MEASURES OF ASSOCIATION ON A BIBLIOGRAPHIC DATA BASE

MEASURES OF ASSOCIATION
ON A
BIBLIOGRAPHIC DATA BASE

By

DONALD ALLAN FOX, B.Sc., B.L.S.

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Master of Science
McMaster University
June 1977

MASTER OF SCIENCE (1977)
(Computation)

McMASTER UNIVERSITY
Hamilton,Ontario

TITLE: Measures of Association on a Bibliographic Data Base

AUTHOR: Donald Allan Fox, B.Sc. (McMaster University)
B.L.S. (University of Toronto)

SUPERVISOR: Professor R.A. Rink

NUMBER OF PAGES: x, 63

ABSTRACT

A critical survey of research done in automatic indexing and classification and statistical linguistics, of importance to the study of bibliographic data bases, is given. A theory of measures of association in vector form is presented and applied using the FAMULUS system for storage and retrieval, and a data base in the social sciences constructed using that system. Certain conclusions are drawn regarding the usefulness of the various measures of association employed, and some areas of future research are given.

ACKNOWLEDGEMENTS

The author wishes to express his gratitude to Dr. R.A. Rink for his patient help and counsel, to the National Research Council of Canada for supporting this work, to Mr. Chris Bryce and other friends for their many kindnesses, to Mrs. Susan Chang and Dr. G.S. Kenyon, Faculty of Human Kinetics and Leisure Studies, University of Waterloo for permission to use a portion of the SIRLS data base, and, last but not least, to his wife, who has endured the inconvenience of having an absentee husband on many occasions while this work was being done.

DEDICATION

to those who struggle to understand

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER I INTRODUCTION

(a) The nature of the challenge

In the last twenty years, and especially in the last ten years, libraries and other information centres have been amassing large computer-readable files of bibliographic information. Much work has been done on efficient storage and retrieval of these files from the machine standpoint, e.g. answers to such questions as how one structures the file for rapid retrieval [cf. Climenson (1966)], how data can be compacted [e.g. Heaps (1975)], or what type of hardware is best for a particular application. A great deal of research has resulted from attempts to employ computerized systems to aid in the indexing of documents (see section (b) following). Studies have also been done in the area of measuring user satisfaction with the results of retrieval from data bases*, but this has concentrated mainly on the determination of such user-dependent measures as relevance (precision) and recall ratios, e.g. Lancaster (1968) or King and Bryant (1971). While these measures are of great importance, since the final test of any service must be the degree of satisfaction produced in its users, they lack any obvious relation to factors influencing their values.

*All references to data bases in the present work will refer to bibliographic data bases.

This means that, especially in the context of a particular search, it is often difficult or impossible to suggest means for improvement based on solid evidence. Specifically, it is hard to say what part of bad precision or recall is due to incorrect choice of a data base and what part to an inappropriate use of the data base(s) employed. Considerable knowledge of the data base(s) developed during many hours of costly experience may be necessary to correctly diagnose problem situations, and to enable a searcher to take corrective measures.

The indexer (data base builder) also has problems of a related nature. He or she would like to find a cost-effective indexing procedure for a given subject area and within a given budgetary framework. This may be regarded as the problem of finding a point lying on a spectrum extending from no indexing at all (free vocabulary) on the one hand, to a highly structured subject heading and/or numerical classification controlled situation on the other, which minimizes cost and maximizes effectiveness. Another problem is whether it is better to continue indexing policies which are no longer suitable for some end users, or to break radically with the old rules and risk confusing users who are content with past practices. To solve these problems considerable user input is necessary. For this reason, and also in order to search more effectively as mentioned previously, the user must have considerable understanding of the present structure and contents of the data base, as the indexer has designed and built it.

From the above arguments, it follows that means must be found for communicating in some detail and with great exactness the properties of data bases. The remainder of this thesis is a contribution to that end, particularly in the area of data base growth and change.

(b) Research in automatic indexing and classification

Two excellent summaries of research in this area are Stevens (1965) and Fangmeyer (1974). The former work is very comprehensive and covers virtually all important studies prior to 1965. Fangmeyer's report may be regarded as an update of Stevens', although the emphasis is slightly different; the former believes that the human indexer may be aided by the machine to produce an index superior to that which either human or computer might compose separately. He lists the advantages of "fully automatic indexing techniques" as:

(1) "consistency is attained as the computer assigns index terms directly from the natural language text of the document, applying the same algorithm for each document"

(2) "simplicity of re-indexing, which is important, because a scientific library is a living thing and classification schemes must always change according to either the aims of the library, or the developments in science"

(3) "accuracy, which is guaranteed by the ability of the computer to select, transfer and re-arrange data reliably without making typographical errors"

(4) "economy, achieved by large-scale processing and computing speed"

(5) "facility for editing".

He points out, however, that fully automatic indexing, while it has achieved results comparable with human indexing in some cases, falls short in other cases because statistical occurence measures are used "to overcome lack of knowledge in linguistics"; that is, the computer is not capable of the type of linguistic analysis a human indexer can do. More specifically, he gives the following advantages of machine-aided or semi-automatic indexing. "Indexers:

- are able to make discriminations as to the relative importance of technical concepts as they appear in an abstract or document,
- have access to the entire document and
- can go beyond the document itself to reference books, to consultation with experts, or other sources as deemed appropriate, to aid in properly indexing the document at hand,
- can apply inductive reasoning to formulate and index concepts which are implied by the document but not expressly stated (assignment indexing),
- become familiar with the requirements of the users of the system by participating in search request analysis, search strategy formulation and search screening". What he has not mentioned, however, is the quantitative measures which the indexer may employ to decide what is the "relative importance of technical concepts" or what the value is of applying "assignment indexing" in terms of the "requirements of the users of the system". Nevertheless, despite his tendency to ignore these important questions, Fangmeyer provides an excellent summary to various man-machine systems and to the techniques they apply to index documents.

Stevens (1965), on the other hand, comes very honestly to grips with problems of evaluation of indexing, "whether applied by man, machine, or man-machine combinations". She correctly notes that these core problems are not so much of technique as of evaluation of the quality of indexing. The first core problem is our lack of a precise knowledge of language as a means of communication. The second is the difficulty in specifying the vagaries of natural language in such a way that a machine can cope with them. The third problem is that no-one has accurately formulated the amount of information lost by representing documents in a condensed fashion. The fourth problem is that no means exists of estimating precisely "whether the benefit to users is worth the cost". Scientists continue to obtain the majority of their information from other than indexing and abstracting publications [cf. King Research (1976)]. The fifth and final problem she isolates is the difficulty of establishing an objective criterion using relevancy. Not only do judgements of the degree of relevance of a document to a query differ from person to person, but they also differ from time to time. That is, documents originally judged irrelevant may later become relevant, and vice versa. She quotes test results which show that these effects appear to be far from negligible.

Both Stevens and Fangmeyer tend to confuse the terms indexing and classification, an all too common error in this field. Unfortunately it is necessary to decide from context which is actually meant at any given point.

One other recent work deserves to be mentioned in the present context; that of Salton (1975). Despite the title, "A Theory of Indexing", this paper is much more empirical than theoretical, and quotes the results of many experiments performed using the SMART system, developed by Salton and his students. It does present a good cross section of work in this area, however, and concludes with the following three questions which the author feels "are the most important for a practical application of the theory":

(1) "To what extent can one justify the replacement of the complicated discrimination value computations by the simple document frequency model?"

(2) "Can the computation of term values obtained from a static model of a given document collection be maintained in a dynamic environment where old documents are removed, and new ones added? If not, how often must one recompute term values?"

(3) "Can the term values obtained from a collection in a given subject area be used for collections in different subject areas?" With suitable modifications to enable them to conform to different goals, these concerns can be applied to the present work. Salton's work will be returned to later.

(c) <u>Research in computational and statistical linguistics</u>

Since the work of Yule (1944) the seminal problem of statistical linguistics has been the determination of disputed authorship. While only two studies have been judged successful, Ellegård (1962) and Mosteller and Wallace (1964), much research has been done to discover the statistical indicators of style, and to distinguish these indicators from the deeper statistical properties of language itself. An excellent historical summary of this work has been given by Bailey (1969). Bailey gives five questions to whose answers he feels statistical methods can contribute: "Who wrote the work? In what directions did this writer develop? What are the constraints imposed on the writer by his language? How does the selection of the mode of presentation influence the shape of this work? How well do critics agree in their judgement of a piece of writing?" Underlying this belief (i.e. that statistical methods can contribute to answering these questions) is the knowledge that statistical tools can be used with confidence to measure and compare language usage.

Probably the most significant compilation and codification of useful statistical techniques and the results of applying them to language has been in the works of Herdan (1964, 1966). In both of these books, and in others preceding them, the author has given impressive amounts of experimental evidence drawn from many sources to support his arguments. He is particularly compelling in his attacks on certain widely-held fallacies, perhaps the best example being the deflation of "Zipf's law". One short quote will suffice. "Mathematicians believe in it because they think that linguists have established it to be a linguistic law, and linguists believe in it because they, on their part, think that mathematicians have established it to be a mathematical law. As can be shown in a few lines, it is not a law at all...." [Herdan (1966)].

Not content with merely attacking the work of others, however, this gifted and experienced statistician has proposed a more suitable model, the Waring distribution, which unlike the "Zipf law", does fit the data to a remarkable degree as is also illustrated in a more recent article by Muller (1969). In addition there are a number of other functions that contend to represent the distribution of words in text, e.g. the decapitated negative binomial distribution. [Yule (1944), Simon (1955)] or the beta function based law of "cumulative advantage" [Price (1976)]. So far, however, none have enjoyed the success achieved by Herdan and his followers.

Despite this success, however, Dillon (1972) writes of the "failure of linguists generally to produce a satisfactory model of language behavior". It is his contention that "the potential of quantification as a means for representing language processes has not been fully realized, though interest in exercising this approach has grown along with computer technology." He goes on to give a fairly comprehensive survey of research on text handling and analysis by computers, though he purposely ignores most of the work done on data base systems. His bibliography is useful nevertheless, particularly since many of the items are annotated from the viewpoint of a statistical linguist. Also of interest is the section on future research, which tries to lay down ground rules for fitting quantitative linguistic research into a coherent picture. Among other questions he asks is whether it is "possible to integrate the variety of statistical models being pursued into some theoretical frame". It is challenging to learn that foundational theory is as lacking in computational linguistics as it is in the study of data bases.

(d) Impact of research in automatic indexing and classification and in computational and statistical linguistics on the study of data bases

It is perhaps not very surprising that research in automatic indexing and classification has made more impact on the study of data bases than has research in computational and statistical linguistics. If, as claimed by Bailey (1969), the combination of expertise in statistics and linguistics is rare, the additional requirement of an interest in data bases must make the resulting set of individuals so oriented even smaller. There is, on the other hand, a very practical connection between automatic indexing and classification, and data bases, the latter providing the material upon which the former must operate. Hence just as a carpenter should study woods, and a medical doctor the human body, so it is reasonable that an indexer or classifier will study data bases. Despite the seeming obviousness of this conclusion, however, the set of articles and books where indexers/ classifiers have reported a serious study of data bases is considerably smaller than the number of all indexers and classifiers. This situation will doubtless change as the practice of using computerised data bases becomes more common and widespread.

One sign of the increasing interest in this area is the recent article by Sparck Jones and Van Rijsbergen (1976) which lists six requirements that a data base should have to make it useful for study. The authors also list fifteen questions which such an "ideal collection" might be used to answer. Many of these questions are irrelevant to the present study, but we will return to the others in Chapter 3.

We return now to the work of Salton (1975), to deal with it in more detail, especially as regards what it suggests for the study at hand. Proceeding from his previous work [Salton (1968), Salton (1971)], Salton has concentrated largely on the application of clustering to automatic indexing. Since his test collections are not large (200-500 documents) and since he is generally more interested in analyzing the results of searches than in examining the basic properties of data bases, his experimental work is of little interest in the present application. The small amount of theory he presents is, however, of some importance in this context. He exhibits the basic difference between the binary and non-binary (weighted) vector approach, and discusses briefly many different measures of association most of which are either equivalent to or similar to those discussed in Chapter 2. The work of Salton and his associates therefore may be viewed as seminal to that presented here, despite our differing aims.

