and extracts a set of factors from the term-term correlation matrix which may be interpreted as the categories of the classification. Each category may be characterized by a vector of weights of each index term for that category. In the 1961 experiments, 10 factors were derived from a set of 618 psychological abstracts. For the factor labeled "academic freedom," as an example, the following terms had the highest weights:

girls,	0.74
boys,	0.73
school,	0.30
achievement,	0.20
reading,	0.18

Baker's latent class analysis (*31*) also begins with the document vectors of the matrix C and with the assumption that the documents can be divided into a number, e.g., p, of mutually exclusive subclasses, latent classes, or categories. The technique is concerned not just with term co-occurrences but with all possible term patterns. If documents have been assigned r index terms each, there are 2^r possible term patterns. Using observed frequencies as estimates of the probability of obtaining each of the various patterns, p ordering ratios are associated with each pattern of terms. The ordering ratio is interpreted as the probability that a particular pattern of terms will be possessed by documents of a particular latent class. The class which has the highest probability of generating the term pattern of a document is the one to which it is assigned. Baker feels his technique is superior to factor analytic techniques in that it is based on the whole pattern of index terms rather than co-occurring pairs. Retrieval is performed by considering the request as a vector of index terms, computing the probability that the term pattern will be possessed by each latent class, and selecting the one with the largest probability. In a later paper (*32*) Winter shows that a computer solution is feasible if Baker's method is slightly modified.

Another multivariate technique for document classification is Williams' discriminatory analysis (*33*). A user starts with a sample set of documents and assigns them to various subject categories of interest to him. Discriminating words for each category are selected on the basis of their consistency in occurring in and only in the documents of that category. Each discriminating word is assigned a weighting coefficient with respect to each category. Discriminant functions based on the mean frequencies, variances, and covariances of the discriminating words in the sample are used in conjunction with observed frequencies in a document to determine the distance of the document, considered as a point in multidimensional space, from the centroid of each category (see Figure 6). The original number of dimensions, or discriminating words, may be reduced, in the case of Williams' sample from 48 to 3, without loss of discriminating ability, and distances may be computed in the reduced space. The unique feature of this approach is that it allows classification of the same set of documents from many divergent points of view.

Doyle, an early investigator in associative retrieval, has in recent years turned to statistical technique of document classification as a more efficient approach to mechanization (*34*). He suggested that an automatically derived classification scheme will not be useful unless an organization and description of the classes is provided to the searcher. His approach has been through a grouping program, originally proposed by Ward, which satisfies these two requirements. Documents are represented by lists of terms; the program forms groups with maximum term list similarity in a series of "passes," each pass forming one group from two. Of the $(n!)^2/n(2^{n-1})$ possible paths for bringing the n items together, the Ward procedure selects the one which will bring together items of the greatest similarity the most rapidly. Labels are assigned on the basis that as far as possible they should describe all and only all items in the group. A hierarchy is imposed on the

