Information Transfer Limitations of Titles of Chemical Documents*

ROBERT T. BOTTLE† and CATHERINE R. SEELEY
Syracuse University, School of Library Science, Syracuse, N. Y.

Received April 1, 1970

Some methods of estimating the minimum amounts of information in a document not retrievable through its title are discussed. An analysis of the information transferred by different types of keywords is helpful in planning search strategies, e.g., 30% of chemical substances mentioned in journal articles are not discernable in their titles even when broad class names are used as synonyms. Patents have considerably less informative titles than journal articles. In nuclear science, report titles are also less informative than those of journal articles, but the proportion of reports with completely uninformative titles is now only 10% of the 1957 value. Titles in chemistry are more informative than those in most other fields, but the use of alerting and other services based on titles requires a good understanding of the underlying information transfer principles.

The title of a document or paper is probably the first thing about that document or paper that really registers in our minds—unless one of the authors has an exceptionally well-known name. This is obviously true when we scan a list of titles that are separated from the remainder of a document as in the contents pages of a journal or in a titles index. On the basis of the title, often in conjunction with an author's name, we frequently make decisions on whether or not to make further reference to the document—such as reading its abstract or skimming through the full paper. Information transfer in a positive sense will occur when the title contains words or word combinations that alert us to the possibility that the document may contain information of interest to us.

Our rejection of titles on the grounds of lack of relevance or interest constitutes a filter through which passes the flood of documents that we do not require. Thus for the majority of documents, the titles of which we scan, information transfer in a negative sense occurs through their titles. How reliable is this process? How often will a title fail to transfer information contained in a document that would be of interest if one but knew of it? The answers to these questions are important factors affecting the decisions one makes about new research directions. Since the interception of research reports by competitors will frequently inhibit one from pursuing a particular line, one needs to be able to assess the probability that one has failed to intercept relevant information. Data from which to make such estimates is not readily available, but this paper describes a simple method for obtaining such data for title indexes.

Several methods are available for assessing the average amount of information in a document, which is not deducible from its title, other than by making a detailed analysis of a large number of documents. They involve using a collection of documents whose contents have already been analyzed. (1) The proportion of annotations providing additional scientific (as distinct from bibliographical, geographical, etc.) information in an annotated bibliography such as the *Bibliography of Agriculture* can be estimated directly. (2) In an indexing service such as *Index Medicus* where titles are arranged under standard headings, the proportion of titles having little or no relation to their specific headings can be deduced. (3) There are several abstracting or indexing services that select for inclusion only papers containing information relevant to their often quite narrow field, e.g., *Mass Spectrometry Bulletin*. The proportion of titles in these that do not suggest the selection criteria can be estimated directly. (4) Randomly selected subject index entries from an informative abstracting service can be compared with the titles of the corresponding abstracts. This last method is described in more detail later. The first three methods are normally quicker to use, and results for several services are shown in Table I. Method 2 is essentially the one used by Montgomery and Swanson[3] and by O'Connor.[4] It was called "synonym inclusion" for a subject heading–title pair by the latter who critically reviewed the former's experiment.[4]

The figures given in Table I represent minimum values for the information that is irrecoverable from titles. Because of considerable differences in indexing quality across the document collections surveyed in Table I, no conclusions about the variation of information content of titles with subject can be made. Method 3, services which select papers on the basis of a particular narrow topic being present, probably represents the greatest depth of indexing. Here 52% of the document titles noted in the December 1968 *Mass Spectrometry Bulletin* contained

* Presented at the 5th Middle Atlantic Regional Meeting, ACS, Newark, Delaware, April 1, 1970.

† Present address, School of Chemistry, University of Bradford, Bradford 7, England.

Table I.

Index	% Information not retrievable from Title: Annotations	Standard Subject Headings	Subject Area of Index
Bibliography of Agriculture, 1969	22[1]		
Mass Spectrometry Bulletin, 1969	51[2]		52
Engineering Index, 1967		29[2]	
Applied Science & Technology Index, 1967		42[2]	
British Technology Index, 1967		38[2]	
Index Medicus, 1967		26[3]	
Index Handbook of Cardiovascular Agents, **2**, 1951–55		30*	25*

* Determined from O'Connor's samples[4]. See text for details.

