STATISTICAL TOOLS FOR DATA BASE STUDY (LES OUTILS STATISTIQUES POUR L'ÉTUDE DES BANQUES D'INFORMATION)

D.A. Fox and R.A. Rink
McMaster University
Hamilton, Ontario
L8S 4K1

ABSTRACT

In this paper statistical tools are developed for the analysis of bibliographical data base growth. A theoretical basis is given and the results applied in determining properties of an actual on-going data base, the SIRLS file of the Department of Human Kinetics and Leisure Studies at the University of Waterloo. Certain conclusions are reached concerning the degree of indexing and the continuity of coverage of the subject, particularly relating to the behaviour of so-called "free" and "controlled" vocabulary terms. Applications of these and similar results are discussed, with an emphasis on practical problems of workers in the fields of bibliographic data base management and searching. Suggestions are put forward for the direction of future research in the area. (Nous développons, dans cet article, des outils statistiques pour analyser la croissance des banques d'information bibliographiques. Nous donnons une base théorique et les résultats nous nous adressons à la resolution des qualités d'une banque actuellement employée; c'est à dire la banque "SIRLS" du département "Human Kinetics and Leisure Studies" à l'université de Waterloo. Nous achévons certaines conclusions à l'égard du degré de répertoiriant et de la continuité dans la mise à jour des indexes, surtout comme elles se rapportent au comportement du "vocabulaire libre" et du "vocabulaire reglé", soi-disant. Nous discutons l'usage de ces résultats et leurs pareils en accentuant les problèmes pratiques pour les cadres employés a l'administration et la récherche des banques d'information bibliographiques. Nous avançons quelques suggestions à la direction de la récherche future.)

TOOLS FOR DATA BASE STUDY

INTRODUCTION

The customary method of evaluating information retrieval systems has been through user-dependent measures such as relevance (precision) and recall (Lancaster 1968, and King and Bryant 1971). While these are of obvious importance, since the usefulness of any service must be determined in the end by its users, they lack the ability to point directly to the reasons for their values. Usually these reasons can be found in the data base itself, or the particular version of it available through a given system. Considerable knowledge of the data base developed during many hours of costly experience may be necessary to correctly diagnose problem situations, and to enable a searcher to take corrective measures.

Suppose for example that a user wishes to perform a comprehensive search on bubble memory research for the last five years. One could choose a system which accesses title terms only, with a broad multi-facetted search strategy, or choose a system which searches abstracts at greater cost, making use of high precision strategies to reduce the number of false drops. A third course might be to search title and index terms with the help of a thesaurus. Clearly, deciding on any particular procedure must involve knowledge of the data base, as well as considerable familiarity with the literature it represents. For the beginner the best (most costly) course would be to try all three, but in practice only one or two are actually carried out.

The indexer (data base builder) also has problems of a related nature. He/she would like to know what is the most cost-effective indexing procedure for a given subject area and within a given budgetary framework on the spectrum extending from free vocabulary on the one hand to a highly structured subject heading and/or numerical classification controlled situation on the other. Another problem is whether it is better to continue indexing policies which are no longer of value to some end users, or to radically break with the old rules and risk confusing users who are happy with past practices. To solve these problems considerable user input is necessary. It is also evident that the user must have some knowledge of the structure of the data base for effective searching. For the above reasons, communication must occur between the builder and the user (indexer and searcher). But what form should this communication take? How can vagueness, misunderstanding and error best be avoided?

Science has always employed mathematics as a language for such communication. Its advantages are many: it is a universal tongue, logically simple and lacking in the confusions resulting from multiplicity of choice. In the last one hundred years a large number of tools of great generality have been developed, and are finding many applications in the social sciences. The purpose of the remainder of this paper is to employ some tools from applied statistics, and to suggest further developments.

TOOLS FOR DATA BASE STUDY

A minimal amount of theory will now be presented, and later applied.

