ELSEVIER

Information Processing and Management 40 (2004) 365–377

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Bibliometric analysis of the automatic indexing literature: 1956–2000

Antonio Pulgarín [a,*], Isidoro Gil-Leiva [b]

[a] *Faculty of Library & Information Science, University of Extremadura, La Alcazaba, 06071 Badajoz, Spain*
[b] *Faculty of Computer Sciences, Polytechnic University of Valencia, 46022 Valencia, Spain*

Received 1 August 2002; accepted 7 November 2002

Abstract

We present a bibliometric study of a corpus of 839 bibliographic references about automatic indexing, covering the period 1956–2000. We analyse the distribution of authors and works, the obsolescence and its dispersion, and the distribution of the literature by topic, year, and source type. We conclude that: (i) there has been a constant interest on the part of researchers; (ii) the most studied topics were the techniques and methods employed and the general aspects of automatic indexing; (iii) the productivity of the authors does fit a Lotka distribution ($D_{max} = 0.02$ and critical value $= 0.054$); (iv) the annual aging factor is 95%; and (v) the dispersion of the literature is low.

Keywords: Automatic indexing; Scientific output; Bibliometric analysis; Bradford's law; Obsolescence

1. Introduction

Indexing, the procedure applied to documents and queries to select their essential concepts, has the function of allowing the storage in databases as well as the later retrieval (two sides of the same coin).

As indexing is an intellectual process (reading, comprehension, analysis, representation), one of its characteristics is having a strong subjective component, and indeed it can be said that subjectivity is inherent to indexing. One way to detect the subjectivity in indexing is to study its consistency, either between different indexers in their analysis of the same document, or between analyses of the same document by the same indexer at different times. The complexity of indexing

* Corresponding author. Tel.: +34-924-286406; fax: +34-924-260680.
E-mail addresses: apulgue@alcazaba.unex.es (A. Pulgarín), isgil@har.upv.es (I. Gil-Leiva).

0306-4573/$ - see front matter
doi:10.1016/S0306-4573(02)00101-2

is very clearly presented in the recent reviews (Anderson & Pérez-Carballo, 2001a, 2001b), and this difficulty becomes even greater when the aim is to make indexing automatic.

At the beginning of the 1970s, a debate began about whether or not automating the process of indexing was worthwhile. Apart from this debate, the process has aroused clear interest amongst researchers, given the quantity of literature produced in the last half century (about a thousand papers).

This quantity of research literature has led to the appearance of new techniques for both the treatment of the information and its later retrieval. The changes took place thanks to the first generally available computers which made it possible to carry out rapid repetitive mechanical operations, and, in this field in particular, to extract keywords from the texts.

In reviewing the scientific literature, one finds a wide variety of terms used to designate concepts similar to what we know as automatic indexing, including such expressions as: ''Automated assisted indexing'', ''Automated indexing'', ''Automated support to indexing'', ''Automatic support to indexing'', ''Computer aided indexing'', ''Computer assistance in indexing'', ''Computer assisted indexing'', ''Computer indexing'', ''Computerized indexing'', ''Indexing by computer'', ''Indexing program'', ''Indexing software'', ''Machine aided indexing'', ''Machine indexing'', ''Machine-assisted indexing'', ''Mechanical indexing'', ''Mechanized indexing'', ''Microcomputer-based indexing'', ''Semi-automatic indexing'', ''Automatic indexing''. The last of these expressions is the most often used in the literature. This terminological variety reflects three different concepts: (i) Computer programs that aid in storing the indexing terms after they have been extracted intellectually. Such systems provide help screens giving explanatory notes on the use of a term or related terms, and allowing terms to be assigned without having to key them in. They even allow any aspect of the process to be checked by on-line consultation of previously indexed documents (computer assisted indexing during storage). (ii) Systems which automatically analyse the documents, but the proposed indexing terms are then checked and edited if necessary by a professional (semiautomatic indexing). (iii) Programs without any kind of validation, i.e., the proposed terms are stored directly as keywords or descriptors of the given document (automatic indexing; Gil-Leiva, 1999).