With regard to computational linguistics, only the work of Yule (1944) and Herdan (1964, 1966) has had any considerable impact on what is presented here, this impact being primarily in the area of interpretation of results. The reason is simple: most computational and statistical linguistics relates to long texts by individual authors, whereas a data base is really a large number of short texts by many different authors. In fact, it may be considered surprising there are any cross-applications at all!

(e) Thesis outline

Chapter 2 will be a brief presentation of sufficient theory to support the remainder of the present work. In Chapter 3 the main experimental results will be given and discussed. Following this, Chapter 4 will draw conclusions and give connections between theory and experiments. Chapter 5 will suggest future research and References will give the bibliographic information for documents referred to in the preceding chapters. Finally, Appendices A and B will briefly discuss programs used in the analysis of the information presented herein, while Appendix C will give additional statistical background.

CHAPTER II THEORY

(a) Aggregates: Zero - one vectors

In order to exhibit tools necessary for the data base study which follows, we must develop some theoretical structures within which these tools may be fitted.

Definition: An aggregate is a piece of text (assemblage of words in some natural language) consisting of at least one term (significant word).

E.g. the above definition is an aggregate since it contains a significant word (text) as well as some which are not significant (of, at), i.e. do not contribute much to understanding.

In the remainder of section (a), capital letters will denote vectors.

Consider a binary term - aggregate vector

(a.1) $X = (x_1, x_2, \dots, x_n)$

where $x_i = 1$ if the ith term is present in the aggregate, and

$= 0$ otherwise.

e.g. X may represent a set of index terms ordered alphabetically. In this case, a "universal" vector.

(a.2) $U = (1, 1, \dots, 1)$

corresponds to the term authority list.

A period of growth in a bibliographic data base may be represented using the above formalism as follows:

$$\text{(a.3)} \qquad X = (x_1, x_2, \ldots, x_n) \rightarrow \bar{X}(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n, \bar{x}_{n+1}, \ldots, \bar{x}_{n+m})$$

Here, two types of changes have occurred. First, some of the terms have disappeared while others have appeared, i.e. some $\bar{x}$'s $= 0$ where the corresponding x's = 1, and some $\bar{x}$'s = 1 where the corresponding x's = 0. Second, m new terms have been introduced which were not previously present. These are represented by $\bar{x}_{n+1}, \bar{x}_{n+2}, \ldots, \bar{x}_{n+m}$.

An important special case is that in which the initial term-aggregate vector X is a universal vector as given in (a.2). In other words:

$$\text{(a.4)} \qquad x_i = 1 \text{ for all } i = 1, 2, \ldots, n$$

This represents the initial state of a total data base, or of the totality of index terms for such a data base. In this connection, it is natural to assume:

$$\text{(a.5)} \qquad \bar{x}_i = 1 \text{ for all } i = n+1, n+2, \ldots, n+m$$

These m conditions represent the totality of terms in the expanded data base or index term list.

If we now assume that k terms have disappeared from the first n terms, we have:

$$\text{(a.6)} \qquad \sum_{i=1}^{n} \bar{x}_i = \sum_{i=1}^{n} x_i - k = n - k \quad \text{by (a.)}$$

If we denote the total number of unique terms in any aggregate by t, then in this case the change in the total number of terms is:

$$\text{(a.7)} \qquad \Delta = \sum_{i=1}^{n+m} \bar{x}_i - \sum_{i=1}^{n} x_i = n-k+m-n = m-k$$

(b) Aggregates: Non-negative integer vectors

A simple and natural extension of the above ideas is to represent not only occurrence or non-occurrence of a term, but the number of times a term occurs as well. In this case, the term-aggregate vector takes the form:

(b.1) $$X = (x_1, x_2, \ldots, x_n)$$

where x_i = r if term i occurs r times in the aggregate, and

= 0 otherwise.

Unfortunately the universal vector concept cannot be used for representing the initial state in this case, and clearly equations (a.4) to (a.7) do not hold. The additional information available from (b.1) can be used however, as will be seen in what follows.

(c) Measures of association

A number of functions for measuring similarity in attributes of two sets have been suggested and/or used in various contexts, (see discussion in Appendix C) . The four measures most commonly used in document retrieval systems are the following, where X and Y are term-aggregate vectors.

(c.1) $$D = 2\frac{(X,Y)}{(X,X) + (Y,Y)}$$ (Dice's coefficient)

(c.2) $$J = \frac{(X,Y)}{(X,X) + (Y,Y) - (X,Y)}$$ (Jaccard's coefficient)

(c.3) $$C \quad \frac{(X,Y)}{\sqrt{(X,X)\cdot(Y,Y)}}$$ (Cosine coefficient)

(c.4) $$O = \frac{(X,Y)}{\min[(X,X),(Y,Y)]}$$ (Overlap coefficient)

The similarity of these various measures is obvious; all have as their numerators the inner or dot products of the two vectors, and all are normalized so they range between zero and one. (In the case of the overlap coefficient, this is only true for binary vectors.) The zero value is only achieved for othogonal vectors, while unity is the condition of maximal association, attained when X = Y. Set theoretic definitions of these measures are given in Van Rijsbergen (1975). Jaccard's coefficient is also called the Tanimoto measure and was (and is) used extensively by the Cambridge Language Research Unit, e.g. Sparck Jones and Needham (1968). Note that the denominators which normalize these functions all include vector lengths or squares of vector lengths. One may regard these as attempts to render such measures independent of the number of terms, but as will be seen, these attempts are not entirely successful.

(d) Yule's characteristic

The desire to obtain a measure of association which is independent of the number of terms has a parallel in mathematical linguistics. In the widely cited volume by Yule (1944) the author carries through an analogy between the distribution of accidents in a population at risk, and the distribution of words in text. The degree of similarity of the two situations may be questioned, but it is certainly true, as Yule points out, that not all men have equal liabilities to accidents and similarly not all words have equal probabilities of appearing in a particular text. In any event, based on the assumed similarity, and the use of the binomial distribution and, its limiting case, the Poisson distribution, both of which had previously been applied by Yule and others to accident distributions with considerable success, the author deduces that the following value, K, is independent of text length considered.

$$\text{(d.1)} \qquad K = k\,\frac{S_2 - S_1}{S_1^{\,2}}$$

where k is a positive constant chosen to give K a convenient magnitude (Yule uses k = 10,000), S_1 is the total number of words (including repeats) in the sample of text being analyzed, and S_2 is the sum of the products of the squares of the number of occurrences of words and the number of different words which occur that many times.

To apply this function to the term-aggregate vectors discussed above, we define

$$S_1 = \sum_{i=1}^{n} x_i \quad \text{and} \quad S_2 = \sum_{i=1}^{n} x_i^{\,2}$$

where the x_i's are defined in (b.1). In vector notation:

$$S_1 = (X,U) \quad \text{and} \quad S_2 = (X,X)$$

where X is defined by (b.1) and U by (a.1). Hence (d.1) may be written:

$$\text{(d.1)}' \qquad K(X) = k\,\frac{(X,X) - (X,U)}{(X,U)^2}$$

and may be seen in this form to differ considerably from (c.1) through (c.4) when U is substituted for Y in the latter. K(X) ranges between zero and k. It is zero whenever X is a binary vector, while for large values of elements in X it comes arbitrarily close to $k\frac{(X,X)}{(X,U)^2}$. It is possible to define a measure of association involving the characteristic as follows:

$$(d.2) \qquad A = 1 - \frac{|K(X) - K(Y)|}{k'}$$

where k' is a positive constant chosen to give A a convenient magnitude (i.e. not too close to unity over the data being compared), but such that $0 \leq A \leq 1$. This is true if $k' \geq \max|K(X) - K(Y)|$ over the data being considered.

Yule's characteristic has been criticized by linguists because its derivation assumes that the vocabulary has a binomial or Poisson distribution. This objection has been removed by Herdan (1966) who has shown that the "coefficient of variation of the sampling distribution of means", which for large vocabularies approximates to none other than the limiting case of K(X) defined above, is independent of text length by definition, when text length is sufficiently large. This function, which may be called Herdan's characteristic, is given by:

$$(d.3) \qquad K^*(X) = k\frac{(X,X)}{(X,U)^2} - \frac{k}{(X_B,U)}$$

where X is as given above and X_B is the corresponding binary vector as described in (a). As mentioned, the second fraction on the right-hand side of (d.3) becomes negligibly small with large vocabularies. For further discussion see Appendix C.

(e) Herdan's law of relative growth

On the basis of some fairly simple assumptions regarding the factors affecting growth of vocabulary, analogous to the growth of biological systems, Herdan (1960) derives the following relation between size of vocabulary, V, and text length, N.

$$(e.1) \qquad V = N^{\gamma}$$

where γ is a constant which gives the ratio of the growth rates of vocabulary and text length. Since both vocabulary and text length increase, but the latter more rapidly than the former, $0 < \gamma < 1$. This relationship may be rewritten in vector notation as:

$$(e.1)' \qquad (X_B, U) = (X, U)^{\gamma}$$

γ in (e.1) and $K^*(X)$ in (d.3) are regarded as "characteristics of style" by mathematical linguists. That is to say they represent properties of particular vocabularies. In the present work we wish to regard them as representative of vocabulary properties of a particular data base.

It may be noted that equation (e.1)′ relates information given by a non-negative integer term-aggregate vector with similar information from a binary term-aggregate vector. It states that the sum of the elements of the latter is equal to the sum of the elements of the former raised to a power, whose value lies between zero and one. As we shall see in the remainder of this thesis, a recurring theme of some practical importance is deciding when sufficient information can be obtained from the binary vector and when, on the other hand, one must determine the integer vector.

One deduction of interest that can be made from the preceding definitions and equations is that the cosine measure (equation (c.3)) is not well suited for use with large data bases. For simplicity we take Y = U in (c.3); this corresponds to comparing a particular term-aggregate vector to an initial state as given by (a.4). In that case:

$$(e.2) \qquad C = \frac{(X,U)}{\sqrt{(X,X)\cdot(U,U)}}$$

but for large vocabularies using (d.3),

$(X,X) = \alpha\cdot(X,U)$ where α is constant.

substituting this value in (e.2):

$$C = \frac{(X,U)}{\sqrt{\alpha.(X,U)\cdot(U,U)}} = \frac{1}{\sqrt{\alpha.(U,U)}}$$

i.e. for a given U, C is constant and hence of no value in discriminating between term-aggregates. It is easy to show that the same is not true for the other three measures of association given in (c), nor for that defined by (d.2)

Since C (like all other measures of association) is symmetric in X and Y, it follows that, for large data bases, wherever one of the term-aggregate vectors being compared is close to a scalar multiple of a universal vector, C will not be useful in discriminating between this and another term-aggregate vector, regardless of the form of the other vector. The increasingly lengthy tails of distributions such as Waring's (see Chapter 1, part c) make such vectors extremely likely to occur for large data bases. This phenomenon may very well have contributed to the poor clustering observed by Gottlieb (1974) when using the cosine coefficient, since he remarks that the Tanimoto (i.e. Jaccard) coefficient performed better.

CHAPTER III EXPERIMENTAL RESULTS AND DISCUSSION

(a) The purpose of the experiments

The following experimental work was begun with a view to studying the alterations produced in a data base over a period of years by changes in terminology. The phrase "changes in terminology" is used here in a rather restricted sense to mean the tendency of an aggregate to drop some terms and to acquire others. We do not consider the meaning of these terms per se, only their physical form.

Emerging from the prime objective given above, a secondary goal arose. It became important to determine the properties of measures of association which were used to compare time segments of the data base being studied. So far as the author is aware, neither of these specific objectives has been pursued in this way in any previous work, although linguists have shown some interest in the way language changes [Fodor (1965), Herdan (1966)].