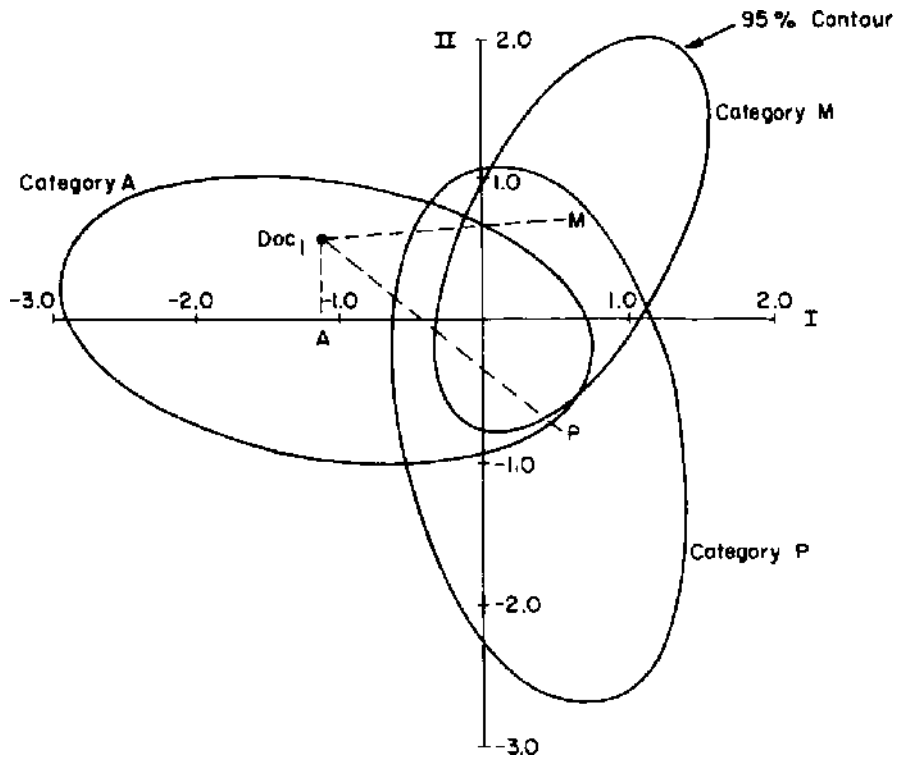

FIGURE 6. *Document 1 as a point in two-dimensional space. The broken lines indicate distances from the centroids of categories* A, M, *and* P. *(Courtesy of Mr. J. H. Williams, Jr.)*

classification as part of this procedure, since at each pass smaller groups are combined into larger ones. In a later paper Doyle and Blankenship (*35*) noted as a weakness of most proposed classification procedures that they use computer time in proportion to the square of the number of items. Such an approach is out of the question for a library of several hundred thousand documents, however well it may perform with small experimental files. He described a procedure in which the Ward program is applied to only a subset of the file and in which other less time-consuming algorithms are then used to modify the initial classification on the basis of the remaining documents.

Statistical Techniques Assessed

To summarize, statistical techniques, whether applied to associative retrieval or automatic classification, will be used primarily to improve recall of relevant documents by providing term or document association trails. In some cases a second motive will be the need to order documents as to probable relevance. Whether they actually perform these two functions depends on the validity of the models chosen.

Statistical associative retrieval is based on the assumption that term co-occurrences in text or in indexes imply a semantic relationship, and, conversely, a semantic relationship implies that the terms will co-occur frequently. The assump-

tions are further refined if one distinguishes first- and second-order associations, as suggested by Stiles and Guiliano. Term co-occurrences then are assumed to imply a contiguous (real-world) relationship and a real-world relationship implies frequent co-occurrence. Similar profiles of first-generation associates imply at least partial synonymy and partial synonymy implies similar profiles.

The validity of these assumptions has received little attention, with the exception of the Rubenstein and Goodenough tests (*36*) of the assumption that the greater the meaning similarity of two words, the greater the number of first-order associates they have in common. Ordinary English words were used and pairwise similarity judgments were made by two groups of college students. Another set of students wrote the sentences (100 per term) which constituted the contexts; sentences for each member of a pair were written by different subjects. The association measure used was

$$r_{AB} = \frac{N(AB)}{\min [N(A),N(B)]}$$

Plots of mean synonymy (measured on a scale from 0 to 4) versus r supported the hypothesis that the more similar the words in meaning, the more similar they are in contextual distributions. However, this relationship was strongest for highly synonymous pairs; for intermediate values the r values were almost constant.

Following Deese's notion of a purely associative meaning, one can ignore the semantic classification of the associated words and postulate simply that the associated terms do provide a key to relevant answers, and most tests of statistical associative retrieval have been of this nature. However, this approach begs the question: Do all relevant answers contain associated terms? One encounters here the entire problem of systems evaluation methodology, which is discussed elsewhere. O'Connor's test of automatic subject recognition (*37*) is to the point here. He found that the assumption that any subject relevant to a document can be identified by a few dozen clue words, including statistically associated terms, was not borne out by his study of documents relevant to the subject "toxicity."

