XI. An Experiment in the Use of Bibliographic Data as a Source of Relevance Feedback in Information Retrieval

M. Amreich, G. Grissom, D. Michelson, E. Ide

Abstract

In order to determine whether bibliographic information provides a useful tool for relevance feedback, parallel retrieval and feedback searches are made with and without such information within the framework of the SMART retrieval system. Bibliographic material operating as a sole source of feedback is shown to be comparable in efficiency to subject material alone. Some recommendations are made for further investigation.

1. Introduction

If the processes of storing documents for future use and of retrieving the appropriate stored documents in response to a given inquiry are to be automated using today's type of computer, storage capacity and search time dictate that some short but meaningful representation of the document be preferred to storage of the full-length natural language text.

Such representations generally consist of subject indicators, classifications of document content, comparable to the subject card in a library catalog. These subject indices may be assigned on the basis of human judgment of the relevance of each document to the subject, or on the basis of some mechanical procedure approximating such a classification. One such mechanical procedure is the assignment of indices based on key words

mechanically judged indicative of content by their relative frequency, either directly, each keyword having its own index, or through the intermediary of a thesaurus, the same index being assigned to words sharing some common property, e.g. synonyms, words differing only by suffix, and so on. In practice, these sets of "concepts" often turn out to be either insufficiently refined, or sensitive to minute differences in the wording of the user's request. This motivates the present attempt to evaluate the utility of a second type of information, bibliographic data, as an additional guide to document retrieval in a keyword classification system.

2. The Bibliographic Assumptions [1]

The underlying assumptions implicit in the use of bibliographic information as a classificatory device are the following:

1. A document cites a work if and only if that work has contributed information (facts, methodology, etc.) to the document, and, as a corollary, since both the work and the document deal with that contributed information, they must share at least that much common subject matter.
2. By extension, two documents citing the same work frequently share some common subject matter or tend to deal with the same subject. Note that no claim is made for the converse of this: it is not assumed that all documents dealing with a given subject need be linked bibliographically.
3. Documents by one author tend to deal with the same or related subjects. Note that this is a weaker claim.

3. The Problem and Method

It should be emphasized that the present experiment, consistent with common practice, does not attempt to classify documents uniquely by bibliographic information, but rather uses this only as a supplementary indication of content. Thus, the method here employed is a form of feedback starting from documents already known to be relevant. The basic test applied here to the assumptions is the comparison, within a controlled framework, of the retrieval capacities with and without the use of bibliographic data. [2]

The tripartite framework consists of a retrieval system, a document collection, and a set of requests for information from that collection, along with human judgments of the relevance or non-relevance of each document to each query, against which to check the machine's decisions. Chosen by reason of availability, the framework used for the present experiment consists of the following:

a. The retrieval system known as SMART, as implemented on the Control Data 1604 at Cornell University.

b. The ADI document collection (already coded on SMART except for bibliographic information), consisting of 82 abstracts from short papers given at the 1963 meeting of the American Documentation Institute [3].

c. The set of 35 queries associated with the ADI collection.

The specific structure of this framework results in certain restictions on the method of investigation. A detailed consideration of that structure is therefore in order.

Although initial queries in a keyword classification system could contain bibliographic information (e.g. "What has X written on the subject of Y?"), the queries available with the ADI collection contain only subject information. To introduce bibliographic information into the system, a procedure called relevance feedback is used. [2] After an initial retrieval operation, the user's request is modified to reflect the contents of those documents which were retrieved and declared relevant by the user. In the present experiment, this modification includes bibliographic information from the documents declared relevant. Therefore, the value of bibliographic information for retrieval is judged here by its usefulness in a relevance feedback environment, rather than in initial retrieval. Relevance feedback provides an equally valid test of the assumptions of part 2.

The ADI collection, the only document collection on SMART with readily available bibliographic information, is atypical in two respects. Firstly, since the collection includes only short conference papers, there is a paucity of bibliographic citations. Slightly over 20 per cent of the documents have no bibliography, and several others attach no authors to the works they cite. Secondly, since the entire collection was published simultaneously, there can exist no cases of documents i citing works j with both i and j in the collection.

SMART's representation of the ADI collection is based on the abstracts of the papers, with concept numbers assigned by a thesaurus. Each document is stored as a vector so that the entire collection may be viewed as a matrix of terms (concepts) against documents. Thus, for concept i and document j, the matrix entry, C_j^i, has an assigned value. While, in theory, the term-document matrix may be binary (i.e. $C_j^i = 1$ implies T_i

occurs in D_j; $C_j^i = 0$ implies T_i does not occur in D_j) in the ADI representation, C_j^i takes on larger values to show the greater importance of a term for one document than for another. These values are called "weights". To save space in the actual machine representation, only those values of C_j^i greater than zero are stored. To indicate the position of the weight of C_j^i in the document vector C_j, the code number (i) for the concept called the "concept number" is paired with the non-zero value.

Extending the notion of "concept number" to allow the inclusion of proper nouns, i.e. authors' names and titles of cited works, permits the direct addition of bibliographic data to the document vector in the following manner:

1. Dictionaries are manually constructed to assign concept numbers to the authors (both of documents in the collection and of works cited by documents in the collection) and to the titles of the works cited.
2. The completed dictionaries are used to generate triplets (D,N,W), where D is the number of the document in which author or title concept N occurs with a weight of W.
3. These triplets are sorted by document number D. Then, for each document, the pairs (N,W) are sorted by ascending value of N, since the particular implementation of SMART used in the present experiment requires that the concept numbers be stored in the vector in strict ascending order. These ordered additions of bibliographic information are then added directly to the original vector.

Each document is thus represented by a tripartite vector, in symbols, OAC, where O is the original vector of subject information. A represents

the authors of the document or of the works cited by the document, and C represents the citations.

Note that little information is lost by grouping the authors of the documents in the collection together with the authors of the works which the documents cite, since the following weighting scheme is used: relative weights of 3, 2, and 1 are assigned respectively to authors of documents, citations made by documents, and authors of those citations. (When a citation is listed in the document by title only, the author of the citation is not added to the vector.) The reasoning behind these relative values is as follows: listing the author of a cited work allows his authoring of two related works to link two documents each of which cites only one of those works; however, two documents citing the same work are more likely to be related; finally, the author of the document has presumably contributed more to the document than the cited references. These basic weights are combined by simple addition when necessary. Thus, the author of two works both cited by a given document has a relative weight of 2 in the vector of that document. Note that citing 3 works by an author renders his weight comparable to that of the author of the document. Note also that if an author cites his own works, his weight is strongly increased [1].

This representation further implies that working with authors of documents alone or with authors of citations alone is not possible. A distinct range of concept numbers for each group would not solve the problem, since SMART as it currently operates has no table look-up to identify the same author in his two functions.

Rather than "retrieve" a given subset of the collection in response

to an incoming query, SMART ranks all the documents in order of increasing correlation with the query, thus presenting the user first with the documents which the system deems most appropriate to his stated needs.