(b) The experimental method

The data base used was a large subset of a moderate sized data base, namely the SIRLS file of the Faculty of Human Kinetics and Leisure Studies, University of Waterloo. Specifically, 892 citations with abstracts were studied, ranging in publication date from 1960 through 1975. Many types of publication were represented, e.g. government documents, monographs, periodical articles. The decision to use this file was based on its availability, its size and range of dates, its format, but most of all its subject matter: social science. It was felt that this choice would provide greater changes in terminology than might be observed in a data base taken from one of the physical sciences.

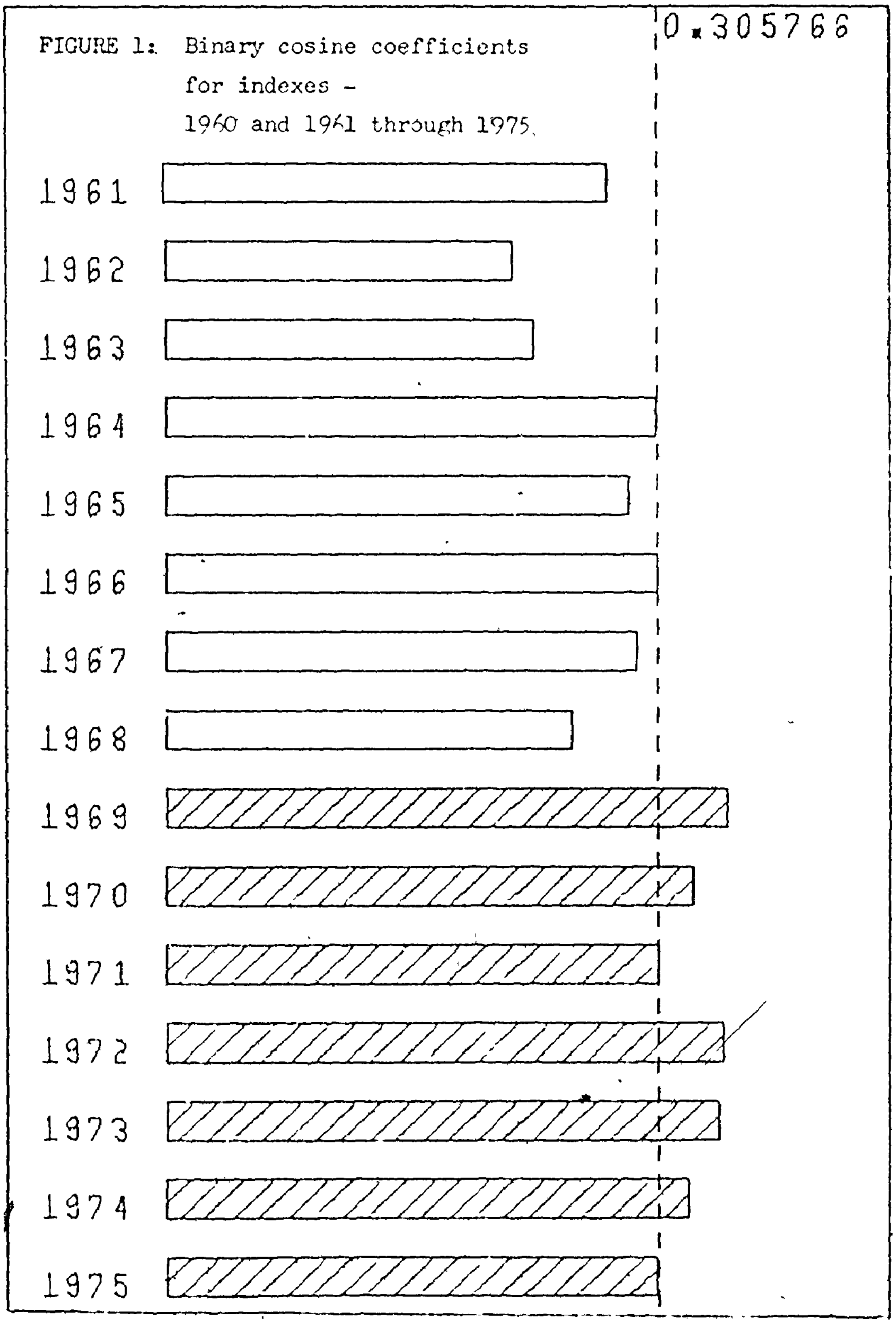
The original file was partitioned into separate years, and descriptor term lists as well as free vocabulary lists were generated in machine readable form, using the FAMULUS programs developed by the U.S. Forest Service (1969) (see also Appendix A). The descriptor terms were controlled vocabulary terms assigned from a total list of 496. The free vocabulary terms were taken from the abstracts and titles, after non-significant terms such as or, a, but, etc. had been removed.

The resulting lists of terms were then compared and the measures of association described in Chapter 2 were calculated (see Appendix B). In the case of the indexes only, term frequencies were also obtained, so that non-binary measures could also be computed. This was not done for the free vocabulary terms since, as will be seen, these tended to be much less accurately representative of the subject changes that were occurring.

(c) The result of the experiments

In Tables 1 and 2, results for indexes and vocabs (abbreviation for "free vocabularies") are given. The years from 1961 through 1975 were compared with the base year, 1960. The resulting cosine measures of association are also illustrated in Figures 1 and 2.

FIGURE 1: Binary cosine coefficients for indexes – 1960 and 1961 through 1975.



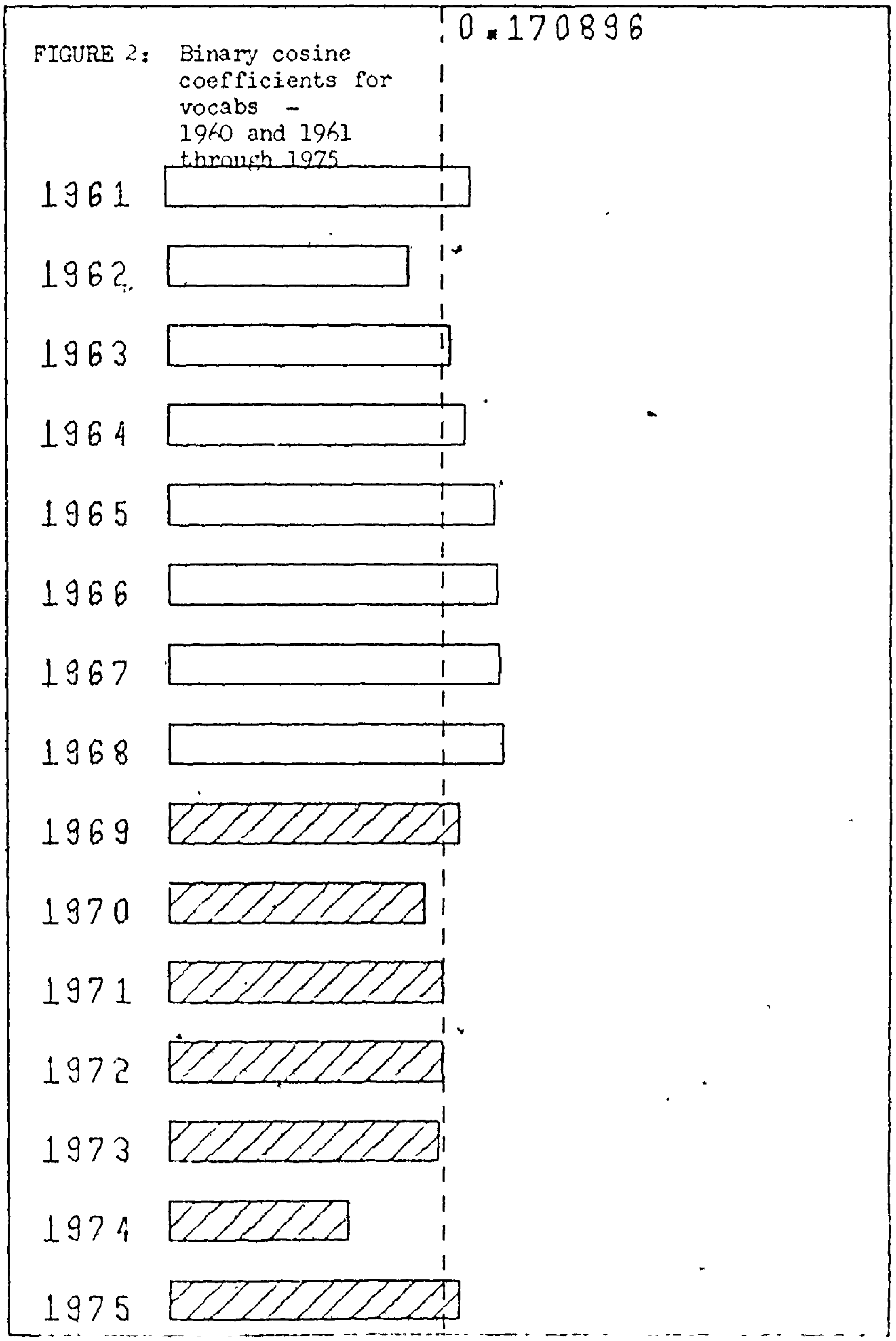


FIGURE 2: Binary cosine coefficients for vocabs – 1960 and 1961 through 1975

TABLE 1 - Measures of Association for Vocabs

YEAR	NUMBER OF TERMS	BINARY COSINE
1960	214	———
1961	369	0.188606
1962	211	0.150592
1963	303	0.176719
1964	667	0.185280
1965	653	0.203306
1966	1198	0.205399
1967	1415	0.207167
1968	1297	0.208793
1969	1977	0.181414
1970	2510	0.159740
1971	3136	0.170896
1972	3125	0.171197
1973	3176	0.168604
1974	2287	0.112924
1975	137	0.181106

As a check against too hasty interpretation, other measures of association were calculated, and their values are also given in Table 2. These values are also illustrated in the bar graph charts labelled Figures 3 through 9. Table 3 gives a breakdown of the number of citations with abstracts per year.

Finally in Table 4 and the accompanying graph labelled Figure 10 is given a picture of how well the data conform to the law of relative growth of Herdan.

(d) Discussion of results

The results presented here are incomplete in some respects. Since only one data base of a moderate size was examined, it is possible that there is some subject bias, and that a larger sample might have behaved differently. Limitations of time, availability and funds have precluded a more extensive study, however. It is clear, therefore, that further research is necessary to determine the reliability and real extent of these results. For this reason discussion based on them must be considered somewhat speculative.

The initial impression created by Tables 1 and 2 and by Figures 1 through 9, is one of considerable complexity. It seems likely that a number of partially opposing forces are at work. For example, despite the large time span of the data base, the similarity of later years with the initial year does not decrease markedly with time in most cases. (Only the non-binary Dice and Jaccard coefficients show such an effect). In fact in Figure 1 there is actually an increase noted, in that the similarity between 1960 and 1975 is significantly (10.2%) greater than that between

1960 and 1961. The same is true of Figures 8 and 9.