Most investigators have now concluded that if a binary measure of association is used, the actual formula employed does not alter output significantly [see Ref. (*38*), for example]. However, whether a binary model is the most appropriate is still open to question. Baker's suggestion that document classifications should not be based solely on pairwise correlations, but on the entire pattern of assigned index terms, appears to have merit. Similarly all patterns of terms appearing with request terms, not just co-occurrences, might be considered in associative retrieval. Edmundson, in a study of mathematical models of synonymy (*39*), noted that synonymy is a ternary rather than a binary relationship; i.e., A is synonymous with B in the sense C. Similarly, from a single term, association trails will radiate in various senses or along various dimensions. Perhaps these senses can be distinguished by context, i.e., by taking more than two words at a time.

A similar criticism, mentioned earlier, is Paisley's that the association trails will have only a normative validity and so be of little value, since there is such a diversity of concept associations among subjects. Although this variability does exist

within the limits of a scientific or academic discipline, much greater uniformity of associative habits might be expected than in a free word association test. If a subject is told to associate a word at random with "pine," he might associate "picnic," recalling a personal experience. However, if a botanist is looking for books about pine trees, it seems unlikely his associations would be so erratic, and would probably confirm closely to the list in the Library of Congress subject headings list, i.e.,

Pine
See also individual species (Loblolly, Longleaf, Yellow pine)
Example under Coniferae, evergreens, timber, trees
—Diseases and pests
See also particular diseases and pests (Blister rust, Pine-sawyer)

No really adequate information exists on the variability of associations from user to user. A number of writers [Paisley (*3*), Maron (*49*), and Curtice (*40*)] have suggested that the system make some attempt to master the user's association, perhaps by a user profile, or that some form of user-system interaction be employed.

Statistical document classification is based on the assumption that the degree of association between two documents implies and is implied by the number of index terms or high-frequency words the two documents have in common. The validity of this assumption obviously depends to some extent on the quality of the initial indexing, both the type and number of assigned terms. But another question to be determined is whether a classification based on similarities between finite lists of document descriptors will provide the necessary document association trails. Human classification is based, not on the set of index terms assigned to a document, but on the totality of the document itself. Again, there is the question of which of the proposed statistical models is to be preferred. Librarians many years ago found a linear, pigeonhole classification inadequate without the multiple access provided by an alphabetic catalog. Many automatic classification schemes are an improvement over this approach in that they allow for the calculation of a set of relevance numbers for a document *vis-à-vis* a set of classes or for the placing of a document in several classes or for classification from various user viewpoints. The unanswered question at this point is whether classification, in any number of dimensions, can best be considered as some linear function of an assigned index term or whether more appropriate models remain to be discovered.

Some writers have felt that word counting is too simple minded an approach to determining term and document connections and that some type of syntactic analysis should be used. Salton includes this as one of the options in his SMART system, but no real evidence has yet been presented as to the effectiveness of syntactic matches. Furthermore, problems appear to exist in the development of adequate computer syntactic analysis techniques.

The second question to be asked about statistical association techniques is whether they are reliable, that is, whether they remain stable or change from collection to collection under random influences. More specifically, are the measures to be considered simply as descriptive for the collection at hand or as sample estimates for population values in some universe of text? If they are purely descrip-

