# Characterization of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts

Edward A. Fox*

83-561
September 1983

*Department of Computer Science
Cornell University
Ithaca, New York 14853

now at:
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061

This work was supported in part by the National Science Foundation, under grant IST-81-08696.

## TABLE OF CONTENTS

# CHARACTERIZATION OF TWO NEW EXPERIMENTAL COLLECTIONS IN COMPUTER AND INFORMATION SCIENCE CONTAINING TEXTUAL AND BIBLIOGRAPHIC CONCEPTS

Edward A. Fox*

**Abstract**

Two new collections are described which are particularly useful for investigating the interaction between textual and bibliographic data in the automatic indexing and retrieval of documents. An extension to the vector space model has been proposed whereby various types of concepts are included in the representation of such documents. Experiments using an enhanced version of the SMART system have shown such an extended model to perform better than simpler schemes. The CACM and ISI collections developed for this research should be of value for future related studies.

The ISI collection has author, title/abstract, and co-citation data for the 1460 most higly cited articles and manuscripts in information science in the 1969-1977 period. The CACM collection contains 7 types of concepts for the 3204 articles published in the *Communications of the ACM* up through 1979. These collections have 76 and 52 queries, respectively, along with relevance judgments.

*Department of Computer Science, Cornell University, Ithaca, NY 14853; now at Dept. of Computer Science, Virginia Tech, Blacksburg, VA 24061. This work was supported in part by the National Science Foundation, under grant IST-81-08696.

## 1. Introduction

In order to retrieve documents relevant to the request of a particular user it is necessary to first index or represent the content of articles and manuscripts. For many years this has been done by trained indexers who assign keyword lists or sets of descriptors from a controlled vocabulary [Borko & Bernier 1978]. Since the early 1960's an alternative method of automatic indexing has been developed whereby word stems, words, phrases, or thesaurus category indicators are selected from the title and abstract and a weighted vector indicating the importance of each is constructed [Salton 1980]. Part of this report deals with the vectors derived in this fashion from two collections in information and computer science.

Another source of data about documents is from their bibliographic references. Citation indexes can be used to locate those entries referred to by an article, or which cite it [Garfield 1964, 1979]. Linkages between documents based on bibliographic coupling [Kessler 1962] and co-citation counts [Small 1973] have been utilized for a variety of analysis and retrieval purposes (e.g., [Bichteler & Eaton 1980], [Garfield 1970], [Kessler 1963a, 1963b, 1965], [Small & Koenig 1977], [Small 1978, 1980, 1981], [Weinberg 1974]. Preliminary experimentation has shown that the vectors produced by automatic indexing of document texts can be usefully supplemented by bibliographic information to produce a representation that can be more effectively searched than if either component were used alone ([Michelson et al. 1971], [Salton 1963, 1971]).

To facilitate exploration of the effects of extending the vector space model to include a variety of types of concepts it was necessary to have test collections containing such concepts. One collection containing 1460 of the most highly cited documents in information science published between 1969 and 1977 [Small 1981] was developed based on citation and co-citation data provided by the Institute for Scientific Information. This ISI collection contains three types of concepts: author names, word stems from the title and abstract sections, and co-citations between each of the articles. A second collection, of the 3204 articles published in *Communications of the ACM* in the years up through 1979, contains the above three types of concepts plus: categories assigned from a hierarchical subject classification scheme, bibliographic coupling connections, date information, and direct references between articles. The CACM collection also includes 52 user supplied queries in 3 different forms and relevance judgements indicating which documents relate to which queries. The ISI collection has a total of 76 queries.

This report describes the CACM and ISI collections in detail. It supplements the theoretical and experimental descriptions of [Fox 1983b] and the elaboration on implementation issues contained in [Fox 1983a]. Ample discussion has been included to enable interested researchers to understand the characteristics of these collections and to determine if they might be of use in related research investigations.

The rest of this introductory section provides useful background for subsequent sections. The notion of extended vectors is introduced, and some overview tables are shown indicating how these collections relate to other test collections discussed in [Fox 1983b].

### 1.1. Extended Vectors

When only the terms of documents are considered, a simple collection representation scheme results. A dictionary can be formed to contain the $T$ distinct word stems and so textually based concepts can be numbered from 1 through $T$. Each of the $N$ documents $D_i$ is represented by a vector of length $T$,

$$\vec{D_i} = ( tm_{i1}, tm_{i2}, ..., tm_{iT} ) \tag{1-1}$$

so the entire collection is a $N$x$T$ matrix,

$$C = \{ tm_{ij} \} \quad 1 \leq i \leq N, 1 \leq j \leq T. \tag{1-2}$$

If other types of information are provided in addition, then each term vector (1-1) can become a subvector of a more complete document vector. Similarly, the matrix (1-2) becomes the term submatrix. It makes sense to have a separate author submatrix indicating which authors contributed to which articles based on a second dictionary of author names. When some subject classification scheme is adopted, such as the category system used for *Computing Reviews*, then a dictionary for those entries can also be constructed. Thus, the CACM collection has term (tm), author (au), and *Computing Reviews* category (cr) submatrices, along with others.

More data about articles comes from the references present in the bibliographies of each publication. An $N$x$N$ matrix can be built indicating which articles refer to which others, or which ones are cited by others. Similar matrices indicate the degree of bibliographic coupling or the number of co-citations received by pairs of articles. Each of these matrices then becomes a submatrix of a large $N$ row collection matrix.

The extended vector model is thus based upon the idea of having multiple concept types. Each document vector has subvectors, one for each concept type included in the representation. Further discussion of the model, and experimental evidence of its utility, can be found in [Fox 1983b]. More detail regarding how the CACM and ISI collections are represented according to this model is given in sections 2 and 3 below.

### 1.2. Contrast with Other Collections

In [Fox 1983b], 1 small and 4 moderate to medium size test collections were employed to test various hypothesis. The CACM and ISI collections were two of those utilized. To provide a suitable contrast, the following subsections present summary information about all 5 of those collections.

### 1.2.1. Document Collections

Table 1 summarizes essential data about each of the document collections. All collections contain information about documents such as monographs or journal articles. At the very least, in almost every case, a title and abstract were available originally. They deal with a a number of subjects, from the "soft" social science like material that makes up a fair proportion of the ISI collection, to the terse, medical articles used in Medlars studies.

(1) The small ADI collection is about librarianship, microforms, and other topics in documentation and information science as of 1963.

(2) The CACM documents include all articles in issues of the *Communications of the ACM* from the first issue in 1958 to the last number of 1979. A considerable range of computer science literature is covered by those 3204 entries in the publication that for many years served as the premier

Table 1: Document Collections Summary

| Short Name | No. of Docs. | No. of Terms | Av.No. Terms | Subvectors Included | Subject Matter | Years Covered |
|---|---|---|---|---|---|---|
| ADI | 82 | 886 | 27.1 | tm | Documentation | 1963 |
| CACM | 3,204 | 10,446 | 40.1 | au,bi,bc, cc,cr, ln,tm | Computer Science | 1958 -1979 |
| INSPEC | 12,684 | 14,683 | 35.4 | tm | Electrical Engineering | 1979 |
| ISI | 1,460 | 7,392 | 104.9 | au,cc,tm | Information Science | 1969 -1977 |
| Medlars | 1,033 | 8,750 | 55.8 | tm | Medicine | to 1969 |

periodical in the field.

(3) INSPEC, which stands for Information Services in Physics, Electrotechnology, Computers and Control, covers three *Science Abstracts* publications, *Electrical and Electronics Abstracts, Computer and Control Abstracts,* and *Physics Abstracts.* The content focuses mainly in electrical engineering and computer science subjects.

(4) The 1460 ISI entries were selected based on co-citation information relating to a study conducted by Dr. Henry Small of the Institute for Scientific Information © (ISI ©). They are the items that could be located at the Cornell University library out of a total of 1627 names listed in the field of information science. Each was published between 1969 and 1977 and received at least five citations.

(5) The 1033 Medlars articles were selected out of a large medical collection available at the National Library of Medicine.

Thus, the five collections are from various sources, have different sizes, and deal with a number of subjects.

### 1.2.2. Query Collections

Since considerable effort was made to study the characteristics and behavior of various query formulations, it is worthwhile to examine the five different query collections and all the versions present for each. Table 2 gives statistics (for the 4 larger collections) relating to the queries and the number of relevant documents for each query.

Table 2: Query and Relevant Document Characteristics

| Collection Name | No. of Queries | Av. Length Cos Query | Rels Per Query Av.No. | Rels Per Query Av. % | Rels in Top 10 Av.No. | Rels in Top 10 Av. % |
|---|---|---|---|---|---|---|
| CACM | 52 | 11.4 | 15.3 | 0.5 | 1.9 | 19 |
| INSPEC | 77 | 16.0 | 33.0 | 0.3 | 2.0 | 9 |
| ISI | 35,76 | 8.1 | 49.8 | 3.4 | 1.7 | 4 |
| Medlars | 30 | 10.4 | 23.2 | 2.3 | 3.8 | 18 |

To give some idea as to the average length of each query, that value is given for the cosine version based on the original natural language (NL). The ISI value is low, because only the 35 queries for which both vector and Boolean logic (BL) forms are available were considered, and those queries are rather short. The INSPEC queries are much longer; users were not very experienced and so described their interests with entire paragraphs, filled with many unimportant words.

Query generality, that is the number of relevant documents per query, is illustrated in the next two columns of the table. The first number is given in absolute terms and the second is a percentage of the total number of documents. CACM queries have very few relevant documents, both in actual numbers and as a percentage. INSPEC questions have roughly twice as many. Medlars queries have slightly fewer relevants, but the percentage of documents that are relevant is much higher. And the ISI collection seems to have too many relevant documents per query, both in absolute numbers and as a percentage; those queries were much too vague to give small, sharply defined relevant sets.

The final two columns of Table 2 illustrate the retrieval behavior of the queries using cosine correlation by considering the top 10 ranked documents. For Medlars, almost 4 of the first 10 are relevant on average, indicating that the queries are likely to have high precision. Further, almost 20% of the relevant documents are already retrieved, indicating that recall is probably good as well. For CACM, recall seems high, but precision is reduced to roughly half. In the other two collections, around 2 relevant documents are found in the top 10 retrieved, so precision is low as well. And there is even a lower percentage of the number of relevant documents that are identified. For ISI, in particular, only 4% of the relevant documents are found in the top 10, indicating that recall will be rather low. With so many relevant documents per query, such behavior could be expected.

Table 3 gives descriptive details about the five different groups of query collections.

Table 3: Query Collections Summary

| Coll. Name | NL Queries No. | NL Queries Origin | BL Queries No. | BL Queries Origin |
|---|---|---|---|---|
| ADI | 35 | Written by 2 Harvard computer science students | 35 | author- 4 forms librarians- 1 form each |
| CACM | 52 | Cornell & other computer personnel | 52 | 2 comp. sci. grad. students- 1 form each |
| INSPEC | 77 | Students etc. at Syracuse Univ. | 77 | 7 searchers at Syracuse |
| ISI | 76 | ADI, ISPRA & SIGIR Forum abstracts | 35 | ADI part only (see ADI coll. above) |
| Medlars | 30 | NLM files | 30 | NLM searchers, then expanded using MESH |

There is one set of natural language queries for each of the five, but in some cases there are several Boolean forms based on each natural language query.

(1) The ADI collection originally had 35 natural language queries. Three searchers then each rephrased the natural language questions into a Boolean representation.

(2) 52 CACM queries were submitted from a variety of sources, and two students (in the Cornell graduate level information retrieval course) responsible for creating the query collection each proposed Boolean forms.

(3) One set of INSPEC queries was selected from the various types of representations devised by seven different searchers working at Syracuse University.

(4) For ISI, the same 35 natural language and Boolean queries used for ADI were employed. Multiple concept type testing, however, required more natural language queries, so 41 additional ones were selected from a set of queries for the ISPRA collection and from some new queries based on abstracts published in issues of the *ACM SIGIR Forum.*

(5) Medlars natural language queries were as provided from the National Library of Medicine (NLM) files and Boolean forms were subsequently constructed at Cornell. The latter were based on Boolean expressions employed by searchers. An expansion process utilizing the Medical Subject Headings of the MESH thesaurus [National Library of Medicine 1968] enabled all category names to be replaced by words that might occur in document texts.

## 2. Multiple Concept Types

To supplement the discussion of documents in section 1.2.1, it is necessary to consider the multiple concept model proposed in [Fox 1983b]. The word "concept" is used to indicate a basic item of indexing information so that a collection of concepts identifies the context of the article or monograph. Concepts can be "terms" – really approximations to stems of words from the title, abstract, or supplied list of keywords – or they may be indexing categories or bibliographic connection markers (e.g., a pointer to a document co-cited with the one being indexed).

Table 1 indicates how many distinct terms are present in the dictionary of word stems for each collection. The dictionary is produced as part of the automatic indexing process, and is the minimum size required to accommodate all distinct stems after a stop word list has been applied to document texts. Vocabulary size goes up rapidly as the number of documents increases (e.g., see the jumps for ADI and Medlars), but then gradually tapers off (e.g., see values for the large CACM and INSPEC collections). The size is somewhat influenced by the nature of the documents; there are more distinct stems in the 1033 Medlars medical articles which have a specialized terminology than for the 1460 ISI information science records which contain fewer technical names.

The multiple concept type model introduced in section 1.1 calls for additional concepts beside terms. In the ISI collection, two additional types, authors and co-citations, were considered. Author names were entered in the author dictionary, which had a total of 1255 items. Co-citations have no attendant dictionary. Since each document can, a priori, be co-cited with any other, the number of concepts equals $N$, the number of collection documents. The value of the $j^{th}$ co-citation concept for the $i^{th}$ document, namely $cc_{ij}$, is the number of times the $i$ and $j^{th}$ documents are co-cited.

Table 4 gives statistics for ISI subvectors.

Table 4: ISI Subvector Length Statistics

| Statistic Measured | Subvector au | cc | tm |
|---|---|---|---|
| mean | 1.4 | 54.0 | 49.6 |
| median | 1 | 40 | 47 |
| min | 1 | 1 | 8 |
| max | 7 | 276 | 179 |
| stdv | 0.8 | 46.4 | 21.5 |
| Total no. of concepts | 1255 | 1460 | 7392 |

Even though there are only 1255 different authors for 1460 documents, each document has an average of 1.4 authors. The term subvectors are much longer, averaging around 50 concepts, but the standard deviation (stdv) indicates that the distribution is spread rather widely. Co-citation subvectors have about the same average length, but an even larger standard deviation. Since the ISI articles were chosen as ones with many citations, and since the collection covers the field of information science for a number of years, it is not surprising that there are so many entries in the co-citation submatrix. Since the lengths of terms and co-citation vectors are comparable, comparisons between the two as to their relative utility for retrieval should be of considerable interest.

The CACM collection was developed in part to allow testing of other concept types besides those found in ISI documents. *Computing Reviews* categories ($\vec{cr}$) allow a manual indexing system to be mixed in with automatic indexing entries. Bibliographic coupling ($\vec{bc}$) indicates the number of references shared between two documents' bibliographies. Links ($\vec{ln}$) are references to or citations from other articles.

Table 5 gives statistics for the various concept types in the CACM document collection.

Table 5: CACM Subvector Length Statistics

| Statistic Measured | Subvector au | bc | cc | cr | ln | tm |
|---|---|---|---|---|---|---|
| mean | 1.3 | 4.2 | 3.7 | 1.2 | 2.7 | 25.0 |
| median | 1 | 0 | 0 | 0 | 2 | 15 |
| min | 1 | 0 | 0 | 0 | 1 | 1 |
| max | 7 | 183 | 111 | 28 | 74 | 168 |
| stdv | 0.7 | 10.8 | 10.7 | 1.9 | 3.1 | 22.7 |
| Total no. of concepts | 2647 | 3204 | 3204 | 200 | 3204 | 10446 |

Note: The minimum length for $\vec{ln}$ is one since a document is, by definition, linked to itself, and so the diagonal of the submatrix is set to ones.

Since $\vec{bc}$, $\vec{\ln}$, and $\vec{cc}$ subvectors are based only on connections among the chosen CACM articles, their mean length is much shorter than that of the $\vec{tm}$ subvector. The $\vec{cc}$ subvectors for CACM have average length of less than 4 concepts, while those for ISI have over 50. In a larger document collection, with better total coverage of a subject discipline, more of the citations and references would be internal to that collection and so these subvectors would probably be longer. Thus, while the ISI collection has unusually long bibliographic connection subvectors, the CACM collection has abnormally short ones, and any evidence from the CACM tests that these subvectors are useful should be suggestive of their value in a more realistic environment with large numbers of documents.