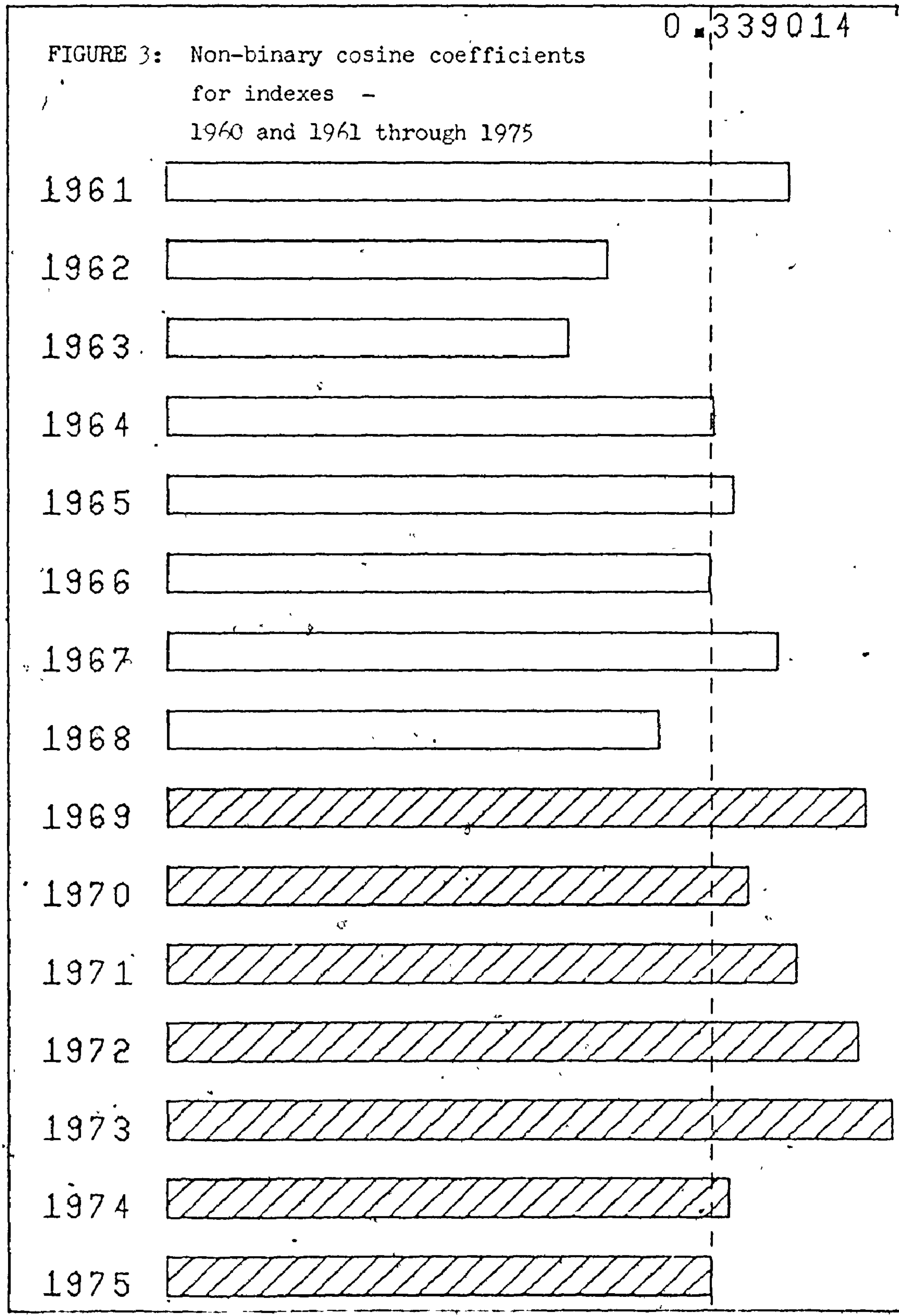
The vocabs, on the other hand, show a different pattern. As shown in Figure 2, their cosine coefficients are considerably smaller than those of corresponding indexes, and are generally more uniform in size. After 1968 the coefficients take a sudden drop in size (hatched bars) whereas the cosine coefficients for the indexes increase at the same point (Figures 1 and [illegible]). Furthermore the change from 1968 to 1969 is only about 16% of the value at the lower side for the vocabs, while in the case of the indexes this change is about 39% of the lesser value. Moreover with two exceptions, 1964 and 1966, all binary cosine coefficients for indexes are less than any after 1968. This is illustrated by the dotted line in Figure 1, which is drawn at the level of the smallest measure of association after 1968 (i.e. 1970). It is also evident that the two deviant years are not above this line by a significant amount. In the case of the vocabs (Figure 2) the dotted line is drawn at the height of the median coefficient after 1968. It illustrates the fact that while all but one (1962) of the coefficients prior to 1969 come above it, none are very far from it. The median of the coefficients before 1969, [illegible], is, however, 19% above the line and this is statistically significant, as can be shown by a chi-square test.

The above described increase of the coefficients for the indexes with time is almost certainly due to the small size of the set of all possible index terms ([illegible]) as compared with the very large size of the set of all possible vocab terms, which is really the union of the sets of working vocabularies of all authors (probably in excess of 10,000 terms). For this reason increasing the size of the number of index terms has a marked effect on the number of repeated terms, given by

TABLE 2 - Measures of Association for Indexes

	NUMBER	BINARY				NON-BINARY			
YEAR	OF TERMS	COSINE	DICE	JACCARD	OVERLAP	COSINE	DICE	JACCARD	A
1960	35	———	———	———	———	———	———	———	———
1961	54	0.276026	0.269663	0.155844	0.342857	0.387985	0.364706	0.223022	0.777635
1962	30	0.216025	0.215385	0.120690	0.233333	0.275334	0.275229	0.159574	0.200380
1963	54	0.230022	0.224719	0.126582	0.285714	0.251397	0.240506	0.136691	0.875594
1964	88	0.306319	0.276423	0.160377	0.485714	0.342193	0.266667	0.153846	0.892576
1965	77	0.288943	0.267857	0.154639	0.428571	0.353798	0.290076	0.169643	0.900783
1966	121	0.307329	0.256410	0.147059	0.571429	0.339056	0.192429	0.106457	0.516827
1967	132	0.294245	0.239521	0.136054	0.571429	0.381712	0.166667	0.090909	0.381712
1968	128	0.253986	0.208589	0.116438	0.485714	0.307339	0.140733	0.075693	0.128644
1969	195	0.351032	0.252174	0.144279	0.828571	0.435668	0.132211	0.070785	0.898937
1970	221	0.329737	0.226562	0.127753	0.828571	0.363208	0.067191	0.034763	0.190126
1971	236	0.308083	0.206642	0.115226	0.800000	0.392531	0.069042	0.035755	0.354659
1972	257	0.347947	0.226027	0.127413	0.942857	0.430704	0.080137	0.041741	0.972370
1973	231	0.314764	0.233083	0.131915	0.885714	0.451990	0.093527	0.049057	0.866504
1974	210	0.326599	0.228571	0.129032	0.800000	0.351103	0.102326	0.053922	0.929965
1975	191	0.305766	0.221239	0.124378	0.714286	0.329014	0.111721	0.059165	0.813741

FIGURE 3: Non-binary cosine coefficients for indexes – 1960 and 1961 through 1975



(X,Y), and hence on the binary cosine coefficient. As data base additions become large, however, and (X,Y) approaches its upper bound of (X,U), where X is given by (a.1) and U by (a.2), it is clear that this phenomenon will be less noticeably, as a kind of saturation point is reached. From that point on the cosine coefficient for indexes will grow steadily smaller. In that case, the number of times particular terms occur in the data base will become important as an indicator of expansion and growth in particular subject areas.

With the vocab terms on the other hand, the increase in repeated terms was compensated for by the overall increase in terms. In other words, because the sample studied was small as compared with the size of the set of all possible vocab terms, the effect of changes in terms is less pronounced, and consequently less can be said with assurance about the behaviour of the data base. For this reason the vocabs were not pursued further, while the indexes were.

The cosine coefficients for the indexes in the non-binary case (Figure 3) confirm the results illustrated by Figure 1 for the binary case, given the effects of random fluctuation. The same cannot be said for either of the two other coefficients which are defined in both the binary and non-binary case, i.e. Dice's and Jaccard's coefficients (Figures 4 through 7).

An interesting result may be noted with these latter two pairs of coefficients. A comparison of the bar graph charts shows that the binary Dice coefficients are very nearly proportional to the binary Jaccard coefficients, and similarly for the non-binary cases, the constant of proportionality being a little less than 2. This phenomenon