The objective of the present work is to describe a bibliometric analysis of the scientific output on automatic indexing from 1956 to 2000. We analyse the volume of information, both overall and by subject, to determine how it evolved and the type of document in which it was published. It is interesting to study authors and their productivity as well as the type of distribution they present as a group. We also study the obsolescence, based on document transfer, to determine the type of subjects involved from a perspective of the dynamics of science. Lastly, focusing on a single source type—scientific journals—we study the dispersion of the works published in them.

2. Method

The material used in the study was a set of 839 bibliographic references covering the period 1956–2000, inclusive. The references used were obtained by means of an exhaustive document search performed in two phases: the first phase, from 1994 to 1997, had already been used in an earlier study (Gil-Leiva, 1999), and the second, from 1998 to 2000, was to cover the publications of those three years.

The literature corpus was constructed in two ways: by extracting the bibliographic references from each article, book, or report that we had read in recent years, and by database searches. The databases used were Library and Information Science Abstracts and Information Science Abstracts—as the only specialized databases in the subject—as well as others not specializing in information science, but which also include relevant references—ERIC, MEDLINE, and SIGLE. We also used the Ph.D. thesis databases Teseo, ThesNet, and Dissertation Abstracts.

All database searches were made on the "descriptor" field, so as to guarantee that the documents retrieved dealt specifically with "automatic indexing". The term "descriptor" means a word, term, or expression, chosen from a set of words or terms considered to be equivalent, to represent an essential unambiguous concept contained in the documents. The field descriptor is a part of the record that represents a document and that contains the descriptors.

This material was used in a bibliometric analysis of the scientific output (distribution of authors and works, etc.), the obsolescence of the scientific literature and its dispersion, and the distribution of the literature by topic, year, source type, etc.

3. Results

3.1. Scientific output by document type

In Table 1, one can observe some of the aspects of the scientific output and document type arranged into five-year periods. Firstly, there is a notable generalized decline in the 1996–2000 period. The causes of this will not become clear until it is possible to analyse the data of the following five-year period. Secondly, beginning in the 1980s, there is a clear increase in the number of Ph.D. theses. Lastly, we found remarkably few patents.

3.2. Annual scientific output by topical classification

The entire set of bibliographic references was arranged into the following thematic classification: *general aspects*, *linguistics*, *automatic indexing vs manual indexing*, *evaluation*, *status of the*

Table 1
Scientific production by document type

Years	Journal articles	Books and book chapters	Congresses	Theses	Reports	Memoires	Patents	Total
1996–2000	81	4	12	14	1	0	0	112
1991–1995	118	12	32	11	7	0	1	181
1986–1990	76	17	25	14	0	2	1	135
1981–1985	48	14	5	12	2	1	0	82
1976–1980	48	10	7	3	2	1	0	71
1971–1975	76	18	9	2	4	0	0	109
1966–1970	64	31	18	1	4	0	0	118
1961–1965	12	5	4	1	1	0	0	23
1952–1960	4	3	0	0	1	0	0	8
Total	527	114	112	58	22	4	2	839

question, *automatic indexing and retrieval*, *automatic indexing software*, and *techniques and methods* (Table 2).

The works with proposals of techniques and methods for automatic indexing were the most numerous (290). There were possibly two reasons for this: the complexity of undertaking this process automatically stimulated many experiments, and the lack of any corresponding methodological consensus amongst the scientific community led each researcher or group of researchers to propose different mechanisms. The second most frequent topic was that of works classified under general aspects (269), with a mix of the theory of automatic indexing, problems to be overcome, and reflections on the theme. Thirdly, there were works dealing with the relationship automatic indexing and retrieval (111).

The analysis of the thematic classification by year showed that there were two peaks in the distribution of the item general aspects, the first between the years 1969 and 1975, and the second, more marked than the other, between 1989 and 1999. The same was the case with linguistics, with one of the two peaks in the early 1970s and the other in the early 1990s. There was little change between 1969 and 2000, however, for techniques and methods, automatic indexing vs manual indexing (the debate on whether or not automating the process was worthwhile), and status of the question. Lastly, there was a notable peak from 1990 to 1997 corresponding to automatic indexing and retrieval.