It should be noted that links are relatively well bounded in number since a given article rarely has a very long bibliography and the quantity of citations to an article is limited by how many entries in the collection deal with that specific topic. Co-citations are somewhat more numerous, especially in such a homogeneous collection, since many later articles may refer to a given article as well as others appearing in the same journal. Bibliographic couplings are slightly more plentiful, perhaps because many articles cite "classics" in their sub-area, and because there is a good deal of referencing of CACM articles by CACM authors.

A final note about the various concept types is that in both the ISI and CACM collections, the total number of terms exceeds the number of documents. For very large collections, the opposite would be true. However, there is no reason to believe that such a consideration would affect retrieval behavior, whereas subvector length is likely to be an important consideration, relating to the specificity of indexing.

## 3. CACM

## 3.1. Documents

Robert Dattola of Xerox Corporation provided a magnetic tape containing the title, abstract, author list, keywords, *Computing Reviews* categories, and date of publication of articles published in the *Communications of the ACM* from the earliest issue, in 1958, to the last number in 1979. The collection format was changed, document numbers (dids) were assigned, and editing was done to correct spelling and typographical errors. Some duplicates were eliminated and missing articles added, and a final renumbering took place.

Carol Fox and Jill Warner looked through printed copies of each article to locate the bibliography. For each article, a list of the "dids" representing articles referenced in the CACM collection was eventually obtained. Many articles were located using a chronologically ordered list of all articles in the collection. However, since there were many errors in bibliographies, a special search program was written to identify the nearest matches to a supplied reference – to compensate for errors in year, month, author name, and title.

The purpose of obtaining that data was to form bibliographic subvectors. Based on the above mentioned lists, a relational form was produced:

Raw_data ( citing, cited )

which contained pairs of identifiers for the citing article and the one contained in the article's bibliography. Figure 1 shows the steps required, in QUEL like notation [Stonebraker et al. 1976], to produce from Raw_data first normal form [Date 1982] versions of the desired relations for $\vec{bc}$, $\vec{\ln}$, and $\vec{cc}$ subvectors.

Figure 1: Relational Processing to Obtain
CACM Bibliographic Subvectors

Given: Raw_data (citing, cited)

Desired:

```
BC (citing1, citing2, coupling_no)
LN (link1, link2)
CC (cited1, cited2, co-citing_no)
```

QUEL Like Statements:

```
1. For BC
     Modify Raw_data to hash on cited
     Range of bc1 is Raw_data
     Range of bc2 is Raw_data
     Retrieve into bc_entries
            ( citing1 = bc1.citing,
              citing2 = bc2.citing,
              counter = bc1.cited )
            where bc1.cited = bc2.cited
     Range of bce is bc_entries
     Retrieve into BC
            (bce.citing1, bce.citing2,
             coupling_no = count ( bce.counter ) )

2. For LN
     Retrieve into LN
            ( link1 = bc1.citing,
              link2 = bc1.cited  )
     Append to LN
            ( link2 = bc1.citing,
              link1 = bc1.cited  )
```

Figure 1 continued: Relational Processing to Obtain
CACM Bibliographic Subvectors

```
3. For CC
    Modify Raw_data to hash on citing
    Range of cc1 is Raw_data
    Range of cc2 is Raw_data
    Retrieve into cc_entries
            ( cited1 = cc1.cited,
              cited2 = cc2.cited,
              counter = cc1.citing )
            where cc1.citing = cc2.citing
    Range of cce is cc_entries
    Retrieve into BC
            (cce.cited1, cce.cited2,
             co-citing_no = count ( cce.counter ) )
```

---

The actual processing was similar. Vectors were eventually formed from the relations after appropriate sorting and compression to unnormalized lists.

Since the CACM collection is the first one studied with so many different concept types, and since the data is available for a number of years, various histograms, scatter plots, and charts are given in the next subsection. The intention is to illustrate the form, content, and distribution of the different types of bibliographic concepts.

### 3.2. Illustrative Charts and Figures

Table 5 gave statistics on the CACM subvector lengths. However, a much more detailed understanding can be gleaned from the graphical presentations below. To begin with, Figures 2 through 6 are histograms with mean values given for various statistics for each year in the period 1958-1979.

Figure 2 shows the number of articles each year. Clearly, the publication grew in size during the early years, and then there were changes in editorial policies at various later times. For example, in the early years, computing algorithms in fairly large numbers were published, each one counting as an article. Subsequently this practice was discontinued and algorithms were collected for a separate publication.

Figure 3 shows the average number of citations to an article for each year. Lower values in later years can be explained by the fact that there were few articles in the collection that were published subsequently, and so citations were not possible. Early articles were not highly cited, perhaps because the first volumes had many methodological or other reports that were not of great interest afterwards. The peak in citations in the middle years is attributable to the fact that CACM was a key publication in computer science during that time, and many important developments were described there, especially after 1962. One might expect that the three bibliographic subvectors would vary in length depending in part on the distribution just described of citations.

Figure 4 shows the average number of bibliographic couplings per year. In general, the number increases as time goes by, since there are more available prior articles in the collection that can be referred to. Thus, the number of couplings for years 1976 through 1978 are very high – bibliographic references can be to any previous article, even back to 1958. All in all, the distribution is a fairly uniform one, perhaps explaining why there is such a large standard deviation in $\vec{bc}$ subvector lengths.

Figure 5 gives the average number of links for each year. Since links include both references and citations, it is not surprising that the curve is a relatively flat one. Only for the early years, when there were not many prior articles to refer to and when the subject matter was not such as to elicit many later citations, is the level of links fairly low.

Figure 6, showing average number of co-citations, has a very different form than that of the other figures shown. A high peak appears near the middle, and lower values are at either tail. Apparently, 1966 was a good year, since articles then were co-cited with many others. Referring back to Figure 3

