Indexing from Abstracts of Documents

GUS J. CARAS
State Services Section, Pesticides Program, National
Communicable Disease Center, Atlanta, Ga.

Received July 27, 1967

Two types of abstracts, informative and indicative, as well as entire documents, were investigated as to their content of significant words for the selection of index terms. The results indicate that abstracts compare favorably with entire documents as sources of index terms. In the case of informative abstracts, approximately 71% of the terms selected from documents were also contained in their abstracts.

Automatic indexing, a term usually applied to machine extraction of index terms from text, has been the subject of considerable study in the past decade. The primary aim in automatic indexing is to derive index terms directly from the text with a minimum of human intervention. The text may consist of the entire document, the abstract, the title, or any combination of these.

Computer indexing has been proposed not only in an effort to cope with the continuously increasing volume of material that must be indexed, but also to overcome some of the inadequacies of human indexing. Studies of the consistency of human indexing indicate that the agreement among a group of indexers who index the same document (inter-indexer consistency), and even the agreement within the same indexer who indexes the same document at different times (intra-indexer consistency), are

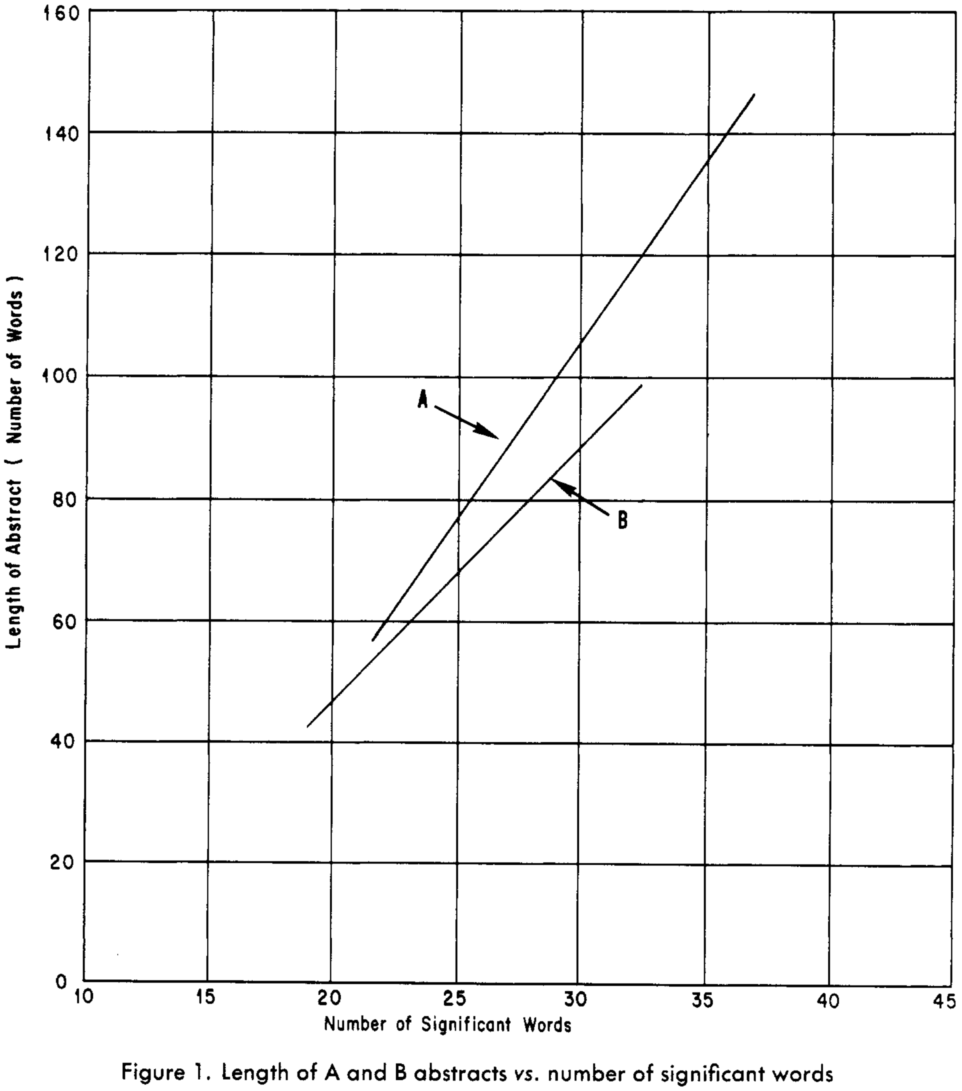

Figure 1. Length of A and B abstracts *vs.* number of significant words

rather low. Zunde (1), for example, reports inter-indexer consistency coefficients ranging from 0.158 for six indexers to 0.453 for two indexers, and an intra-indexer consistency coefficient of 0.661. These values are typical of those found in the literature.

