VIII. Bibliographic Data as an Aid to Document Retrieval

J. W. McNeill and C. S. Wetherell

Abstract

The hypothesis of this project is that bibliographic data added to a SMART style document collection will improve retrieval effectiveness. Two uncommon kinds of bibliographic data, authors and place of publication, are used to build concept matrices. These matrices are used with associated queries in a typical retrieval environment involving relevance feedback. With the aid of a new statistic, it is found that these matrices actually do aid retrieval.

1. Introduction

Intuitively, bibliographic information is one of the most valuable tools available for a search of technical reference material. This fact is recognized by several authors. Salton [10,11] describes a general method for incorporation of bibliographic data into a SMART style retrieval system. Garfield [2] writes of the importance of citation indices for literature searches in the sciences. His work has had some practical effect, as the Science Citation Index published by the organization he heads is now a standard library item after only six years of publication. Recently, Garfield has announced automated search techniques of current literature, available for a modest fee. These too are based on the citation index. Finally, Kessler [4,5,6,7,8,9] discusses the use of bibliographic coupling as a retrieval method. He also notes the importance of the journal of publication as a clue to the content of technical documents.

In report ISR-12, Amreich, Grissom, Michelson, and Ide [1] follow Salton's suggestion and attach a citation index to the concept matrix of the ADI document collection. They showed that retrieval is about as efficient using the index alone as it is with the original matrix, and that combining the two types of terms results in a significant improvement in retrieval effectiveness. This result is perhaps not typical since the ADI collection, although small, is heavy in cross-reference.

The present project originally expected to check the results of Amreich, _et al._, with another less heavily cross-referenced document collection and then to extend the work toward that of Kessler. However, the collection chosen has so few cross-references that its citation index is, for practical purposes, null. This fact made it advisable to use other easily available bibliographic data.

Several other thoughts reinforced this decision. First, although bibliographic data in all its forms have long been recognized as a tool for document retrieval, only citation indices and their derivatives, coupling indices, have actually been tested for use in a mechanized system. Two other major sources of information, author and place of publication, have been neglected in the literature.

Second, bibliographic information has a double practical importance. It is easy to obtain such information through automated methods. Also, bibliographic data, even more than keywords or subject indices, reflect the position which the author feels the document holds in the present literature. Kessler discusses this point and his examples make it very clear that bibliographic data can be used to chart the mainstreams of physics. Although subject indices certainly have their uses, by their very construction, they cannot portray this implicit but usually accurate evaluation of a paper's standing in the literature.

The hypothesis upon which the project is based is the following: Bibliographic information other than that embodied in the citation index will, when added to a conventional concept matrix, improve retrieval effectiveness. This improvement may be demonstrated through the use of a statistic developed in this paper. The project is a test of this hypothesis.

2. The Experiment

The document collection used to test the experimental hypothesis is the small MEDLARS collection. This is a set of 273 documents concerning various aspects of medicine and 18 associated queries. The concept matrix which describes the collection is a word-form matrix consisting of approximately 500 concepts and a basic concept of weight 12. The practical weight range is 12 to 60.

The 18 queries each consist of two parts. The first is a query concept vector which is constructed exactly like a document vector, except that it is usually much shorter. In addition, each query has associated with it a set of relevant documents. These relevance judgments are used to calculate the efficiency of a retrieval method applied to the collection.

The basic direction of this project is the compilation of six new concept matrices describing the collection. These matrices are numbered 1 through 6. The number 0 matrix is the original word-form matrix. Along with the new matrices, new query sets are constructed. These query sets operate with the same sets of relevance judgments as the originals, but have concept vectors drawn from their respective concept matrices.

The bibliographic data upon which the matrices are based are the following:

Matrix	Data
1	Author of document
2	Author of citing document
3	Author of original or citing document
4	Original place of publication
5	Place of publication of citing document
6	Place of publication of original or citing document

For matrix 1, the authors of all the documents in the collection are listed. Authors of two or more documents are given concept numbers. Concept vectors are generated for documents in the standard way, with each concept having a weight of 12. For matrix 2, a similar procedure is followed using a list of all the authors who cite a document in the collection.

However, authors may cite, in different papers, a given document more than once, so that while the basic weight remains 12, the actual weight for a concept reflects the number of times the document in question has been cited by the author associated with the concept.