Figure 2: Mean Number of CACM Articles Published Each Year

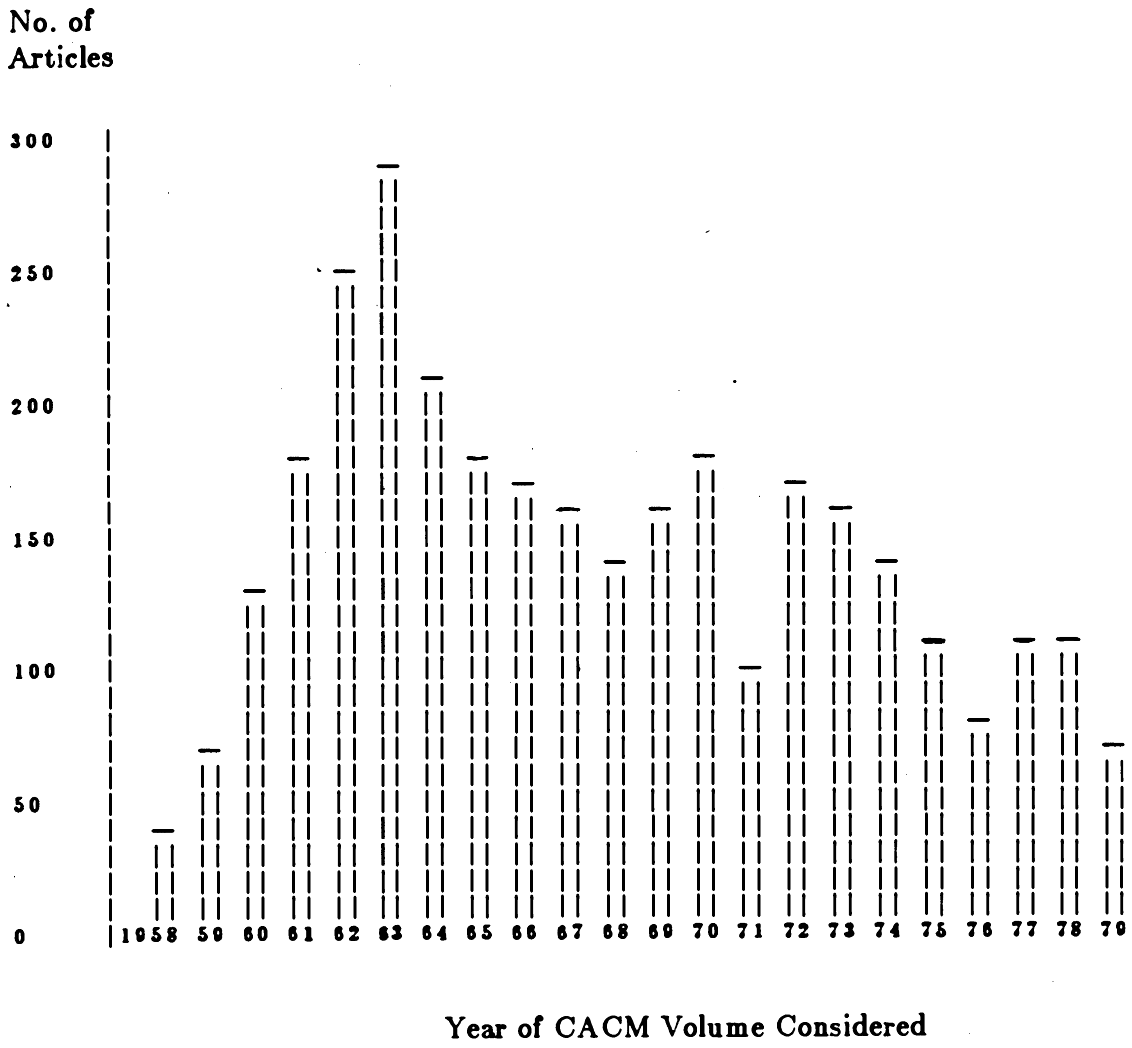


Figure 3: Mean Number of Citations Per Article Per Year

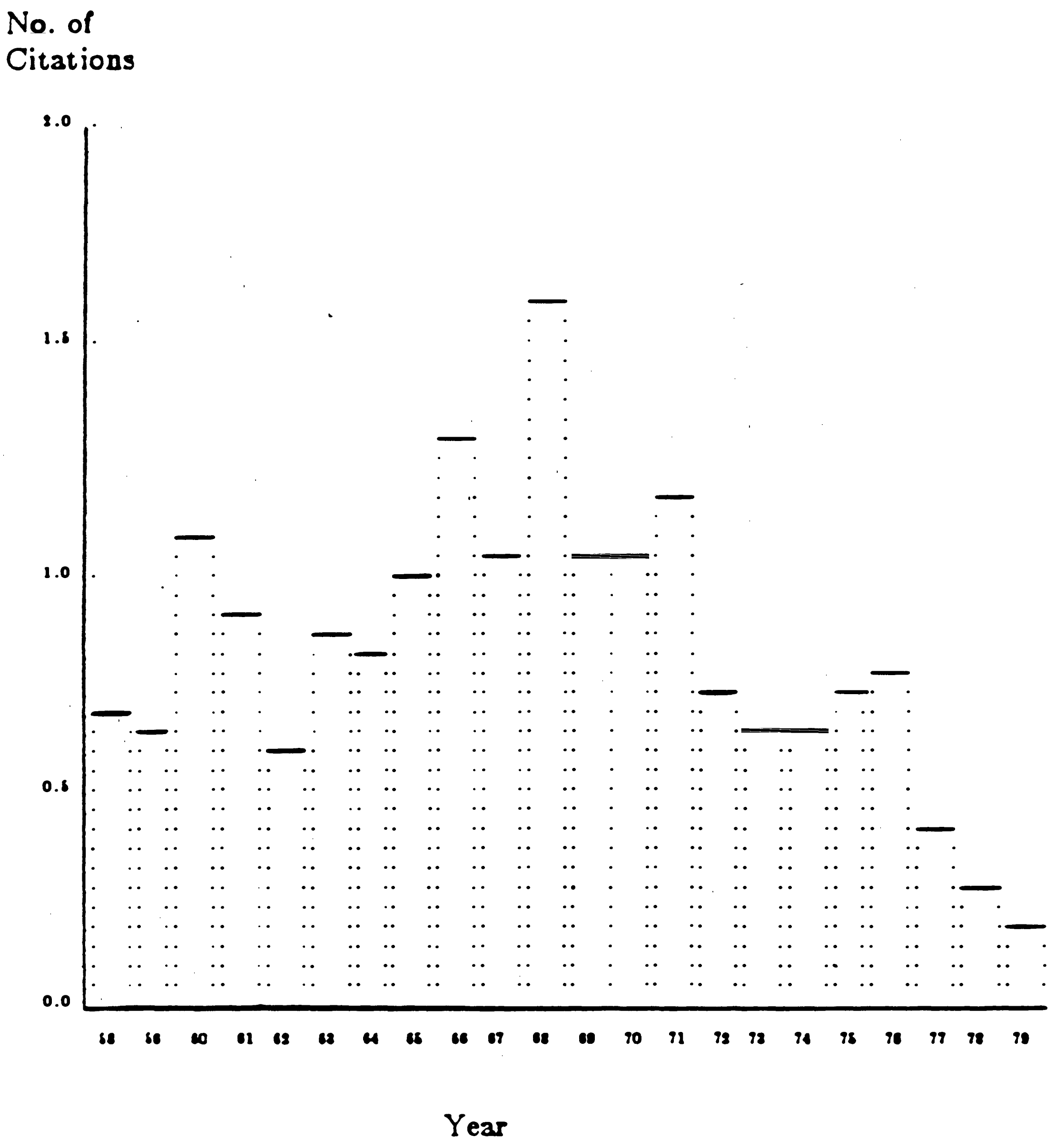


Figure 4: Mean Number of Bibliographic Couplings
Per Article By Year

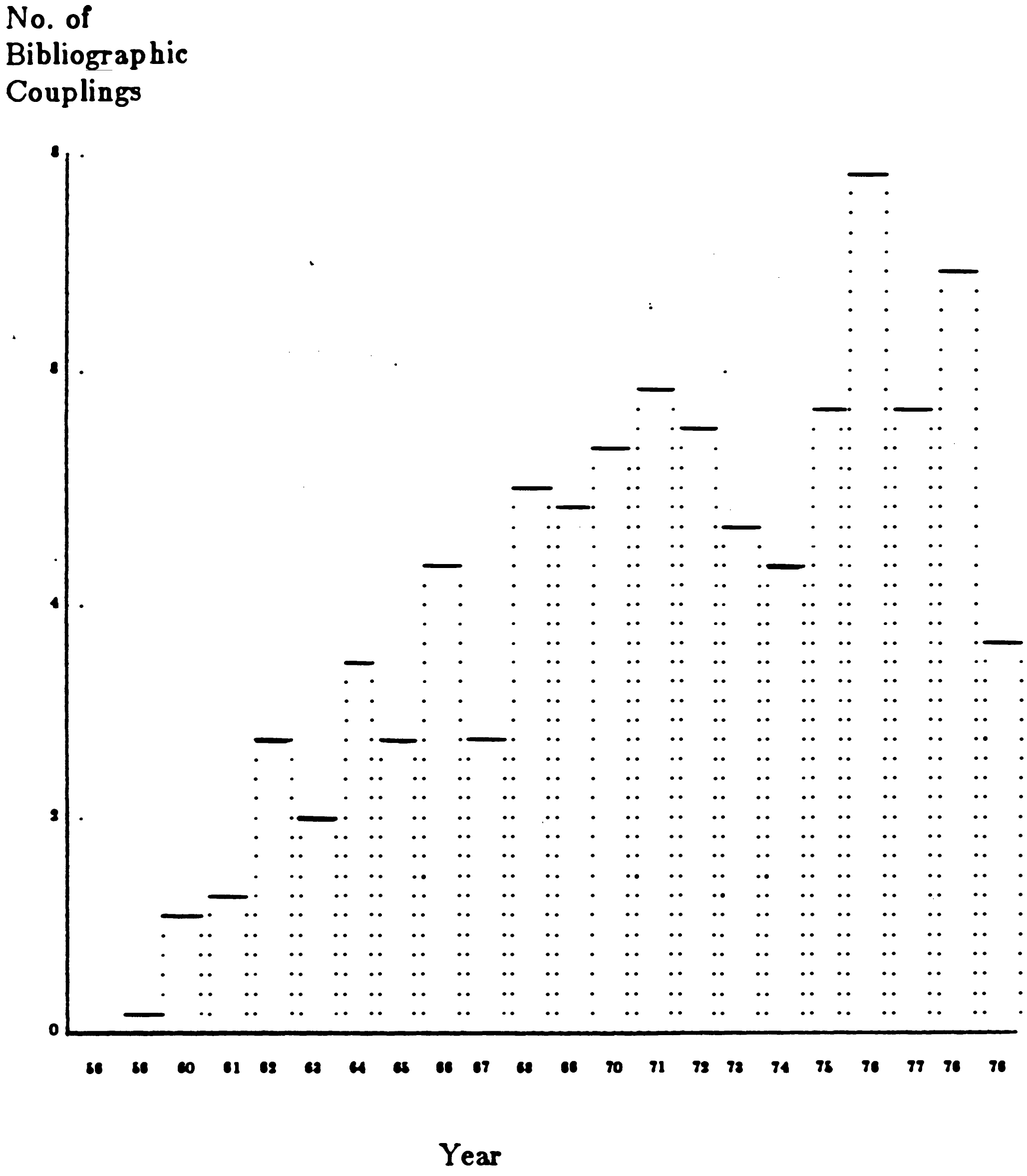
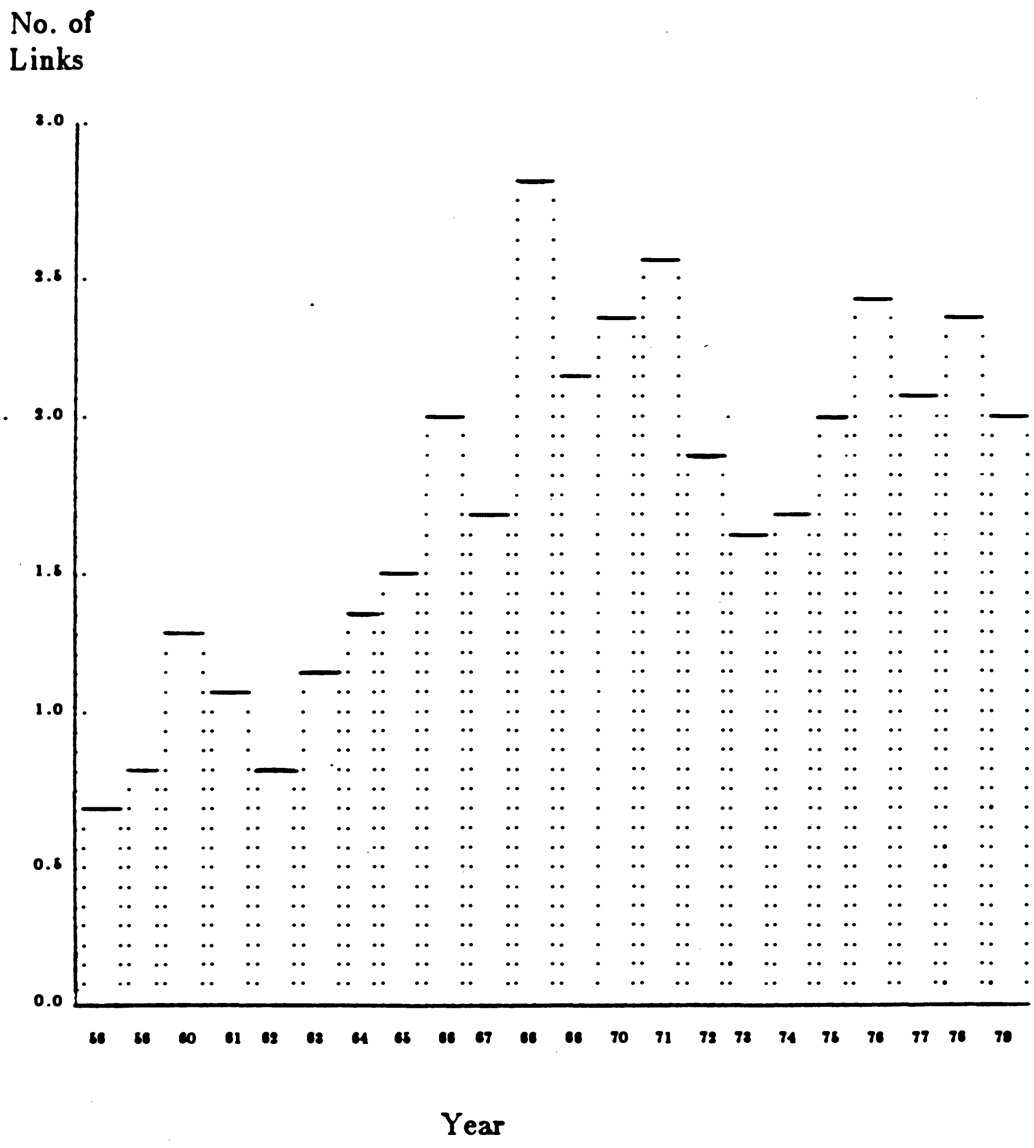


Figure 5: Mean Number of Links Per Article By Year




one sees that the most citations, after those to 1968, were to articles in 1966. Indeed, an examination of articles in that year does reinforce the conviction that there were many important articles that year, and that they related to a considerable number of other articles. The number of co-citations could serve as another indication of merit, to go along with the number of citations, but probably the emphasis on number of related articles rather than number of citing articles would reduce the appeal of such a proposal. In any case, one would expect the average number of co-citations to be greatest during the middle years of the range, since there are more articles later that can cite both the given article and one other (which could be either earlier or later).

Figure 7 highlights the difference in the distribution of $\vec{bc}$ and $\vec{cc}$ lengths over the years. Bibliographic couplings are greatest towards the end of the range, and co-citations peak between the beginning and middle of the period. Hence, when mean $\vec{bc}$ length is divided by mean $\vec{cc}$ length, the ratio is around 1-2 for middle years, lower beforehand, and higher after. An obvious strategy is to somehow add together the effects of bibliographic coupling and co-citations, since then the overall depth of bibliographic connection based indexing would be fairly uniform over the complete range of years of interest.

Turning next to distributions of frequencies for the various subvectors, one sees the type of curves that are expected, in Figures 8 through 11. The number of documents for each subvector length is shown. Almost every article has fewer than 10 *Computing Reviews* categories assigned or 10 links to other articles. There are a goodly number of documents, however, with fairly long $\vec{bc}$ and $\vec{cc}$ subvectors, as can be seen in Figures 10 and 11.

Most articles have less than 4 entries in their $\vec{cr}$ subvector, since categories are usually fairly broad and few articles deal with a much wider diversity of subjects. Usually at least a few categories are assigned to an article, so Figure 8 actually shows a slight initial increase before the usual descending portion of the curve is reached.

Figure 6: Mean Number of Co-Citations Per Article By Year

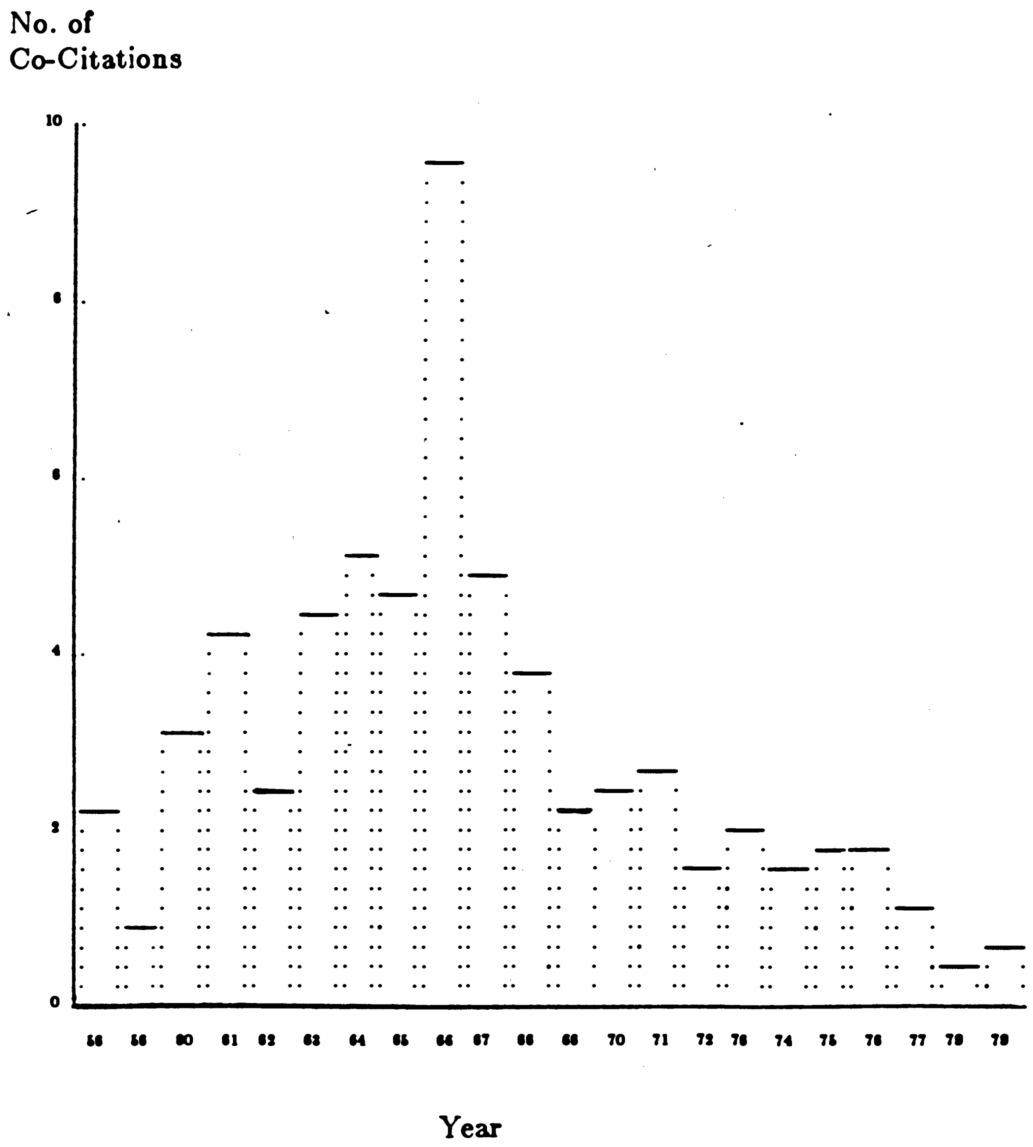


Figure 7: CACM Bibliographic Connection Ratio for 1958-79
(For Each Year, Mean Value of BC/CC, Counted When
Both Values are Positive for an Article)