FIGURE 4: Binary Dice coefficients for indexes – 1960 and 1961 through 1975

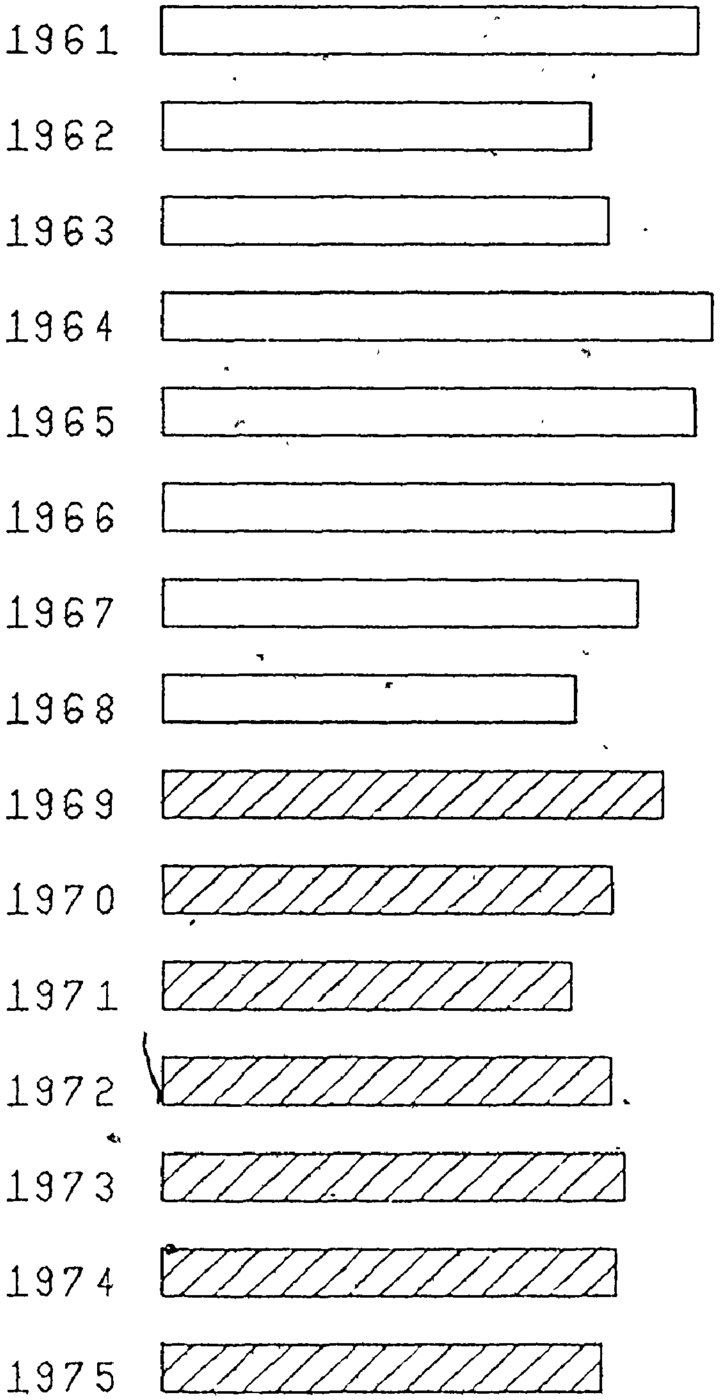


FIGURE 5: Non-binary Dice coefficients for indexes – 1960 and 1961 through 1975

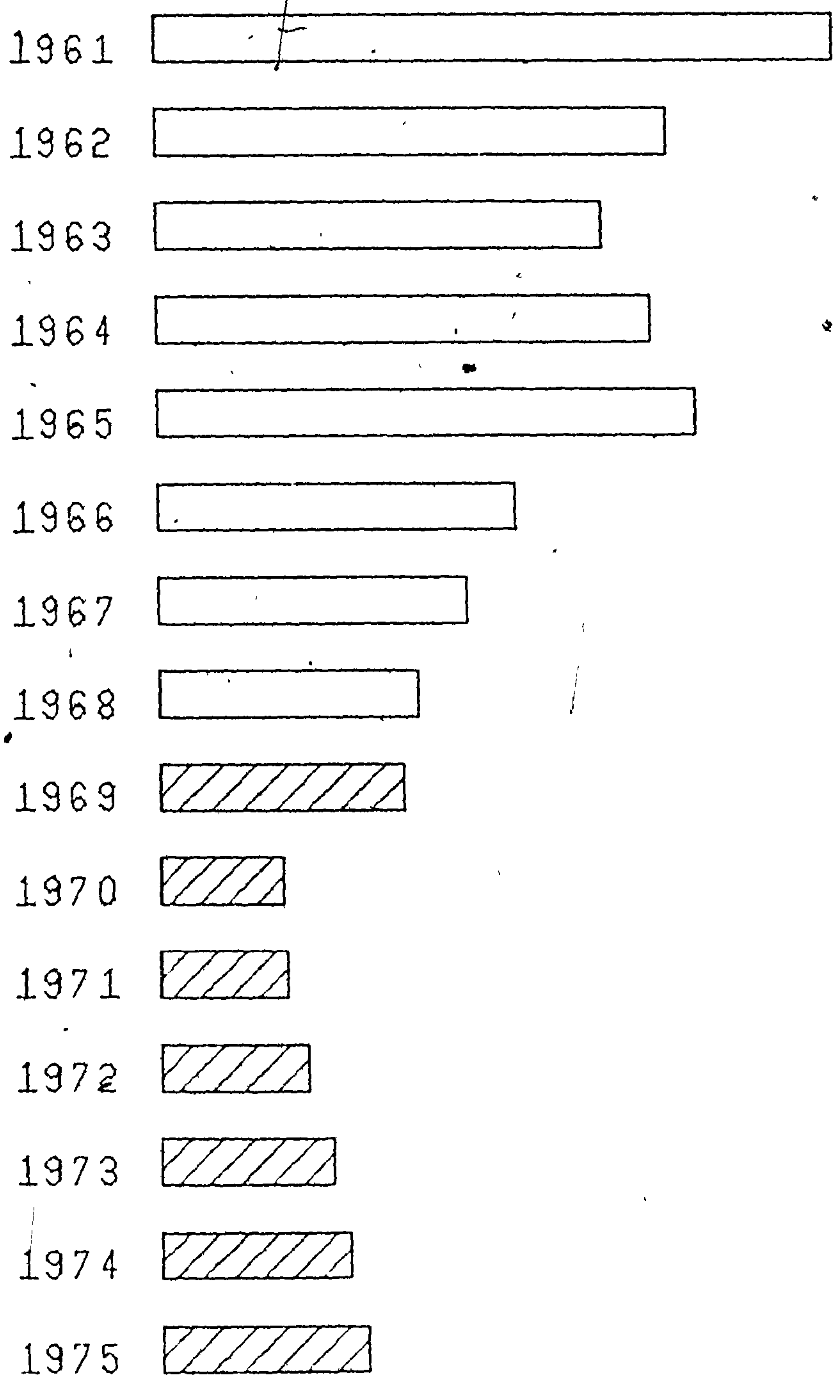


FIGURE 6: Binary Jaccard coefficients for indexes - 1960 and 1961 through 1975

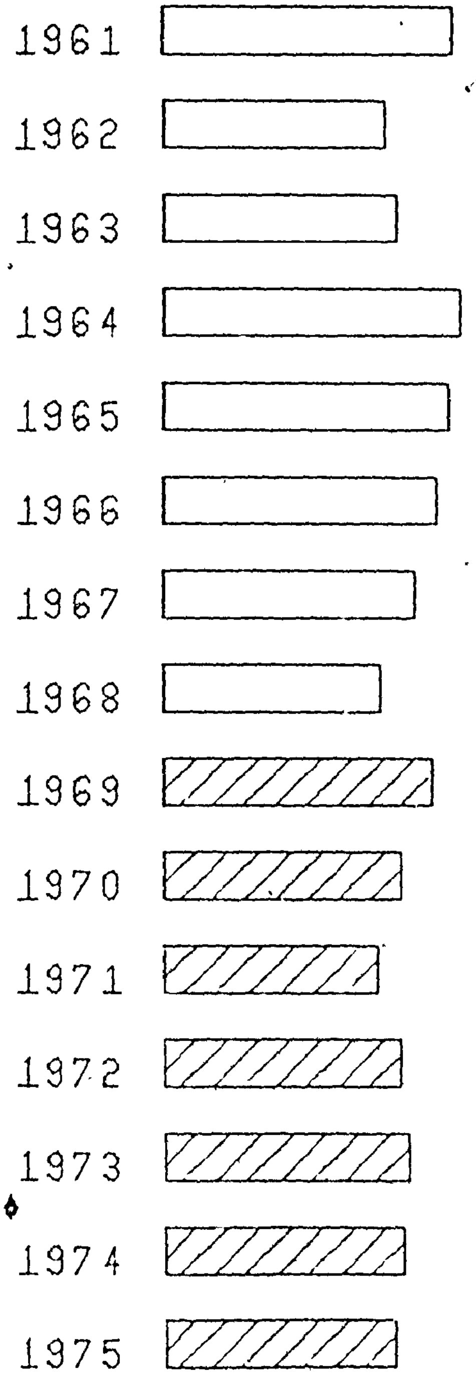


FIGURE 7: Non-binary Jaccard coefficients for indexes – 1960 and 1961 through 1975

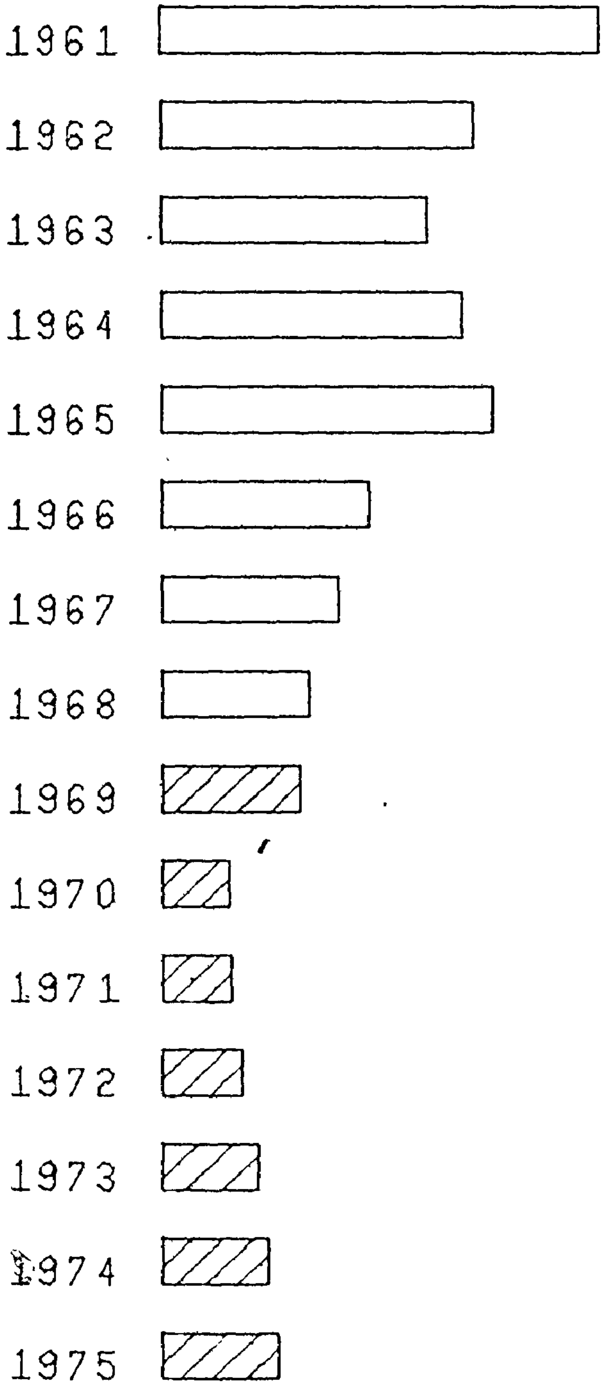


FIGURE 8: Overlap coefficients for indexes scaled by 1/2 - 1960 and 1961 through 1975

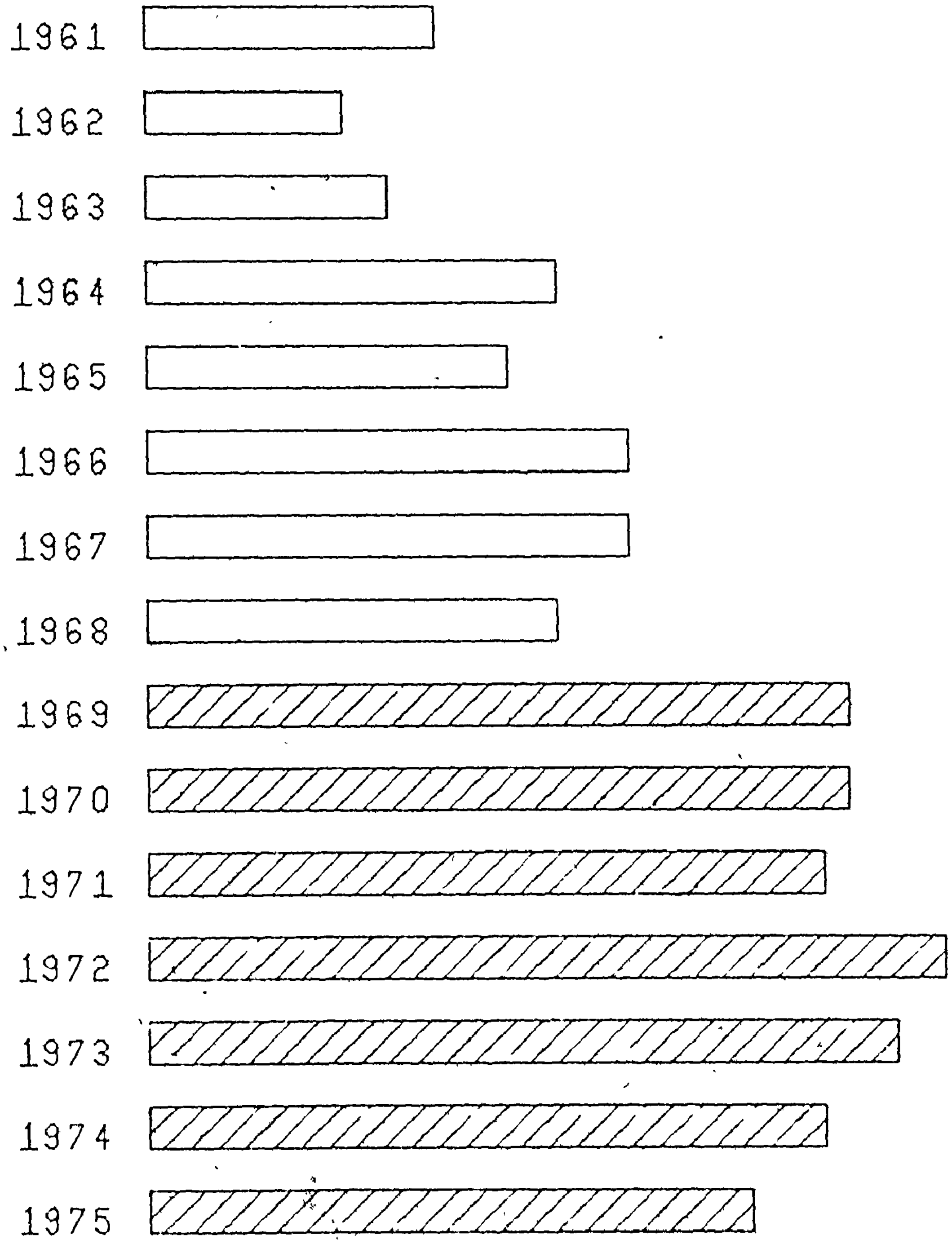


FIGURE 9: Measures of Association based on Yule's characteristic for indexes scaled by 1/2 - 1960 and 1961 through 1975