Theory

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

Consider a term-aggregate vector [1]

$$X = (\chi_1, \chi_2, \ldots, \chi_n),$$

where χ_i = 1 if the i-th 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

(1) $U = (1, 1, \ldots, 1)$

corresponds to the term authority list.

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

(2) $$X = (\chi_1, \chi_2, \ldots, \chi_n) \rightarrow \bar{X} (\bar{\chi}_1, \bar{\chi}_2, \ldots, \bar{\chi}_n, \bar{\chi}_{n+1}, \ldots, \bar{\chi}_{n+m})$$

Here two types of change have occurred. First, some of the terms have disappeared and others have appeared; i.e. some $\bar{\chi}$'s = 0 where the corresponding χ's = 1, while some $\bar{\chi}$'s = 1 where the corresponding χ's = 0. Second m new terms have appeared which were not previously present.

For the purpose of examining the overall change that has occurred we now define a similarity coefficient [2] (Salton 1962):

(3) $$c(X, \bar{X}) = \frac{(X_e, \bar{X})}{\|X_e\| \; \|\bar{X}\|}$$

where (a) X_e is the extended X vector:

$X_e = (\chi_1, \chi_2, \ldots, \chi_n, 0, 0, \ldots, 0)$, ending in m zeros.

(b) $(X_e, \bar{X})$ is the scalar (inner) product of X_e with $\bar{X}$, and

(c) $\|X_e\|$ and $\|\bar{X}\|$ are the lengths of X_e and $\bar{X}$ (standard euclidean distance between their end points).

---

1. In what follows capital letters denote vectors.

2. The reference cited (Salton 1962) gives an excellent discussion of the appropriateness of applying this coefficient to vectors of ones and zeros. For further discussion of similarity coefficients see Duran and Odell (1974), and Sneath and Sokel (1973).

For our present analysis, let us examine a special case. Assume that X is a universal vector as given in (1). This is equivalent to selecting a base vector, in a sense to be made clear below. Then

(4) $\gamma_i = 1$ for all i = 1, 2, ..., n .

Equation (3) can be rewritten as follows

(5) $$c(X, \bar{X}) = \frac{nrep}{\sqrt{n\,\bar{n}}}$$

where $nrep = (X_e, \bar{X})$ is the number of ones common to X and $\bar{X}$ (i.e. occuring in the same positions), and $\bar{n} = \|X\|^2$ is the number of ones in $\bar{X}$. In other words, nrep is the number of terms in the initial aggregate which are repeated in the new aggregate, while $\bar{n}$ is the total number of terms, old and new, occuring in the new aggregate.

Experimental Results and Discussion

In order to apply the above theory a data base in the social sciences was used. This type of data base provides an example where changes in terminology are more prevalent than a data base in the physical sciences. It should be noted that we are using the phrase "changes in terminology" in a rather restricted sense. By this phrase we mean the tendency of an aggregate to drop some terms and acquire others. We do not consider the meaning of these terms per se, only their physical form.

The data base employed in these studies is a large subset of the SIRLS file of the Department of Human Kinetics and Leisure Studies, University of Waterloo. Specifically, 892 citations with abstracts were studied, ranging in publication date from 1960 through 1975. This 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). The resulting lists were then compared and similarity coefficients were computed.

While the results which follow may seem highly theoretical, they have the objective of obtaining a better understanding of a rather complex entity, the bibliographic data base. Indeed, it is doubtful whether the builders of data bases understand them completely, while those who interact with them occasionally have less chance of comprehensive knowledge. Much of the following discussion is speculative, and the need for further research to ascertain the reliability and real extent of the results is obvious.

In Table 1, results for indexes and vocabs (free vocabulary, i.e. title terms and abstract terms after being passed through a stop list) are given. These are also illustrated in Figures 1 and 2.

The initial impression created by these results is one of

Table 1 - Numbers of terms and similarity coefficients for base year 1960 compared with years 1961 through 1975.