The correlation of query vector with document vector may be done by various different formulae. Of the several correlation coefficients programmed into SMART, two in particular invite consideration for present purposes: the "overlap" and the "cosine".

$$\text{Cosine:}\quad R_j^i = R_i^j = \frac{\sum_{k=1}^{m} c_k^i \; c_k^j}{\sqrt{\sum_{k=1}^{m} (c_k^i)^2 \sum_{k=1}^{m} (c_k^j)^2}}$$

$$\text{Overlap:}\quad R_j^i = R_i^j = \frac{\sum_{k=1}^{m} \min (c_k^i, c_k^j)}{\min \left(\sum_{k=1}^{m} c_k^i , \sum_{k=1}^{m} c_k^j \right)}$$

where m is the highest concept number in the system. Each of these coefficients, has a serious drawback. As indicated by the formulae, the cosine's denominator is sensitive to document vector length (or, more precisely, to the number of non-zero weights regardless of match with request), and the overlap's numerator is insufficiently sensitive to varying weights above the

minimum. Thus, in view of the reliance on a relative weighting scheme, described above, the overlap is distinctly undesirable. Furthermore, the earlier relevance feedback experiments on SMART (using only subject material) do not use overlap, and therefore, cosine provides a better basis for comparison. The choice of the cosine for this study, however, is not made without reservation. The ADI collection is typical in having a varying number of non-zero weights from one document to another; further, the revised document vectors (OAC) are all longer than the original document vectors (O), but not by any fixed length or fixed proportion of the length of O. Since the present experiment works only with the feedback capabilities of bibliographic information, the set of initial queries associated with the old ADI collection remains unaltered and still contains only subject (O) information. The initial query can match only on O information; nevertheless, because of the changed length, the cosine correlation value of a given query matched against the OAC representation of a document will be lower than that of the same query matched against the O representation of the same document.

Because this lowering of the cosine coefficient could affect the document ranks, two parallel classes of initial and feedback retrieval searches are made. In one class (denoted by the symbol K) all query sets, initial and feedback, are matched with a constant set of document vectors containing all types of information (OAC), regardless of the type(s) of information in the query set. In the other class (denoted by V), the set of document sub-vectors searched is allowed to vary appropriately, each type of query set being matched only with those sections of the document vector containing the same type(s) of information as the query set.

4. Query Alteration in Feedback

The standard procedure in feedback is to expose the user to a limited number n of documents which the system suggests after an initial search (in SMART, the n top ranking documents) and to alter the query in favor of those documents the user indicates are relevant. How limited a number the user examines is an arbitrary decision. Any relevant documents above this arbitrary cutoff are considered "retrieved" and become the source of feedback information. It follows that if no relevant documents are retrieved on the initial search, no feedback procedures may be employed. To minimize such cases, a cutoff of 15 is employed in the present experiments. This cutoff increases the risk of retrieving all relevant documents, in which case the application of feedback techniques can produce no additional relevant documents. However, the cutoff at 15 leaves 3 out of 5 of the queries associated with the collection in the ideal range for feedback, retrieving some, but not all, of the relevant documents.

The results of this study are presented as average Quasi-Cleverdon curves [4] of precision plotted at each 0.05 of recall. Recall is the ratio of relevant documents retrieved to the total number of relevant documents in the collection and precision is the ratio of relevant documents retrieved to the total number of documents retrieved. For purposes of comparison with other studies, all 35 queries are averaged in the results, although not all of these queries can be improved by relevance feedback.

5. Evaluation

Three iterations of standard relevance feedback are applied to both

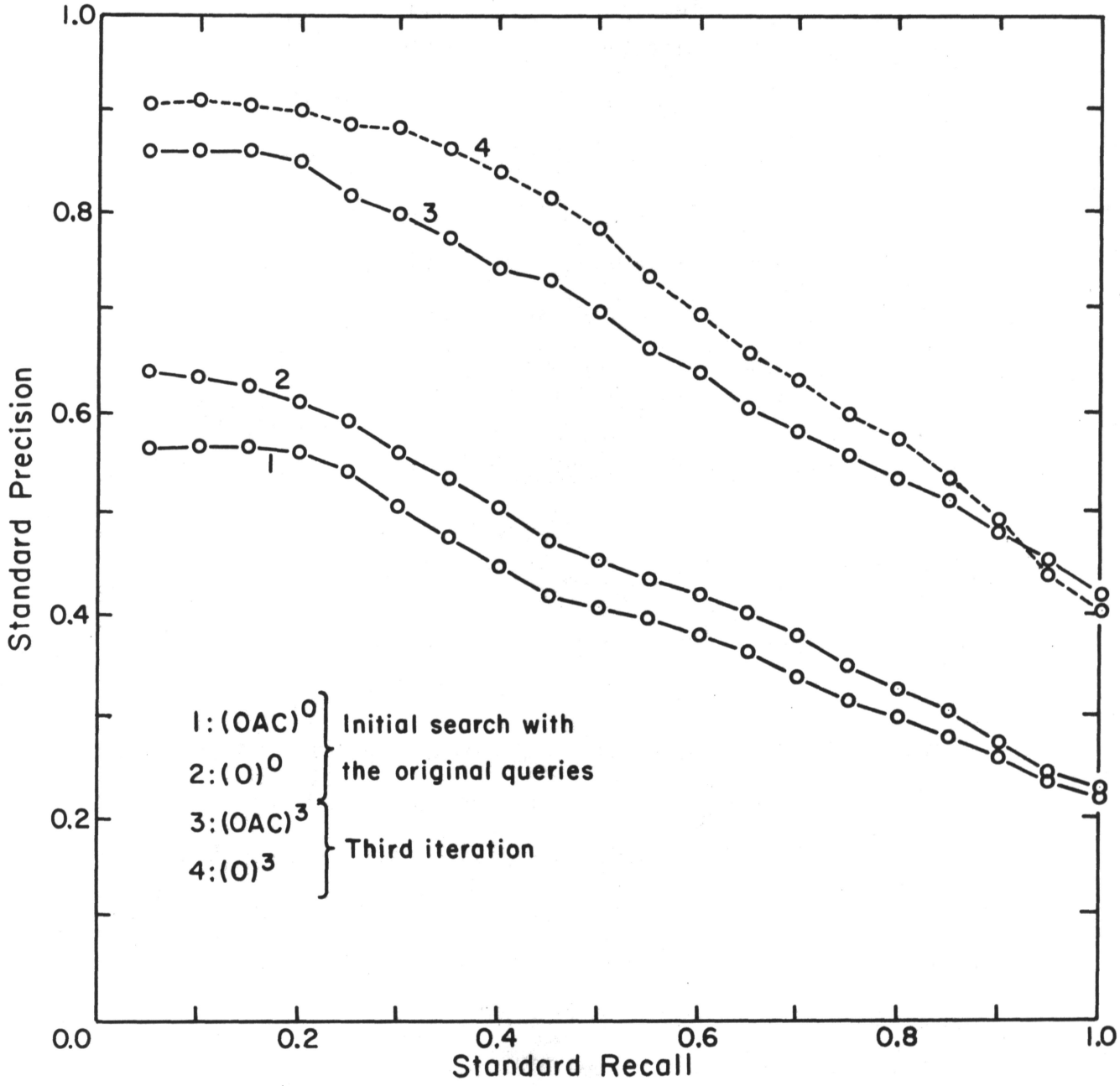
the original (O) vectors, which have only subject information, and to the augmented vectors (OAC) with added bibliographic information. Both sets of vectors are searched starting with the same set of 35 queries, which contains only subject information.