tive, then nothing can be said about their reliability. However, when investigators indicate, even implicitly, that the larger the collection, the more useful the measures derived from it, they are regarding association measures as predictive rather than descriptive statistics, whose reliability depends on the size of the sample. Work is still to be carried out in determining the amount of variation to be expected in association coefficients derived from various size samples. As with any statistic, r values will be subject to sampling error; and until more is known about the underlying distributions of the proposed measures, it is not possible to say much about their reliability. In addition, one must ask how well the sample represents the parent population. Most statistical inference is based on random samples, and, as Doyle has pointed out, few libraries or document collections are random. Borko determined the reliability of factor analysis classes by deriving them from three different samples of documents. He found that the factors which emerged from the analysis were closely related to the data used in the study and that to the extent that the data base was an adequate sample of the total document collection, the factor-derived categories were reliable. In general it might be concluded that the reliability of a classification is affected by the number of documents used to set up the classes, the representativeness of the documents, the number of classes in the scheme, the number of index terms assigned to the documents, as well as the particular algorithm chosen.

The third question concerning statistical association is whether the techniques are practical. The major problem here is the time and cost involved in manipulating large matrices. Most of the suggested methods require a comparison of each term with every other term or a comparison of each document with every other document. This means $n(n-1)/2$ comparisons for a file of n terms or documents. Another problem, mentioned by Baker, is the computational accuracy of manipulations with large matrices. Special techniques for sparse matrices (where most entries are zero) may offer some solutions here [see, for example, Seidel (*41*)], but it seems likely that further investigators will turn, as Doyle has done, to the problem of reducing computer time involved in determining both term and document associations. The possibility of a special purpose device for handling associative processing has been made by Guiliano and implemented on an experimental scale in the analog electrical network devices, called ACORN devices, at Arthur D. Little. Also, Doudnikoff and Connor (*42*) described an optical coincidence scanner and hole counter which may be used to determine frequencies for collections up to 10,000 terms.

The use of English language terms as computer input also presents problems. If the full text or significant words in the title are used, considerable editing must be carried out, either manually or automatically, so that, for example, words with the same root but differing prefixes or suffixes, will be grouped together. Sanders (*43*) suggested replacing English with a formal language which would be more suitable for machine processing. The problem of which words may be considered nonsignificant also arises. Wallace found that common words, in addition to the unique vocabulary of a field, distinguishes subject matter (*44*). If assigned index terms are used as input, how much additional information is gained by associative

procedures? If human indexers assign terms, would not they be able to map associations as well as or better than a machine? The answer to this question perhaps depends on the size of the collection. Machines, having more consistent memories, would map more consistently in large collections, if practical algorithms could be evolved for their doing so.

Nonstatistical Methods

Automated mapping of association trails in a file of documents has, to the present time, been predominantly statistical, probably because of the suitability of computers for counting operations. However, the unresolved questions of the validity and reliability of the measures thus obtained from small files and the practical difficulty of deriving such measures from large files are likely to reawaken interest in nonnumerical methods. One such method, which has already proved itself operationally, is citation indexing. Here the association is from document to document, but documents are related not on the basis of matching term lists but on the basis of their bibliographic citations. A citation trail is followed by starting with a document known to be relevant and continuing through referring and referred to documents until a document not of interest is reached. Then a second trail is begun, and so on. The compilation of citation indexes can be automated beyond the editing of author and journal names to assure consistency. As developed by Garfield, a citation index gives, for each included paper, the later papers which have referred to it. Garfield suggested (*45*) that implicit references, i.e., references omitted fortuitously or otherwise, may eventually be handled by computer analysis. Price (*46*) has pointed out that citation indexes tend, at least directly, to join new papers to relatively recent ones, but by following the trail of associations back through successive generations of citations, the earlier literature can be reached. Another disadvantage is that papers will not be linked unless a publication explicitly links them. Thus a citation index will not discover connections not already known in the field.