	YEAR	NUMBER OF TERMS	SIMILARITY COEFFICIENT
(a.) Indexes:	1960 (base)	35	--------
	1961	54	0.276026
	1962	30	0.216025
	1963	54	0.230022
	1964	88	0.306319
	1965	77	0.288943
	1966	121	0.307329
	1967	132	0.294245
	1968	128	0.253986
	1969	195	0.351032
	1970	221	0.329737
	1971	236	0.308083
	1972	257	0.347947
	1973	231	0.344764
	1974	210	0.326599
	1975	191	0.305766
(b.) Vocabs:	1960 (base)	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	0.159640
	1971	3136	0.170896
	1972	3125	0.171197
	1973	3176	0.168604
	1974	2287	0.112924
	1975	1975	0.181506

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 years does not decrease markedly with time. In fact, in the case of the index terms (Fig. 1) there is actually an increase noted, in that the similarity between 1960 and 1975 is significantly (10.8%) greater than that between 1960 and 1961. The vocabs, on the other hand show a different pattern (Fig. 2). Their similarity coefficients are considerably smaller than

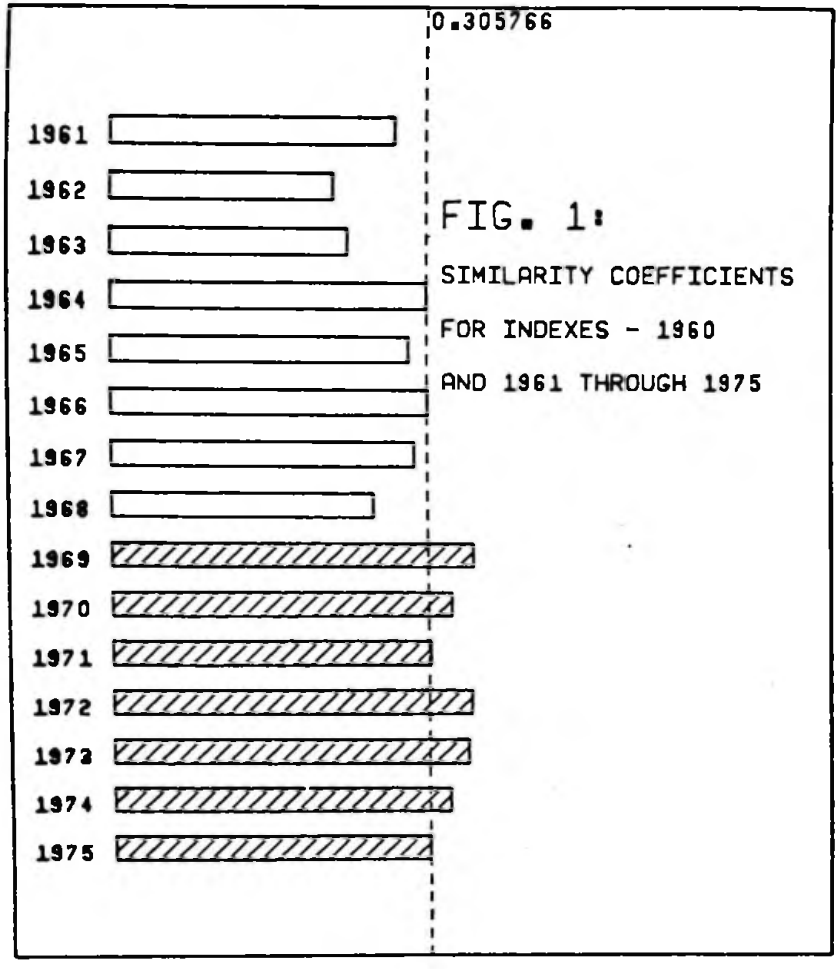

FIG. 1:

SIMILARITY COEFFICIENTS

FOR INDEXES - 1960

AND 1961 THROUGH 1975

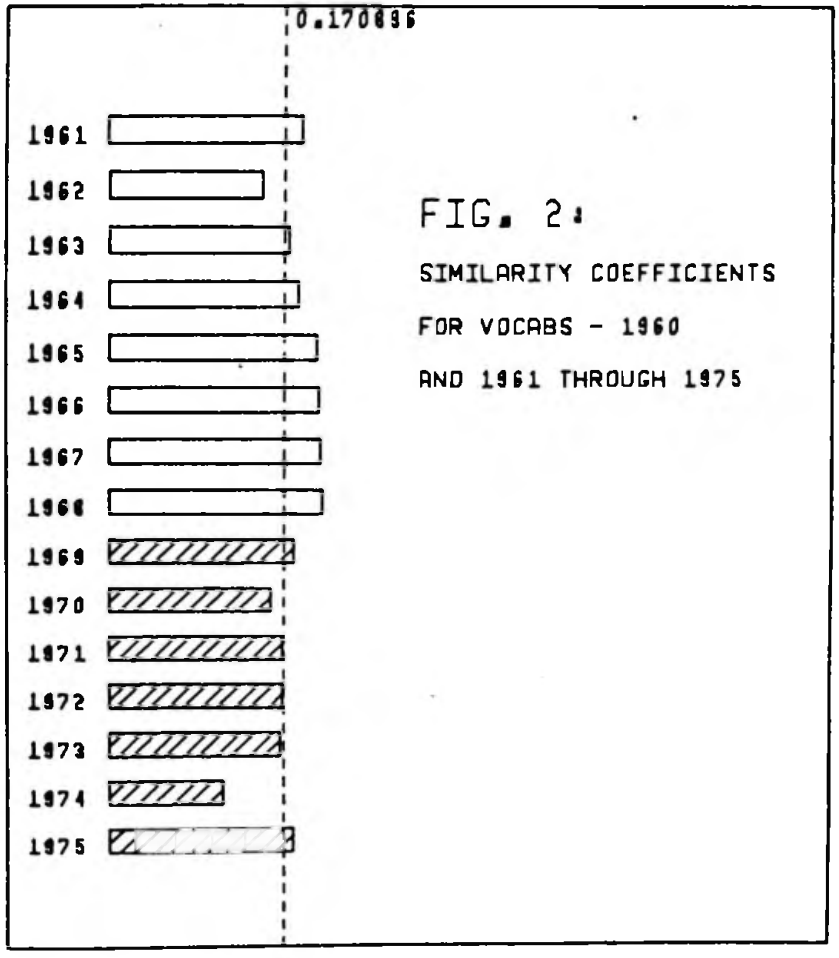

FIG. 2:

SIMILARITY COEFFICIENTS

FOR VOCABS - 1960

AND 1961 THROUGH 1975

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 coefficients for the indexes increase at the same point. 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%. Moreover, with two exceptions (1964 and 1966), all similarity coefficients prior to 1969 are less than any after 1968. This is illustrated by the dotted line in Fig. 1, which is drawn at the level of the smallest similarity coefficient after 1968 (i.e. 1975). This also shows that the two deviant years are not above the line by a significant amount. In the case of the vocabs the dotted line is drawn at the height of the median similarity 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, 0.203306, 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 similarity coefficients of the indexes with time is almost certainly due to the small size of the set of all possible index terms (496) as compared with the very large size of the set of all possible vocab terms. For this reason increasing the size of the number of index terms has a marked effect on the number of repeated terms, nrep, and hence on the coefficient. It seems likely, however, that as data base additions become very large this phenomenon will be less noticeable, as a kind of saturation point is reached. In this 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 respect to the vocab terms, the increase in repeated terms was compensated for by an overall increase in terms. Here it is doubtful whether term frequency counts would be of value as indicators of expansion, since the complete text of the document is not present, and since there is no synonym control.