As shown in figure 1, the initial search with this query set yields different results on the two vector representations of the same document collection, differing only in length (see the discussion of the cosine coefficient in Part 3). The application of the three feedback iterations produces for each initial search comparable improvement in performance. From this result alone, it is unclear to what degree the bibliographic data furnishes material for feedback. In particular, it is uncertain if the improvement in the case of the augmented vectors can be attributed to the subject (O) material alone, or in part to the author (A) and citation (C) material.

To answer this question it becomes necessary to abandon the standard relevance feedback technique of treating the entire vector uniformly, and to feed back only certain types of information. Two new feedback strategies are employed:

a. Only author and citation information are added to the original query (symbolized as $O^0\ A^1\ C^1$ for the first iteration, $O^0\ A^2\ C^2$ for the second).

b. The bibliographic feedback is further examined in complete isolation from the initial query. That is, the initial query O^0 is discarded and $A^1\ C^1$, and even A^1 and C^1 separately, are used as feedback query sets.



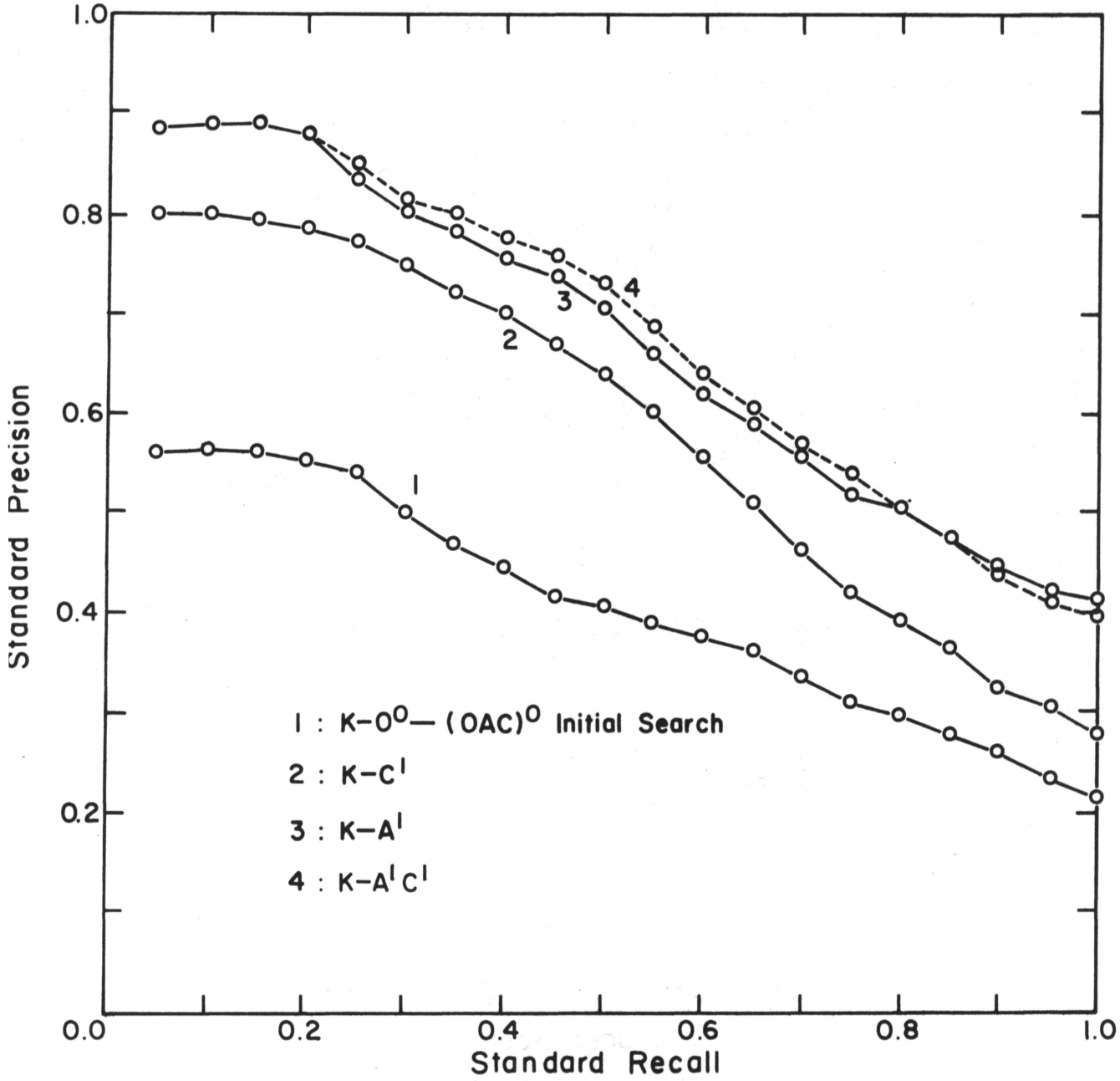
Standard Feedback Process
(initial search and third iteration)
on Original and Augmented (OAC) Vectors

Fig. 1

As a control for the sensitivity of the cosine coefficient to vector length (Section 3), parallel K and V searches are made using each of the feedback strategies mentioned above. K indicates that the full length (OAC) document vector is used consistently, while V indicates that for each initial or feedback search, the query is correlated only with those document vector segments with which a match is possible. For example, the initial query (O^0) in a V search is correlated with only the O segment of the document vectors, and A^1 C^1 queries (using the b strategy above) are correlated only with the bibliographic segment AC. Results of the searches described above are shown in Figs. 2 to 5.