One type of automatic indexing which uses the titles of documents as a source of index terms is the keyword-in-context (KWIC). Although this type of indexing has been used extensively for current awareness purposes, it is not considered suitable for retrospective searching. As a compromise between the title, which does not normally contain sufficient terms, and the full text, which contains too many, abstracts of documents have been considered as a source of index terms. Slamecka and Zunde (2), in an experiment using documents and their abstracts from NASA's *Scientific and Technical Aerospace Reports,* found that 80.4% of the index terms assigned by human indexers were also contained in the abstracts.

Assuming that abstracts are suitable sources of index terms, how does one select which abstract to use, particularly when several abstracts of the same document are available? Are indicative abstracts of the type generally found in *The Engineering Index*, for example, suitable, or is it necessary for the abstracts to be of the informative type, as those found in *Chemical Abstracts*, in order to be used successfully for indexing purposes? The purpose of this study was to compare the indicative abstract with the informative abstract as a source of index terms.

EXPERIMENTAL PROCEDURE

Because there is not a clear distinction between indicative and informative abstracts, it was decided to consider the abstracts from *The Engineering Index* indicative and those from *International Aerospace Abstracts* informative. One hundred and ninety-nine abstracts were randomly taken from several volumes of the *International Aerospace Abstracts* (designated as the A abstracts) and were paired with abstracts for the same set of documents from *The Engineering Index* (designated as the B abstracts). A pair of abstracts, thus, consisted of one A and one B abstract, both of which referred to the