The question of whether the similarity coefficients depend on the size of the aggregate may be raised at this point. A negative correlation would cast doubt on the value of the statistic, since it would presumably be due to the occurrence of the sizes of the aggregates in the denominator of the right-hand side of equation (5), thus indicating that the variation in the number of repeats, nrep, was of little significance. This is not the case however. The results of performing rank correlation tests are $p = 0.84375$ for the indexes and $p = -0.36429$ for the vocabs. In the former case this represents positive correlation with a rejection of the null hypothesis at the 0.01 level of significance; in the latter case the null hypothesis cannot be rejected. Positive correlation between number of terms and similarity coefficients for the indexes is of course reasonable since, as already mentioned, changes in nrep are of so much importance to the value of the similarity coefficient

in this case.

Apart from the above considerations which lead to a better understanding of the behaviour of index and vocab terms, other more immediately practical facts appear. For example in both sets of similarity coefficients, that of the year 1962 appears small relative to its neighbours. It is reasonable to suspect that either the sample for 1962 is unrepresentative of the data base, or that something has happened to the data base in that year. In the latter case either the literature was sparse for that year or (as is more likely) some articles have not been abstracted. Similarly 1974 may be looked at, since the vocab similarity coefficient is small whereas the index coefficient is about normal. This is more likely to be a case of random variation however, since the two coefficients are not both small.

Finally returning to the change in both sets of coefficients which occurs between 1968 and 1969, we may postulate that there was a 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.

Conclusions

Since this is only a preliminary study of a single data base it is difficult, if not foolhardy, to draw concrete conclusions. Some interesting facts have come to light however; it appears that index terms behave differently from free vocabulary (vocab) terms in growing data bases. Their role is to preserve the continuity of terminology, and the extent to which they fulfill this role can be made mathematically precise as seen. On the other hand, an indication of the extent to which the vocab terms might be able to fill the role of index terms is also given; evidently in the case of this data base, vocab terms do not come near to exhibiting the continuity of terminology of the index terms.

In summary we list seven areas of possible application of the above type of analysis.

1. Spotting possible gaps in data base coverage
2. Discovering the extent of changes in terminology
3. Keeping track of continuity of indexing
4. Studying the effects of changes in the field on the data base
5. Seeing how close the free vocabulary is to an indexing vocabulary
6. Estimating the advantage to be gained from creating inverted files in terms of reduction in the number of terms that must be checked in a search (the more repeated terms, the more reduction)
7. Accounting for recall, precision etc. properties of data bases.

The last two possibilities especially will require more research, and are really outside the scope of this study. The remainder are suggested by the results, but corroborative evidence is clearly necessary.

TOOLS FOR DATA BASE STUDY

In the past decade we have been busy amassing large machine - readable bibliographic files. Perhaps it is not too early to attempt to understand the nature of these creations in an effort to employ them more effectively.

REFERENCES

DURAN, B. S. and ODELL P. L. 1974 Cluster analysis, a survey. Berlin, Springer-Verlag. 137 p.

KING, D.W. and BRYANT E. C. 1971 The evaluation of information services and products. Washington, D.C., Information Resources Press. 306 p.

LANCASTER. F. W. 1968 Information retrieval systems: characteristics, testing and evaluation. New York, Wiley. 222 p.

SALTON, G. 1962 The use of citations as an aid to automatic content analysis, pp. III - 1 to III - 51 in Information storage and retrieval. Scientific report no. ISR - 2, the Computation Laboratory, Harvard University, Cambridge, Mass.

SNEATH, P. H. A. and SOKAL R. R. 1973 Numerical taxonomy. San Francisco, W. H. Freeman. 573 p.

U. S. FOREST SERVICE 1969 FAMULUS: a personal documentation system ... Users' manual. Berkeley, Pacific Southwest Forest and Range Experiment Station, Forest Service, U.S. Department of Agriculture, National Technical Information Service, P B 202 534. 40 p.