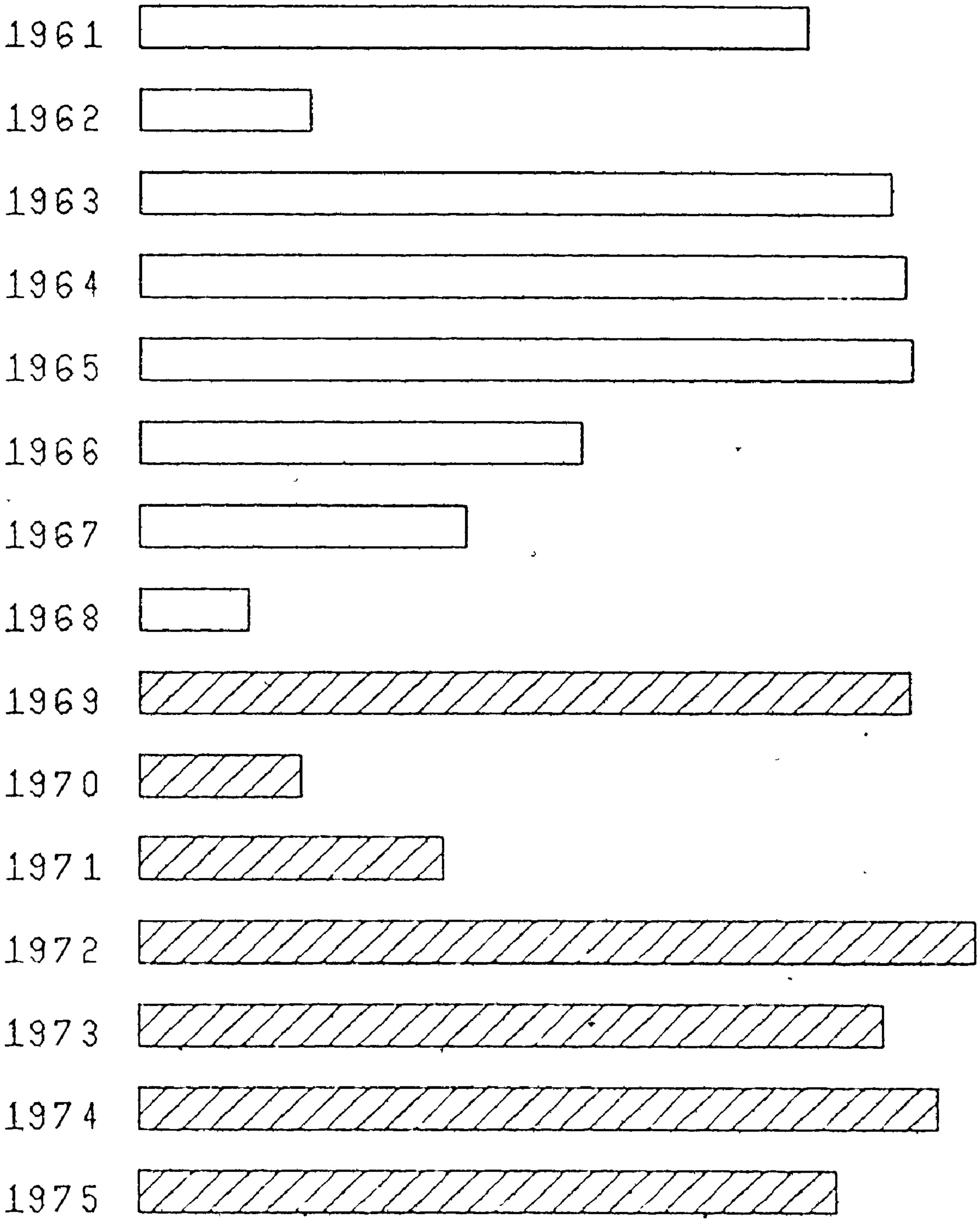
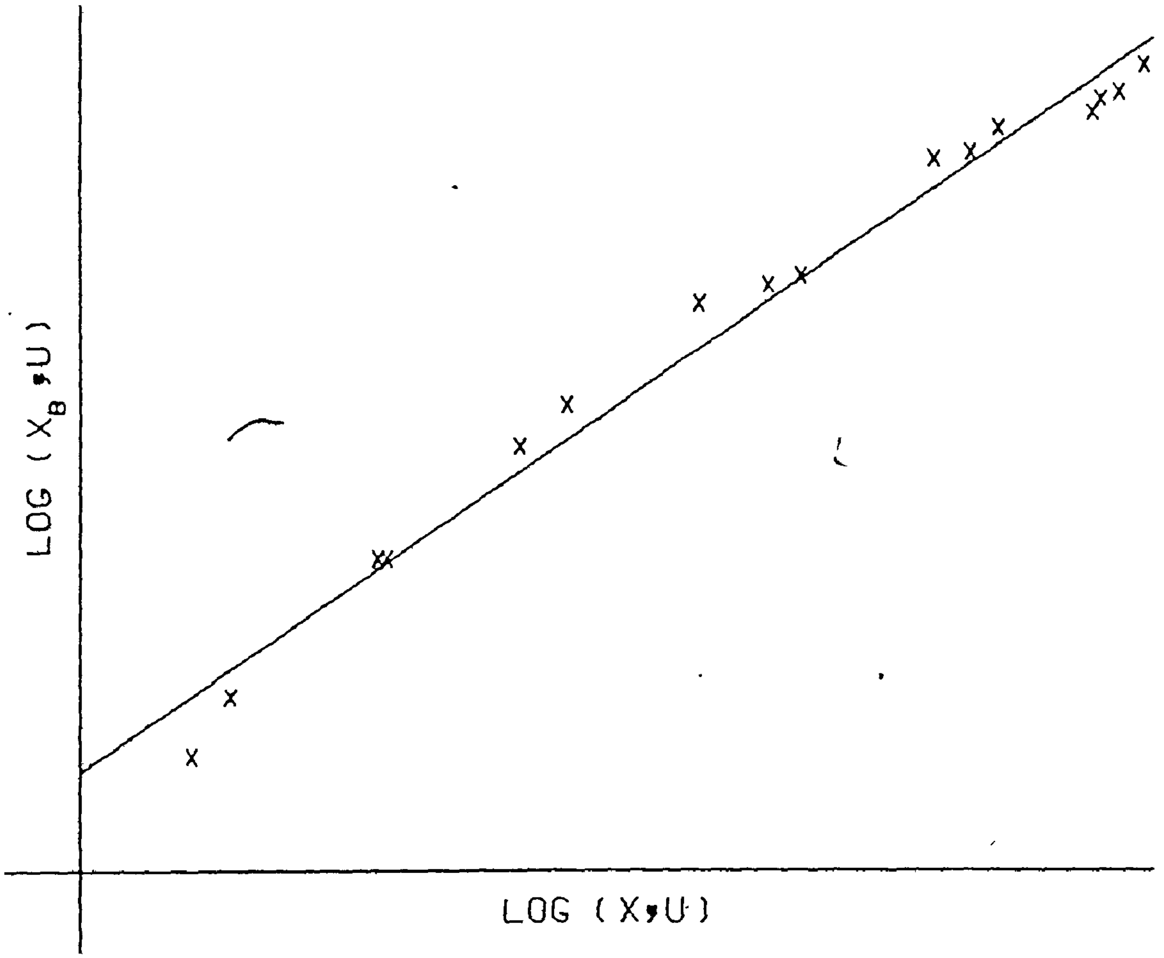


FIGURE 10: Bi-logarithmic plot of vocabulary vs. text length



is due to the nearly negligible size of (X,Y) as compared with (X,X) + (Y,Y). It should also be noted that the contrast between years 1968 and before as against years 1969 and after is not as marked in Figures 4 and 6 as it is in Figures 5 and 7. In all four cases it is in any event anomolous as compared with Figures 1, 3 and 8. Statistical considerations discussed in the following paragraphs indicate that the anomoly renders the Dice and Jaccard coefficients unsuitable for use in the present context.

A little thought reveals that in order for a measure of association such as those given in equations (c.1) through (c.4) and (d.2) to be useful, it must not exhibit a negative correlation between its value and the size of samples being compared. For if such a correlation exists, there is the danger that it is due to the normalization of these measures, rather than to any similarity between the samples. Strong positive correlation is also slightly suspect, since this could indicate that the normalization is unsuccessful in preventing sample size from having an effect on the result. For these reasons Spearman's rank order correlation coefficient [Conover (1971)] was calculated for the coefficients illustrated in Figures 1 through 9, with the following results:

(1.)	Binary cosine for indexes	(Figure 1)	0.843750
(2.)	Binary cosine for vocabs	(Figure 2)	0.364286
(3.)	Non-binary cosine for indexes	(Figure 3)	0.458036
(4.)	Binary Dice for indexes	(Figure 4)	-0.27[illegible]39[illegible]
(5.)	Non-binary Dice for indexes	(Figure 5)	-0.3[illegible]321

(6.)	Binary Jaccard for indexes	(Figure 6)	-0.275893
(7.)	Non-binary Jaccard for indexes	(Figure 7)	-0.934821
(8.)	Overlap coefficients (binary) for indexes	(Figure 8)	0.954464
(9.)	Measure based on Yule's characteristic for indexes	(Figure 9)	0.156250

The null hypothesis can be rejected at the 0.001 level in cases (1.), (5.), (7.) and (8.) above, at the 0.01 level in cases (1.), (3.), (5.), (7.) and (8.); i.e. in one additional case. Considering sample sizes used it cannot be safely rejected in the other cases; i.e. in cases (2.), (4.), (6.) and especially in case (9.). These results were checked using Kendall's Tau [Conover(1971)]. It may be seen that the non-binary Dice and Jaccard coefficients are immediately ruled out as being too negatively correlated. Moreover, the behaviour of the binary and non-binary cases for these two coefficients are seen to be quite different, the importance of which will be dealt with in the next chapter. The binary and non-binary cosine coefficients, on the other hand, once again show fairly good agreement, although as mentioned above, their positive correlation with sample size is somewhat suspect. Even more suspect, however, is the much higher positive correlation of the overlap coefficients.

Finally, the measure of association based on Yule's characteristic is, as might be expected from the theory given in section (d) of the last chapter, least correlated to sample size. Unfortunately this coefficient does not behave well for small samples for the following reason. As indicated in the aforementioned section of Chapter 2, without assuming a Poisson-like distribution for vocabulary, an error will occur in the calculation of the characteristic K, due to the use of small samples [cf. Herdan (1966)]. This error can be estimated with the second term on the right-hand side of equation (d.3), $\dfrac{k}{(X_6,U)} = \epsilon$, say. In other words, errors inversely proportional to vocabulary size will occur in the calculation of K. Hence from the definition of A given in equation (d.2) this will result in errors in the values of A of the order of

$$\frac{1}{k'}\left|\epsilon - \epsilon'\right| \quad \text{where } \epsilon' = \frac{k}{(Y_8,U)}$$

and Y_s is the base term-aggregate vector, (i.e. that corresponding to the year 1960) against which the X_n vectors are being compared. This results in a mean error estimate of 4.08759 over the data being considered, which is considerably larger than any value attained by A. It is an understatement to say that systematic errors of this order are not acceptable.

In the present case, therefore, it appears that despite misgivings over the positive correlation of the cosine coefficients with sample size, these are the best of the coefficients studied in their overall behaviour. To reinforce this statement, several practical observations can be made, based on a more detailed examination of Figures 1, 2 and 3.

Firstly, in all three sets of coefficients, that for the year 1962 appears small relative to most of its neighbours. (The smallness of the coefficient for 1963 in Figure 3 may be attributed to statistical fluctuation, in all likelihood). It is reasonable to suspect that something unusual occurred in the development of the data base during that year. Either the production of literature in the field was sparse, or, as is more likely, some articles have not been abstracted. Similarly 1974 may be looked at, since the vocab coefficient is small (Figure 2), whereas the index coefficients (Figures 1 and 3) are about normal. This is more likely to be a case of random variation however, since all coefficients are not small for 1974. Finally, returning to the change in the coefficients which occurs between 1968 and 1969, we may postulate some sort of change in the literature during that period. In fact discussions with workers in the field reveal that this was the case. At about this time a great deal of new ground was broken by a number of researchers.

We return now to a detailed consideration of the questions raised by Sparck Jones and Van Rijsbergen (1976), mentioned in section (d) of Chapter 1, which are relevant to the present study. The first is "Exactly how competitive is natural language indexing derived from titles with controlled language indexing from abstracts, for retrospective searching, and for on-line searching?" More will be said on this subject in the next chapter, but suffice it to note that Figure 1 represents controlled language indexing while Figure 2 represents the natural language situation. The next question is "How dangerous is document retiring?" This question relates to subject growth and change, i.e. to the first objective of the present study. Another relevant question is "How reliable is automatic assignment indexing?" Since most such schemes depend on measures of association such as those defined herein, the previous remarks regarding our secondary objective may be of some interest to researchers in this area. More will also be said in the next chapter on this subject. Finally, "How are statistical associations best exploited?" While this question may be too general to have a very complete answer, or at least to be anewered in the present context, a postulate of this thesis is that the statistical measures of association given here are of great value in studying the behaviour of data bases, at least for some of its aspects.