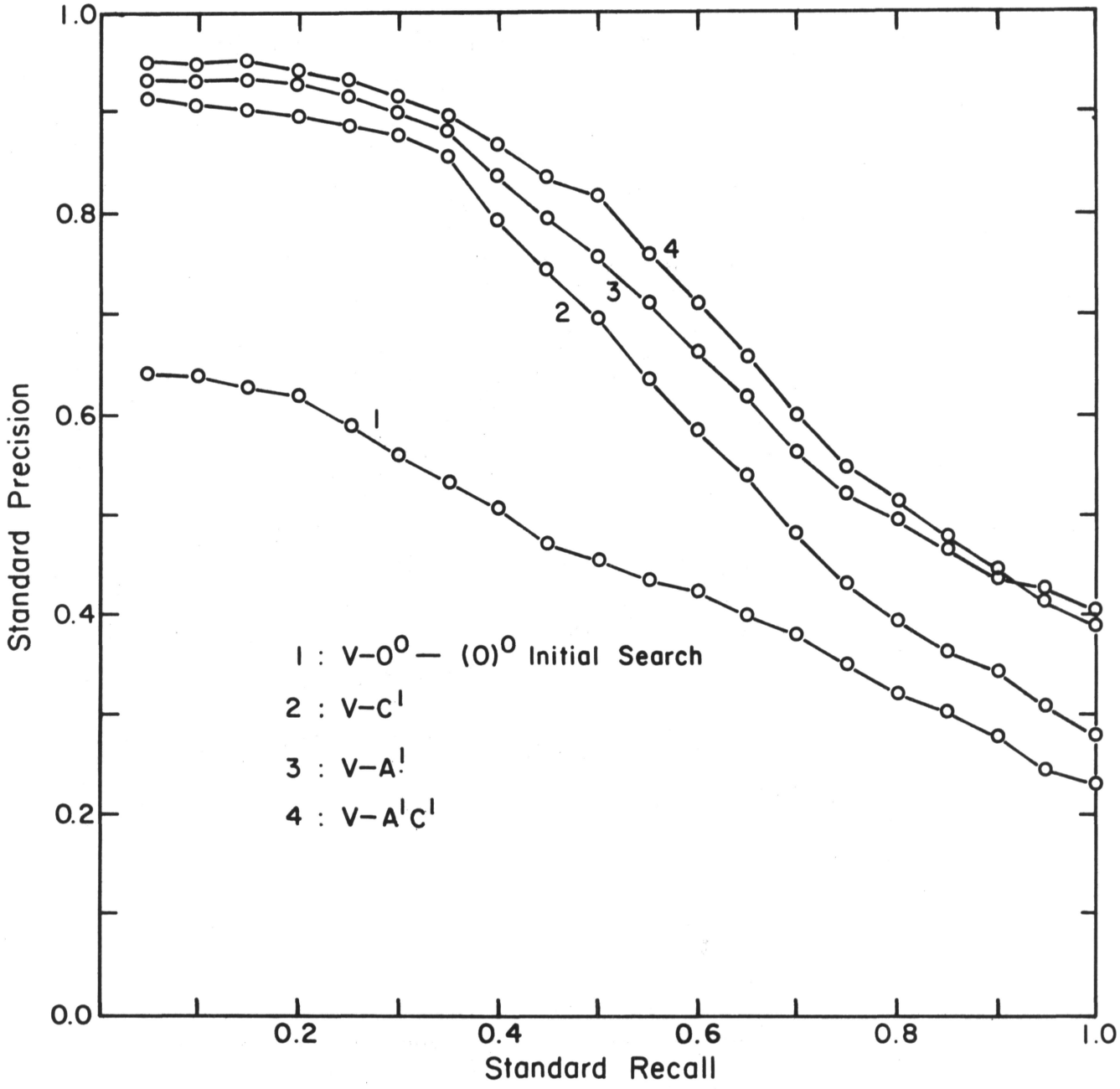
Comparison of Fig. 1, where the original (O) subject information is used for feedback, with Figs. 2 to 5, where no O data is fed back, shows that the same type of improvement over the initial searches occurs with each of the various query sets. The runs using citation information (C^1) alone show the least improvement, since over twenty per cent of the documents include no citations. This firmly establishes that bibliographic data alone, even in a citation-poor collection, can be used as a successful instrument of relevance feedback, either in complete isolation from subject classi i-cations, or as modification of the original query.

Direct comparison of the feedback results in Fig. 1 with those in Figs. 2 to 5 is misleading, because Fig. 1 shows the result of three iterations, while Figs. 2 to 5, except for one second iteration O^0 A^2 C^2 , show only one iteration. Fig. 6, however, which presents selected comparisons at the same iteration level, shows no differences greater than four percent



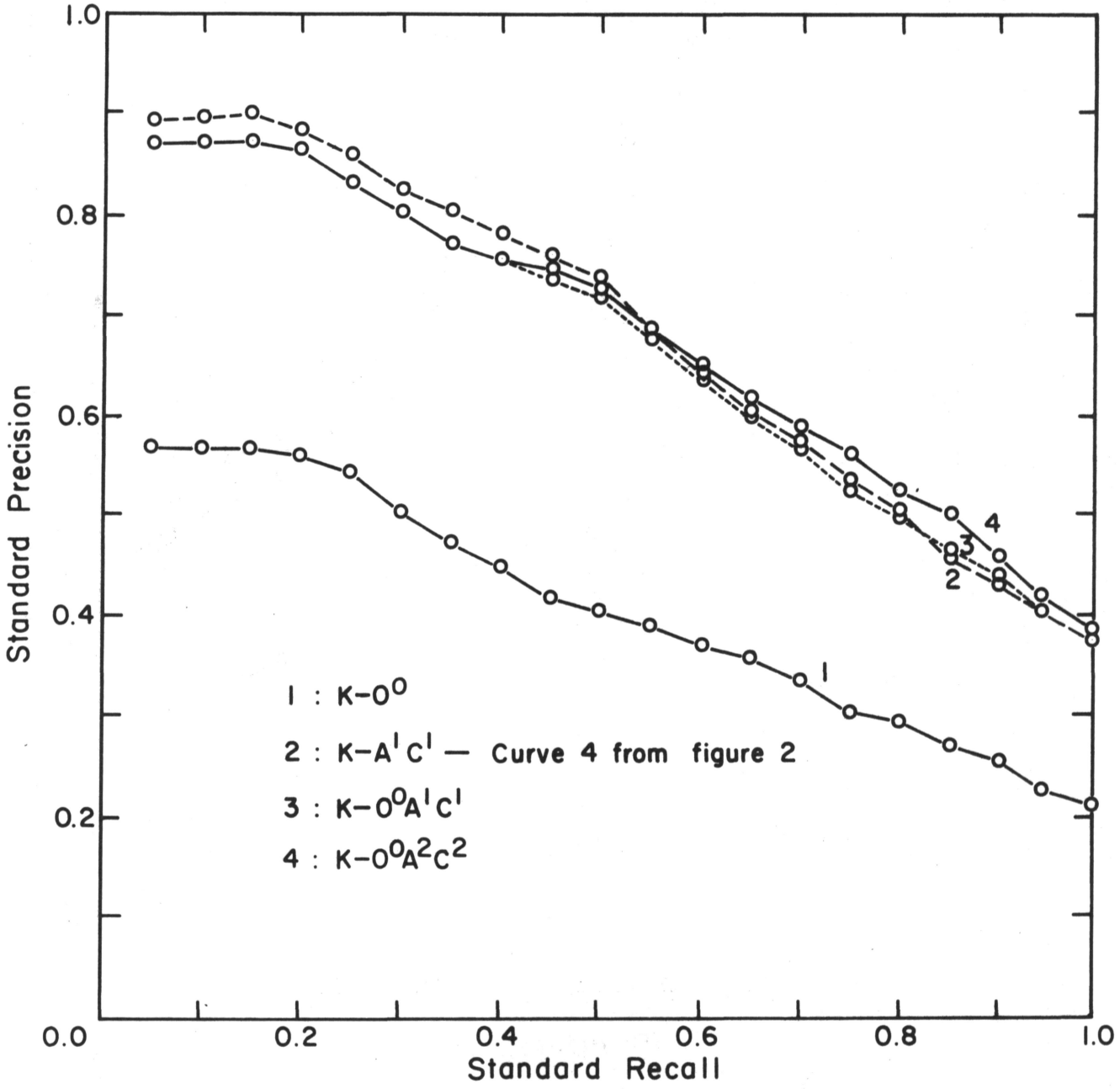
First iteration feedback of citation (C)
and author (A) information alone,
with full length document vectors (K)