In matrix 3, the above two lists are combined. Every author who is associated with two separate documents is assigned a concept. The basic weight for a citation author is 6 and for an original author 12. The weight for a concept is computed by summing basic weights for all of the author's contacts with the document in question.

Matrix 3 is larger than a simple union of matrices 1 and 2. If author A wrote only document 23 and referenced only 211, he will not appear in either matrix 1 or 2, but he will appear as a concept in matrix 3. This does, in fact, add a great deal of information to matrices 3 and 6. Matrices 4, 5, and 6 are constructed in the same manner from place of publication information.

Query sets are needed to operate with these matrices. Each set contains 18 queries, corresponding to the original 18, and each of the 18 has the same set of associated relevant documents as the corresponding original. Construction of query i for matrix j proceeds as follows:

1. The union is taken of the document vectors in relevance set i included in matrix j to form a list of concepts relevant to query i.
2. If the list has less than four elements, enough random concepts are added to make four.

3. If the list has four or more elements, concepts are deleted using a random binary distribution. (deleting concepts corresponding to the zeros of the distribution). At least four of the original list are left in each reduced list. Finally, two concepts are added at random.

4. If the list is null, the query is assumed to be null.

5. Weights for all concepts are set to twelve.

6. The random concepts are chosen from the matrix j.

The rationale for this scheme is simple. It is assumed that queries will contain some concepts which are relevant to the documents which it is to select from the collection, and that the query will also contain some concepts which are not relevant (i.e., are noise concepts). Authors of queries are likely to know some of the authors (or publications or subjects) in the field they are investigating and they are also likely to make some mistakes in their query construction. It is also assumed that a difference in weights will not alter the query operation after four iterations, so that it is safe to set all query weights to 12.

One point should be emphasized concerning the construction of these matrices. An author or place of publication which has relevance to only one document has not been assigned a concept. Thus, retrieval for many queries of the form "I only know it was written by J. J. Smith" or "It appeared in the Czech Journal of Sedimentology" will not operate successfully. In this small data base, several queries for the constructed matrices vanish because of this requirement. However, in a large data base, this problem is less acute, since there are then fewer authors of only one document who are not referenced or do not reference, and few journals with only one relevant document. On the other hand, the same reasoning implies that adding concepts for single authors or single documents would probably not be very expensive.

The experiment is conducted in a "typical retrieval environment". Unfortunately, there is no experience to indicate what such an environment is. The assumption is made that this environment would include a standard SMART retrieval system, utilize the cosine measure of document similarity, and use some type of relevance feedback. The relevance feedback equation used is

$$q_{i+1} = q_i + r_1 + r_2 - n_1$$

where q_i is the old query, q_{i+1} is the new query, r_1 and r_2 the first two relevant documents, and n_1 the first nonrelevant document retrieved. Only documents in the top five retrieved are used in this equation. Each query is iterated four times (the original query constitutes the first iteration).

The matrices and query sets have been designed so that they may be combined by concatenation of corresponding document and query vectors. The experiment consists of running various combinations of matrices and query sets against one another and measuring the retrieval effectiveness of each pair. The original word-form matrix and query set is used as a control group. A matrix-query set pair will improve retrieval only if it does better than the control group. If none of the matrices which include bibliographic data do better than the control group, the hypothesis will have been shown to be false.

The matrix-query set combinations run are

Matrix i vs query set i, i = 1,...,6.
Matrix 0i vs query set 0, i, 0i, i = 1,...,6.
Matrix 36 vs query set 3, 6, 36.
Matrix 036 vs query set 0, 03, 06, 36, 036.

Concatenated matrices and query sets are denoted by listing their elements in order. Thus, matrix 02 is the matrix made by combining matrix 0 and matrix 2.

The nature of the document vectors, and the query vectors after several iterations is illustrated with the following example (concept) numbers are shown followed by weights):

Document 268 (data set 2)
6102 12 6119 12 6123 24

Query 8 for data set 2 (Iteration 1)
6102 12 6110 12 6119 12 6129 12

Query 8 for data set 2 (Iteration 4)
6102 36 6110 12 6116 12 6119 60
6125 24 6129 24

3. The Statistical Measure

The hypothesis proposed can be considered validated only if it can be shown that the expanded data bases actually produce better retrieval than the original data base. To this end some measure of retrieval effectiveness is needed. The measure chosen is a sign test based on rank recall.