A word association or thesaurus approach which recalls Bush's Memex has been described by Reisner and others in the collection *Some Problems in Information Science* (*47*). The optimal indexing language varies from user to user in the sense that whether a given pair of words is redundant (near synonyms) or ambiguous (homographs) depends on the point of view of the reader or speaker. Thesauruses are created by experts who attempt to guess which words a user will consider near synonyms, but it is not generally known how adequate the guess will be in actual operation. Reisner suggested as a solution a growing, man-machine thesaurus, which the users themselves build in the course of using a man-machine system (the AMNIP system). The user chooses from a composite thesaurus those relationships which correspond to his viewpoint. If the machine's store of associations proves inadequate, the user inserts his own, which are added to the system, leaving a trail for future users and thus enlarging the thesaurus. The system designer builds in the small initial thesaurus, but in time it becomes a composite list of near synonyms for the users of the system.

Another paper in this collection, by Abraham (*48*), deals with techniques for organizing a thesaurus which resemble Hillman's model of term relationships. The thesaurus is represented by a graph whose vertices correspond to the terms in the system and whose edges correspond to term-term relationships. The model assumes that pairwise synonymous and hierarchical relationships are defined. Then a single directed edge represents a hierarchical relationship and two edges with opposite directions between two vertices represent a synonymic relationship. Properties of leaf and lobe decomposition of a graph are used to (1) form clusters of mutually synonymous terms, (2) establish consistent hierarchies in the thesauruses, (3) and detect differences between thesauruses constructed by different methods.

The growing thesauruses appear to be in an experimental if not purely speculative stage, and the practical problems which may subsequently arise, as they have arisen with statistical association methods, cannot yet be defined.

Bush's Memex seems further from practical realization now than it did in 1945. The problem still remains whether any equipment can map association trails which will be valid for an individual user or whether diversity of interests among users preclude any but the most superficial common trails. If all diversity is built into the system, as with the growing thesaurus, the user may well be overwhelmed with a multiplicity of associations. Thus some form of man-machine interaction will likely be the next step in automated association mapping. However, the form of the interaction must be considered. That it requires less time and effort, on the part of the user, to select the subsets of term or labeled document classes which interest him than to select subsets of relevant documents from an actual document collection cannot be argued. However, it also cannot be argued that the reliability of full-text documents in determining relevancy is much greater than a single index term or combination of terms or classification label, considered as a document surrogate. The user's time will best be spent in evaluating sets of documents, not sets of terms or class labels. From this point of view, man-machine interaction, if it means the user will attempt to construct his own association trails without reference to the documents involved, will not necessarily improve performance. Future developments should lead to systems which will provide for the user the associative mapping capability of an open stack library where the search proceeds on the basis of the searcher's recognition of clues in the author, title, and subject catalog, the shelf classification, and the documents themselves.

REFERENCES

1. V. Bush, "As We May Think," *Atlantic Monthly*, 176, 101–108 (1945).
2. H. B. English and A. C. English, *A Comprehensive Dictionary of Psychological and Psychoanalytic Terms*, McKay, New York, 1958, p. 45.
3. W. J. Paisley and E. B. Parker, "Information Retrieval as a Receiver-Controlled Communication System," *Education for Information Science*, Spartan, New York, 1965, pp. 23–31.
4. J. Locke, *Concerning Human Understanding*, Chap. 25, 1689.
5. P. E. Jones, "Historical Foundations of Research on Statistical Association Techniques for Mechanized Documentation," Symposium on Statistical Association Methods for Mechanized Documentation, Washington, 1964, *Proceedings*, National Bureau of Standards Miscellaneous Publication 269, pp. 3–8.

