EXHIBIT J

JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE

asis

PUBLISHED BI-MONTHLY BY
AMERICAN SOCIETY FOR INFORMATION SCIENCE

SEPTEMBER, 1976, Vol. 27, No. 5

OCTOBER, 1976, Vol. 27, No. 6

Published bi-monthly by the American Society for Information Science. Second class postage paid at Washington, D.C. and at additional mailing offices.

An *n*-Dimensional Retrieval Model*

This paper reports a technique which expands W. Goffman's Indirect Method search strategy by using means other than index terms to reflect document content. The four basic measures of document relatedness were: (1) Index terms, (2) Journals in which the documents appeared, (3) Closeness of the authors of the documents and (4) Closeness of citations. In the experiment a distance function between documents is defined, based on the properties of the documents themselves. The proximity between pairs of documents is then determined by calculating the Euclidean distance.

Based on the results of this experiment, authors or authors with journals are the best measures of document relatedness. This aspect should be explored further.

Donald B. Cleveland
Case Western Reserve University
Cleveland, OH 44106

• Introduction

Search strategies for automatic document retrieval systems are usually based on Boolean functions, although it has been shown that such functions are inadequate (*1, 2, 3*). In these models the assumption is made that a relation between each document and the query is a sufficient condition for optimal retrieval results. It is also assumed that each document in a file is independent of other documents.

Goffman (*4*) has shown clearly that documents are *not* independent of each other. In fact, a search strategy which takes account of document relatedness and mutual dependency results in a retrieval model that is superior to traditional models.

The experiment reported in this paper presents a technique of combining Goffman's search strategy with means other than index terms for reflecting document content. The results suggest that such a combination may be used to construct effective retrieval systems.

*This work was supported in part by U.S. Office of Education Grant O.E.G.-0-72-7511(320)

• Document Relatedness

The concept of the inter-relatedness of documents in a file has been discussed for a number of years. One example is *probabilistic indexing* which associates a probability of relevance with index terms. Maron and Kuhns (*5*) proposed this technique, which essentially works as follows: A request for information is given to a computer which makes a statistical inference and derives a number called the relevance number for each document. This relevance number is a measure of the probability that the document will satisfy the request. The result of the search is an ordered list of those documents that satisfy the request, ranked according to their probable relevance.

Also, given a request consisting of one or more index terms, the computer can elaborate on the request to increase the probability of selecting relevant documents that would not have been selected otherwise. An *a priori* statistical distribution connects related documents in a file.

Another idea was suggested by Doyle (*6*). Concerned with the idea of the interdependence of documents in a file, Doyle proposed that the documents in a file be dis-

played in an *association map* which would express the relationship of the subject matter in the documents. The user is presented first with a map of concepts which are expressed in the file. He then chooses specific direction according to interest and need. The available documents are then displayed so that he can decide which to investigate.

One of the classical proposals for structuring retrieval files independently of index terms is *bibliographic coupling,* attributed to M.M. Kessler (*7*) who worked with citations in the early 1960's at M.I.T. Documents are linked together by the references the authors include in the paper to form networks of papers on a subject.

Another idea, called "context clues," was proposed by a group of researchers at the Institute of Library Research, under M.E. Maron (*8*). Context clues are those items of information that describe various objective properties and relationships which hold for individual documents; *e.g.:* authors, reviewers, professional societies and the like.

Likewise, there is a "context" that surrounds each library user who comes in for service. Indeed, there seems to be a total, complex environment in which librarian, documents, and users interact—and this environment can give "clues" to documents that are relevant.

These ideas, and others like them, are attempts to break out of the constraints imposed by classical indexing approaches.

• Goffman's Method

Basically, the Goffman *Indirect Method* model reflects the concept that the exposure of a user to any given document modifies for him the relevance of succeeding documents. Thus, a file of documents can be classed (or ordered) on the basis of conditional probabilities between each pair in the file in terms of one being relevant to a stated query, given that its *predecessor* is relevant to that query. Content relatedness between each pair is the basis of the file structure.