Fig. 2



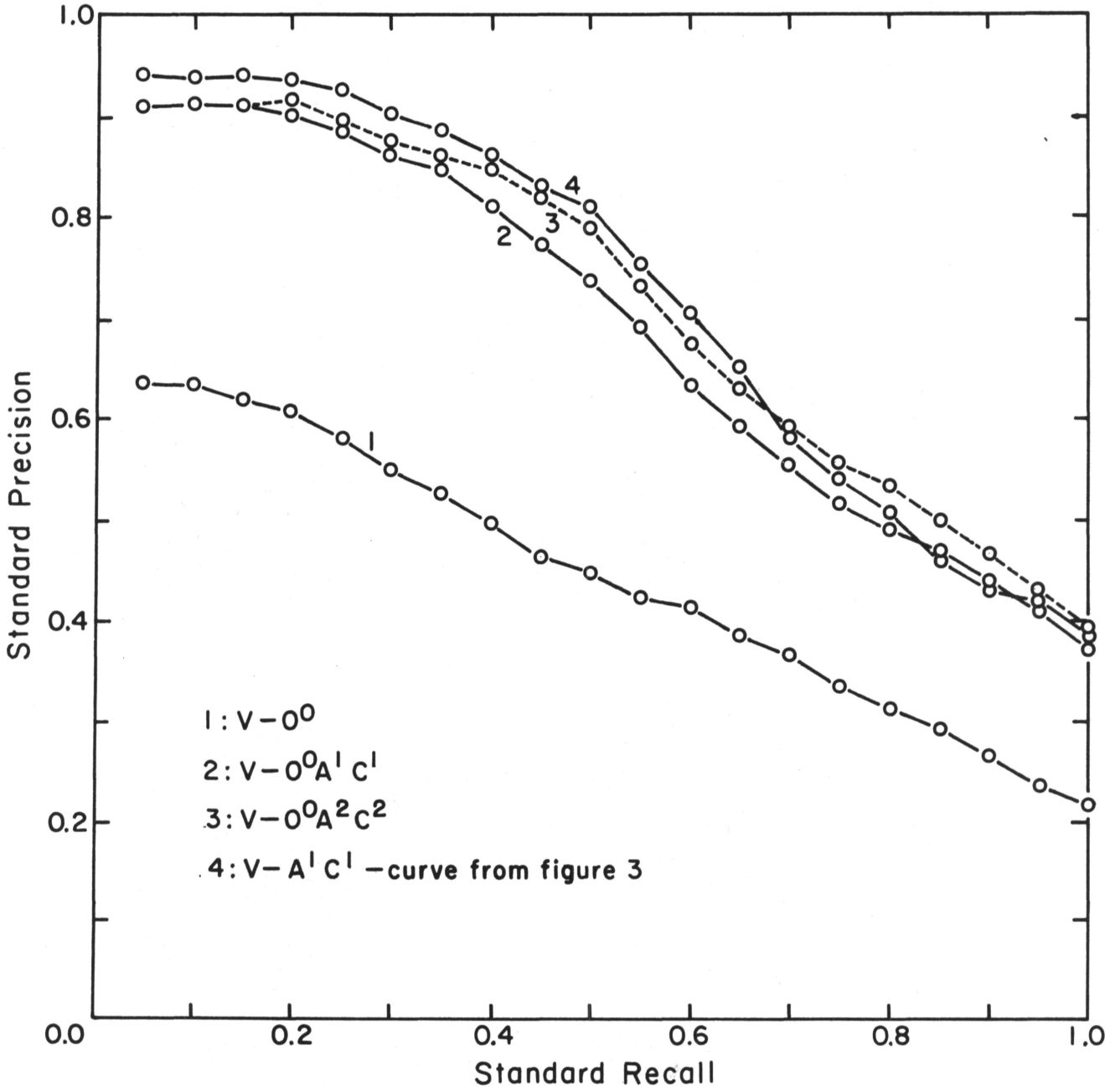
First iteration feedback of citation (C) and author (A) information alone, with variable document vector segments (V)