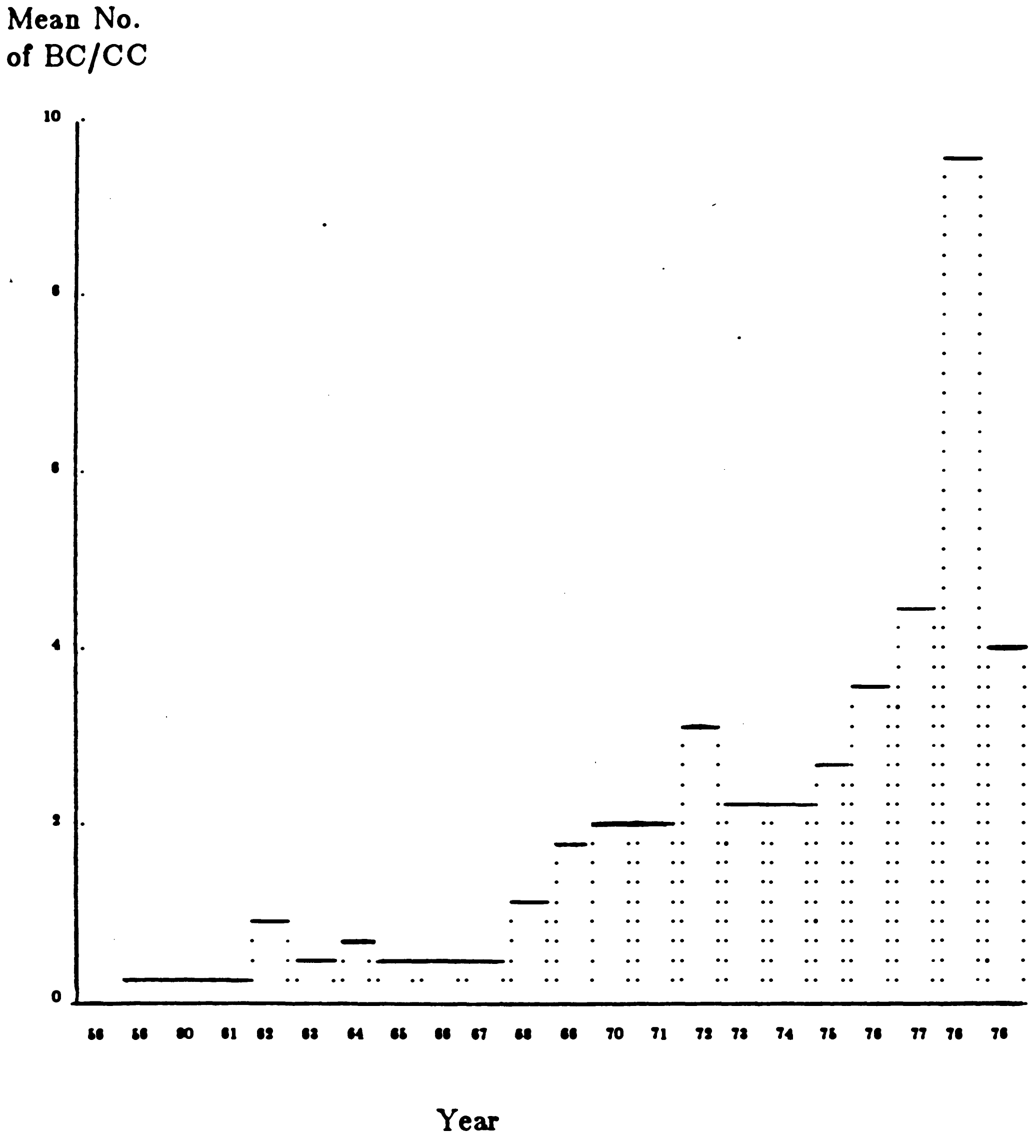


Figure 8: CACM Frequency Distribution for $\vec{cr}$ Subvector Length

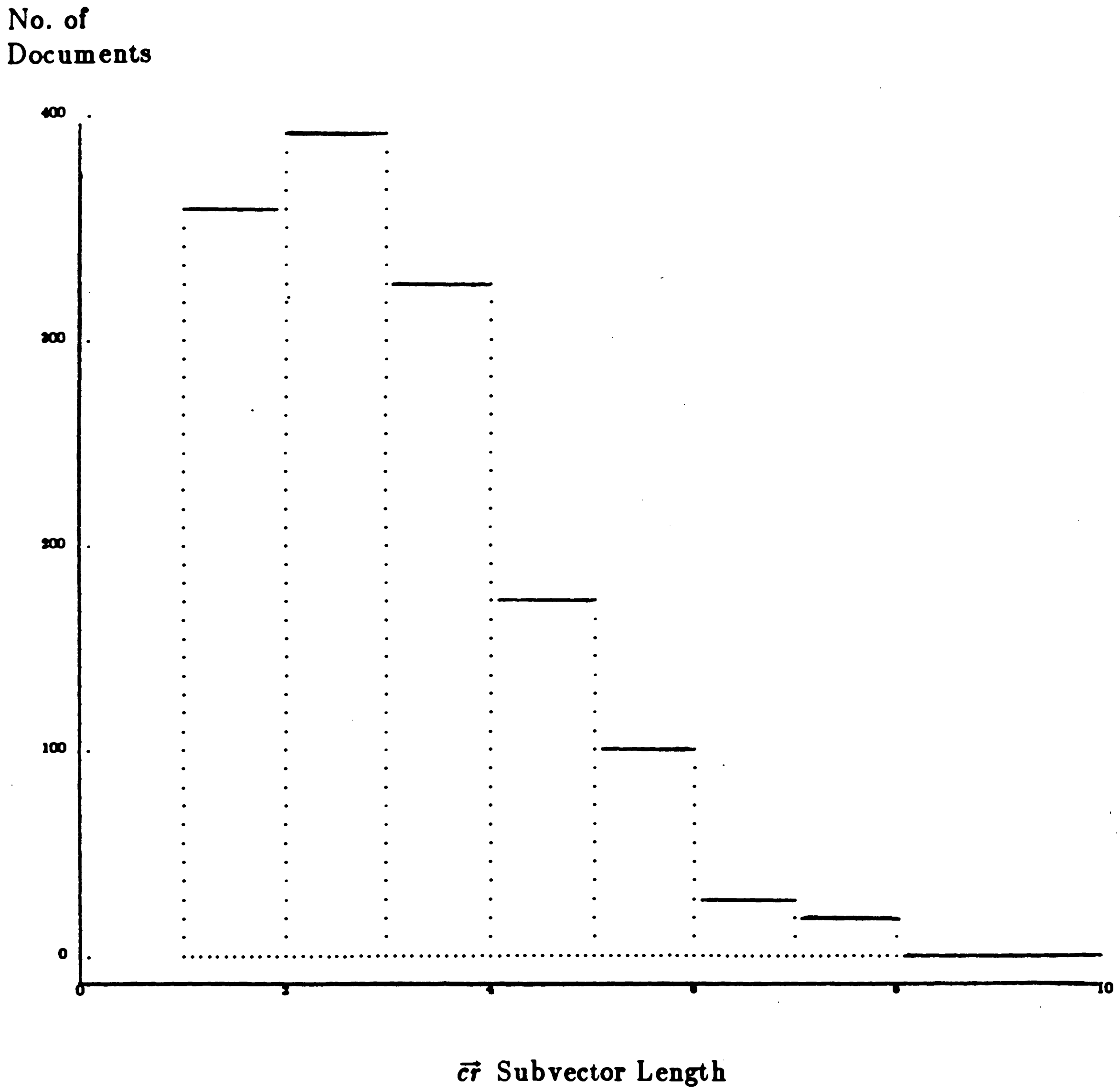


Figure 9: CACM Frequency Distribution for $\vec{\ln}$ Subvector Length

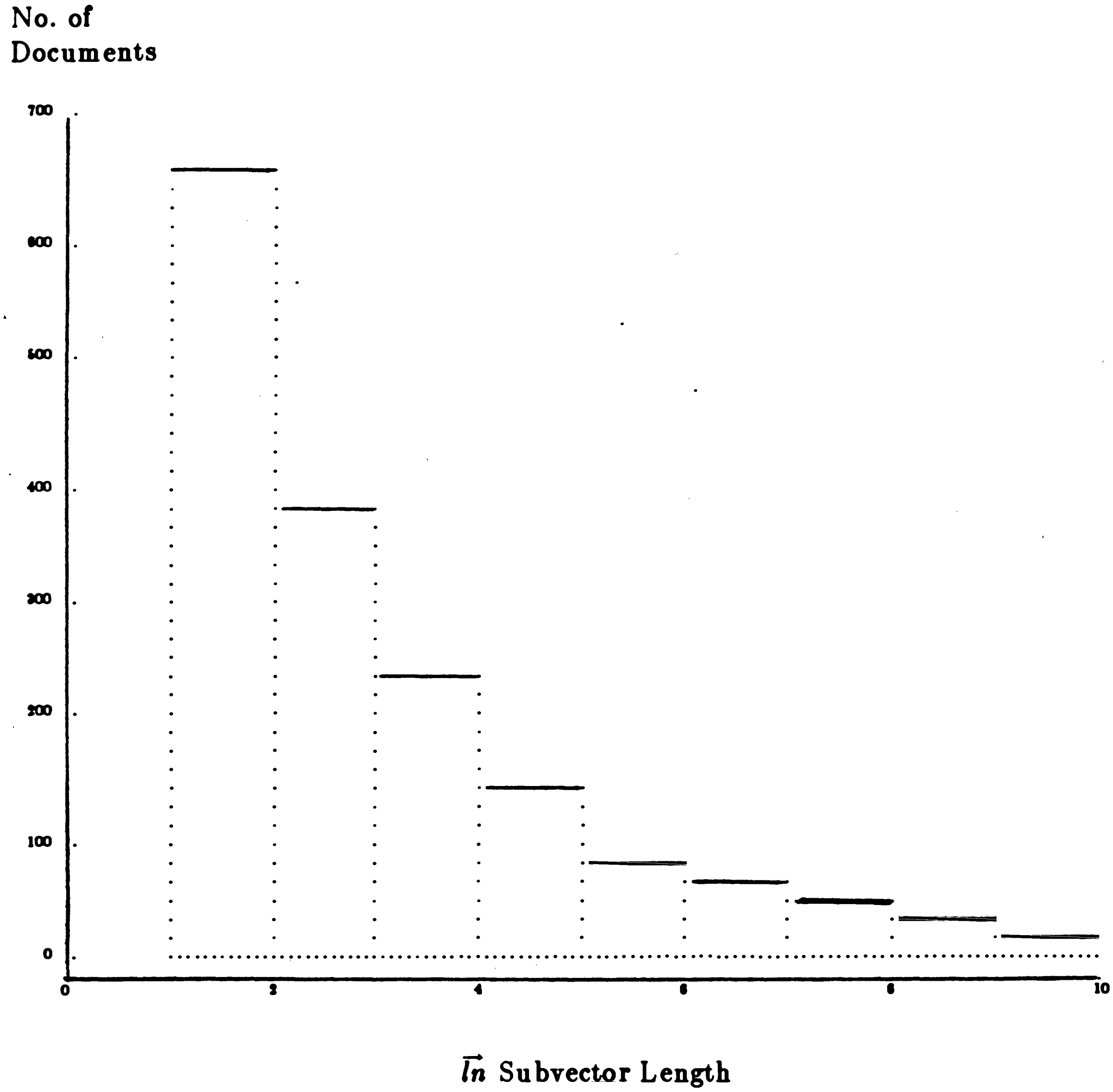


Figure 10: CACM Frequency Distribution for $\overrightarrow{bc}$ Subvector Length

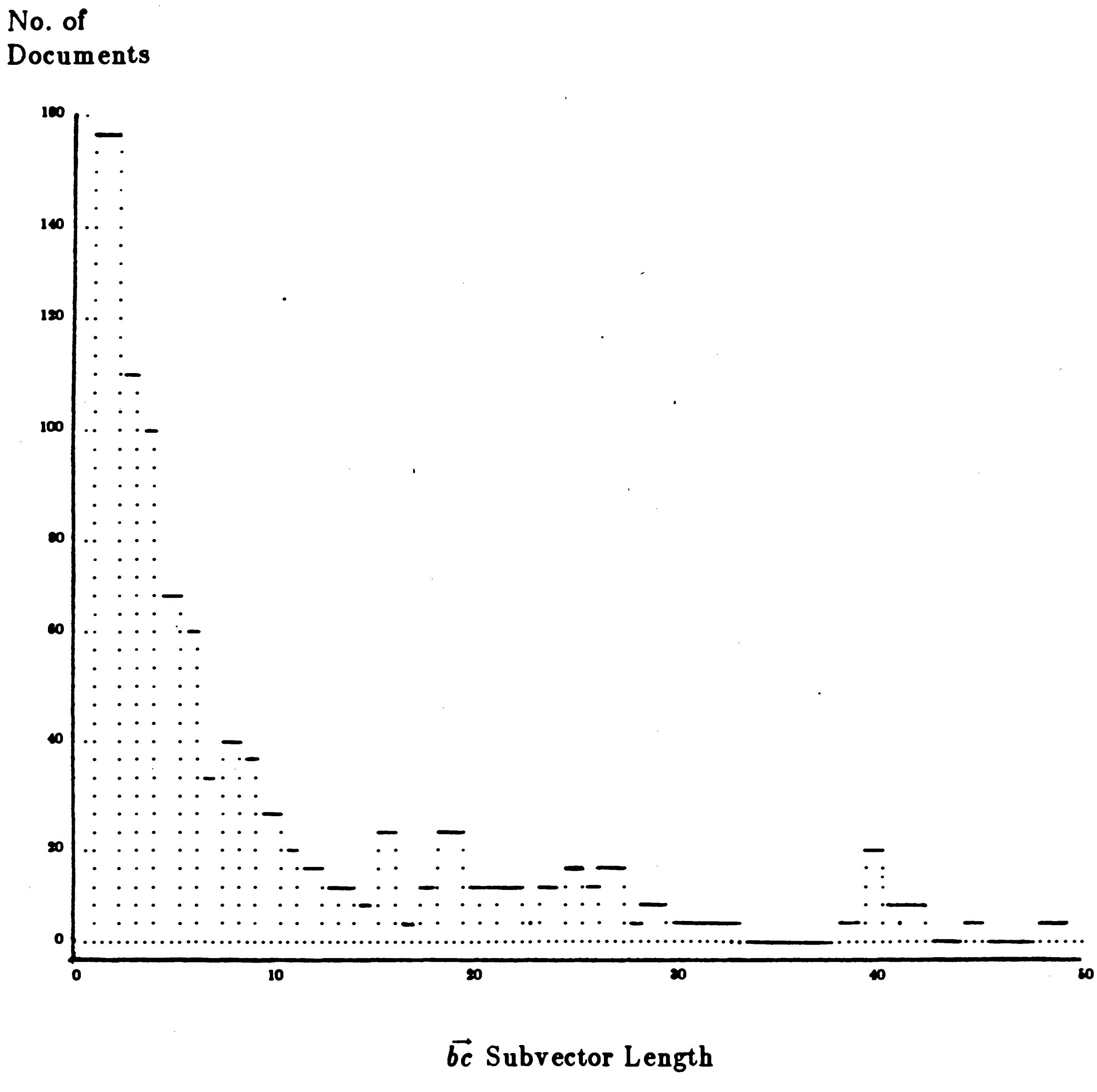


Figure 11: CACM Frequency Distribution for $\vec{cc}$ Subvector Length

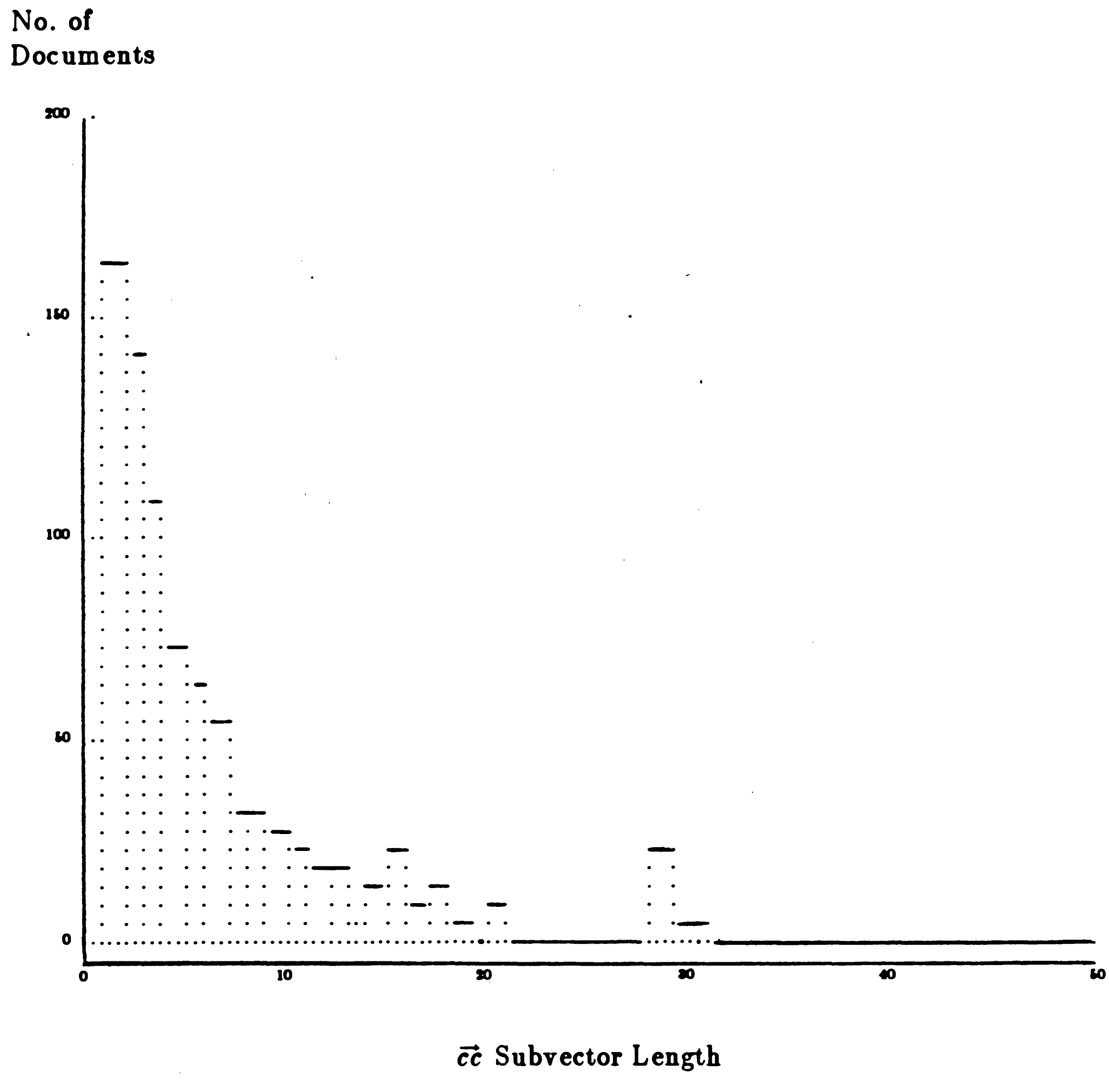


Figure 9, dealing with links, appears to follow the distribution predicted by Zipf's law [Salton 1975 page 169]. Figures 10 and 11 do also, except as one moves beyond frequencies of around 15. After that, those two curves seem to level off, showing occasional peaks randomly distributed.

To better compare the subvector length distributions, scatter plots are given in Figures 12 through 15. Figure 12 shows bibliographic couplings against citations. There is no obvious linear relationship, but values are concentrated in the region close to the origin. For one to ten citations, there are usually less than 50 to 100 bibliographic couplings. There are some articles with many citations but few bibliographic couplings – perhaps they are referred to by various authors, but, possibly because they are original or independent studies, they do not refer to many other works.

Figure 13, comparing links and citations, does follow a linear relationship. As expected, since citations are one type of link, when citations increase, there are longer $\vec{ln}$ subvectors.

Figure 14, showing co-citations versus citations, has separated concentrations that would best be modelled by two straight line relationships. For 0-40 co-citations, the number of co-citations gradually increases with number of citations – these are for normal articles which as they are cited more are accordingly co-cited with a greater number of other articles. The other concentration is of articles which have more than 50 co-citations, and which follow a different relationship with the number of citations. Roughly, the number of co-citations again goes up with number of citations, but begins at a higher level. Perhaps these are articles referred to by review papers – they are co-cited with many others, but need not be highly cited. One might expect, then, that there is a linear relationship between citations and co-citations, but that the number of co-citations is also affected by other phenomenon, such as by whether the given article is referenced in a review paper.