For a given query this search strategy selects answer sets, chained together on the basis of a quasi-distance between each pair of documents. The chain represents the logical reading order of the documents.

The file is structured independent of any query. Obviously, a major key to the success of this model lies in determining, or measuring, the relatedness property between pairs of documents. Goffman used the co-occurrence of index terms as the measure. The present experiment extended the *Indirect Method* by representing documents in an n-dimensional Euclidean space with n measures of relatedness.

• The Notion of Document Distances

In the literature dealing with statistical indexing approaches, especially the associative retrieval methods, there are many references to index spaces, vectors and distances. Such discussions are usually launched without any clearly defined spaces or any indication that the conditions for a distance exist. The space is not defined and an essential ingredient is usually missing—an axiomatic definition of distance based on the properties of the objects involved. A distance function is often implied without a properly defined unit or any valid assignment rule.

In the experiment reported here, a distance function between documents is defined based on the properties of the documents. It is then shown that these documents can be thought of as existing in n-dimensional Euclidean space, with n representing n measures of the content relatedness between pairs of documents. The proximity between pairs of documents is then determined by calculating the Euclidean distance.

• The Mathematical Model

The Mathematical Model is the simple, ancient one of a distance in Euclidean space. The use of index terms as a measure of document-relatedness might be represented as a directed line segment between document X_i and document X_j. Suppose a second measure is added so that two measures are used to determine the relatedness of each pair of documents in the file. Now there is a relatedness measure between document X_i and document X_j expressed in two dimensions. Using the formula for distance, we can express the distance between X_i and X_j as

$$D(X_i,X_j) = \sqrt{(X_j-X_i)^2 + (Y_j-Y_i)^2}$$

where $(X_j - X_i)$ is the first measure and $(Y_j - Y_i)$ is the second measure. In general, the measures could be N and the distance would be expressed as

$$D(X_i,X_j) = \sqrt{(X_j-X_i)^2 + (Y_j-Y_i)^2 + \ldots + (N_j-N_i)^2}$$

It is assumed that the measures are linearly independent and mutually orthogonal, creating an orthonormal reference system in n-dimensional Euclidean space.

This is a true distance function between documents based on the *properties* of the documents. The proximity between pairs of documents is determined by calculating the Euclidean distance.

• The Experimental Model

The objective of the experiment was to test relative retrieval effectiveness using different measures of document relatedness. In order to carry out such a test, certain ingredients are needed. These are:

1. One or more queries,
2. A document file which contains answers to the queries,
3. Some way of measuring the content-relatedness between the documents,
4. A means of determining relevance.

Two randomly selected scientific papers were taken to be queries for the system. That is, the topics in those papers were considered as areas about which information was to be sought.

Usually, scientific papers have a set of references which the author dealt with when he was writing the paper. These were the items he felt were relevant to his topic. They include the authorities and the previous works he used to write the paper, and those out of a probably much larger set, he chose as being the most important or relevant. We can consider these references as being answer sets to the query, which is the paper itself.

Thus, we have queries, files of documents, various means of measuring relatedness (to be tested) and authoritative judgments of relevance. Taken together, these elements may be considered as a simulated information retrieval system for testing a search strategy and its relatedness measures.

• Measures

The measures of document proximity used to structure the file in order to carry out the *Indirect Method* search strategy can be expressed in terms of Cartesian coordinates as follows:

1) *X-axis*–Keyword co-occurrence between the documents in the file. This is the measure used in the original *Indirect Method* experiment and is, of course, the most obvious measure. Documents with similar index terms probably have similar information content.
2) *Y-axis*–The relatedness between the journals in which individual documents appear. If the articles in journal i cite articles in journal j more than any other, then it is reasonable to assume that the subject matter of j is more like the subject matter of i than is the subject matter of any other journal *in the sample*. This is an indication of the relatedness of the articles in the two journals.
3) *Z-axis*–The relatedness between the authors of the documents. The relatedness among the authors in the data set was established by determining the extent of common authorship on various topics. If the author of document X_i usually writes about the same topics as the author of document X_j, there is a high probability that document X_i and X_j are related in information content.
4) *W-axis*–The commonality of citations between the documents. It is assumed that closely related documents will have closely related citations.