3.3. The scientific output of the authors

In 1926, Alfred J. Lotka from his investigations into the frequency distribution of the scientific output of physicists and chemists, formulated a relationship between the frequency of authors and their publications.

In his article, Lotka says: "In the cases examined it is found that the number of persons making 2 contributions is about one-fourth of those making one; the number making 3 contributions is about one-ninth, etc.; the number making x contributions is about $1/x^2$ of those making one; and the proportion, of all contributors, that make a single contribution, is about 60 per cent."

In other words, for every 100 authors with one article as output, there will be 25 with two articles each, about 11 with three, approximately 6 with four contributions, and so on. Lotka (1926) found, for the two data sets he analysed, an exponent of 2.02 ± 0.017, and 1.888 ± 0.007. Consequently: "The general formula for the relation thus found to exist between the frequency of persons making x contributions is $x^n y = \text{constant}$ ($y_x = c \times x^{-n}$)."

The methods for the calculation of the values of the constant 'c' and the slope 'n' were also defined (Nicholls, 1986; Pao, 1985). The calculation of 'c' (the value corresponding to the number of authors with a single work in Lotka's equation) requires the prior determination of the value of the slope of the distribution 'n', which in turn requires having decided on the number of data pairs to be used in its calculation.

While the two calculations, of 'c' and of 'n', have been resolved methodologically, the last step mentioned, the choice of the number of data pairs to use in calculating the slope, has been the object of most proposals, without any of them being accepted by the scientific community up to now. We shall determine the parameters excluding that part of the data representing the more prolific authors (at $y_x = 1$; Table 3). We calculated the author productivity using the "normal counts" method, which gives full credit to all contributors (Egghe & Rosseau, 1990; Lindsey, 1980; Rousseau, 1992).

Table 2
Annual scientific production by topical classification

Year	General aspects	Linguis-tics	Automatic indexing vs manual indexing	Evalua-tion	Status of the ques-tion	Automatic indexing and retrieval	Automatic indexing software	Techniques and methods	Total
2000	1	0	0	0	1	1	2	8	13
1999	11	0	1	0	1	2	0	4	19
1998	12	1	0	1	0	4	0	9	27
1997	9	2	1	0	0	9	0	5	26
1996	12	1	1	1	2	3	0	7	27
1995	6	0	0	2	0	14	1	8	31
1994	13	2	0	1	1	5	1	12	35
1993	13	3	0	0	2	4	1	16	39
1992	9	6	3	3	2	3	2	11	39
1991	12	8	0	0	1	1	0	15	37
1990	11	9	1	1	2	8	3	12	47
1989	10	1	1	0	0	4	1	5	22
1988	7	3	2	1	1	4	3	10	31
1987	4	0	0	0	0	3	3	9	19
1986	2	2	0	1	0	3	1	7	16
1985	2	1	0	1	1	0	1	8	14
1984	7	1	1	1	0	2	0	12	24
1983	3	1	0	0	0	0	1	12	17
1982	4	2	0	0	1	0	0	4	11
1981	3	3	0	0	0	1	2	7	16
1980	4	1	0	0	0	0	0	9	14
1979	3	2	1	0	0	1	0	5	12
1978	6	0	0	2	0	1	0	8	17
1977	4	1	2	0	2	0	0	8	17
1976	2	2	0	1	0	0	0	6	11
1975	8	0	0	2	2	1	0	8	21
1974	9	0	0	0	3	2	0	5	19
1973	6	1	0	1	2	1	0	9	20
1972	4	4	1	0	2	4	0	6	21
1971	12	1	0	0	1	6	1	7	28
1970	11	4	0	4	0	9	0	16	44
1969	21	2	3	2	1	9	0	11	49
1968	6	1	0	2	0	1	0	2	12
1967	5	0	0	0	0	2	0	0	7
1966	3	1	0	0	1	0	0	1	6
1965	2	0	0	0	1	1	0	3	7
1964	2	0	0	1	0	0	0	1	4
1963	3	0	0	0	1	0	0	1	5
1962	2	0	0	0	0	0	0	0	2
1961	2	1	0	0	2	0	0	0	5
1960	0	0	0	0	0	1	0	0	1
1959	0	0	0	0	0	0	0	1	1
1958	2	0	0	0	0	1	0	1	4
1957	0	0	0	0	0	0	0	1	1
1956	1	0	0	0	0	0	0	0	1
Total	269	67	18	28	33	111	23	290	839