6. R. S. Woodworth and H. Schlosberg, *Experimental Psychology*, rev. ed., Holt, New York, 1955, Chap. 3.
7. G. A. Miller, *Language and Communication*, McGraw-Hill, New York, 1951, Chap. 9.
8. B. C. Vickery, *On Retrieval System Theory*, 2nd ed., Butterworth, London, 1965, pp. 37–40.
9. J. W. Perry and A. Kent, *Tools for Machine Literature Searching*, Wiley (Interscience), New York, 1958, pp. 278–279.
10. L. S. Papier and E. H. Cortelyou, "Use of a Technical Word Association Test in the Preparation of a Thesaurus," *J. Doc.*, **18**, 183–187 (1962).
11. C. E. Osgood, G. Suci, and P. H. Tannenbaum, *The Measurement of Meaning*, Univ. of Illinois Press, Urbana, 1957.
12. J. Deese, "On the Structure of Associative Meaning," *Psych. Rev.*, **69**, 161–175 (1962).
13. B. E. Garskof and J. P. Houston, "Measurement of Verbal Relatedness: An Idiographic Approach," *Psych. Rev.*, **70**, 277–288 (1963).
14. G. U. Yule, "On Measuring Associations between Attributes," *J. Royal Stat. Soc.*, **75**, 579–642 (1912).
15. L. A. Goodman and W. H. Kruskal, "Measures of Association for Cross Classification," *Am. Stat. Assn. J.*, **49**, 732–764 (1954); **54**, 123–163 (1959).
16. M. Taube et al., "Storage and Retrieval of Information by Means of the Association of Ideas," *Am. Doc.*, **6**, 1–18 (1955).
17. M. E. Maron and J. L. Kuhns, "On Relevance, Probabilistic Indexing and Information Retrieval," *J. Assn. Computing Machinery*, **7**, 216–244 (1960).
18. H. E. Stiles, "The Association Factor in Information Retrieval," *J. Assn. Computing Machinery*, **8**, 271–279 (1961).
19. L. B. Doyle, "Indexing and Abstracting by Association," *Am. Doc.*, **13**, 378–390 (1962); "Semantic Road Maps for Literature Searchers," *J. Assn. Computing Machinery*, **8**, 553–578 (1961).
20. J. Becker and R. M. Hayes, *Information Storage and Retrieval: Tools, Elements, Theories*, Wiley, New York, 1963, pp. 370–376.
21. V. E. Guiliano and P. E. Jones, "Linear Associative Information Retrieval," in *Vistas in Information Handling*, Vol. 1 (P. W. Howerton and D. C. Weeks, eds.), Spartan, Washington, D.C., 1963, pp. 30–34.
22. V. E. Guiliano, "The Interpretation of Word Associations," in Symposium on Statistical Association, *Proceedings*, pp. 25–32 [see Ref. (5)].
23. G. Salton, "Some Hierarchical Models for Automatic Document Retrieval," *Am. Doc.*, **14**, 213–222 (1963); G. Salton and M. E. Lesk, "The SMART Automatic Document Retrieval System," *Comm. Assn. Computing Machinery*, **8**, 391–398 (1965).
24. G. Salton, "Associative Document Retrieval Techniques Using Bibliographic Information," *J. Assn. Computing Machinery*, **10**, 440–457 (1963).
25. J. Spiegel and E. M. Bennett, "A Modified Statistical Association Procedure for Automatic Document Content Analysis and Retrieval," in Symposium on Statistical Association, *Proceedings*, pp. 47–60 [see Ref. (5)].
26. D. J. Hillman, "Studies of Theories and Models of Information Storage and Retrieval;; Report No. 7, Graphs and Algorithms for Term Relations," Lehigh Univ., Center for the Information Sciences, Bethlehem, Pa., 1964; ". . . Report No. 8, The Structure of Document Relations," (1965).
27. J. L. Kuhns, "The Continuum of Coefficients of Association," in Symposium on Statistical Association, *Proceedings*, pp. 33–40 [see Ref. (5)].
28. T. T. Tanimoto, *An Elementary Mathematical Theory of Classification and Prediction*.. IBM, New York, 1958.
29. A. F. Parker-Rhodes and R. M. Needham, *The Theory of Clumps*, Cambridge Language Research Unit, Cambridge, England, 1961; R. M. Needham and K. Sparck Jones, "Keywords and Clumps," *J. Doc.*, **20**, 5–15 (1964).
30. H. Borko, "The Construction of an Empirically Based Mathematically Derived Classifi-

cation System," in *Proc. Spring Joint Computer Conf.*, **21**, 279–289 (1962); "Studies on the Reliability and Validity of Factor-Analytically Derived Classification Categories," in Symposium on Statistical Association, *Proceedings*, pp. 245–258 [see Ref. (*5*)].