Figure 15 interrelates bibliographic coupling and co-citations. Though there is a concentration of articles near the origin with fairly low values of each count, the distribution does seem rather uniform

Figure 13: CACM Bibliographic Connections Scatter Plot for $\overrightarrow{ln}$
Comparing Citations Versus Links

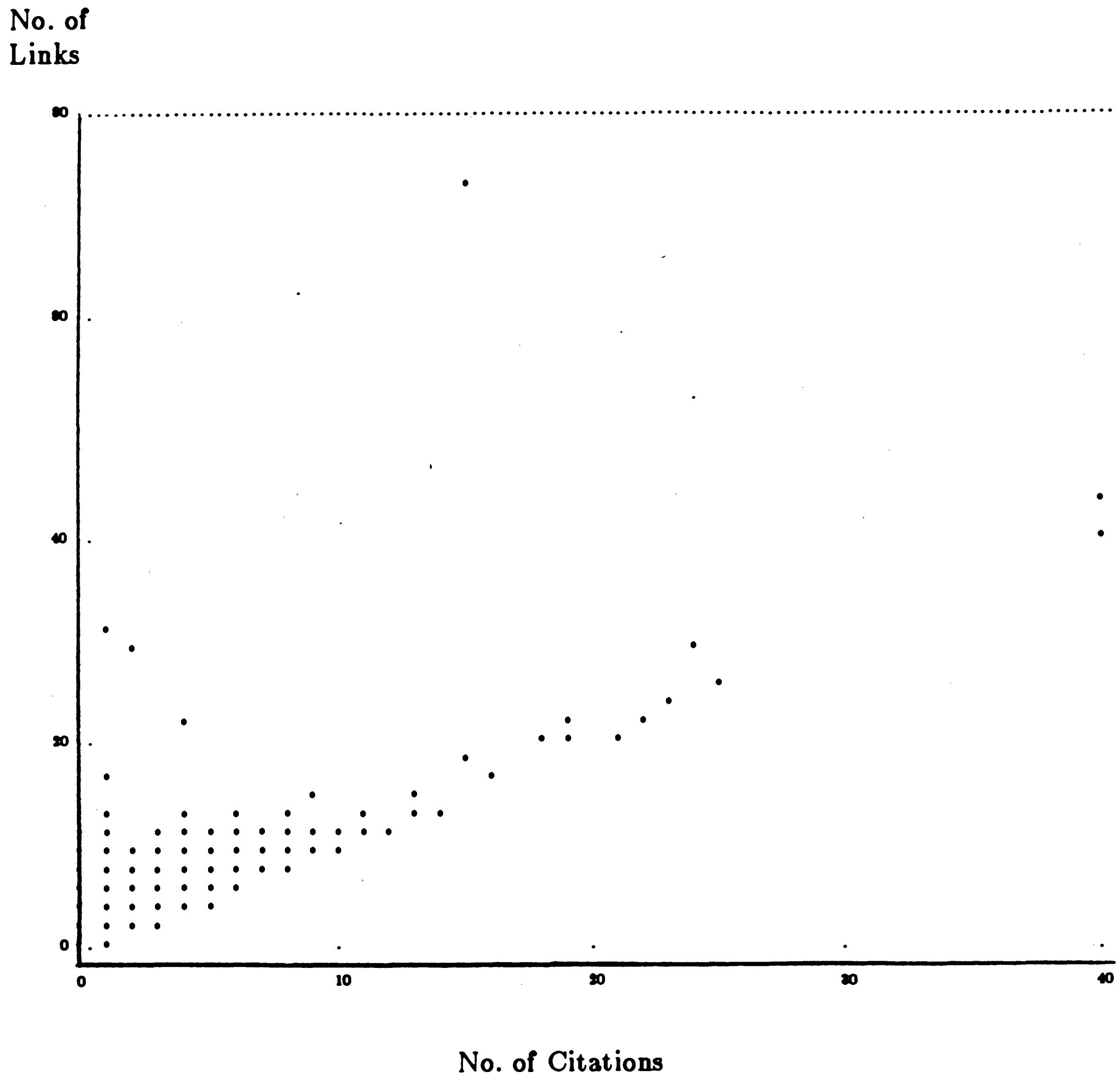


Figure 12: CACM Bibliographic Connections Scatter Plot for $\vec{bc}$
Comparing Citations Versus Bibliographic Coupling

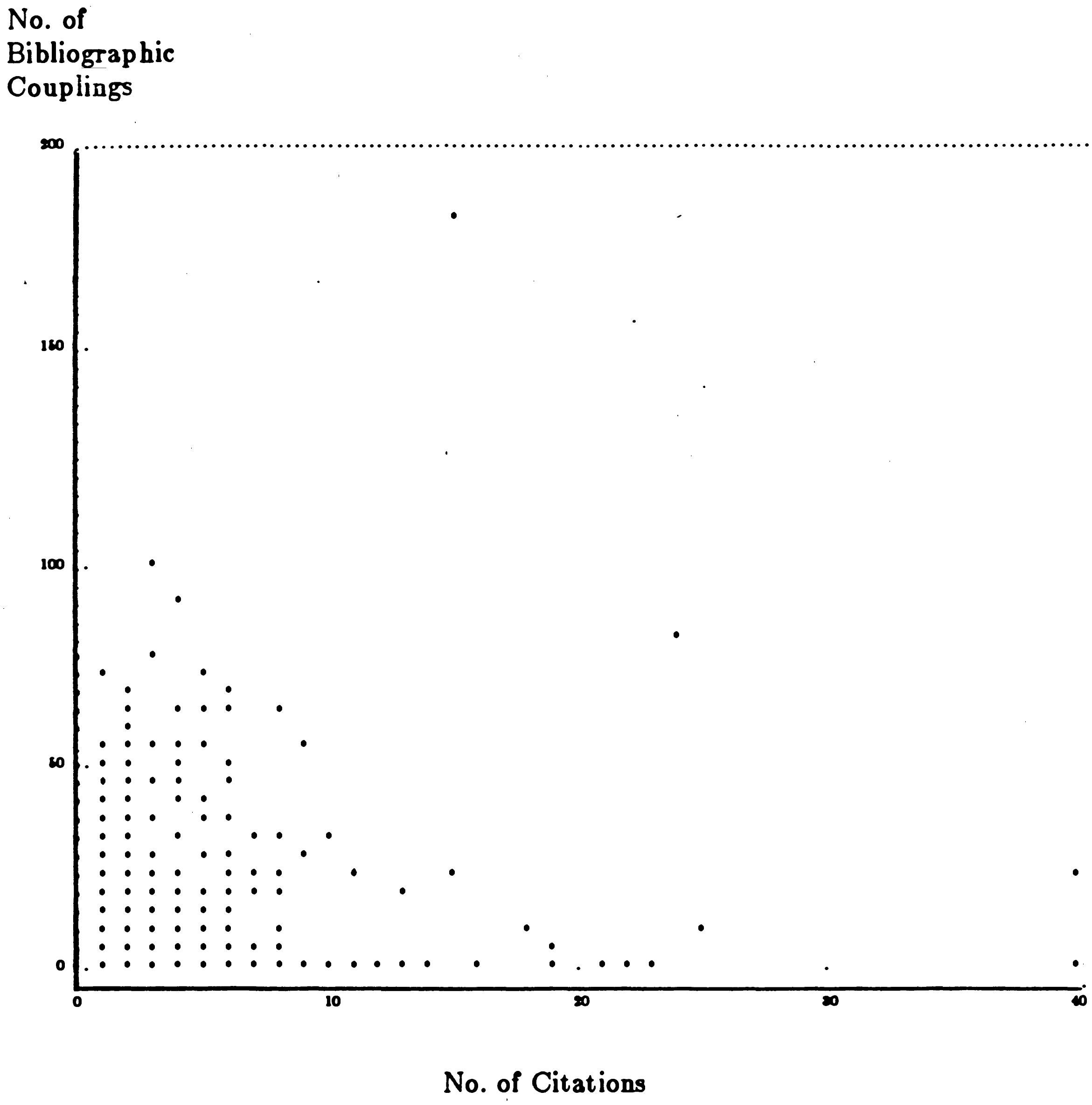


Figure 14: CACM Bibliographic Connections Scatter Plot for $\vec{cc}$
Comparing Citations Versus Co-Citations

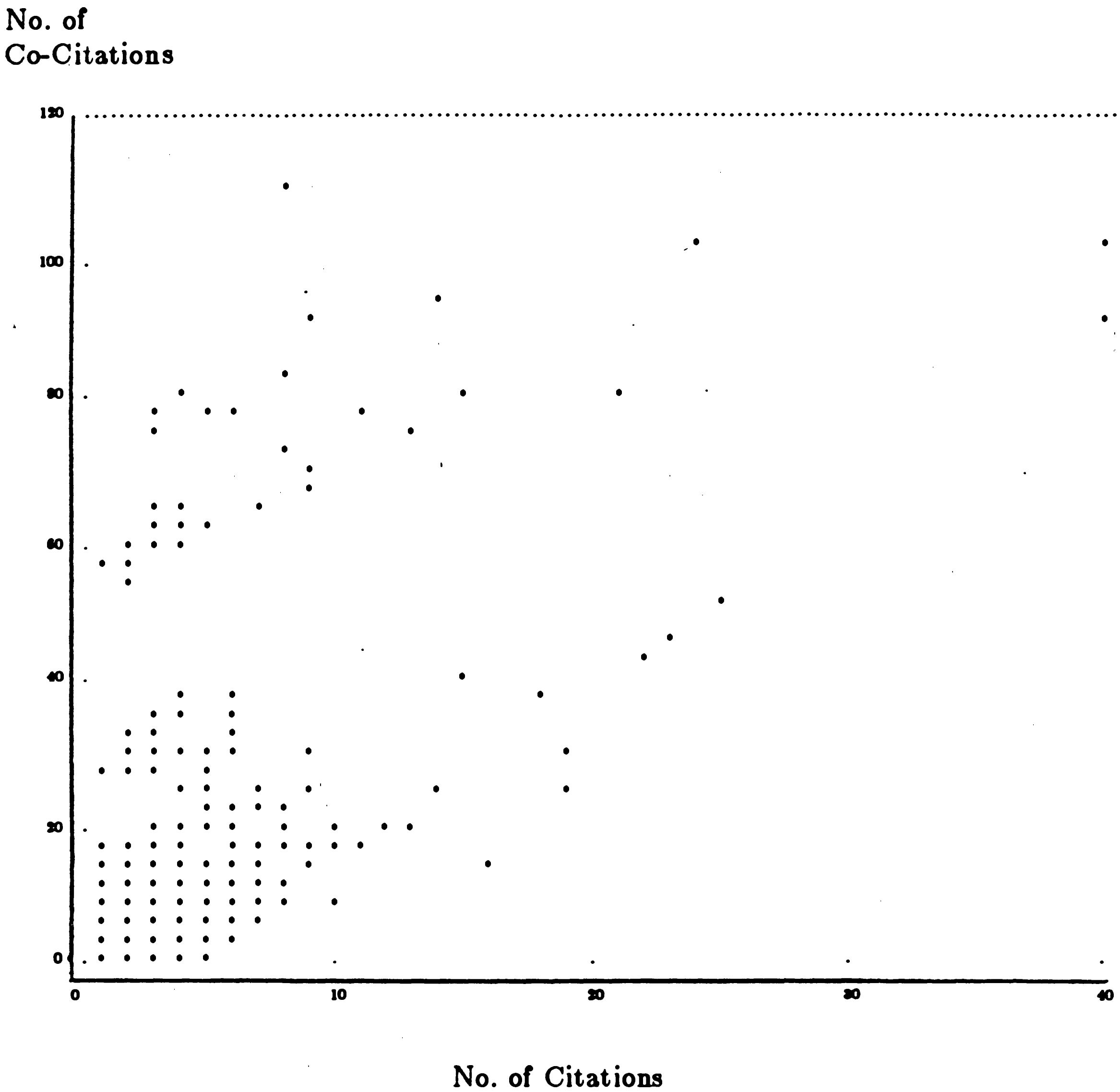
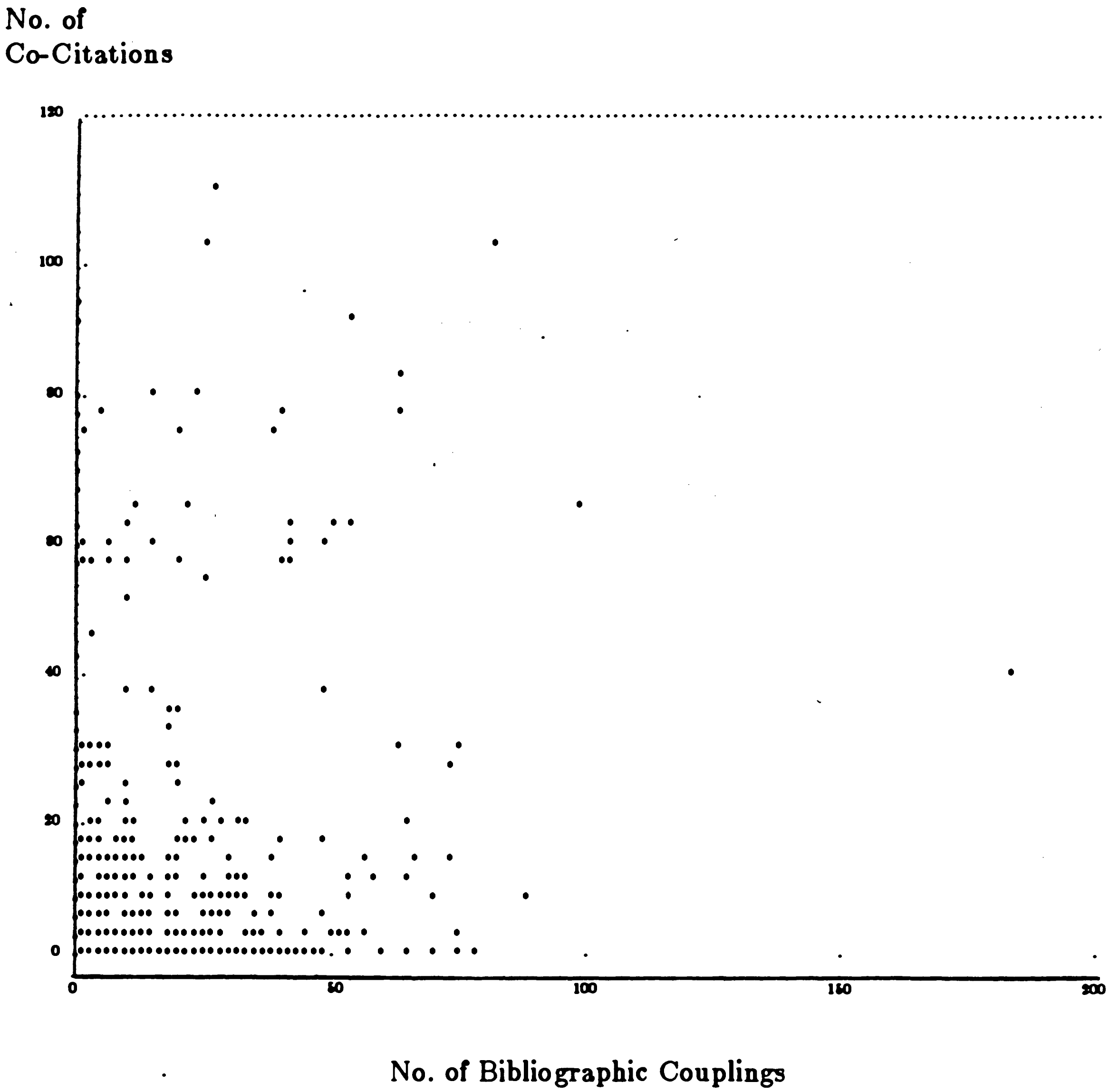


Figure 15: CACM Bibliographic Connections Scatter Plot for $\vec{bc}$, $\vec{cc}$
Comparing Bibliographic Coupling Versus Co-Citations



or random elsewhere. Once again, it seems reasonable to assume that the subvector lengths of $\vec{bc}$ and $\vec{cc}$ are relatively independent, and so combining their effects might be of value.

Having seen ample graphical evidence, it is now appropriate to examine several tables of information about the CACM bibliographic subvectors. In Table 6, categories have been selected to separate low, medium, and high levels of one variable. Accordingly, one is shown the corresponding average number for a second variable, for each category of the first. In all examples, as levels go up there is a corresponding rise in the second variable. Thus, for various levels of citations, the number of references goes up, since articles that are cited more tend to have more references. Likewise, articles with more references tend to receive more citations, though the variation between medium and high levels of references is probably not significant. In other words, articles with few references receive few citations, but one cannot predict the number of citations when the number of references is medium or high.

A more dramatic interrelationship is shown between quantities of citations and of bibliographic couplings. As the number of citations goes up, the number of couplings does too. Possibly, articles that tend to have wide appeal are more tightly coupled with other articles – they are in densely interrelated areas of the collection. A similar but not quite as dramatic effect can be seen when number of co-citations is related to level of bibliographic coupling. Possibly due to articles that are highly connected to others, the amount of co-citation roughly doubles as one proceeds to each higher level of bibliographic coupling.

Tables 7 through 9 show the percentages of the documents in the collection that are in various categories relating to levels of subvector lengths. Table 7 considers a single subvector at a time, while the others deal with pairs or triples, respectively.