Table 3
Author productivity

X	Y_x	$X = \lg x$	$Y = \lg y$	XX	XY	$Y_x/\sum Y_x$	$\sum(Y_x/\sum Y_x)$	f_e	$\sum f_e$	D
1	703	0	2.8469	0	0	0.7961	0.961	0.867	0.867	0.094
2	112	0.3010	2.0492	0.0906	0.6168	0.1268	0.229	0.169	0.036	0.193
3	35	0.4771	1.5440	0.2276	0.7366	0.0396	0.625	0.383	0.419	0.206
4	11	0.6020	1.0410	0.3624	0.6267	0.0124	0.749	0.173	0.592	0.157
5	8	0.6989	0.9030	0.4885	0.6311	0.0090	0.839	0.094	0.686	0.153
6	7	0.7781	0.8450	0.6055	0.6575	0.0079	0.918	0.057	0.743	0.175
7	3	0.8450	0.4771	0.7141	0.4031	0.0034	0.952	0.037	0.780	0.172
9	1	0.9542	0	0.9105	0	0.0011	0.963	0.017	0.797	0.166
10	1	1.0000	0	1.0000	0	0.0011	0.974	0.013	0.810	0.164
14	1	1.1461	0	1.3136	0	0.0011	0.9985	0.005	0.815	0.170
34	1	1.5314	0	2.3454	0	0.0011	0.9996	0.004	0.819	0.177
$\sum$	883	8.33	9.71	8.03	3.67[a]					

x: number of works. y: number of authors. $Y_x/\sum Y_x$: frequency of authors with a single work, with two, with three, etc. (the frequencies observed in the distribution of authors on automatic indexing). $\sum(Y_x/\sum Y_x)$: cumulative frequency of authors with one, two, etc. works. f_e: expected frequencies, calculated by Lotka's formula (the value of the first cell corresponds to the value of 'C'). $\sum f_e$: cumulative expected frequencies. $D = D_{max}$: differences between the columns of the observed and expected cumulative frequencies.
The value of n is calculated by the least squares method:

$$n = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$$

where N is the number of data pairs considered, X is the (decimal) logarithm of x (x = number of works), Y is the (decimal) logarithm of y (y = number of authors).
In the present case, with the data of this table, the slope will be the following:

$$n = \frac{7(3.67) - (3.7 \times 9.71)}{7(2.489) - (3.7)^2} = -2.75$$

[a] Totals excluding the data $y_x = 1$.

If x and y follow an inverse power law, the resulting (log–log) plot will be a straight line of negative slope n. The value of n is calculated by the least squares method (see Table 3).

The calculation of 'c' starts from Lotka's law, $y_x = c \times x^{-n}$ (see Appendix A). This represents the authors with a single work in the theoretical distribution, i.e., in the expected frequencies. From this datum, and applying Lotka's law $y_x = c \times x^{-n}$, one completes column 9 of Table 3.

In order to verify that the observed distribution of the productivity of the authors fits the theoretical distribution, we subjected the data to the non-parametric Kolmogorov–Smirnov test. To this end, we used the data in the last column of Table 3 (D_{max}), obtained as the absolute value of the difference between columns 8 and 10 of the same table. The greatest value of this column (D_{max}) will be taken as reference for comparison with the "critical value" (c.v.), obtained by the asymptotic formula

$$\text{c.v.} = \frac{1.63}{\left(\sum y_x + (\sum y_x/10)^{1/2}\right)^{1/2}}$$

For our case, we shall use a significance level of 0.01, so that the expression will be

$$\text{c.v.} = \frac{1.63}{\left(883 + (883/10)^{1/2}\right)^{1/2}} = 0.054$$

with $\sum y_x$ being the total number of authors (in our case 883).