Another postulate is that some relationships discovered by the statistical linguists may be applied in the study of data bases. An example is given in Figure 10, which shows that Herdan's law of relative growth appears to apply with remarkable accuracy in this case, as the regression line shows. All points lie within the limits of experimental error from this line.

TABLE 3 - Breakdown of Data Base

YEAR	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975
NUMBER OF CITATIONS	7	10	5	11	18	20	32	41	37	71	111	124	141	121	76	67

TABLE 4 - Relative Growth Data

YEAR	(X_o,U)	(X,U)
1960	35	42
1961	54	70
1962	29	37
1963	54	68
1964	88	126
1965	77	108
1966	121	194
1967	132	270
1968	128	243
1969	195	469
1970	221	699
1971	236	762
1972	257	828
1973	231	717
1974	210	514
1975	191	417

CHAPTER lV CONCLUSIONS

(a) Specific conclusions

Since only one data base of moderate size was examined, the most concrete conclusions based on experimental evidence are of a very specific nature, i.e. specific to this sample. More general conclusions will be put forward in the next section, but these must be regarded as speculations and heuristic based in part on the theory of Chapter 2.

As indicated in the last chapter, there are strong statistical indications that a change of terminology of some magnitude occurred between 1968 and 1969, and that this altered the character of the data base from that time on. All measures of association point to this in some degree, some more markedly than others.

Secondly, based on outside knowledge of the field from which items in the data base arose, one can assert with some confidence that the cosine coefficients gave the best statistical picture of changing terminology of the measures of association studied. Furthermore, the use of the cosine coefficient led to the recognition of 1962 as a year when there was an unusually small representation of data. Also there was a close correlation between the behaviour of the binary and non-binary cosine coefficients. This last consideration is of some practical importance.

To establish the value of a particular element in a binary term-aggregate vector as unity, it is only necessary to find one occurrence of the particular term involved which usually will not mean an extensive search. For all values of elements in non-binary term-aggregate vectors, and for zero values in binary vectors, on the other hand, it is necessary to examine the entire aggregate. Hence there is some saving in search time in using the binary vectors when these convey sufficient information for the purpose at hand. Moreover since, on the average, increasing the size of the aggregate will make the occurrence of a particular term more likely, this saving will increase with larger data bases. Of course, if the file has been previously inverted on the field being considered, non-binary measures do not present a problem.

Finally, focussing on the different behaviour of the cosine coefficients for indexes and vocabs, it appears that the 1968-1969 change was considerably greater for the latter than it was for the former. This is in keeping with the role of index terms which is to preserve continuity of terminology over time. Thus the variation in these coefficients gives a measure of how well or badly this goal is achieved.

(b) General conclusions

It appears very likely that for aggregates of fewer than 1000 terms the binary cosine coefficient is the most effective of those tested in estimating similarity. Likewise, for such aggregates the binary cosine coefficient gives sufficiently accurate comparisons without the necessity of the more arduous calculation of non-binary coefficients.

Unfortunately, the cosine coefficient loses efficacy with larger aggregates for two reasons. Firstly, experiment has shown that there is a positive correlation between the cosine coefficient and aggregate size. Since the cosine coefficient (like all other measures of association) is bounded above by unity, this means that a reduction in the ability to discriminate between large aggregates must occur. Secondly, theory (Chapter 2, section (e)) predicts that for larger aggregates there is a strong likelihood that many term-aggregate vectors will have nearly equal elements leading to a lack of discrimination by the cosine coefficient on this account. This is further verified by the paper of Van der Meulen and Janssen (1977). The theory of statistical linguistics comes to the rescue in this case, however, with Yule's characteristic which is free of both the above defects, but is not usable for small aggregates, due to an error term which becomes negligible for large aggregates (over 1000 terms), but predicts a large error for small aggregates (fewer than 1000 terms). One regrettable point is that the coefficient A, based on Yule's coefficient, is not defined for binary vectors. Hence in this case nothing can be saved in search time as noted above.

(c) Summary of applications

In summary we list six areas of application of the foregoing analysis:

(1) Spotting possible gaps in data base coverage:

Cases where the measure of association is unusually small, particularly if several such measures agree, indicate that there is a smaller overlap of terminology between the samples being compared. This could mean the absence of literature relevant to the data base.

(2) Discovering the extent of changes in terminology:

Abrupt changes in terminology are represented by lasting increases or decreases in the measures of association, as opposed to individual short-term changes which occur due to gaps.

(3) Keeping track of continuity of indexing:

The smaller the changes in measures of association on the indexes, the greater is the continuity of indexing.

(4) Studying the effects of changes in the field on the data base:

Changes in the field often introduce large numbers of new terms which will cause the measures of association to decrease as seen.

(5) Seeing how close the free vocabulary is to an indexing vocabulary:

If the free vocabulary shows sufficient continuity of indexing, the cost of indexing may be unnecessary.

(6) Accounting for recall, precision etc. properties of data bases:

The last possibility especially will require more research and is really beyond the scope of this study. The need for more research will be examined in detail in the next chapter.

CHAPTER V FUTURE RESEARCH

(a) Theoretical research

One apparent need pointed to by the results presented herein is for a single measure of association which will be usable for any size of aggregate. It must be independent of aggregate size, of course, and free from systematic errors. Preferably, it should also be simple to calculate, having a binary vector form, if possible, whose values do not differ too greatly from the non-binary form. All this may be impossible to achieve, but then the mathematical problem of showing that such a coefficient does not exist presents itself. This problem will require a more intensive study of the underlying statistical theory.

(b) Experimental research

A comparison of recall and precision in actual retrieval experiments with measures of association within the data base may show that where lack of continuity of terminology exists, precision is low and recall tends to be inadequate. It should also reveal cases where free vocabulary searches are adequate for certain retrieval purposes and where they are not. Of course changing terminology is not the only culprit upon which inadequate retrieval may be blamed, but it is one which has received scant quantitative attention.

Another area of importance is the bibliometric study of different subject data bases to see what properties of the field of interest might appear in measures of association between years, or between sub-fields of the subject area. This would lead not only to a better knowledge of the history of the field, but also to the possibility of custom-designing a retrieval system especially suited to the peculiarities of the data base. Presumably certain basics such as boolean search logic or variable print fields would be standard, but whether or not to search abstracts, or to include man-made and/ or machine-generated index terms might be decided partly on the basis of such studies.

Finally a confirmation or denial of the conjectures put forward here with regard to the use of the coefficient A with large data bases would be welcomed by the author. Certainly if time and financial resources can be made available, the large data bases are not lacking. In this connection, it would be useful to continue the task of verifying what results of statistical linguistics are applicable to data bases (see section (c) of Chapter 1). In particular, discovering whether the Waring distribution gives the best fit to frequency distribution of terms, or some of the other contenders, would be of considerable advantage in several types of applications, e.g. data base compression [Heaps (1975)], or identifying the effect of deletion of low or high frequency index terms [Svenonius (1972)]. A promising beginning in this area of study has been made by Houston and Wall (1964).

REFERENCES

Bailey, R.W. (1969) Statistics and style: a historical survey. In Statistics and Style, edited by L. Dolezel and R.W. Bailey. New York, Elsevier. p. 217-236.

Climenson, W.D. (1966) File organization and search techniques. Annual review of information science and technology, v.1, p. 107-135.

Conover, W.J. (1971) Practical nonparametric statistics. New York, Wiley.

Dillon, M. (1972) The quantitative analysis of language: preliminary considerations. Computer studies in the humanities and verbal behaviour v.3, p. 191-207.

Ellegård, A. (1962) A statistical method for determining authorship: the Junius letters, 1769-1772. Gothenburg, University of Gothenburg.

Fangmeyer, H. (1974) Semi automatic indexing, state of the art. London, Advisory Group for Aerospace Research and Development, North Atlantic Treaty Organization.

Fodor, I. (1965) The rate of linguistic change. The Hague, Mouton.

Gottlieb, T.Z. (1974) A pattern recognition approach to automatic classification and retrieval of documents. Master's thesis. Toronto, Department of Electrical Engineering, University of Toronto.

Heaps, H.S. (1975) Data compression of large document data bases. Journal of chemical information and computer sciences, v. 15 no. 1, p. 32-39.

Herdan, G. (1960) Type-token mathematics. 'S-Gravenhage, Mouton.

Herdan, G. (1964) Quantitative linguistics. London, Butterworths.

Herdan, G. (1966) The advanced theory of language as choice and chance. Berlin, Springer - Verlag.

Houston, N. and E. Wall (1964) The distribution of term usage in manipulative indexes. American documentation, v. 15, p. 105-114.

King, D.W. and E.C. Bryant (1971) The evaluation of information services and products. Washington, D.C., Information Resources Press.

King Research, Inc. (1976) Statistical indicators of scientific and technical communication (1960 - 1980) Springfield, Virginia, National Technical Information Service.

Lancaster, F.W. (1968) Information retrieval systems: characteristics, testing, and evaluation. New York, Wiley.

Mosteller, F. and D.L. Wallace (1964) Inference and disputed authorship: the Federalist. Reading, Massachusetts, Addison-Wesley.

Muller, C. (1969) Lexical distribution reconsidered: the Waring - Herdan formula. In Statistics and Style, edited by L. Dolezel and R.W. Bailey. New York, Elsevier. p. 42-56

Price, D. de S. (1976) A general theory of bibliometric and other cumulative advantage processes. Journal of the Americal Society for Information Science, v. 27, p. 292-306.

Salton, G. (1968) Automatic information organization and retrieval. New York, McGraw-Hill.

Salton, G. (1971) The SMART retrieval system; experiments in automatic document processing. Englewood Cliffs, New Jersey, Prentice-Hall.

Salton, G. (1975) A theory of indexing. Philadelphia, Society for Industrial and Applied Mathematics.

Simon, H.A. (1955) On a class of skew distribution functions. Biometrika, v. 42, p. 425-440.

Sparck Jones, K. and C.J. Van Rijsbergen (1976) Information retrieval test collections. Journal of documentation, v. 32, p. 59-75.

Sparck Jones, K. and R.M. Needham (1968) Automatic term classifications and retrieval. Information storage and retrieval, v. 4, p. 91-100.

Stevens, M.E. (1965) Automatic indexing: a state-of-the-art report. Washington, D.C., National Bureau of Standards, N.B.S. monograph 91.

Svenonius, E. (1972) An experiment in index term frequency. Journal of the American Society for Information Science, v. 23, p. 109-121.

U.S. Forest Service (1969) FAMULUS: a personal documentation system...Users' manual, Berkeley, California, Pacific Southwest Forest and Range Experiment Station, Forest Service, U.S. Department of Agriculture, National Technical Information Service, PB 202534.

Van der Meulen, W.A. and P.J.F.C. Janssen (1977) Automatic versus manual indexing. Information processing and management, v. 13, p. 13-21.

Van Rijsbergen, C.J. (1975) Information retrieval. London, Butterworths.

Yule, G.U. (1944) The statistical study of literary vocabulary. Cambridge, Cambridge University Press.

APPENDIX A THE FAMULUS PROGRAMS

This suite of programs was originally designed as a personal bibliographic storage and retrieval system for researchers. It is extremely portable, being mainly written in FORTRAN, and is used by a number of institutions throughout North America, such as the Freshwater Institute in Winnipeg, Manitoba, and the University of Waterloo, for a variety of purposes. It executes under a batch operating system and makes efficient use of files in excess of 10,000 documents. The version of FAMULUS used in the present work was implemented at McMaster in 1973 by Dr, B.P, Guru of the Computing Centre on the CDC 6400. A description of the purpose of each of the main programs follows:

1. EDIT:

This program is used for input correction and deletion of records. It also prints records which are being manipulated in one of these ways.