Table 6: No. of Bibliographic Connections for Various Levels of other Measures

No. of References vs. No. of Citations

| | Number of Citations | | |
|---|---|---|---|
| | Low (0-1) | Medium (2-4) | High (5+ ) |
| No.Refs. | 0.69 | 1.36 | 2.18 |

No. of Citations vs. No. of References

| | Number of References | | |
|---|---|---|---|
| | Low (0-1) | Medium (2-4) | High (5+ ) |
| No.Cits. | 0.67 | 1.59 | 1.70 |

No. of Bib.Couplings vs. No. of Co-Citations

| | No. of Co-Citations | | |
|---|---|---|---|
| | Low (0-1) | Medium (2-4) | High (5+ ) |
| No. BC | 2.33 | 5.11 | 10.31 |

No. of Co-Citations vs. No. of Bib.Couplings

| | No. of Bib.Couplings | | |
|---|---|---|---|
| | Low (0-1) | Medium (2-4) | High (5+ ) |
| No. CC | 2.01 | 3.93 | 8.57 |

Table 7: Distribution of CACM Articles for Various Levels of BC, LN, CC Lengths (Key: Low=0-1, Medium=2-4, High=5+ ) Single Subvector Breakdowns

| % Low | % Med. | % High |
|---|---|---|

| | | | |
|---|---|---|---|
| For BC | 69 | 12 | 19 |
| For LN | 45 | 40 | 15 |
| For CC | 69 | 13 | 18 |

In Table 7 one sees that the distribution is about the same for $\vec{bc}$ and $\vec{cc}$ – over two thirds of the articles fall in the low group and roughly the same numbers fall in the other categories. For links, there are about the same number in the low and medium groups, but about one sixth of the total left in the high group. Thus, while all 3 subvector types are common at low levels, links frequently are found in moderate numbers as well.

Table 8 gives pairwise breakdowns. Articles that have no such entries have been removed from consideration. In the first part, one sees a fairly symmetric relationship between $\vec{bc}$ and $\vec{cc}$ length. The emphasis in number of articles, however, is toward higher levels of both variables.

Table 8: Distribution of CACM Articles for
Various Levels of BC, LN, CC Lengths
(Key: Low=0-1, Medium=2-4, High=5+ )
Pairs of Subvectors Breakdowns - Zeroes Removed

| | Co-Citations (CC) | | |
|---|---|---|---|
| | Low | Med | High |
| BC Low | 2.6 | 5.4 | 3.4 |
| BC Med | 5.6 | 8.9 | 12.5 |
| BC High | 4.6 | 15.7 | 41.2 |

| | Co-Citations (CC) | | |
|---|---|---|---|
| | Low | Med | High |
| LN Low | 7.0 | 8.4 | 1.1 |
| LN Med | 10.2 | 18.4 | 4.4 |
| LN High | 6.6 | 21.4 | 22.5 |

| | Bib.Couplings (BC) | | |
|---|---|---|---|
| | Low | Med | High |
| LN Low | 14.8 | 0 | 0 |
| LN Med | 0 | 24.3 | 32.4 |
| LN High | 1.0 | 4.0 | 23.5 |

| | BC + CC | | |
|---|---|---|---|
| | Low | Med | High |
| LN Low | 8.9 | 11.5 | 8.2 |
| LN Med | 4.4 | 13.7 | 31.2 |
| LN High | 0.1 | 0.9 | 21.2 |

When links and co-citations are compared, the emphasis is on higher levels of the $\vec{ln}$ length. When links and bibliographic couplings are charted, the emphasis is on higher levels of $\vec{ln}$ length followed by medium and higher levels of both. Finally, when $\vec{ln}$ level is compared to level of the combined measure $|\vec{bc}| + |\vec{cc}|$, the emphasis is on higher levels of that combined value. Apparently, for medium or high levels of $\vec{ln}$, the combined measure must also be medium or high. One would expect, then, that using the $\vec{ln}$ subvector, or using both the $\vec{bc}$ and $\vec{cc}$ subvectors, will each give roughly the same fairly good coverage of the set of documents.

Table 9 shows the three way contingency chart for $\vec{bc}$, $\vec{ln}$, and $\vec{cc}$ length levels.

Table 9: Distribution of CACM Articles for
Various Levels of BC, LN, CC Lengths
(Key: Low=0-1, Medium=2-4, High=5+ )
Three Way Breakdown - Zeroes Removed

| | Low LN | | |
|---|---|---|---|
| | Low CC | Med CC | High CC |
| BC Low | 1.4 | 0.2 | 0 |
| BC Med | 0 | 0.7 | 0 |
| BC High | 0 | 0 | 0 |

| | Med LN | | |
|---|---|---|---|
| | Low CC | Med CC | High CC |
| BC Low | 2.4 | 2.7 | 1.4 |
| BC Med | 2.2 | 10.3 | 6.8 |
| BC High | 2.0 | 6.4 | 7.5 |

| | High LN | | |
|---|---|---|---|
| | Low CC | Med CC | High CC |
| BC Low | 0.0 | 0.8 | 0.5 |
| BC Med | 0.8 | 2.2 | 6.8 |
| BC High | 1.4 | 8.3 | 35.3 |

Very few articles have low $\vec{ln}$ lengths, since zeroes have been removed. When a medium number of $\vec{ln}$ entries are present, the breakdown is fairly even, though there is more emphasis on having a medium level of the other two measures, or of having a high level of one or both. Finally, when a high level of $\vec{ln}$ is achieved, roughly a third of the whole collection has high levels of the other two measures, and the remainder has high levels of one and medium levels of the other.

Finally, Table 10 gives the correlation values (on the usual scale of -1.0 to + 1.0) between selected pairs of subvector lengths, after pairs with at least one zero have been eliminated.

Table 10: Correlations between CACM Bibliographic Subvector Lengths, with Zeroes Trimmed

| Case | Variable 1 | Variable 2 | Correlation |
|---|---|---|---|
| 1 | BC | LN | .49 |
| 2 | CC | LN | .43 |
| 3 | BC | CC | .33 |
| 4 | BC+ CC | LN | .60 |

The lengths of $\vec{bc}$ and $\vec{cc}$ do not correlate as well as the rest – this is in accord with earlier observations about different distributions for these two. One would expect $\vec{ln}$ length to correlate better with $\vec{bc}$ and $\vec{cc}$, and indeed that is the case. However, adding together the two, as in $|\vec{bc}| + |\vec{cc}|$, gives the best correlation of all to $\vec{ln}$. The conclusion then is that for best coverage, one should either use the $\vec{ln}$ subvector, or a combination of $\vec{bc}$ and $\vec{cc}$.

Many other figures, tables, and charts could be presented about the CACM subvectors, but the above provides a good overview of the salient characteristics, and along the way illustrates some of the capabilities of the S statistical package (see the general description in [Becker & Chambers 1981] and a discussion of its value for retrieval studies in [Fox 1983a]). Since the raw bibliographic linkage data was obtained with only a moderate amount of effort, it is reasonable to say that a similar task of examining references internal to a given publication could be undertaken in other fields, with probable similar results.

### 3.3. Queries

In the spring of 1982, two students in CS632, the graduate information retrieval course at Cornell University, assumed responsibility as part of their class project of obtaining a reasonably large set of queries and relevance judgments to use with the CACM documents. Faculty, staff, and students at Cornell, and members of other U.S. computer science departments as well, were asked to submit short natural language paragraphs stating a specific interest in computer science.

Working with the teaching assistant, Chris Buckley, and this author, they obtained a set of 52 queries along with relevance judgments. Selected natural language queries are given in Table 11.

---

Table 11: Selected CACM Natural Language Queries

(1) What articles exist which deal with TSS (Time Sharing System), an operating system for IBM computers?

(5) I'd like papers on design and implementation of editing interfaces, window managers, command interpreters, etc. The essential issues are human interface design, with views on improvements to user efficiency, effectiveness and satisfaction.

(9) Security considerations in local networks, network operating systems, and distributed systems.

(19) Parallel algorithms

---

The two students were also obligated to prepare Boolean forms of the queries. Table 12 shows, for the examples given in Table 11, the Boolean expressions (in a special prefix notation) chosen by each of the two searchers.

Table 12: Boolean Forms of Selected CACM Queries
(Actual Version Interpreted by SMART)

Searcher 1:

```
#q1  =  #or ('tss', #and ('ibm', #and ('time', 'sharing')));

#q5  =  #and ('editing', #and (#or ('human', 'user'),
                         #or ('satisfaction',
                             'efficiency')));

#q9  =  #and ('security', #or ('network', 'distributed'));

#q19 =  #and ('parallel', 'algorithms');
```

Searcher 2:

```
#q1= #or (#and('ibm', 'tss'),
       #and('ibm','time','sharing','system'));

#q5= #and( #or( 'design', 'implementation'),
        #or( 'human', 'satisfaction', 'user'));

#q9= #and( 'security',
        #or( #and( 'local', 'network'),
            #and( 'network', 'operating', 'system'),
            #and( 'distributed', 'system')));

#q19= #and( 'parallel', 'algorithms');
```

### 3.4. Relevance Judgments

With 52 queries and 3204 documents, it would be a formidable task to obtain exhaustive relevance judgments. The users who provided the questions were not likely to be willing to make such an effort and even if they were, the time and expense required would have been prohibitive. Since the users' judgments were deemed much more important than those of a third party not totally familiar with the intent of the original question, a compromise solution was sought whereby the query authors would be asked to examine all articles "likely" to be relevant.

To implement this scheme, a fairly complex processing sequence was called for. The basic outline was that of doing an initial search for 30 documents, having the users submit relevance judgments on those, doing a number of different searches to obtain 70 more documents with high similarity, and having the users make relevance judgments on that set as well. From an original set of over sixty queries, 52 were finally obtained, each with user relevance judgments on the 100 documents considered most likely to be relevant. Though exhaustive relevance information was not available, it was assumed for the sake of retrieval evaluation that the judgments provided were a good approximation to having complete information. Thus, techniques of using relative recall measures and the like were not needed.

The initial search for 30 retrieved per query was made using standard vector methods. After automatic indexing, document and query terms were all given weights which included document and collection wide frequency components (specifically, "{term frequency} * {inverse document frequency}" values). Cosine correlation was used, and the 30 documents with greatest similarity to each query were identified. As mentioned, these were given to the user to obtain relevance judgments.

Next, a number of runs were made to obtain 70 more possibly relevant documents for each of the original queries. Indeed, the 70 were selected by merging together the results of 7 different search runs, according to the rules:

(1) The best 7 for each run were guaranteed to be included in the 70 chosen.

(2) The others were chosen so that documents appearing in the greatest number out of the 7 retrieved sets were favored.

The 7 different search runs were planned based on the idea that there is not a great deal of overlap between the documents retrieved by different query forms [Katzer et al. 1982]. Hence, using a variety of strategies would tend to cover the relevant set more completely. Each Boolean form by different searchers tends to be based on a different strategy, so the two available versions were utilized. The other 5 forms were based on term relevance feedback methods employed with different combinations of subvectors [Fox 1983b]. Since no obvious guidelines were available to determine the best conditions, five different ones with relative weightings that seemed reasonable (i.e., weighted equally, but to reduce the effect of having terms in each version, terms should be weighted less than other subvectors). The seven cases are:

(1) Boolean queries by searcher 1.

(2) Boolean queries by searcher 2.

(3) Feedback queries using $\vec{tm}$ subvector.

(4) Feedback queries using $\vec{tm}$, $\vec{\mathrm{ln}}$ subvectors, weighted as 1:4.

(5) Feedback queries using $\vec{tm}$, $\vec{bc}$, $\vec{cc}$ subvectors, weighted as 1:3:3.

(6) Feedback queries using $\vec{tm}$, $\vec{au}$, $\vec{cr}$ subvectors, weighted as 1:3:3.

(7) Feedback queries using $\vec{tm}$, $\vec{au}$, $\vec{bc}$, $\vec{cc}$, $\vec{cr}$, $\vec{\mathrm{ln}}$ subvectors, weighted equally.

The 70 documents identified were placed in pseudo-random order and given to the users. Most queries were very specific and so had few relevant articles; consequently it seems likely that the procedure followed did approximate fairly well having exhaustive relevance judgments.

### 3.5. Retrieval Performance

The 52 CACM queries did not yield very good precision or recall values. The users submitting queries were all involved in computer science, many focusing on current research or implementation issues. The collection, however, only went up through 1979, so numerous contemporary topics were scantily represented. In addition, users expected very specialized answers and so their relevance decisions were typically based on a very narrow set of criteria. Thus, an average of about 15 documents, 0.5% of 3204, were relevant for each question – for ISI the percentage figure is 3.4%!

In addition, the CACM queries began as natural language statements. They were put in Boolean form by two students who were inexperienced in the practice of Boolean query construction. Based on this fact, then, it is likely that the Boolean queries would not be able to surpass the performance of the original, longer, probably more detailed natural language queries that were converted to vector forms.

## 4. ISI

### 4.1. Documents

#### 4.1.1. Background

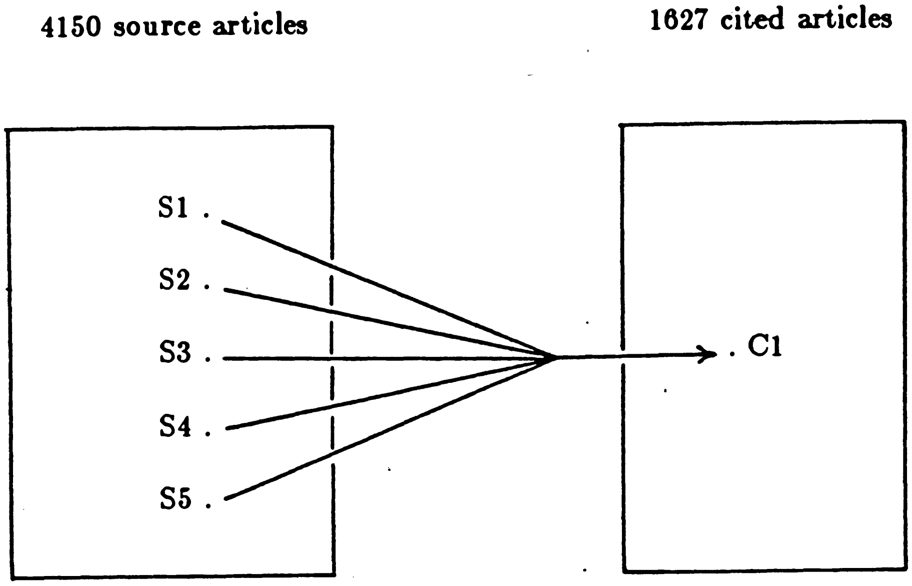
In recent years, the use of co-citation strength as a measure of document similarity has received a good deal of attention. In some studies, clustering is done based on this co-citation measure and a perspective on the structure and organization of a given area of study is presented. Alternatively, given an article that is of interest, one can find similar articles that are co-cited with it.

In order to study the interrelationship of similarities based on terms and on citations, a collection was needed containing both types of data. Dr. Henry Small of the Institute of Scientific Information © (ISI ©) had done some clustering of entries in a nine year information science file (1969-1977). There were 321 documents cited 10 or more times, and 1627 cited at least 5 times; he provided the titles and some bibliographic and citation information for the 1627 requested.

### 4.1.2. Tape Information Provided

Dr. Small's information science study [Small 1981] involved 4150 source articles which referred to the 1627 articles chosen. This situation is illustrated in Figure 16, where a document C1 is referred to by the 5 source articles designated S1 through S5.

Figure 16: ISI Source / Cited Sets




Source articles were simply information science reports in the group initially chosen. Each one was then paired with articles referred to in its bibliography, resulting in a group of source-cited pairs. File 2 on the magnetic tape contained a record for every one of the 13,075 such pairs, according to the format shown in Figure 17. Descriptive identification is given for both elements of the pair, and then the citation frequency (i.e., total number of times cited by all source articles) of the cited article is provided.

Figure 17: File 2 on ISI Citation Tape - 13,075 Records

Record Format

| Source Article | | Cited Article | |
|---|---|---|---|
| Field | No.Chars. | Field | No.Chars. |
| Journal | 11 | Journal | 20 |
| Volume | 4 | Volume | 4 |
| Page | 4 | Page | 4 |
| Year | 2 | Year | 2 |
| Author | 11 | Author | 18 |
| | | Citation Freq | 5 |

The other magnetic tape file contained information about each of the 88,294 pairs of cited articles that were co-cited. The format of each such tape record is shown in Figure 18. Besides identifying both documents in a pair, the number of times they were co-cited (co-citation frequency) was also included.

Figure 18: File 1 on ISI Citation Tape – 88,294 Records

Articles 1 and 2 each have entries according to format:

| Field Name | No. of Characters |
|---|---|
| Journal | 20 |
| Volume | 4 |
| Page | 4 |
| Year | 2 |
| Author | 18 |
| Citation Freq | 5 |

The pair has entry:
co-citation frequency – 4 characters

### 4.1.3. Collection Preparation

Based on the data provided by ISI, a number of relations were created to facilitate processing. Table 13 gives some derived summary information, showing the number of journal-volumes and journals involved. It should be noted that the source articles were taken from a relatively small number of journals (64), but referred to a much larger number (691).