Fig. 3



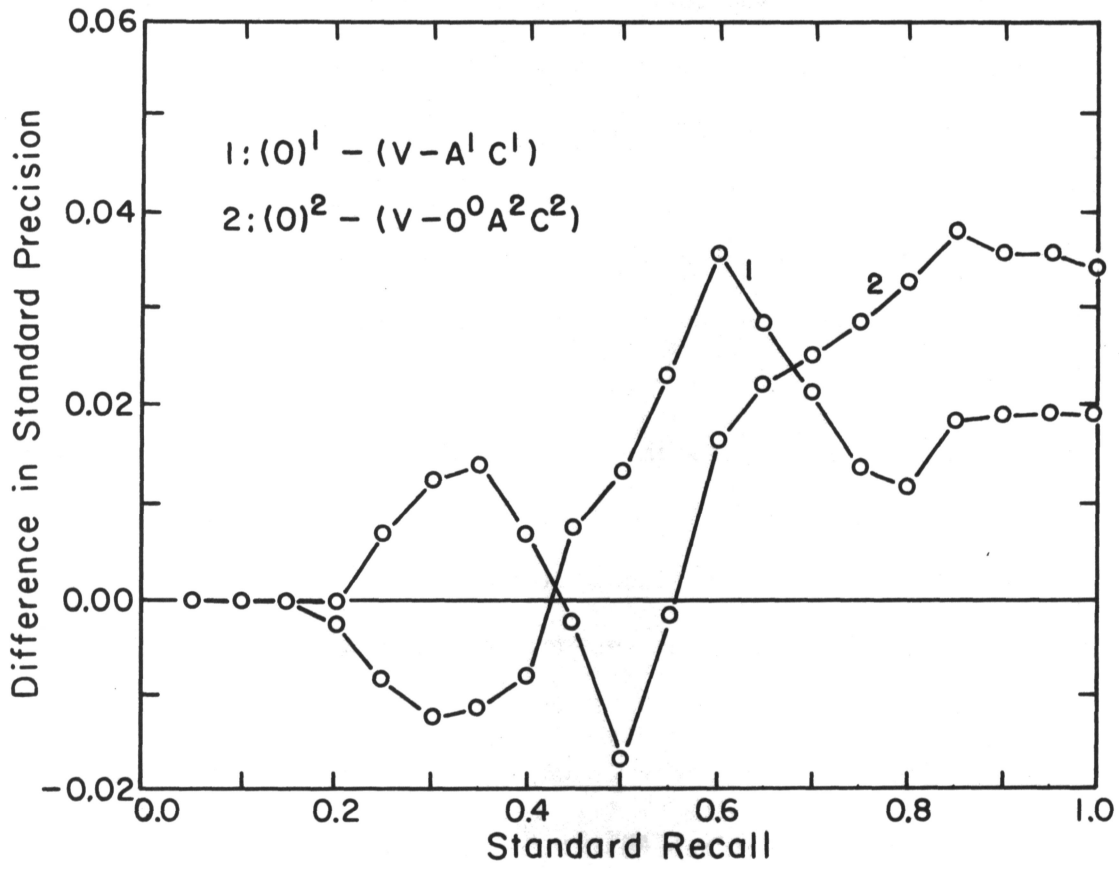
First and second iteration feedback of author (A) and citation (C) information retaining the original (subject) query, with full-length document

Fig. 4



First and second iteration feedback of author (A) and citation (C) information retaining the original query, with variable document vector segments (V)

Fig. 5



First and second iteration comparisons of feedback of subject (O) and bibliographic (AC) information

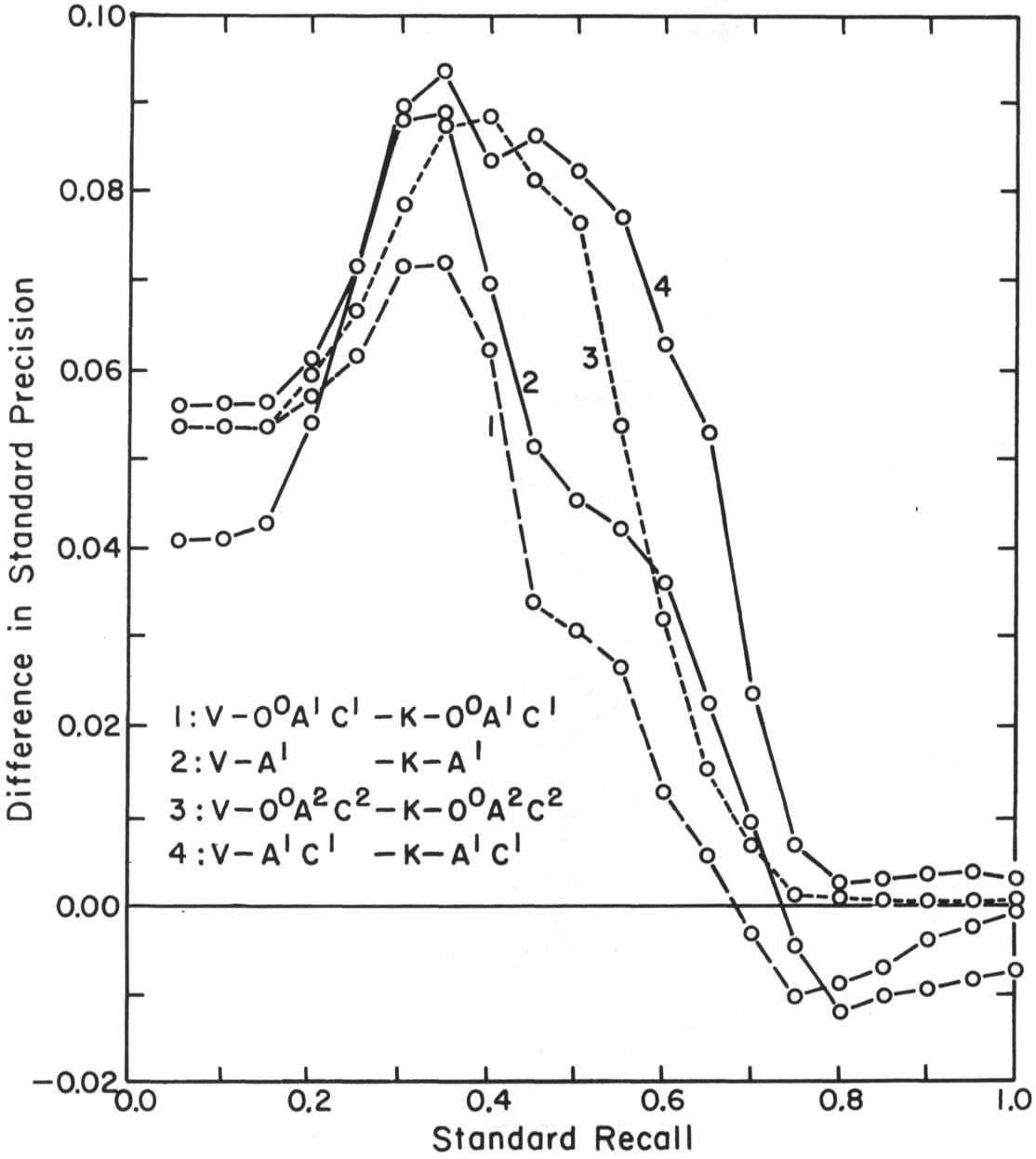
Fig. 6

between the results of feeding back only subject data (O) and those of feeding back only bibliographic data (AC). This implies that the usefulness of bibliographic data for feedback is of the same order as that of subject descriptors.

The difference between the $(OAC)^3$ and $(O)^3$ results in Fig. 1 is therefore not due to any lack of merit of the bibliographic information. There is evidence, in fact, that this difference is primarily attributable to the difference in initial search results, which is caused by the sensitivity of the cosine coefficient to vector length. Since the initial searches are different, different documents are available for first iteration feedback; the discrepancy in performance is similarly propagated to the third iteration.