31. F. B. Baker, "Information Retrieval Based on Latent Class Analysis," *J. Assn. Computing Machinery*, **9**, 512–521 (1962); "Latent Class Analysis as an Association Model for Information Retrieval," in Symposium on Statistical Association, *Proceedings*, pp. 149–156 [see Ref. (*5*)].

32. W. K. Winter, "A Modified Method of Latent Class Analysis for File Organization in Information Retrieval Work," *J. Assn. Computing Machinery*, **12**, 356–363 (1965).

33. J. H. Williams, "Results of Classifying Documents with Multiple Discriminant Functions," in Symposium on Statistical Association, *Proceedings*, pp. 217–224 [see Ref. (*5*)].

34. L. B. Doyle, "Is Automatic Classification a Reasonable Application of Statistical Analysis of Text?" *J. Assn. Computing Machinery*, **12**, 473–489 (1965).

35. L. B. Doyle and D. A. Blankenship, "Technical Advances in Automated Classification," in American Documentation Institute, Annual Meeting, 1966, *Proceedings*, Adrianne Press, Santa Monica, Calif., 1966, pp. 13–72.

36. H. Rubenstein and J. B. Goodenough, "Contextual Correlates of Synonymy," *Comm. Assn. Computing Machinery*, **8**, 627–633 (1965).

37. J. O'Connor, "Automatic Subject Recognition in Scientific Papers," *J. Assn. Computing Machinery*, **12**, 490–515 (1965).

38. J. Tague, "An Evaluation of Statistical Association Measures," in A.D.I. 1966, *Proceedings*, pp. 391–398 [see Ref. (*35*)].

39. H. P. Edmundson, "Mathematical Models of Synonymy," Systems Development Corporation, SP-1976/000/01, 1966.

40. R. M. Curtice, "Experiments in Associative Retrieval," in A.D.I. 1966, *Proceedings*, pp. 373–384 [see Ref. (*35*)].

41. M. Seidel, "Threaded Term Association Files," in Symposium on Statistical Association, *Proceedings*, pp. 167–176 [see Ref. (*5*)].

42. B. Doudnikoff and A. N. Conner, "Statistical Vocabulary Control with Optical Coincidence," in Symposium on Statistical Association, *Proceedings*, pp. 177–180 [see Ref. (*5*)].

43. J. Sanders, "Document Association and Classification Based on L-Languages," *J. Assn. Computing Machinery*, **12**, 249–253 (1965).

44. E. M. Wallace, "Rank Order Patterns of Common Words as Discriminators of Subject Content in Scientific and Technical Prose," in Symposium on Statistical Association, *Proceedings*, pp. 225–229 [see Ref. (*5*)].

45. E. Garfield, "Can Citation Indexing be Automated?," in Symposium on Statistical Association, *Proceedings*, pp. 189–192 [see Ref. (*5*)].

46. D. Price, "Networks of Scientific Papers," *Science*, **149**, 510–515 (1965).

47. P. Reisner, "Semantic Diversity and a Growing Man-Machine Thesaurus," in *Some Problems in Information Science* (M. Kochen, ed.), Scarecrow, New York, 1965, pp. 117–130.

48. C. T. Abraham, "Techniques for Thesaurus Organization and Evaluation," in *Some Problems in Information Science* (M. Kochen, ed.), Scarecrow, New York, 1965, pp. 131–150.

49. M. E. Maron, "Mechanized Documentation: the Logic behind a Probabilistic Interpretation," in Symposium on Statistical Association, *Proceedings*, pp. 9–13 [see Ref. (*5*)].

JEAN TAGUE