2. SORT:

This program permits rearrangement of the file in alphabetical order by any field.

3. GALLEY:

This program allows the user to produce printed output in a variety of formats. Together EDIT and GALLEY give quite a wide range of possibilities in this regard.

4. SEARCH:

This enables the user to retrieve a sub-set of a file by constructing an appropriate search formula. It is a powerful facility, offering complex nested Boolean expressions.

5. VQCAB:

This program allows one to list all significant words from any desired field or set of fields, in alphabetical order. It was very useful in estimating binary measures of association. It optionally provides card as well as printed output.

6. MERGE:

This offers the user the possibility of combining two files with similar record structure into one file.

7. INDEX:

For every keyword in the descriptor field, INDEX lists the number (assigned by the system) of the record in which that keyword is found. This facility was used in determining the non-binary measures of association.

8. OSSIFY:

This allows the user to reproduce a file on punch cards as back-up to the magnetic tape copy.

In compiling the raw data for various samples, the FAMULUS SEARCH program was altered by the author to operate interactively via INTERCOM. This enabled the author to sequentially search records on selected fields and to complete the collection of data in less time.

APPENDIX B OTHER PROGRAMS

Two additional types of programs were also used in this present study. Firstly, a number of programs were written in interactive BASIC on the HP2000 for estimating various statistical parameters mentioned previously. These were of a standard nature, and are not described any further.

The second type of program written employed SNOBOL IV. This program accepted card output from the FAMULUS VOCAB program (see Appendix A) and compared two such outputs to calculate the number of common terms (i.e. terms the two aggregates had in common). The latter was then used in the BASIC programs mentioned previously to compute the measures of association. The algorithm employed in this analysis is given in Figure 11.

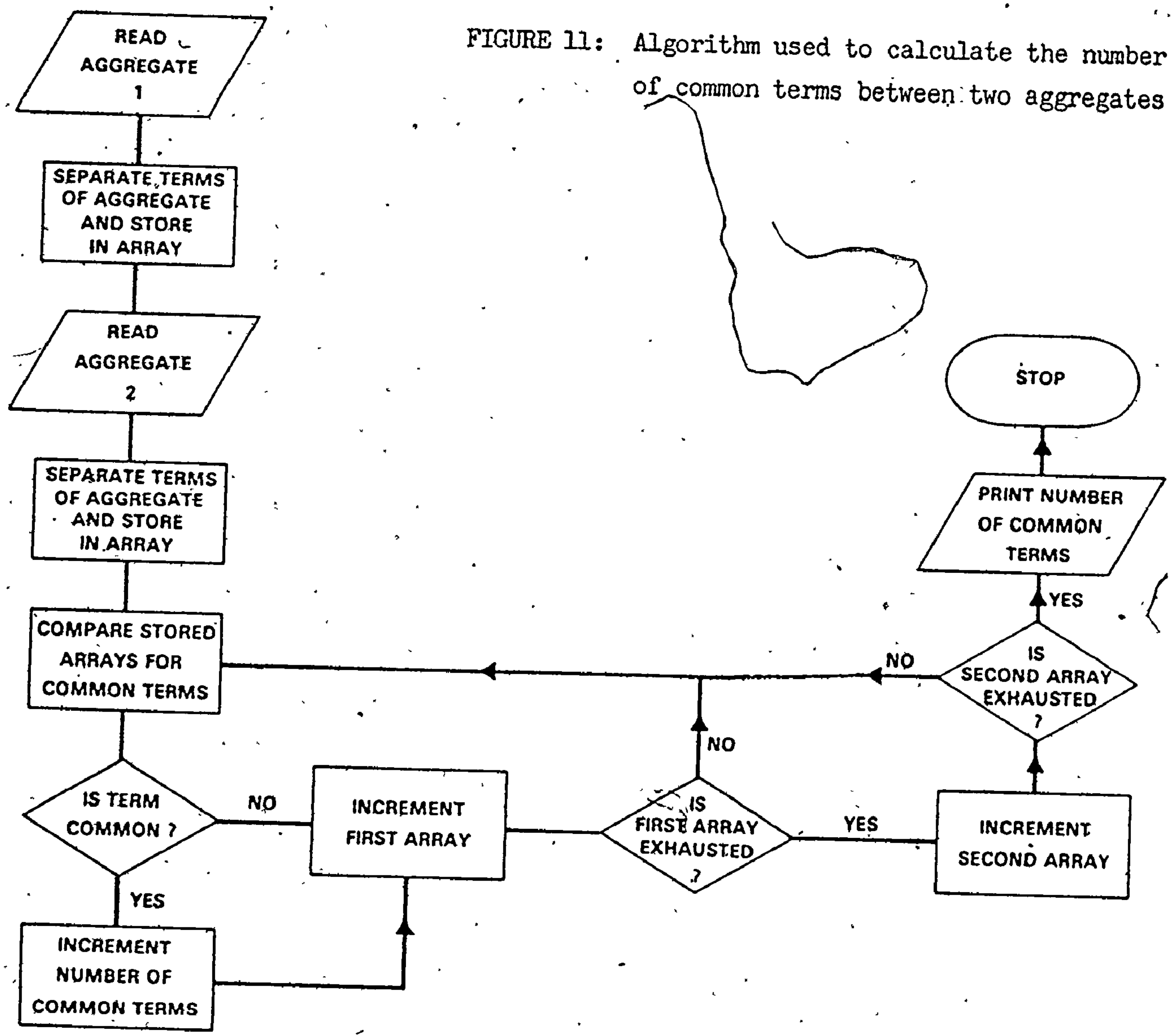

READ AGGREGATE 1
SEPARATE TERMS OF AGGREGATE AND STORE IN ARRAY
READ AGGREGATE 2
SEPARATE TERMS OF AGGREGATE AND STORE IN ARRAY
COMPARE STORED ARRAYS FOR COMMON TERMS
IS TERM COMMON ?
YES
INCREMENT NUMBER OF COMMON TERMS
NO
INCREMENT FIRST ARRAY
IS FIRST ARRAY EXHAUSTED ?
NO
YES
INCREMENT SECOND ARRAY
IS SECOND ARRAY EXHAUSTED ?
NO
YES
PRINT NUMBER OF COMMON TERMS
STOP

FIGURE 11: Algorithm used to calculate the number of common terms between two aggregates

APPENDIX C SOME STATISTICAL CONSIDERATIONS

In Chapter 2, section (d), it is stated that the measure of association designated A is independent of sample size for large samples. A proof of this, due to Yule, follows.

First of all, by the definition of A given in equation (d.2), it is sufficient to show that K(X) is independent of sample size for any integer vector X, and large enough sample size. We now introduce more conventional statistical notation.

Consider a set of m distinct terms distributed in an aggregate according to the Poisson series

(ac.1) $$me^{-\lambda}(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots)$$

where λ is the product of the probability that a given term is used and the total number of occurences of all terms in the aggregate being considered. The number of terms not occurring in the sample is $me^{-\lambda}$; the number occurring n times is $\frac{1}{n!} me^{-\lambda} \lambda^n$, for n = 1,2,...

It is a well known property of the Poisson distribution that its mean and variance are equal. In this case, they are each equal to λ. By the above definition of λ, it is directly proportional to the total number of occurrences of all terms in the aggregate.

Now the distribution for all aggregates together (e.g. a file made up of files for each year) is compounded of a number of such components (ac.1) with different values of λ. Let the mean of the λ-distribution of these components be $\bar{\lambda}$ and its standard deviation be σ_λ. Then the mean of the complete term-distribution, M_c say, is given by:

(ac.2) $$M_c = \bar{\lambda}$$

and since the variance of the complete term-distribution is the mean variance of the component distributions, i.e. $\bar{\lambda}$, plus the variance of the means, which is σ_λ^2,

(ac.3) $$\sigma_c^2 = \bar{\lambda} + \sigma_\lambda^2 = M_c + \sigma_\lambda^2$$

But as remarked above, all λ's are directly proportional to the total number of occurrences. Therefore, doubling the latter will double the former and also double σ_λ. Hence the coefficient of variation of λ or

(ac.4) $$v_\lambda = \sigma_\lambda / \bar{\lambda}$$

is independent of the number of occurrences. But

(ac.5) $$v_\lambda^2 = \sigma_\lambda^2 / \bar{\lambda}^2 = \frac{\sigma_c^2 - M_c}{M_c^2}$$

~~from~~ (ac.2) and (ac.3).

Thus the fraction on the right hand side of equation (ac.5) is independent of the number of occurrences, that is, of the size of sample.

Now if N is the total number of terms (including repetitions) in the complete term-distribution we can write

(ac.6) $$M_c = \frac{S_1}{N}$$

and

(ac.7) $$\sigma_i^2 = \frac{S_2}{N} - \frac{S_1^2}{N^2}$$

where S_1 and S_2 are defined in chapter 2, section (d) (p.16). Hence the right hand side of equation (ac.5) may be re-expressed as

(ac.8) $$\frac{N^2}{S_1^2}\left(\frac{S_2}{N} - \frac{S_1^2}{N^2} - \frac{S_1}{N}\right) = \frac{S_2 - S_1}{S_1^2}N - 1$$

But N is independent of the number of occurrences. Hence since the expression on the right of (ac.8) is independent of the number of occurrences, so must the following expression be:

(ac.9) $$\frac{S_2 - S_1}{S_1^2}$$

But this is just Yule's characteristic, apart from a constant multiplier, as defined in equation (d.1), and re-expressed in vector form K(X) in equation (d.1)$'$. This completes the proof.

As mentioned this proof has the disadvantage of assuming a Poisson distribution for terms. The following proof, due to Herdan, shows that this assumption is unnecessary, however.

Let the number of distinct terms in an aggregate (the vocabulary) be V; let v_x be the coefficient of variation of the term-distribution, and M_x and σ_x be its mean and standard deviation, respectively. Then

(ac.10) $$\frac{v_x^2}{V} = \frac{\sigma_x^2/V}{M_x^2} = \frac{S_2/V^2 - M_x^2/V}{M_x^2}$$

$$= \frac{S_2 / V^2}{(S_1 / V)^2} - \frac{1}{V} = \frac{S_2}{S_1^2} - \frac{1}{V}$$

Since V cancels out of the first term on the right, and 1/V can be neglected for large V, it follows that v_1^2 / V is independent of V, i.e. of sample size.

But the right side of (ac.9) is also identical to the right side of (d.3) apart from a constant factor, which means v_1^2 / V is directly proportional to Herdan's characteristic given by that equation. As mentioned in chapter 2(d), Herdan's characteristic and Yule's characteristic become arbitrarily close to

$$k \frac{S_2}{S_1^2} = k \frac{(X,X)}{(X,U)^2} .$$

This completes the proof.

More generally, one may ask what is known about the properties of measures of association which might allow a better choice of measure than either A or the cosine coefficient. Unfortunately the answer to this question is at present, very little. There are two standard references in cluster analysis, the field in which discussions of such measures occur most frequently, namely Sneath and Sokal (1973), and Duran and Odell (1974) . Both survey the literature, the former somewhat more extensively than the latter. The latter gives a more theoretical approach than the former, but neither comes to grips with the problem encountered here, the dependence of results on sample size. A theory which would allow the user of such measures to minimize this dependence is lacking in the literature, but Yule's results given above would tend to indicate the likelihood that such minimization would require in most cases a knowledge of the underlying statistical distribution of the elements used in the comparison, in this case the terms. This knowledge is lacking in the present case, and, judging from literature previously cited (see chapter 5, section b), will require much more research to establish with the desired degree of exactness.

In the absence of a general theoretical structure of the kind described in the last paragraph, one is left with the option of trying many possibilities and comparing the results. This, in part at least, is what has been done in the present work.

Duran, B.S. and P.L. Odell (1974). Cluster analysis; a survey. New York, Springer-Verlag.

Sneath, P.H.A. and R.R. Sokal (1973) Numerical taxonomy; the principles and practice of numerical classification. San Francisco, W.H. Freeman.