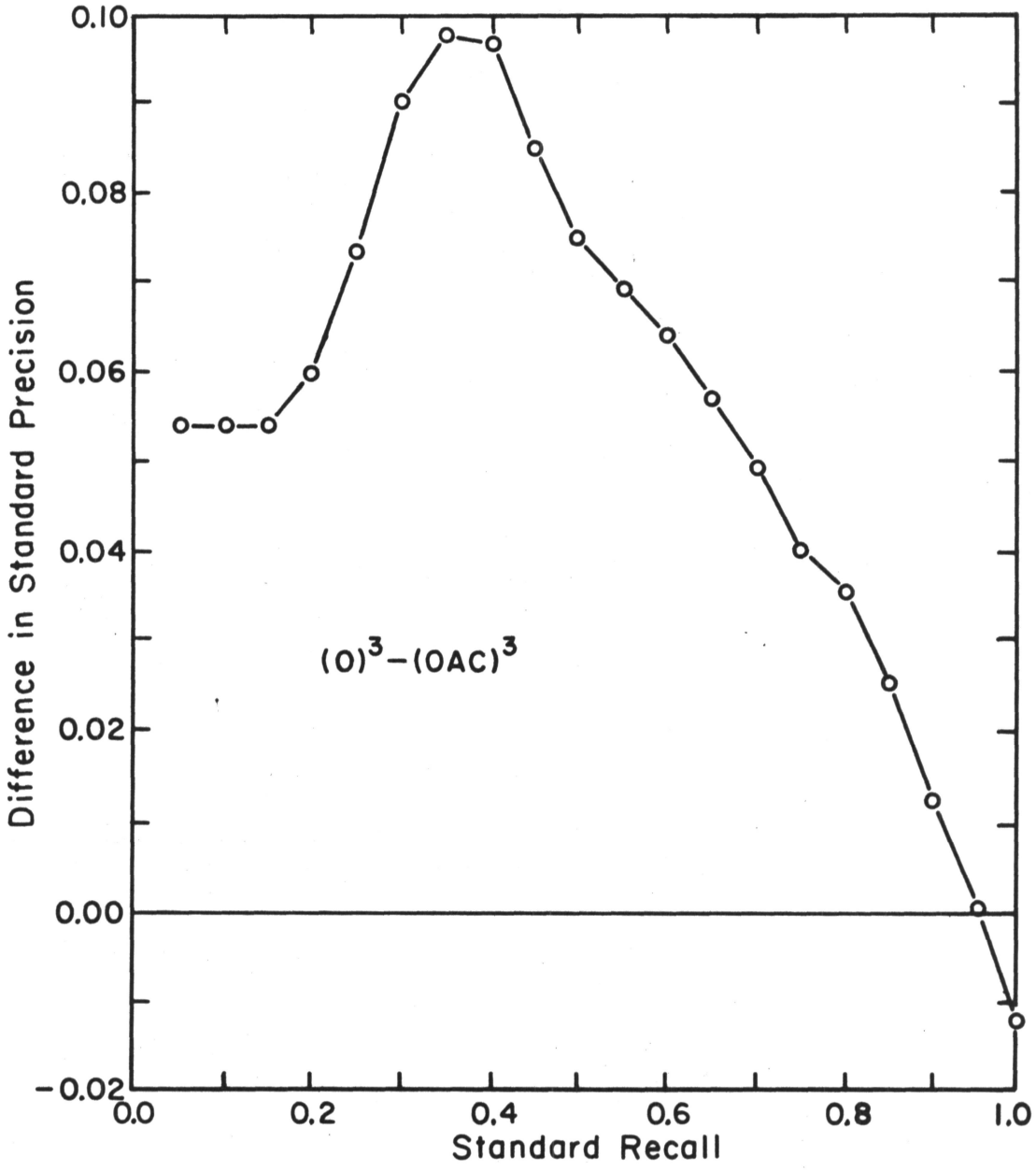
Fig. 7 shows difference curves which compare results from Figs. 2 and 4 (K) with those from Figs. 3 and 5 (V). It is obvious that the V runs produce consistently better results than the K runs regardless of the type of feedback information used. The forms of these difference curves are so similar that a single cause is indicated. The only constant difference between all V and K runs is the initial search, which for all K runs is $(OAC)^0$ and for all V runs is $(O)^0$.

Fig. 8 shows the difference curve for $(O)^3$ minus $(OAC)^3$ from Fig. 1. This curve has the same shape and magnitude as the difference curves in Fig. 7. This similarity indicates that the difference between OAC and O feedback is also attributable to the difference in initial search results.



Differences between corresponding V and K curves (V - K)

Fig. 7



Difference between feedback based on original vector and on augmented vectors (OAC)

Fig. 8

An investigation of the effect of added vector length on results in a keyword system using the cosine coefficient is obviously necessary before the addition of a second type of information to the classification can be recommended. If the added bibliographic vectors were to vary randomly in length from document to document, the changes in the cosine correlation would be random, and the resulting changes in ranking would give the same average retrieval performance. In fact, the shorter (O) document vectors give significantly better initial search results on the initial query set than do the full (OAC) vectors, so that it is clear that the added vector length lowers the ranks of some documents that are relevant to some queries. One would guess from this observation alone that there is a direct relationship between the length of the bibliographic sub-vector and the relevance of the document to the query set.

Fig. 9 shows that in fact, in the ADI collection, documents relevant to four or more queries have on the average longer bibliographic vectors and more non-zero bibliographic concepts than do documents relevant to three or fewer queries. The twelve highly relevant documents thus have lower relative cosine coefficients when the full length document vectors are used. Twenty-three out of the thirty-five queries name at least one of these twelve documents as relevant. Therefore, roughly two thirds of the requests are affected by the length-relevance relationship shown in Fig. 9. The up to ten per cent differences in feedback performance observed in Figs. 7 and 8 are no longer surprising.

Two important questions remain concerning the combining of different types of information in a keyword retrieval system. First, to what extent

	In documents relevant to	
	0 - 3 queries	4 - 7 queries
Average number of non-zero bibliographic concepts	8.0	11.3
Average length of bibliographic vector	71.0	80.7
Average number of non-zero subject concepts	16.2	18.1
Number of documents involved	70	12