Table 13: Source-Cited Article Statistics

| | Source Articles | Cited Articles |
|---|---|---|
| No. of Articles | 4150 | 1627 |
| No. of Journal-Vols. | 344 | 1009 |
| No. of Journals | 64 | 691 |

It was noted that 13 journals had between 20 and 72 cited articles, 12 had between 10 and 19, and 584 others had 1. Lists were obtained of cited journal articles, in alphabetical order by title and then ordered by year, volume, and number. Other lists were of cited monographs. Each distinct cited work was assigned a document identifier (did) for future identification.

In actuality there were only 1538 items cited, since the 1627 counted articles repeatedly when they had several authors. Other errors were discovered, such as having the same article appear several times with different spellings or other minor differences. All of the articles were looked up, using the lists prepared, at Cornell University libraries. Finally, 1460 were located and relevant parts were photocopied by Elena Siefrid.

It had been decided earlier that the title, author list, and abstract should be typed to obtain machine readable copy. When no abstract was available, the beginning of the preface, or some other key location was chosen instead. Articles in other languages were translated into English. These texts were then automatically indexed using SMART programs [Fox 1983a].

To obtain co-citation subvectors, the data shown in Figure 18 was reduced, replacing each of the two article descriptions in a co-citation pair by the appropriate document identifier. From the resulting

( did1, did2, co-citation_frequency )

relation, the $\vec{cc}$ subvectors were obtained by sorting and converting to unnormalized list format.

## 4.2. Subvectors

Table 4 gives the basic statistics about subvector lengths for the ISI documents. To understand the distributions better, however, Figures 19 through 21 are given.

Each figure shows how many of the 1460 documents have a given number of concepts in a subvector. Figure 19 is for the term histogram, 20 is for authors, and 21 is for co-citations. For terms there seems to be almost a normal distribution, though the left tail, which is bounded by an abrupt limit (i.e., the 0 point), does evince a bit of skewness. For the sake of presentation, the right tail is trimmed off, though it is clear that it is decreasing as expected. The mean value, around 40-50 terms, shows that generally vectors are of adequate length for reasonable retrieval.

Table 14 is a list of documents, along with the length of their $\vec{tm}$ subvector, where that length is smaller than 20. Only 8 have 10 or less terms. It is probable that these were indexed without an abstract, since such probably was not available. There are relatively few short $\vec{tm}$ subvectors – another collection with titles only might have an average of say 7 terms per document.

Referring back to Figure 20, about authors, one notes that almost all documents have exactly one author. There are far fewer articles with 2, 3, or 4 authors – perhaps this field does not have a high degree of collaboration reported. Regarding the high end of the curve, one would not expect more than 4 authors to be very common, so the curve is trimmed at the top end to omit values of 5 or more. In actuality, the maximum number of authors is 7.

Figure 21, giving numbers of documents for each value of length of the $\vec{cc}$ subvector, shows a distribution that has almost no left tail. Since all chosen articles have at least 5 citations, and all are con-

Figure 19: Top Trimmed Histogram of ISI $\vec{tm}$ Lengths

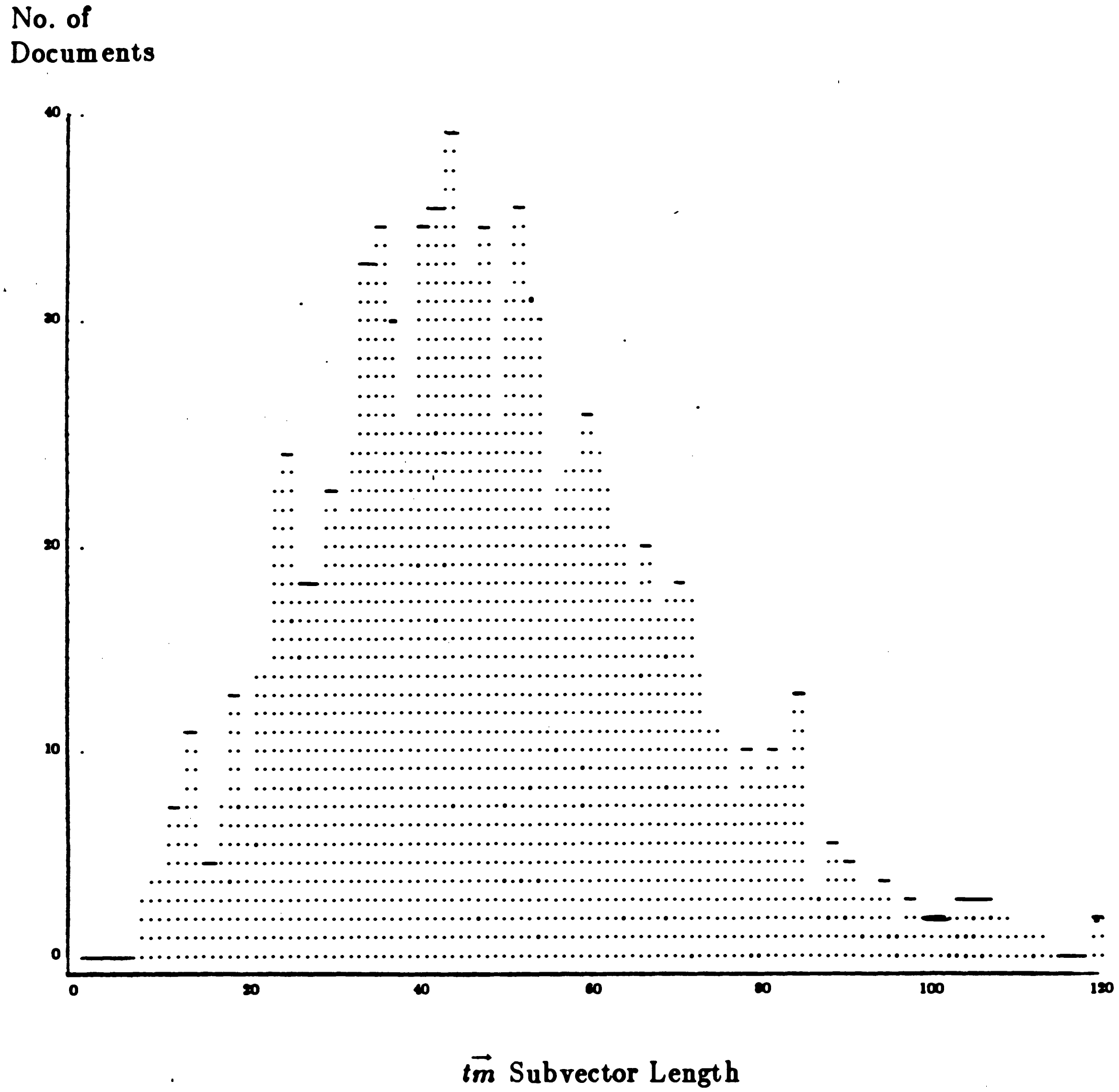


Figure 20: Top Trimmed Histogram of ISI $\vec{au}$ Lengths

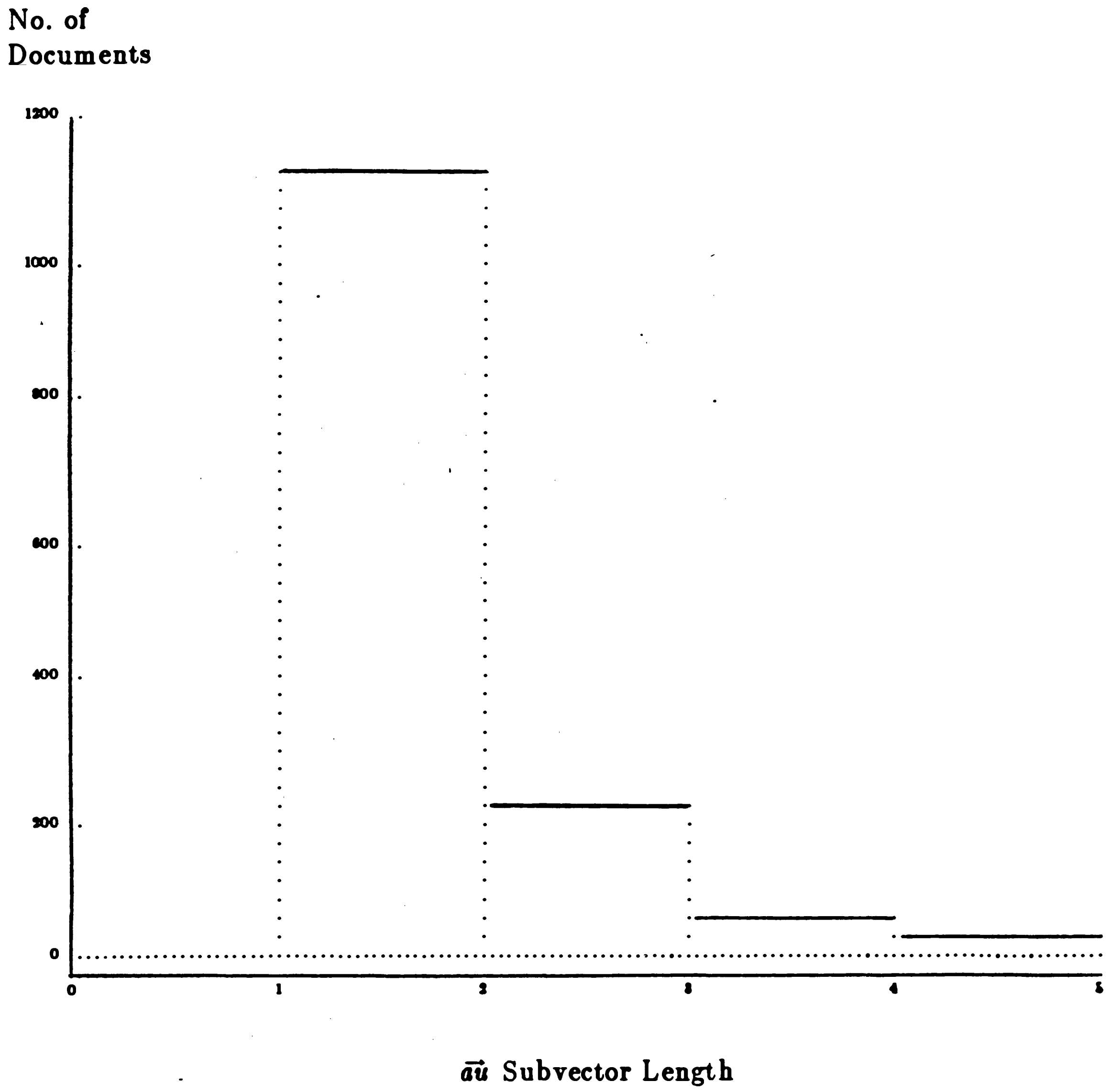


Figure 21: Top Trimmed Histogram of ISI $\vec{cc}$ Lengths

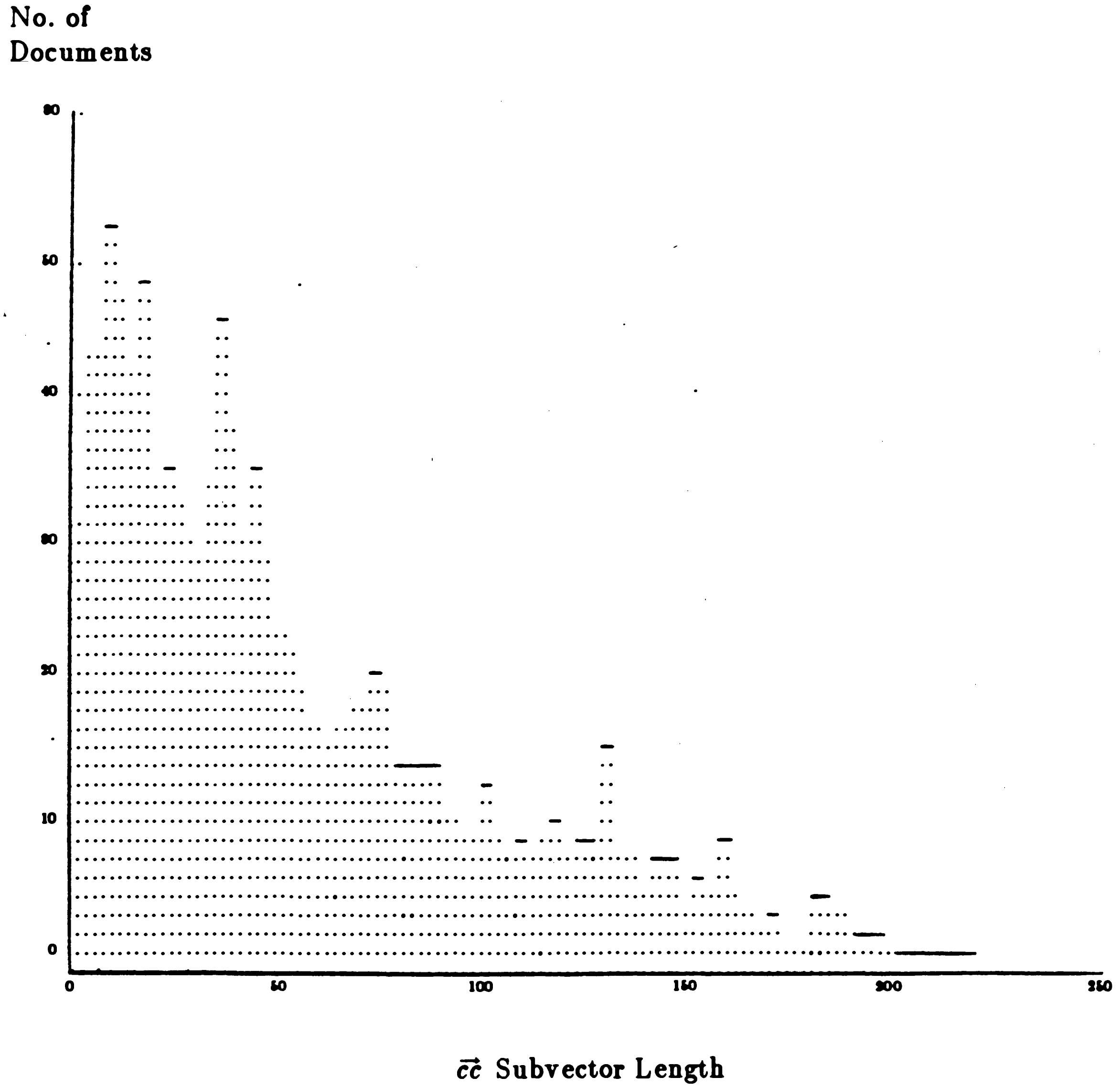


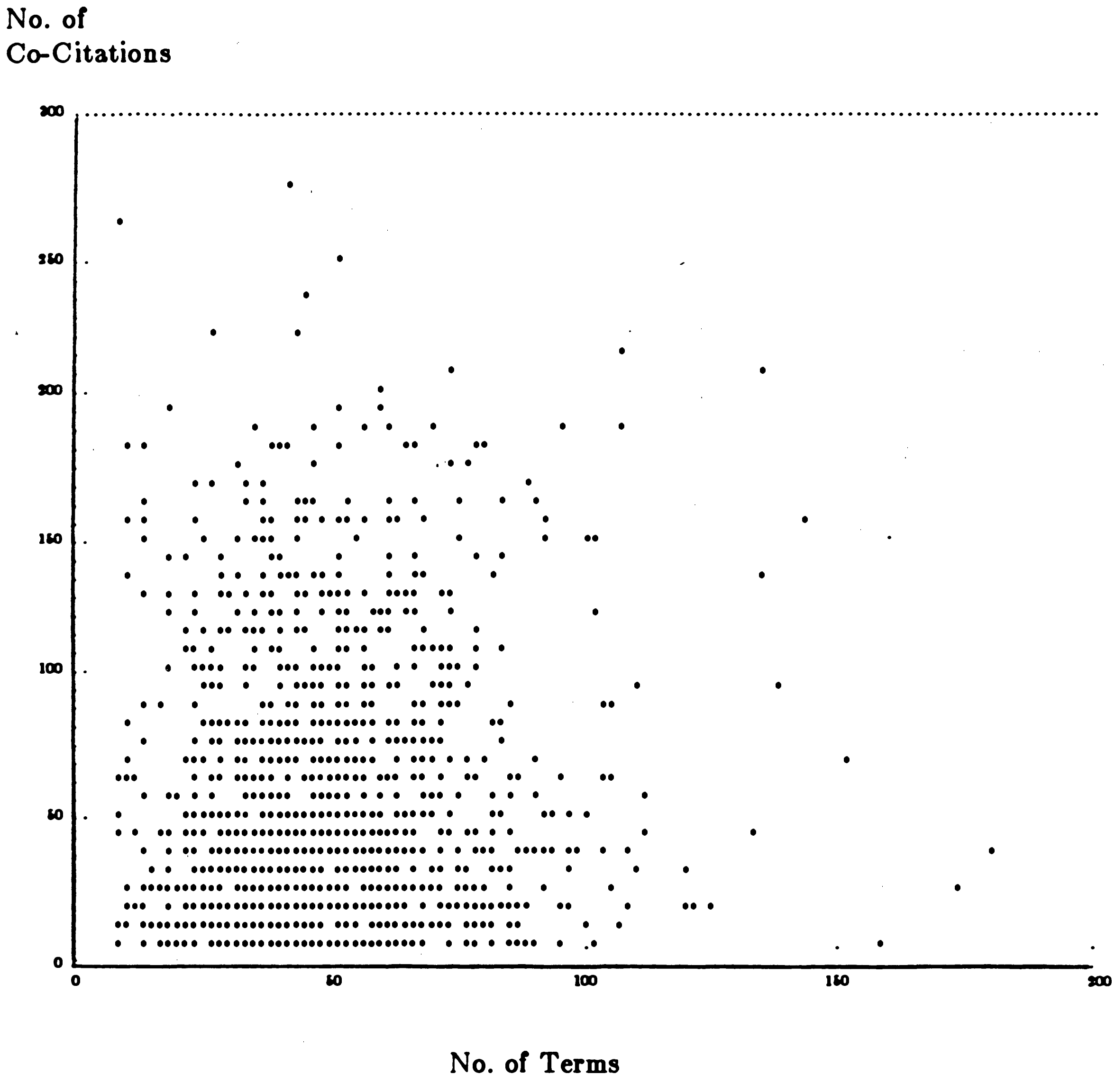
Table 14: ISI Short Vector Document List, Showing
(Did, No._of_terms) for Documents with < 20 Stems
Sorted Ascending on No.Terms and then on Did