The data of Table 3 give a value of $D_{\max} = 0.02(*)$, and the critical value is 0.054.

Since the value of the present distribution is smaller than the critical value ($0.02 < 0.054$), the null hypothesis that the data follow a Lotka distribution has to be accepted.

3.4. The obsolescence of the scientific literature

By "obsolescence" one understands the temporally declining utility, or use, or validity of information or measurements (Line & Sandison, 1974).

There are two possible approaches in studying obsolescence: (i) a *diachronous study*, which takes a certain moment in time as the starting point, and follows the impact that a body of literature has on the surrounding science as measured by the citations it receives in the years following publication; and (ii) a *synchronous study*, which analyses the antiquity of the references that the body of literature has cited and on which its own contribution is based.

Burton and Kebler (1960) introduced the concept of "half-life" into the field of information science, finding that the half-life of the references in the journals of various sciences depends on the topical area concerned. Brookes (1970) established the mathematical law describing the temporal loss of utility of a set of documents.

The half-life h is an indicator of obsolescence, and represents the age at which the utility (number of references or citations) has fallen by a half. We shall start with Brookes' formulation $a^h = 0.5$ to calculate the obsolescence of the "automatic indexing" literature.

The references used are from 1952 to 2000. Table 4 gives the references from 1962, leaving out the 13 references published earlier. The year with most references is 1969, with 49, followed by 1990 with 47. Separated into decades, the most productive was clearly that of the 1990s.

From Brookes' equation, with appropriate operations, one has:

$$a = e^{(\ln 0.5)/h}$$

where a is the annual aging factor.

Table 4 shows that the half-life, h, the age at which the utility is reduced by half, is between 14 and 15 years. For its exact calculation, we interpolate between columns 2 and 5 of Table 4:

$$\frac{0.509 - 0.489}{14 - 15} = \frac{0.509 - 0.500}{14 - h} = 14.47 \quad \text{years} = h$$

Substituting this value in the equation for $a = e^{(\ln 0.5)/h}$, one has

$$a = e^{(\ln 0.5)/14.78} = 0.95$$

This is an annual aging factor of 95%, or an annual loss of currency of 5%.

Considering either the value calculated for h or the annual aging factor, one observes that this literature is very stable from the perspective of its use, since a half-life of 14.47 years means that it takes about 15 years for the utility of this literature to be reduced by 50%.

Table 4
The obsolescence of the scientific literature

Years	Age, *t*	References/year	Cumulated references	Utility
2000	0	13	839	1.000
1999	1	19	826	0.984
1998	2	27	807	0.961
1997	3	26	780	0.929
1996	4	27	754	0.898
1995	5	31	727	0.866
1994	6	35	696	0.829
1993	7	39	661	0.788
1992	8	39	622	0.741
1991	9	37	583	0.695
1990	10	47	546	0.650
1989	11	22	499	0.595
1988	12	31	477	0.568
1987	13	19	446	0.531
1986	14	16	427	0.509
1985	15	14	411	0.489
1984	16	24	397	0.473
1983	17	17	373	0.444
1982	18	11	356	0.424
1981	19	16	345	0.411
1980	20	14	329	0.392
1979	21	12	315	0.375
1978	22	17	303	0.360
1977	23	17	286	0.340
1976	24	11	269	0.319
1975	25	21	258	0.307
1974	26	19	237	0.282
1973	27	20	218	0.258
1972	28	21	198	0.234
1971	29	28	177	0.210
1970	30	44	149	0.177
1969	31	49	105	0.124
1968	32	12	56	0.066
1967	33	7	44	0.052
1966	34	6	37	0.043
1965	35	7	31	0.037
1964	36	4	24	0.028
1963	37	5	20	0.024
1962	38	2	15	0.018

3.5. Dispersion of the scientific literature (Bradford's law)

One of the milestones in the development of bibliometrics and information science was the evidence for the regularity in the distribution of scientific journals, known as Bradford's law (Bradford, 1934, 1948).