An automatic word frequency technique was used to get the index terms measure. This technique has been used successfully in documentation studies at Case Western Reserve University for several years. Its basic form is described by Goffman (*4*).

The resulting lists of index terms were used to construct a matrix of relatedness between each pair of documents in the file. The numerical value was calculated as follows:

$$p_{ij} = \frac{m(X_i \Lambda X_j)}{m(X_i)} ,$$

where $m(X_i \Lambda X_j)$ is the number of index terms common to document X_i and document X_j. $m(X_i)$ is the total number of index terms for document X_i.

The second measure was based on the journals in which the documents appeared. There were 16 different journals in the data set. Approximately 30,000 citations, all the citations for a one year period, were examined. The result was a frequency list of citations for each of the 16 journals, giving the total citations to other journals in the data set.

For the purpose of constructing the measure, the top 3.5 percent of each distribution was arbitrarily chosen as representative of the most cited journals for each particular journal. Thus, connected with each journal was its journal citation profile. The measure between journal J_i and journal J_j was defined to be

$$Q_{ij} = \frac{n(J_i \Lambda J_j)}{n(J_j)} ,$$

where $n(J_i \Lambda J_j)$ is the number of cited journals common to the profiles of journal J_i, and journal J_j and $n(J_i)$ is the number of journals representing the journal citation profile of J_i.

Authors were the next concern. From *SCIENCE CITATION INDEX®* a list consisting of other authors

who cited each author in the data set was compiled. The number of authors commonly citing authors A_i and A_j for a given topic suggested a profile of relatedness between A_i and A_j. The measure between each pair of authors in the data set was then determined as follows:

$$R_{ij} = \frac{\emptyset(A_i \Lambda A_j)}{\emptyset(A_i)}$$

where $\emptyset(A_i \Lambda A_j)$ is the number of citing writers common to the profiles of author A_i and author A_j and $\emptyset(A_i)$ is the number of writers representing the author profile of author A_i.

Finally, the measure of citations was calculated as follows:

$$S_{ij} = \frac{\Theta(C_i \Lambda C_j)}{\Theta(C_i)} \quad ,$$

where $\Theta(C_i \Lambda C_j)$ is the number of citations common to document C_i and C_j,, and $\Theta(C_i)$ is the number of citations representing document C_i.

• Data Sources

The articles for the experiment were picked in the following way: A "random" walk was made through the open periodicals stacks of the Health Sciences Library at Case Western Reserve University and ended when a pleasing color of binding was detected. The nearest volume was pulled down and flipped open twice. This selection yielded Volume 123 of *The Journal of Infectious Diseases* and the two articles, "Precipitin Responses to Rubella Vaccine RA 26/3" by George L. Bouvier and Stanley A. Plotkin, and "In-vitro and In-vivo Studies of Resistance to Rifampin in Meningococci" by Theodore C. Eickhoff.

All the references in the two articles were used as the experimental data file, except one which was a monograph not readily available.

There were 26 articles, 13 from each of the two query articles. The 26 articles, plus the two query articles, represented 16 different journals and included 69 different authors.

The other data sources included a one year's run of each of the 16 different journals, and for the 69 authors, one year of the *SCIENCE CITATION INDEX®*.

Using these measures, four basic matrices resulted, showing the relatedness between the 26 documents in the file in terms of:

1. Closeness of index terms between the documents;
2. Closeness of the journals in which the documents appeared;
3. Closeness of the authors of the documents; and
4. Closeness of citations.