| Did | No.Terms | Did | No. Terms |
|---|---|---|---|
| 1284 | 8 | 1312 | 14 |
| 1288 | 8 | 578 | 15 |
| 1302 | 8 | 1307 | 15 |
| 931 | 9 | 710 | 16 |
| 1100 | 9 | 1110 | 16 |
| 1279 | 9 | 1122 | 16 |
| 1296 | 9 | 1137 | 16 |
| 1086 | 10 | 1138 | 16 |
| 1306 | 10 | 753 | 17 |
| 1320 | 10 | 864 | 17 |
| 55 | 11 | 1125 | 17 |
| 1282 | 11 | 1127 | 17 |
| 1287 | 11 | 1199 | 17 |
| 1289 | 11 | 1222 | 17 |
| 1295 | 11 | 1316 | 17 |
| 1301 | 11 | 36 | 18 |
| 1303 | 11 | 118 | 18 |
| 856 | 12 | 281 | 18 |
| 968 | 12 | 316 | 18 |
| 1059 | 12 | 372 | 18 |
| 69 | 13 | 574 | 18 |
| 444 | 13 | 787 | 18 |
| 827 | 13 | 915 | 18 |
| 969 | 13 | 1004 | 18 |
| 1087 | 13 | 1083 | 18 |
| 1166 | 13 | 1220 | 18 |
| 1225 | 13 | 1278 | 18 |
| 1280 | 13 | 1285 | 18 |
| 1283 | 13 | 60 | 19 |
| 1290 | 13 | 684 | 19 |
| 1294 | 13 | 717 | 19 |
| 1085 | 14 | 859 | 19 |
| 1088 | 14 | 1022 | 19 |
| 1133 | 14 | 1413 | 19 |
| 1177 | 14 | 1446 | 19 |

centrated in a fairly narrow time period in a specific subject area, almost every document has a good many co-citations. And the distribution, even though top trimmed, has a good many entries for values of $|\vec{cc}|$ exceeding 150.

To see how terms and co-citations relate, consider the scatter plot of Figure 22. The graph is fairly uniformly covered, except that the distributions of each subvector seem to be followed. Thus, more cases appear close to the line for 40-50 terms, and there is also a concentration for 1-100 co-citations. There is no apparent direct relationship between lengths of the two subvectors. Rather, both seem to vary independently, but as hoped for in this study, the $\vec{tm}$ and $\vec{cc}$ subvectors are of suitable length to allow comparisons between their utility.

Figure 22: Scatter Plot of Terms Versus Co-Citations For ISI Documents



## 4.3. Queries

Since the ISI collection is in the field of information science, and the ADI queries are in the subarea of documentation, the ADI queries were selected to use for ISI as well. Those 35 queries, therefore, were available in natural language and Boolean forms.

However, there were portions of the ISI collection dealing with topics not considered when the ADI queries were made. Hence, some of the queries from another collection, the ISPRA[1] set, were also added. Finally, to cover more modern interests, it was decided to adapt some entries from the Abstracts section of some issues of the *ACM SIGIR Forum.*[2] All told, 41 natural language queries were identified to supplement the 35 ADI ones. See Figure 23 for examples of the new queries.

---

[1]Michael Lesk, at Harvard in 1967, gathered a group of queries to use with some 1268 ISPRA documents prepared by a research group at Euratom. The documents were selected from *American Documentation* in 1963-4, and from other journals of the time. The abstracts constituted another collection in the field of information retrieval.

[2]Specifically, the SIGIR issues were:

Fall 1980 (Vol. XV No. 2)
Spring 1981 (Vol. XV No. 4)
Fall 1981 (Vol. XVI No. 2)

Figure 23: Sample ISI Queries from ISPRA, *SIGIR Forum*

ISI Documents Selected from ISPRA Query Collection:

(44) There are presently fifty to one hundred technical journals being published. On the average, two new journals appear every day. In the many journals published, one to two million articles appear every year. What attempts have been made to cope with this amount of scientific and technical publication in terms of analysis, control, storage, and retrieval?

(52) What are the characteristics of Medlars (Medical Literature Analysis and Retrieval System) project which has been undertaken by the National Library of Medicine? How does it index current medical journals and of what relation is this indexing system to Index Medicus? What are the major components of the Medlars project and its major operating details?

ISI Documents Selected from *SIGIR Forum* Abstracts:

(60) The Combined Use of Bibliographic Coupling and Cocitation for Document Retrieval. Bichteler, J. and Eaton, E.A. III.
A linkage similarity measure which takes into account both the bibliographic coupling of documents and their cocitations (both cited and citing papers) produced improved document retrieval over a measure based only on bibliographic coupling. The test collection consisted of 1712 papers whose relevance to specific queries had been judged by users. To evaluate the effect of using cocitation data, we calculated for each query two measures of similarity between each relevant paper and every other paper retrieved. Papers were then sorted by the similarity measures, producing two ordered lists. We then compared the resulting predictions of relevance, partial relevance, and non-relevance to the user's evaluations of the same papers. Overall, the change from the bibliographic coupling measure to the linkage similarity measure, representing the introduction of cocitation data, resulted in better retrieval performance.

(62) Fuzzy Requests: An Approach to Weighted Boolean Searches. Bookstein, A.
This article concerns the problem of how to permit a patron to represent the relative importance of various index terms in a Boolean request while retaining the desirable properties of a Boolean system. The character of classical Boolean systems is reviewed and related to the notion of fuzzy sets. The fuzzy set concept then forms the basis of the concept of a fuzzy request in which weights are assigned to index terms. The properties of such a system are discussed, and it is shown that such systems retain the manipulability of traditional Boolean requests.

The *SIGIR Forum* issues chosen had a number of abstracts of selected articles of interest. The text of each of those abstracts defines a research interest, and so some were used as queries to aim at retrieving fairly recent articles that were about similar subjects.

### 4.4. Relevance Judgments

It was necessary to have a list of the relevant documents for each of the 76 ISI natural language queries. Exhaustive relevance judgments were desired, since this collection was the first one to be used for combined retrieval experiments. The work was split up between four members of the SMART project (including this author). Since all participants in this work were knowledgeable in the field of information science, the result should be fairly accurate, and it is expected that having several different people make relevance judgments on different queries should have little effect on final results.

### 4.5. Retrieval Performance

The ISI queries gave fairly low precision in most cases. The relatively poor performance can be partially explained by the following three points.

(1) The ADI queries were originally phrased to deal with issue in 1963, while the documents chosen were from between 1969 and 1977. The questions were very general, in terms of the subject matter of more up-to-date literature, and so each had an average of 50 relevant documents (3.4% of 1460). However, many of these relevant documents have no terms in common with each matching but rather general query. To be precise, 26 of 35 queries have relevant documents with zero cosine similarity, and in total, 128 relevant documents have this property – 3.7 per query, 7.3% of all the relevants. Hence, even the cosine run did badly, achieving 15% precision at 50% recall, or an average of 16% based on averaging the results at .25, .50, and .75 recall levels. It is hoped that feedback with extended vectors [Fox 1983b] would show improvements based on using co-citation information.

(2) For the 35 Boolean ADI queries there was little amplification of the terminology employed originally in the natural language statement, to adapt to new trends in jargon and language use. The searchers did not have access to postings information on terms or to retrieved sets from partial

queries, and so the Boolean queries were constructed in conditions far worse than optimal.

(3) Because of the evaluation scheme employed, and the fact that there were so many relevant documents per query, retrieval performance was quite low. The nonretrieved relevant documents were randomly positioned among ranks not filled by prior feedback runs or by currently retrieved articles. This was particularly bad for strict Boolean queries. Thus, for example, if there are 50 relevant documents, then even if 5 relevant documents are found in a retrieved set of 10, that only means that 50% precision will occur up to recall level 0.1. The other 45 relevants would be scattered among the rest of the 1460 possible slots, causing low precision values for most of the recall precision curve. However, the same problem exists for other similarity functions, to a degree.

Nevertheless, the ISI queries were a complete set, and though they have different characteristics than many other collections, do present one real life retrieval situation. Without getting better queries and relevance judgments, there are no easy solutions to the problems mentioned, so using the collection as is, for contrast with other collections, seems worthwhile.

## 5. Conclusion

Five different collections were developed to facilitate the experimental study of [Fox 1983b]. Two of them, the CACM and ISI collections, were specially prepared so as to enable research into the effects of having a variety of types of data available.

The ISI collection is like others that might be developed in different subject areas – one would certainly like to at least have the highly cited articles in any field when a collection is being constructed. With data such as that gathered by ISI © available, similar collections could be produced for various fields.

The CACM collection, on the other hand, shows that one need not have all the data provided by such an organization as ISI in order to benefit from using multiple concept types. Gathering enough references to be of use for a collection even as large as that of CACM is not an overly arduous task. Now that more and more articles are available in machine readable form and with parsing methods possible to identify references in such articles, automatic processing is also not out of the question. Hence, the results of using the CACM collection could probably be generalized, at least to certain other disciplines. The criterion is that "internal" references, among articles appearing only in one or a few key journals, would be common enough to give reasonably long subvectors.

Thus, it is hoped that the collections described in this report are fairly representative and can be used for further work. The explanations contained herein should be adequate to enable a prospective user to see if they are appropriate. If selected, the information should also allow results obtained from use of them to be suitably interpreted.

People interested in applying these collections to future studies should contact this author or the SMART project.

## REFERENCES

[Becker & Chambers 1981]
Becker, Richard A. and John M. Chambers. *S: A Language and System for Data Analysis.* Bell Laboratories, 1981.

[Bichteler & Eaton 1980]
Bichteler, Julie and Edward A. Eaton III. The Combined Use of Bibliographic Coupling and Cocitation for Document Retrieval. *Journal of the American Society for Information Science,* Vol. 31, No. 4, July 1980.

[Borko & Bernier 1978]
Borko, Harold and Charles L. Bernier. *Indexing Concepts and Methods.* Academic Press, New York, 1978.

[Date 1982]
Date, C. J. *An Introduction to Database Systems. Volume I.* Addison-Wesley Publishing Company, Reading, Massachusetts, 1982.

[Fox 1983a]
Fox, Edward A. Some Considerations for Implementing the SMART Information Retrieval System Under UNIX. Technical Report 83-560, Cornell University Department of Computer Science, Ithaca, New York, July 1983.

[Fox 1983b]
Fox, Edward A. Extending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types. Cornell University Ph.D. Thesis, University Microfilms, Ann Arbor, Michagan, August 1983.

[Garfield 1964]
Garfield, Eugene. Science Citation Index: A New Dimension in Indexing. *Science*, Vol. 144, No. 3619, May 1964, pages 649-654.

[Garfield 1970]
Garfield, Eugene. Citation Measures as an Objective Estimate of Creativity. *Current Contents*, No. 34, August 1970, pages 4-5.

[Garfield 1979]
Garfield, Eugene. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities.* John Wiley and Sons, New York, 1979.

[Katzer et al. 1982]
Katzer, Jeffrey et al. *A Study of the Overlap Among Document Representations.* Syracuse University School of Information Studies, 1982.

[Kessler 1962]
Kessler, M. M. An Experimental Study of Bibliographic Coupling Between Technical Papers. M.I.T. Report R-1, June 1962.

[Kessler 1963a]
Kessler, M. M. Bibliographic Coupling Between Scientific Papers. *American Documentation*, Vol. 14, No. 1, January 1963, pages 10-25.

[Kessler 1963b]
Kessler, M. M. Bibliographic Coupling Extended in Time: Ten Case Histories. *Information Storage and Retrieval*, Vol. 1, 1963, pages 169-187.

[Kessler 1965]
Kessler, M. M. Comparison of Results of Bibliographic Coupling and Analytic Subject Indexing. *American Documentation*, Vol. 16, No. 3, July 1965, pages 223-233.

[Michelson et al. 1971]
Michelson, D. et al. An Experiment in the Use of Bibliographic Data As a Source of Relevance Feedback in Information Retrieval. In *The SMART Retrieval System, Experiments in Automatic Document Processing,* edited by Gerard Salton, Prentice Hall, Inc., Englewood Cliffs, N.J., 1971.

[National Library of Medicine 1968]
National Library of Medicine. Medical Subject Headings 1968. *Index Medicus,* Vol. 9, No. 1, Part 2, Washington, D.C., January 1968.

[Salton 1963]
Salton, Gerard. Associative Document Retrieval Techniques using Bibliographic Information. *Journal of the ACM,* Vol. 10, No. 4, October 1963, pages 440-457.

[Salton 1971]
Salton, Gerard. Automatic Indexing Using Bibliographic Citations. *Journal of Documentation,* Vol. 27, No. 2, June 1971, pages 98-110.

[Salton 1975]
Salton, Gerard. *Dynamic Information and Library Processing.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1975.

[Salton 1980]
Salton, Gerard. The SMART System 1961-1976: Experiments in Dynamic Document Processing. In *Encyclopedia of Library and Information Science, Vol. 28,* 1980, pages 1-36.

[Small 1973]
Small, Henry G. Co-Citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science,* Vol. 24, No. 4, July-August 1973, pages 265-269.

[Small & Koenig 1977]
Small, Henry G. and Michael E. D. Koenig. Journal Clustering Using a Bibliographic Coupling Method. *Information Processing and Management*, Vol. 13, No. 5, 1977, pages 277-288.

[Small 1978]
Small, Henry G. Cited Documents as Concept Symbols. *Social Studies of Science*, Vol. 8, 1978, pages 327-340.

[Small 1980]
Small, Henry. Co-Citation Context Analysis and the Structure of Paradigms. *Journal of Documentation*, Vol. 36, No. 3, September 1980, pages 183-196.

[Small 1981]
Small, Henry. The Relationship of Information Science to the Social Sciences: A Co-Citation Analysis. *Information Processing and Management*, Vol. 17, No. 1, 1981, pages 39-50.

[Stonebraker et al. 1976]
Stonebraker, Michael et al. The Design and Implementation of INGRES. *ACM Transactions on Database Systems*, Vol. 1, No. 3, September 1976.

[Weinberg 1974]
Weinberg, Bella Hass. Bibliographic Coupling: A Review. *Information Storage and Retrieval*, Vol. 10, No. 5/6, 1974, pages 189-196.