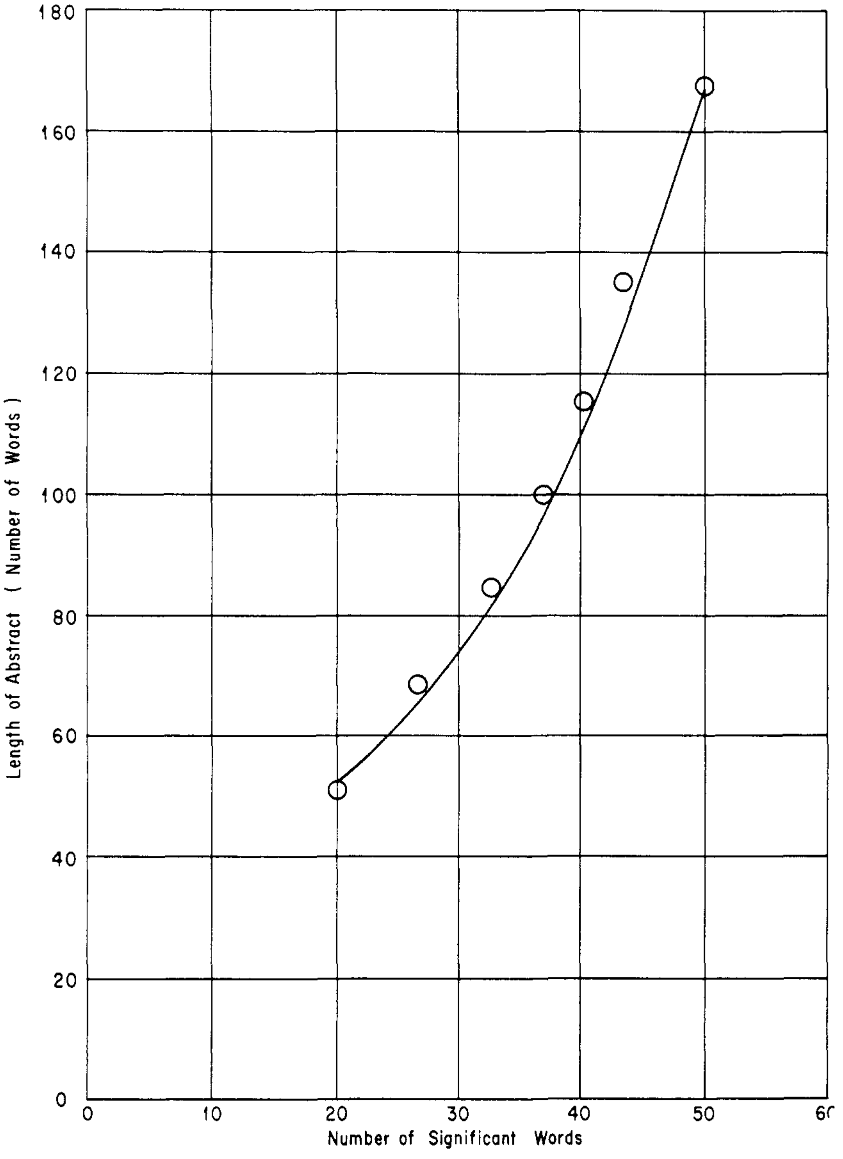

Figure 2. Length of abstracts *vs.* number of significant words

same document. The abstracts, including the titles, were keypunched so that comparisons could be made by a computer.

All of the words in each abstract were counted to determine its length, but all other comparisons were made using only the significant words of each abstract. The term significant words, as used here, refers to those words which survived the following operations:

> Common words, such as conjunctions, prepositions, articles, were eliminated.
>
> Words possessing little discriminatory power were eliminated. Many adjectives and nouns such as *method, study, problem,* which are not used for indexing purposes were included in this group.
>
> Multiple occurrences of the same significant word in an abstract were ignored; each significant word thus was counted only once in any given abstract.

The following variables were measured for each pair of abstracts:

a. The length (total number of words) of the A abstract.
b. The length of the B abstract.
c. Number of significant words in the A abstract.
d. Number of significant words in the B abstract.
e. Number of significant words which appear in both (A and B) abstracts.
f. Number of unique significant words in the pair.
g. Number of significant words in the title.

In addition to the variables listed above, the 199 documents were indexed by human indexers and the index terms assigned by these indexers were compared with both the A and B abstracts for agreement.

RESULTS

The results of the comparison of abstracts are listed in Table I. There were 167 pairs in which the A abstract was longer than the B, and in 32 pairs the A abstract was shorter. The percentages of significant words in the abstracts were 28.5 for the A abstracts and 36.8 for the B abstracts.

Table I. Summary of Average Values of Variables

Variable	Value	Range
a	111.5	37–235
b	71.5	24–111
c	31.7	10–71
d	26.3	11–46
e	17.9	5–39
f	40.1	12–81
g	4.8	1–11

The number of significant words in both types of abstracts was found to increase linearly as the length of abstracts increased (Figure 1). Figure 2 gives the number of significant words as a function of the abstract size without regard to the abstract journal from which the abstracts were taken.

The number of significant words that appeared in both abstracts represents the index terms that would be derived regardless of whether the A or B abstracts were used as a source of terms. The number of unique significant words represents the index terms that would be derived if both types of abstracts were used as sources of terms.

The average number of significant words found in the titles (4.8) is only 15.1% of the significant words of the A abstracts and 18.3% of the B abstracts. Even the inclusion of the indicative abstracts would result in approximately 5.5 times as many index terms as those found in the titles alone.

When the documents themselves rather than their abstracts were indexed by human indexers, the average number of index terms assigned was 22.3 per document. In comparing the terms derived from the abstracts with those assigned by the indexers, it was found that 71.3% of the terms assigned by the indexers were also contained in the A abstracts and 52.6% in the B abstracts. These values are somewhat lower than the 80.4% overlap between documents and their abstracts in NASA's *Scientific and Technical Aerospace Reports* reported by Slamecka and Zunde.

CONCLUSIONS

There is a fairly good agreement between the index terms contained in the two types of abstracts as 18 terms were common to both abstracts. Since 71.3% of the index terms of a document were contained in its informative abstract (*vs.* 52.6% in the indicative abstract), the informative abstract should be preferred as a substitute for the entire document. Over half of the terms found in a document are likely to be contained even in a short, indicative abstract. This type of abstract, therefore, may be an acceptable source of terms for some indexing applications. Obviously, the decision to use abstracts rather than entire documents, and if so which type of abstract, must depend on considerations such as the indexing depth desired, the availability of abstracts and the relative cost of converting to machine readable form.

LITERATURE CITED

(1) Zunde, Pranas, "Automatic Indexing from Machine Readable Abstracts of Scientific Documents," Report **AFOSR 65-1425,** Office of Aerospace Research, U. S. Air Force, Washington, D. C., 1965.

(2) Slamecka, V. and P. Zunde, "Automatic Indexing from Textual Condensations," *Automation and Scientific Communication,* Short Papers, American Documentation Institute, Washington, D. C., 1963.