Bradford found that, on dividing the journals into three zones, each with the same number of articles, the number of journals in each zone grew geometrically. Also, the distribution of journals according to their productivity presented a different model of concentration and dispersion when it was represented as a statistical distribution, with a larger group forming a long tail of less productive journals.

Evidently the implications of these findings went beyond the mere description of the dispersion of the scientific literature in journals. Aspects as wide-reaching as the principles by which the scientific community functions as a stable integrated system, or the universality of the application of Bradford's formulation derive from this regularity.

In presenting the formulation of his empirical law of the literature in scientific journals, Bradford included a graph as illustration. Along the x-axis, he placed the journals $1, 2, 3, \ldots, r$ in decreasing order of productivity of relevant works on a given topic, using a logarithmic scale. The y-axis represented the accumulated total of publications $R(r)$. The resulting semi-log plot began with a curve which, beyond a critical point, became a straight line.

Subsequently, Leimkuhler (1967), Brookes (1969), Rousseau and Leimkuhler (1987), Egghe (1990) and Rousseau (1994) have given mathematical expressions for Bradford's law. Egghe's method is based on the earlier formulation of Leimkuhler:

$$R(r) = a \log_e(1 + br)$$

Using the following notation:

r_0 number of journals in Bradford's first group
y_0 number of articles in each group (all groups are of the same size)
k Bradford's multiplier
Y_m number of articles in the most productive journal (rank 1)
$R(r)$ cumulative number of articles produced by the journals of rank $1, 2, 3, \ldots, r$
a and b constants that appear in Leimkuhler's formulation. Egghe showed that: $a = y_0/\log_e k$, for the present case $a = 132/1.91 = 110.8$, $b = (k-1)/r_0$, for the present case $b = 2.29/3.60 = 0.64$
P number of Bradford groups, and also determined k: $k = (e^{\gamma} \times Y_m)^{1/P}$
γ Euler's number $= 0.5772$; $e^{\gamma} = 1.781$; $k = (1.781 \times Y_m)^{1/P}$
A number of articles in the literature: $y_0 = A/P$
T total number of journals: $r_0 = T(k-1)/(k^P - 1)$

To calculate $r_1, r_2, r_3, \ldots$, one uses the exact value for r_0 and, with rounding off, that of k.

$$r_0 \times 1 = r_0; \quad r_0 \times k = r_1; \quad r_0 \times k^2 = r_2; \quad r_0 \times k^3 = r_3$$

With the data of Table 5, and using Egghe's procedure, we obtain the following results:

We chose $P = 4$ as being the ideal number of zones for the distribution, since the Bradford multiplier, k, is very similar in the three zones, and does not satisfy other values of $P \neq 4$, so that

$$k = (1.781 \times 66)^{1/4} = 3.29 \quad \text{and} \quad r_0 = \frac{183 \times 2.29}{117.2 - 1} = 3.6$$

Table 5
The dispersion of the scientific literature

No. of journals	No. of articles	Cumulative journals	Cumulative articles	*Ln* (cumulative journals)
1	66	1	66	0
1	40	2	106	0.6931
1	36	3	142	1.0986
1	28	4	170	1.3862
1	21	5	191	1.6094
1	18	6	209	1.7917
1	10	7	219	1.9459
1	8	8	227	2.0794
2	7	10	241	2.3025
4	6	14	265	2.6390
5	5	19	290	2.9444
6	4	25	314	3.2188
14	3	39	356	3.6635
27	2	66	410	4.1896
117	1	183	527	5.2094

Bradford's distribution

Zones	Number of journals	Number of articles	*k*
Core	4	170	–
Zone 1	12	105	3.00
Zone 2	39	113	3.25
Zone 3	128	139	3.28

The similarity of the different values of k, and between these and the calculated Bradford multiplier (3.29), clearly shows that the distribution fits a four-zone Bradford law.