Vectors were also calculated showing the direct, or Boolean, relatedness between the two query articles and each document in the file. This was done for each of the four basic matrices.

• Distance Matrices

At this point, the four matrices showed the relatedness between each pair of documents in terms of the four basic measures with values between 0 and 1.

It was now necessary to convert these matrices into distance matrices and combine them, using the Euclidean distance formula. If the measure value between document X_i and X_j was greater than some chosen threshold, then the distance between the pair was defined as being unit distance one. The following tactic was employed to convert each of the four basic matrices into distance matrices:

Step One: Arbitrary thresholds were picked for each matrix in terms of the calculated numerical values. In actual practice the thresholds would depend on whether a fine or a broad scope of retrieval is desired. For purposes of experimentation, it is only necessary that the thresholds be held constant throughout the experiment. The thresholds picked were .14 for the index terms, .50 for the journals, .08 for the authors and .01 for the citations. Any relatedness values that fell below these thresholds were considered zero.

Step Two: Go along the row of document X_i and assign a unit distance of one to each document X_j which is above the threshold.

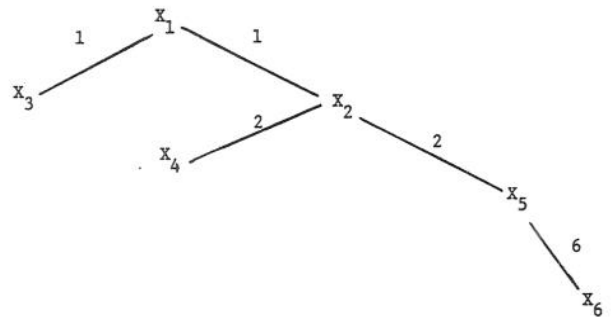
Step Three: For each document X_j that is a distance of one from document X_i, go along the row of X_j and assign a distance of two to each document that is above the threshold, provided it is not already of distance one from document X_i.

Step Four: Continue this procedure until all documents have a distance from document X_i. Those documents with zero relatedness values are considered to be of infinite distance.

Step Five: Repeat the procedure for all i.

Step Six: Repeat the total procedure for all n basic matrices.

The results are links of documents for each n basic matrix. These sequences reflect the smallest communication chain between elements, hence a quasi-distance. Graphically, it looks like this:



Thus, four basic distance measures were created, representing the four basic measures under consideration. Since a quasi-metric space existed, the objective now was to combine these orthogonal measures into various one, two, three and four-dimensional measures, using the Euclidean distance formula to determine the shortest chain between neighborhoods of documents in each dimension. Eleven matrices resulted.

Terms and Journals
Terms and Authors
Terms and Citations
Journals and Authors
Journals and Citations
Authors and Citations
Terms, Journals and Authors
Terms, Journals and Citations
Terms, Authors and Citations
Journal, Authors and Citations
Terms, Journals, Authors and Citations

The rationale behind the testing was this: The document file consists of 26 documents. Thirteen documents form an exhaustive answer set for one test query and the other 13 documents form an exhaustive answer set for the other query, provided of course, that none of the references in set *a* are relevant to query *b* and vice versa for set *b* and query *a*.

Therefore, a test consists of presenting queries to the system, using a particular relatedness measure or a particular combination of measures and observing how close the retrieval results approach the ideal.

Two points were important: (1) Exhaustiveness of retrieval, and (2) Exclusion of non-relevant documents.

The distinction between a "Boolean" search and an *"Indirect Method"* search should be made clear. Some form of Boolean operation is the most basic of techniques. (In the experiment reported here the index terms used to represent the "query" article made up the search vectors.) For a Boolean search, a query is compared with *each document* in the file, using any Boolean operation desired. The relevance of one document is entirely indepent of the relevance of all other documents.