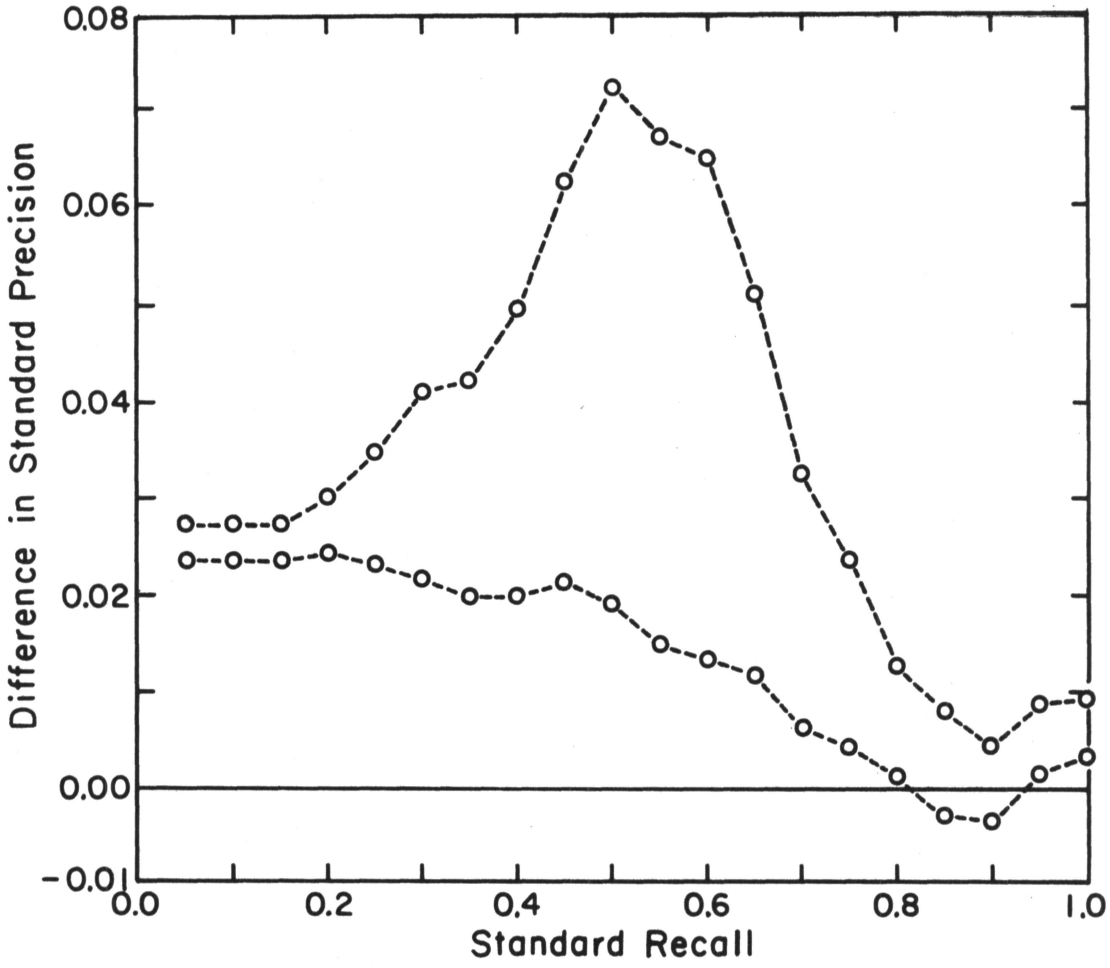
Relevance of a document to the query set as related to bibliographic and subject information

Fig. 9

is the observed length-relevance relationship true of other types of information than bibliographic. Fig. 9 shows a less striking tendency in the ADI collection for the more relevant documents to have more non-zero subject concepts. This fact may explain an otherwise strange result in Fig. 5, that the V curve using the initial query ($V - O^0\ A^1\ C^1$) is lower than the curve that discards the initial query ($V - A^1\ C^1$).

K and V difference curves of $A^1\ C^1 - O^0\ A^1\ C^1$ are shown in Fig. 10. The K curves are essentially the same, with no differences above two per cent. The V curves, however, differ by as much as seven per cent. The length-relevance relationship of Fig. 9 probably causes the difference between the V and K curves. Both K runs and the $V - O^0\ A^1\ C^1$ run use the full length document vectors. However, the $V - A^1\ C^1$ run is correlated using only the bibliographic document subvectors ($A^1\ C^1$). Again the twelve highly relevant documents, identified by longer subject subvectors, have relatively higher cosine coefficients in the $V - A^1\ C^1$ run which does not use the subject subvectors. This effect of adding subject to bibliographic information is similar to but weaker than that observed in Figs. 7 and 8, where bibliographic information is added to subject information.

The second remaining question concerns the likelihood of observing relevance relationships found for the ADI query-document set in other retrieval environments as well. This question deserves more investigation. It seems reasonable, however, that in a typical document collection, works covering a wider subject range, which by the assumptions in Part 2 should also contain larger bibliographies, are likely to be relevant to a wider range of requests than are works dealing with narrower topics. In the ADI



Effect of adding subject information
to a bibliographic vector

Fig. 10

collection, only 12 such documents out of 82 caused significant changes in retrieval performance. In most retrieval environments, a similar or greater proportion of such general documents might be expected. Thus, length-relevance relationships are likely to be important in any attempt to use more than one type of information for retrieval in a keyword system with cosine correlation.

6. Conclusions and Recommendations

This study indicates that bibliographic information can be used effectively in a mechanical keyword retrieval system as a source of feedback information. In fact, bibliographic information used alone seems as valuable as subject information alone in the retrieval environment studied. Since the bibliographic information is useful for relevance feedback, it should also prove valuable for initial retrieval searches.

A direct relationship between document vector length and relevance of the document to the query set is observed in the ADI collection of 82 documents and 35 queries. Documents relevant to more queries tend to have longer bibliographic description vectors and, less strongly, longer subject descriptor vectors. This length-relevance relationship lowers retrieval performance when a query containing only one type of information is correlated (using the cosine coefficient) with document vectors containing both types. This situation can probably be observed in most retrieval environments.

These conclusions support the following recommendations for keyword retrieval systems using the cosine correlation coefficient:

a. Bibliographic information (the author of the document, citation titles, and citation authors, each treated in this study as single concepts) should be used as well as subject descriptors to classify the documents;

b. the user should be permitted to use bibliographic information in his initial request;

c. whenever a request (initial or feedback) contains only subject or only bibliographic information, the request should be correlated only with the appropriate subvector of each document descriptor vector.

Several topics remain for further investigation:

a. Joint subject and bibliographic feedback should be attempted in the present retrieval environment using the retrieval order of the initial subject search ($V - 0^0$) for first iteration feedback. That is, recommendation c. above should be implemented for joint feedback. This procedure will provide a more valid comparison of subject and joint feedback than does Fig. 1;

b. other document-query collections should be investigated to see if length-relevance relationships are observed;

c. the document collection used for this study (ADI) is citation-poor. The value of bibliographic information should be investigated in more normal collections;

d. only one relative weighting of the four types of keywords (subject, document author, citation, citation author) is used in this study. Some other relative weighting schemes might improve retrieval;

e. initial queries using bibliographic information should be utilized.

References

[1] G. Salton, Associative Document Retrieval Techniques Using Bibliographic Information, Journal of the Association for Computing Machinery, Vol. 10, No. 4, October 1963, pp. 445-447.

[2] W. Riddle, T. Horwitz, and R. Dietz, Relevance Feedback in an Information Retrieval System Report No. ISR-11 to the National Science Foundation, Section VI, Department of Computer Science, Cornell University, June 1966.

[3] H. P. Luhn, editor, Automation and Scientific Communication, Part 2. Short papers contributed to the 26th Annual Meeting of the American Documentation Institute, October, 1963.

[4] G. Salton, The Evaluation of Automatic Retrieval Procedures – Selected Test Results Using the SMART System, American Documentation, Vol. 16, No. 3, July 1965, pp. 209-222.