The equation for the Bradford curve in the present case is:

$$R(r) = 110.8 \log_e(1 + 0.64r)$$

4. Conclusions

The main goal of indexing is the storage and retrieval of information. The interest in this process that arose in the 1970s is a consequence of the need to constantly increase the functionality of information systems in order to satisfy the demand from the scientific community, due to the exponential increase in scientific information that occurred in the 1960s.

Interest in the indexing process has given rise to a body of scientific literature that is sufficiently voluminous to merit analysis. The goal of the present study was to determine the structure of this body of literature, its volume and evolution, the subject areas which the documents cover, their dispersion with respect to the sources in which they were published, the distribution of the authors' productivity, the obsolescence of the literature, etc.

Automating the task of indexing has been a theme of constant interest for researchers from the 1950s until the present day, as this study of more than 800 research works published between 1956 and 2000 has shown. In the last five-year period, however, there has been a considerable decline in scientific output with respect to the previous five-year periods, perhaps because of the lag in updating some of the databases. The scientific output corresponding to "Ph.D. theses", however, has not fallen. If it were not for this failure to update the databases, there would probably be even more theses today because of the constant interest on the part of authors in researching into the indexing process. This would imply that there is still a demand from the scientific community for the creation of information systems designed to overcome the problems involved in the storage and retrieval of information. Furthermore, despite the ever-greater number of studies, there is still no consensus on methodology, so that this is still an open door for investigation.

With respect to scientific output by topic, we divided the material into 8 sections. The aim is to determine the number of documents on each topic and thereby be able to analyse the possible significance, motives, or consequences of why some investigate more than others. The most productive was that of "techniques and methods", followed by "general aspects". There were certain variations observed within the topics in some cases. The fact that the most investigated topics are those corresponding to such aspects as "techniques and methods", "automatic indexing and retrieval" or "general aspects", indicates where the research is being concentrated. This is logical since these topics are of the greatest current interest due to the requirements alluded to above of the scientific community. Together, the document output on these three topics is almost 80% of the total scientific output.

The author distribution by productivity was examined following the approach of Pao (1985). The result of applying the Kolmogorov–Smirnov goodness-of-fit test was that it did fit a Lotka distribution ($D_{\max} = 0.02$ vs critical value $= 0.054$). We obtained a steep slope ($n = -2.75$) for the author distribution. This means that, according to the data of Table 3, there exist a great many unproductive and few highly productive authors. In other words, a great proportion of the authors have studied this subject only occasionally or temporarily. This is confirmed not only by the data in Table 3 but also by the value of $C = 0.7867$, i.e., that more than 78% of the authors have only one published work.

We applied the concept of half-life to calculate the obsolescence of the scientific literature, finding an annual aging factor of 95%, i.e., a 5% annual loss of currency. In terms of obsolescence, this percentage means that this literature has a low level of aging, with a half-life close to 15 years (meaning that the use of this literature is reduced by 50% every 15 years).

Finally, the dispersion of the bibliographic corpus that was analysed was found to be low, given the number of journals in which the articles were published. There were 527 articles published in 183 journals. The most productive journal published 66 articles, while 117 journals published only a single article. There therefore exists a concentration of articles in a small number of journals. The journals of the core and of zone 1 (16 journals altogether) accounted for more than 50% of the articles (275).

In sum, our analysis of the scientific literature on "automatic indexing" showed there to exist a high level of scientific output (around 1000 documents) with a broad diversity in both subject investigated and in document type according to the sources in which the documents were published. The distribution of authors showed a major concentration of unproductive authors and very few highly productive authors, indicating an only occasional dedication to the field. The

literature used by the authors presents a low level of obsolescence, since the calculated loss of currency was 5% p.a., meaning that research on indexing is very stable and uses references published over a long period of time. The dispersion of the scientific literature was also found to be low, indicating a concentration of sources in which the documents were published.