The rank recall for a query is calculated by the formula

$$rr = \frac{\sum_{i=1}^{n} i}{\sum_{i=1}^{n} r_i}$$

where n is the number of relevant documents for the query in question, and r_i is the retrieval rank of the i-th relevant document. The measure varies from 1 for the best possible retrieval to 0 for the worst.

The sign test is calculated in the following manner. Rank recall is calculated for each query of the control group (rr_g^i) and of the test group (rr_t^i). The sign value is

$$S = \sum_{i=1}^{18} \mathrm{sgn}(rr_t^i - rr_g^i)$$

where sgn is the signature function. The difference is calculated to within a standard error of 0.005 on either side.

If S is non-negative, retrieval effectiveness is at least as effective for the test group as for the control group. If the absolute value of S is greater than or equal to 6, there is almost certainly a significant difference between the test group and the control group. It is reasonable to assume that is the absolute value of S is greater than or equal to 3, there is probably some difference between the groups. The direction of the difference depends on the sign of S.

A second statistic of interest is obtained when matrix i is run against query set i. In this case, retrieval is not good enough to bring S above -10 or so. This is so because the bibliographic data matrices all have at least 60 null document vectors, and do not contain enough information to compete with a matrix of a size equalling that of the basic MEDLARS matrix. However, the question arises whether these query set-matrix pairs do better than random queries might. The retrieval method used only does better than random if the rank recall for a query is larger than

$$d_j = \frac{\sum_{i=1}^{n} i}{\sum_{i=1}^{n} N_i}$$

where n is the number of relevant documents for the query and N_i is the document identification number of the i-th relevant document. Using the d_j in place of the rr_g, the statistic D is calculated in exactly the same manner as S. Again D must be non-negative if these queries are judged as performing better than random more than half the time.

To facilitate a graphic illustration of the most significant results, average recall precision curves are presented. The average value of recall and precision is computed over all 18 queries for 14 different cutoff levels using the following formulas:

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}}$$

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of documents retrieved}}$$

The cutoff values are chosen to provide results over the entire range of possible recall precision pair values.

4. The Results

The results of the experiment can be summarized using the S and D statistics. Table 1 shows the results of running query set i against matrix i, i = 1,...,6, and 36.

i	S	D
1 vs 1	-18	-2
2 vs 2	-17	6
3 vs 3	-14	11
4 vs 4	-14	15
5 vs 5	-18	14
6 vs 6	-11	16
36 vs 3	-17	10
36 vs 6	-12	16
36 vs 36	- 9	16

Statistics for Bibliographic Matrices Alone

Table 1

In table 2, the results of running matrix 0i against query sets 0, i, and 0i are presented.

i	S for 0	S for i	S for 0i
1	0	-10	Not available
2	-1	- 8	-3
3	0	- 3	1
4	0	1	5
5	-2	- 6	-1
6	1	- 9	4

Statistics for Combined Bibliographic and Original Matrices

Table 2

In table 3, the results of running various query sets against matrix 036 is presented.

Query Set	S
0	2
03	4
06	2
36	-4
036	5

Statistics for Combined Matrix 036

Table 3

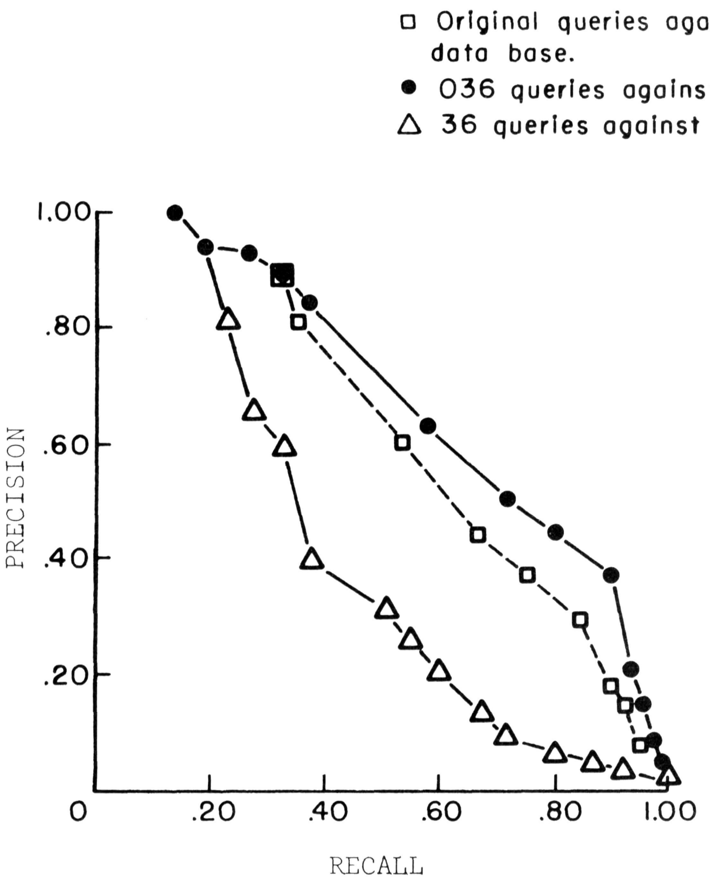
Recall precision curves are presented in Fig. 1 for the following three sets of results:

1. Original queries against original matrix
2. 036 queries against 036 matrix
3. 36 queries against 36 matrix

5. Conclusions

The most important conclusion to be drawn from this experiment is that the hypothesis has been confirmed. A number of entries in Table 2 show that the addition of bibliographic data midly improves retrieval. The last entry in Table 3 shows that using full queries against a full matrix of original and bibliographic data improves retrieval effectiveness signficantly.

This conclusion may be reached on the basis of either the sign test, or from the recall precision curve. Using only the 36 data base, document retrieval is quite effective considering that only 230 concepts are included in this matrix. This amounts to 1/25 of the number in the original matrix. Furthermore, these concepts are scattered fairly sparsely through



Recall Precision Curves for 0, 36, and 036 Data Base and Query Sets

Fig. 1

the documents with about 60 documents having no bibliographic data attached to them at all. Also, the data chosen were not regarded, in advance, as having much value as a retrieval tool. This suggests that the addition of citation index data would strongly improve retrieval once more.

If the document collection is conceived of as a growing structure, new entries will tend to cause bibliographic concepts to be added to the concept matrix. Further, the number of concepts added will probably be linear with regard to the number of documents added. This means that the concept matrix will grow unboundedly as the document collection grows. In a practical system, this may not be allowable. If so, techniques to discriminate between useful concepts and concepts which do not carry their weight will have to be developed.

References

[1] Amreich, M., Grissom, G., Michelson, D., and Ide, E., "An Experiment in the Use of Bibliographic Data as a Source of Relevance Feedback in Information Retrieval", Information Storage and Retrieval, Report ISR-12 to the National Science Foundation, Section XI, Department of Computer Science, Cornell University, Ithaca, New York, June 1967.

[2] Garfield, E., "Citation Indexes for Science", Science, 122, 3159, July 15, 1955, pp. 108-111.

[3] Ide, E. "User Interaction with an Automated Information Retrieval System", Information Storage and Retrieval, Report ISR-12 to the National Science Foundation, Section VIII, Department of Computer Science, Cornell University, Ithaca, New York, June 1967.

[4] Kessler, M. M.,"An Experimental Study of Bibliographic Coupling Between Technical Papers," M.I.T., June, 1962.

[5] Kessler, M. M., Bibliographic Coupling Between Scientific Papers", M.I.T. July, 1962.

[6] Kessler, M. M., "Analysis of Bibliographic Sources in a Group of Physics-Related Journals," M.I.T., August, 1962.

[7] Kessler, M. M., "Bibliographic Coupling Extended in Time: Ten Case Histories," Information Storage and Retrieval, 1, 1963, p. 169.

[8] Kessler, M. M. and Heart, F. E., "Analysis of Bibliographic Sources in the Physical Review (Vol. 77, 1950 to Vol. 112, 1958)," M.I.T., July, 1962.

[9] Kessler, M. M. and Heart, F. E., "Concerning the Probability that a Given Paper Will Be Cited," M.I.T., November, 1962.

[10] Salton, G., "Some Experiments in the Generation of Word and Document Associations", Proc. AFIPS FJCC, Spartan Books, Philadelphia, 1962.

[11] Salton, G., "Associative Document Retrieval Techniques Using Bibliographic Information", JACM 10, 4, October, 1963, pp. 440-457.