With the *"Indirect Method,"* the query simply serves as an entry point to the file. Once a relevant document is found, the remaining retrieved documents are determined by internal file structure, *independently* of the query. Relevance is not a zero or one comparison between the query and each document, but is based on a conditional probability of relevance between the documents in the file.*

Measure	Relevant Retrieved	Relevant, not Retrieved	Non-relevant, Retrieved
Authors	26	0	0
Authors-Journals	26	0	0
Authors-Terms	25	1	0
Authors-Terms-Journals	25	1	0
Citations	22	4	0
Journals-Citations	22	4	0
Authors-Citations	22	4	0
Journals-Authors-Citations	22	4	0
Terms-Journals-Authors-Citations	22	4	0
Terms-Citations	21	5	0
Terms-Journals-Citations	21	5	0
Terms-Authors-Citations	21	5	0
Terms (Boolean)	13	13	0
Terms-Journals	25	1	12
Journals	26	0	26

Fig. 1. Tabulation of Results—Both Queries Combined

• Observations

1. The *Indirect Method* gave better results than a straight Boolean search. This was true both in the repetition of the original *Indirect Method* (using chains) and with the distance technique.
2. The *Indriect Method* distance techinque gave better results than the chaining method with certain combined measures. This was not true of index terms alone.
3. The best retrieval results came from:
   a. Authors alone,
   b. Authors and journals in combination.
4. The next best retrieval results were:
   a. Authors and Terms,
   b. Authors, Terms and Journals.

• Conclusions

Based on the results of this experiment:

1. Authors or authors with journals are the best

*For a full description of this technique, the reader is referred to Goffman's paper.

measure of document relatedness. Measures other than index terms can be used for retrieval.

2. The *Indirect Method* is still best, no matter which form it takes (chains or distances).

3. The distance technique actually worked better than the chaining method, and it is simpler to use.

• Implications

The importance of authors and the combination of authors and journals was the most interesting result of this experiment. It is an area that should be explored further.

For example, it is possible to visualize a library catalog based on author networks. Knowing an author on a subject of interest, a patron might enter the catalog with that author and find other authors who write on the subject. This is not a subjective subject heading cross-reference, but is based on a quantitatively determined author network. Such networks could be machine-constructed from existing data bases.

Another implication lies in the use of computers in information retrieval systems. Automatic indexing has never been completely successful. One reason is the input and processing problem involved with total text manipulation. Why not forget this full text approach and establish retrieval systems based on authors and the journals in which they publish? If desired, another dimension, based on abstracts or manually assigned index terms might be used, but the index terms would play a secondary role.

Obviously, the experiment reported here is only a pilot project, based on a small set of documents. However, the positive results are encouraging enough to warrant a larger investigation, pointing toward possible operational systems.

References

1. **Goffman, W.** 1964. "A Searching Procedure for Information Retrieval." *Information Storage and Retrieval.* 1964; 2: 73-78.
2. **Verhoeff, J.; Goffman, W; Belzer, J.** 1961. "Inefficiency of the Use of Boolean Functions for Information Retrieval Systems." *Communication of the Association for Computing Machinery.* 1961; 4: 557-559.
3. **Goffman, W.** 1964. "On Relevance as a Measure." *Information Storage and Retrieval.* 1964; 2: 201.
4. **Goffman, W.** 1968. "An Indirect Method of Information Retrieval." *Information Storage and Retrieval.* 1968; IV; (4): 361-373.
5. **Maron, M.E.; Khuns, J.L.** 1960. "On Relevance, Probabilistic Indexing and Information Retrieval." *Journal of the Association for Computer Machinery.* 1960; VII: 3.
6. **Doyle, L.B.** 1961. "Semantic Road Maps for Literature Searches." *Journal of the Association for Computing Machinery.* 1961; VIII.
7. **Kessler, M.M.** 1963. "Bibliograpic Coupling Between Scientific Papers." *American Documentation.* 1963 January; XIV.
8. **Maron, M.E.; Shoffner, R.M.** 1969. *The Study of Context: An Overview.* Berkeley, CA: The University of California. 1969.