There is still at least one door open to research, as is shown by the increase in the number of doctoral theses over the last years, despite the lack of currency of some databases. This implies that there is still a great intensity of research into indexing today.

Appendix A

The calculation of 'c' starts from Lotka's law, $y_x = c \times x^{-n}$.

Dividing both terms by $\sum y_x$, the number of authors, $y_x / \sum y_x = (c / \sum y_x)(1/x^n)$ and writing $c / \sum y_x = C$, the fraction of the total sample of authors, one has $y_x / \sum y_x = C(1/x^n)$, and hence $\sum y_x / \sum y_x = C \sum 1/x^n = 1$. Finally $C = 1/(\sum 1/x^n)$.

For fractional non-negative values of n, the sum of the series in its general form $\sum 1/x^n$ can be approximated by a function that calculates the sum of the first P terms. The result, according to Pao (1985), is due to Professor David Singer

$$\sum_{x=1}^{\infty} \frac{1}{x^n} = \left[\sum_{x=1}^{P-1} \frac{1}{x^n} + \frac{1}{(n-1)(P^{n-1})} + \frac{1}{2P^n} + \frac{n}{24(P-1)^{n+1}} \right]$$

For the present case, using the slope 'n' calculated before, 2.75, one has

$$\sum_{x=1}^{\infty} \frac{1}{x^{2.75}} = \left[\sum_{x=1}^{19} \frac{1}{x^{2.75}} + \frac{1}{(2.75-1)(P^{1.75})} + \frac{1}{2P^{2.75}} + \frac{2.75}{24(P-1)^{3.75}} \right] = 1.271$$

and $C = 1/1.271 = 0.7867$.

References

Anderson, J. D., & Pérez-Carballo, J. (2001a). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing and Management, 37*, 231–254.

Anderson, J. D., & Pérez-Carballo, J. (2001b). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort. *Information Processing and Management, 37*, 255–277.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering, 23*(3), 85–88.

Bradford, S. C. (1948). *Documentation*. London: Crosby, Lockwood Sons Ltd.

Brookes, B. C. (1969). Bradford's law and the bibliography of science. *Nature, 22*(5523), 953–956.

Brookes, B. C. (1970). Obsolescence of special library periodicals: sampling errors and utility contours. *Journal of the American Society for Information Science, 21*(5), 320–329.

Burton, E., & Kebler, R. W. (1960). The "half-life" of some scientific and technical literatures. *American Documentation, 11*(1), 18–22.

Egghe, L. (1990). Applications of the theory of Bradford's law to the calculation of Leimkuhler's law and to the completion of bibliographies. *Journal of the American Society for Information Science, 41*(7), 469–492.

Egghe, L., & Rosseau, R. (1990). *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier.

Gil-Leiva, I. (1999). *La automatización de la indización*. Gijón, Spain: Trea.

Leimkuhler, F. F. (1967). The Bradford distribution. *Journal of Documentation, 23*, 197–207.

Lindsey, D. (1980). Production and citation measures in the sociology of science: the problem of multiple authorship. *Social Studies of Science, 10*, 145–162.

Line, M. B., & Sandison, A. (1974). Obsolescence and changes in the use of the literature with time. *Journal of Documentation, 30*(3), 283–350.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*(12), 317–323.

Nicholls, P. T. (1986). Empirical validation of Lotka's law. *Information Processing and Management, 22*(5), 417–419.

Pao, M. L. (1985). Lotka's law: a testing procedure. *Information Processing and Management, 21*(4), 305–320.

Rousseau, R. (1992). Breakdown of the robustness property of Lotka's law: the case of adjusted counts for multiauthorship attribution. *Journal of the American Society for Information Science, 43*(10), 645–647.

Rousseau, R. (1994). Bradford curves. *Information Processing and Management, 30*, 267–277.

Rousseau, R., & Leimkuhler, F. F. (1987). The nuclear zone of a Leimkuhler curve. *Journal of Documentation, 43*(4), 322–333.