Table II. List of Categories

Category	
1	Substances, specific compounds and elements
2	Processes and reactions (nonbiological), including apparatus and equipment
3	Biological activity, systems and conditions
4	Properties including their measurement
5	Theoretical aspects
6	Miscellaneous

no reference to mass spectrometry or to a number of related terms.

Figures for the *Index Handbook of Cardiovascular Data* were determined from the sample used by O'Connor[4]. For Method 2 O'Connor's figure of 32% "synonym inclusion" was augmented by a further 38% for those titles from which one could infer the subject heading when the complete title was scanned. (In Tables III and IV, "b" terms include related terms that are not strictly synonyms. Indeed a few may be antonyms.[2]) Thus the title irretrievable information was 30% (= 100 − 32 − 38). (The standard deviation σ, cf. Tables III and IV, is ±7% for this sample.)

The most useful collection of chemical documents that have been analyzed for information content is *CA* and its subject indexes, well known for its depth of indexing which goes far beyond the keywords of the title. Index entries were selected at random from the *CA* Subject Index to Vol. 66 (1967) and compared with the titles of the appropriate abstracts. (The alternative procedure used by an erstwhile CAS employee[6] of examining all the subject index entries for randomly selected papers was not available to us.) The index entries are long and complex, and it was quickly noticed that very few did not contain at least one keyword which was the same as, or synonymous with, a keyword in the title, but frequently they related to the least important part of the index entry, e.g., *CA* 66:105225t, "Aliphatic heterocyclic polymers;" Index entry: 3,3′Biphenyldithiol, 4,4′diamino, polymer with sebacic acid. While this could be retrieved through the keyword 'polymer'—a very broad term indeed—anyone interested in the reactions, uses, etc. of sebacic acid or thiols could not retrieve this paper via its title.

It was therefore necessary to split the index entry into keyword groups (conceptual entities) before comparing them with the keyword groups in the title. For example the above index entry is split into three groups, viz: 4,4′diamino 3,3′biphenyldithiol, polymer, and sebacic acid.

A number of different categories of keywords can be distinguished, such as those which are the names of compounds or substances (as are all three in the above example), biological systems, properties, etc. As it was observed that some categories—e.g., compounds—were less likely to be retrieved from titles than others, the keyword groups from the index entry were therefore divided into the six categories shown in Table II. The keyword groups from the index entry were then matched against the title of the corresponding abstract, and a judgment was made as to whether they were (with respect to the title keywords) (a) identical or syntactical variants (O'Connor[4] calls these 'inflectional variants'.), (b) synonyms or other related terms, or (c) neither (a) nor (b). Class (c) obviously represents those indexable concepts that are irretrievable from a title index. The results of this analysis are shown in Tables III and IV. Table III refers to the data obtained from Vol. **66** (1967) of *CA* and Table IV to that from Vol. **21** (1967) of *Nuclear Science Abstracts*. Because it was observed that their titles were less informative than those of journal articles, patents are shown separately in Table III, and reports are shown separately in Table IV. (The proportion of reports in Vol. **66** of *CA* was less than 5%, and these are included in with the journal articles, etc.)

In spite of the small sample of documents represented in Tables III and IV (approximately 0.02%), it is nevertheless possible to estimate the sampling error. Since the

Table III. 171 Titles from *Chemical Abstracts*, Vol. 66, 1967

Category	143 Journal articles etc. Number of Concepts: Identical etc. (a)	Synonyms etc. (b)	Neither (a) nor (b), i.e. (c)	% c (irretrievable concepts)	Standard Deviation σ (%)	28 Patents Number of Concepts: Identical etc. (a)	Synonyms etc. (b)	Neither (a) nor (b), i.e. (c)	% c (irretrievable concepts)	Standard Deviation σ(%)
1	45	35	34	30	4	6	15	19	48	8
2	16	16	5	14	6	8	4	1	13	9
3	24	22	0	(0)	...	0	0	3	(100)	...
4	17	10	5	16	6	1	0	1	(50)	...
5	9	6	1	(6)	6	0	0	0	(0)	...
6	7	3	0	(0)	...	6	1	4	36	15
1–6	118	92	45	17.6	2.4	21	20	28	40.6	5.9
Over-all 1–6	139	112	73	22.6	2.3	...	...	...	...	...

Table IV. 100 Titles from *Nuclear Science Abstracts*, Vol. 21, 1967

	74 Journal articles etc. Number of Concepts					26 Reports Number of Concepts				
Category	Identical etc. (a)	Synonyms etc. (b)	Neither (a) nor (b), i.e. (c)	% c (irretrievable concepts)	Standard Deviation σ (%)	Identical etc. (a)	Synonyms etc. (b)	Neither (a) nor (b), i.e. (c)	% c (irretrievable concepts	Standard Deviation σ (%)
1	43	10	23	30	5	6	2	15	65	10
2	47	7	21	28	5	15	9	7	57	8
3	17	12	11	28	7	1	2	1	(25)	22
4	26	5	3	9	5	9	1	9	47	11
5	6	0	4	40	16	1	0	1	(50)	...
6	0	0	0	0	...	0	0	0	0	...
1–6	139	34	62	26.4	2.9	32	14	33	41.8	5.5
Over-all 1–6	171	48	95	30.2	2.6	...	...	...	...	...

documents were randomly selected and because classes (a) and (b) together represent those concepts which are theoretically retrievable from the title when all possible synonyms etc. are used, and class (c) represents those which are irretrievable, the system is analogous to quality control sampling. The standard deviation $\sigma = [C(100 - C)/n]^{1/2}$ where C is the % of defective items found in a random sample of size n. Values of σ are given in the final column of each section of Tables III and IV. A few of the figures in the "% c, title irretrievable" columns of Tables III and IV appear in parentheses. As can be seen from the preceding columns, they are based on very small numbers and/or where $\pm 2\sigma$ would take "% c" outside the limits of 0 – 100. They are therefore probably not very reliable.

An additional check on the sampling procedure was obtained by comparing the percentage of patents in the sample from *CA* with the over-all percentage of patents for that volume. The sample contained 16.35% patents compared with 16.4% for Vol. 66.[7]

Table III clearly shows that patents have much less informative titles than journal articles. (This was not, however, an unexpected result as dissemination of information is not the prime function of a patent.) Because of the large number of distinct compounds which may be mentioned in a single paper, one would perhaps expect that the names of specific substances would be difficult to retrieve from titles, and this category had indeed almost twice the percentage of irretrievable items recorded for the average of journal articles, etc. from *CA*. What perhaps is remarkable, is that such a high proportion (70%) of substances could theoretically be retrieved from titles if all possible synonyms and related terms were used. [These related terms would include nomenclature fragments which corresponded to a substantial portion of the molecule. For example, in the case where the index entry referred to '2-thiophene carboxylic acid, 3-*tert* butoxy ethyl ester' the occurrence in the title of the words "acetyl and carbethoxy chelated 2-hydroxythiophenes" (*CA* 66:75499r) was sufficient for this index term to be placed in class (b). If the words "carboxylic acid," "ethyl ester," etc. had occurred in the title, these would not have been considered sufficient grounds for placing it in class (b)]. Table III also indicates the effect of ignoring related terms even when the most widely used term is considered. If only this term is used in a search of a titles index the proportion of relevant articles retrieved is reduced from 70 to 40% in the case of specific substances and from 82 to 46% for all categories. The equivalent reduction in the over-all retrieval of nonreport literature from 74 to 59% calculated from the data of Table IV is not quite so great, probably reflecting a greater tendency on the part of the indexers of *Nuclear Science Abstracts* to utilize titles more frequently in constructing the (secondary) indexing terms.

Nuclear Science Abstracts was used as a source of analyzed documents because it was known to contain a substantial proportion of report literature, 26% of the sample used to produce Table IV. (Of the 102 documents randomly selected from the Subject Index of Vol. 21, two were patents and these were discarded from the sample since Table III clearly indicated they should not be considered along with the journal articles etc.) Table IV shows just how much more informative are journal article titles than those of reports. Since some reports had titles such as "Quarterly Progress Report" or "Organic Chemistry" etc., which are completely uninformative, the report titles analysis of Table IV was re-examined, and it was found that 2 of the 26 were of this type. If these were eliminated from the sample, the percentage *c* figures are reduced from 65 to 62 for category 1, from 23 to 20 for category 2, from 47 to 41 category 4, and from 41.8 to 37.9 for the over-all average for reports. Thus even "informative" report titles are less informative than the titles of journal articles. It was, however, observed that the proportion of reports with completely uninformative titles in 1967 was very much lower (at 8%) than in a random sample taken from *Nuclear Science Abstracts* 1957 where it was 76%. (The 1957 sample of 100 items consisted of 33 reports of which only 8 had titles which were at all informative.) Reports are generally longer than journal articles and encompass a greater variety of topics within the single report than would a journal article, and this may account, at least in part, for their less informative titles. But whatever the reason, it does not seem likely that an unenriched title index of report literature (at least for nuclear science) would be a useful information retrieval tool.

The analyses presented in Tables III and IV imply that certain search strategies will be more successful than others. For example, if one requires information on a property (i.e., Category 4) of a specific compound (i.e., Category 1) then one's chance of failing to retrieve information is about halved if one searches a title index for the property and its related terms than if one searches for the specific compound and its synonyms. This argu-

ment is just as applicable to manual as to computer searches. The probability of retrieval of information in computer searches involving Boolean *AND* logic will, of course, be the product of the probabilities for each of the parameters involved. Since the probability of retrieving information is (100 – % retrieval failure), Table III predicts the probability of retrieving information about a specific compound as 70%, i.e. (100 – % c) and that for a property as 84%. Thus the probability of retrieving information for the property *AND* the compound is (70 × 84)/100 = 59%, i.e., the probability of failure to retrieve information in such a search is (100 – 59) = 41%.

It must be stressed that the method described in this paper will give a reliable estimate of the minimum amount of relevant material which one will fail to retrieve from an index of unenriched titles, but, because of differing depths of indexing in the services serving differing subject areas, it is unfortunately not possible to compare the information content of titles in different subject areas with a great degree of confidence. Having carried out similar analyses in several fields, our impression is that the titles of chemical papers are much more informative than those in other fields, especially the social sciences.[8,9]

It is of interest to compare the results of this analysis with that made by Ruhl[1] on the 1960 *CA*. She was able to use the CAS Indexing sheets and so used all the index entries for each sample of 84 titles (all which had appeared in *Chemical Titles*). Since she classed these titles according to whether 0, 1, 2, 3, or more indexed "concepts" were omitted, an exact comparison with the present work is not possible. Since the "average number of *CA* Subject Index entries for this sample was four," the total number of "concepts" was 4 × 84, and it can be calculated from her data that the number of "concepts" omitted was 68 or more, the proportion of "concepts" irretrievable from titles (% c) is approximately (68 × 100)/(84 × 4) = 21%. The figures 4 and 68 in this calculation are subject to considerable uncertainty, but by making plausible assumptions about their upper and lower limits one can calculate that the probable maximum value would be 27% and the probable minimum value would be 19%. This lower value is, of course, within the limits of error of the over-all value of 17.6% c recorded for journal articles in Table III. Ruhl's analysis does not, however, provide any information on which concept categories are more likely than others to be omitted from a title.

Analyses such as this paper describes are relatively simple to perform and can provide quantitative data from which to make decisions on search strategies and the research directions suggested by the searches.

REFERENCES

(1) Bottle, R. T., "Title Indexes as Alerting Service in the Chemical and Life Sciences," *J. Amer. Soc. Inf. Sci.* **21**, 16–21 (1970).

(2) Bottle, R. T., "Information Content of Titles in Engineering Literature," IEEE *Trans., Eng. Writing & Speech*, in press.

(3) Bottle, R. T., and L. Murray, unpublished data.

(4) O'Connor, J., "Corelation of Indexing Headings and Title Words in Three Medical Indexing Systems," *Amer. Doc.* **15**, 96–106 (1964).

(5) Montgomery, C., and Swanson, D. R., "Machinelike Indexing by People," *Amer. Doc.* **13**, 359–366 (1962).

(6) Ruhl, M. J., "Chemical Documents & Their Titles: Human Concept Indexing *vs.* KWIC-Machne Indexing," *Amer. Doc.* **15**, 136–141 (1964).

(7) Terrant, S. W., personal communication.

(8) Preibish, C. I., "Information Transfer through the Titles of Psychological Literature," paper presented to S.L.A. Upstate N. Y. Chapter meeting, Syracuse, N. Y., May 24, 1970; or Bottle, R. T. and C. I. Preibish, "The Proposed KWIC Index for Psychology: an Experimental Test of its Effectiveness," *J. Amer. Inf. Sci.*, in press.

(9) Brodie, N. E., "Evaluation of a KWIC Index for Library Literature," *J. Amer. Soc. Inf. Sci.* **21**, 22–28 (1970).