F. W. LANCASTER

Evaluation of Systems by Comparison Testing

This paper contends that the retrieval abilities of four index languages studied in the Cranfield Project are comparable, although many of their respective characteristics differ considerable one from another. The ability of a system to retrieve a high percentage of documents may not, in itself, be meaningful; the total expenditure of effort must also be taken into account. In the case of the Cranfield Project, the four systems, utilizing a common conceptual analysis and given identical entry vocabularies, would have achieved identical recall performance for any given group of requests.

PHYLLIS RICHMOND'S recent article "Systems Evaluation by Comparison Testing" criticizes the early study by the Cranfield Project of the comparative ability of four index languages to retrieve known relevant documents (*i.e.*, the *recall* powers of these index languages). Her main point of criticism is that the test program compared unlike things: that the Universal Decimal Classification, the special faceted classification, the scheme of alphabetical subject headings, and the system of Uniterms were not equally applicable to handle the subject matter of the test collection, namely aeronautics. This argument I believe to be ill-founded.

In the view of Mrs. Richmond, use of the UDC and of alphabetical subject headings represents a "dilute approach" to the indexing of aeronautics documents, whereas the "concentrated" approach is provided by the special faceted classification devised for the Cranfield Project and by the use of Uniterms extracted from the document texts.

Mr. Lancaster is with Herner and Company, Washington, D.C.

Admittedly the Universal Decimal Classification is an example of hierarchical classification (allowing for a certain element of synthesis) designed to organize the whole of recorded knowledge. Unlike the Dewey Decimal Classification, however, UDC is applied much less to the control of general document collections than to the control of collections in fairly restricted subject fields. Indeed the UDC appears to be used more for microdocumentation than macrodocumentation. In England, at least, a principal application of the scheme is for the detailed indexing of reports and journal articles in specialized technical libraries. In many if not most cases, these libraries are centrally interested in only a small segment of the total schedules, as, for example, the aeronautics section. The advantage of the UDC under such circumstances is that, in many subject areas, it has been developed in sufficient detail to cope with the specific indexing of highly specialized collections, while the remainder of the schedules can be drawn upon in a more general way to index the subject areas of peripheral interest. Thus, insofar as application to special collections

is concerned, many documentalists would disagree with the statement that if a particular section of the index language is "selected for special treatment or expansion or realignment, the ramifications are soon felt throughout the rest of the system, which then needs the same kind of attention so that it will continue to function as an organic whole."

Mrs. Richmond's claim that alphabetical subject headings are "generalized-concept index terms" would appear to be naive. She perhaps confuses an indexing method with popular examples of its application. Certainly the subject headings in the authority lists of Sears and of the Library of Congress are somewhat general, but this should not therefore make the whole subject heading principle inapplicable to the indexing of highly specific subject matter. Properly designed, a scheme of alphabetical subject headings can afford an approach to indexing of aeronautics (or whatever other subject you care to name) equally as "concentrated" as an approach through a special faceted classification, Uniterms, or any other type of index language.

The recall performance of a system (*i.e.*, its score in retrieving relevant documents) is not in itself a very meaningful measure of the efficiency of a document retrieval system, since it is obvious that 100 per cent recall can always be obtained by examining the entire document collection. It is to save the time and effort involved in this task that an index to a collection is created. By so doing, the number of documents that need to be looked at is reduced (*i.e.*, *precision* is improved). At the same time, some relevant items tend to be lost (*i.e.*, *recall* deteriorates). It follows, then, that any recall figure for a particular search (*i.e.*, the percentage of the relevant documents that are retrieved) is only meaningful when considered in relation to the precision figure (*i.e.*, the percentage of the total documents retrieved that are in fact relevant) achieved at the same time.

In reviewing Dr. Richmond's conclusions, it is worthwhile considering briefly the principal factors governing recall and precision power of a document retrieval system. Precision is governed primarily by the *specificity* of the index language (*i.e.*, by its ability to define classes uniquely). This is not a direct reflection of the number of terms used to define classes in the system. The five thousand classes that are defined by, say, five thousand distinct subject headings or five thousand notational elements from a traditional hierarchical classification (pre-coordinate) may be uniquely definable by one thousand Uniterms, three hundred Mooers-type descriptors, or as few as one hundred carefully chosen semantic factors.

Recall, on the other hand, is governed by the *exhaustivity* of the indexing. The more concepts we recognize in our analysis of document content, and convert into the terms of some index language, the greater will be the number of requests for which the indexed documents will be retrieved. Maximum recall would be assured if we were able always to foresee all the types of requests for which each document entering the system would provide a relevant response. But it is not enough to recognize indexable concepts and to translate these into the terminology of the index language. We must also create a record to show what particular terms, or combination of terms, we have used to represent some particular idea. In others words, we must create an *entry vocabulary* to supplement the working vocabulary of our index language.

It is important at this point to emphasize the fact that the indexing process consists of two quite distinct steps. The first step we might call "conceptual analysis." It is the intellectual task of determining what a document is about, or more properly, of deciding for what

types of requests the document is likely to provide a suitable response.

The second step involves the translation of the notions identified in this conceptual analysis into the terms of some index language. Once a suitable *entry vocabulary* has been developed to link textual expressions (from the indexing of documents) and verbal expressions (from the indexing of requests) to the working terms of our vocabulary, this translation task can be a purely clerical operation. In fact, with suitable table lookup procedures, it can very well be delegated to a machine. That Mrs. Richmond has failed to recognize the distinction between these two steps is shown in her statement that one "system was used for the initial analysis . . . [and] its result was then matched to the terminological or structural pattern of the other three."

Let us assume that we have a collection of one thousand documents and that we carry out a conceptual analysis of these items. Now we translate these conceptual analyses into the terms of four separate index languages, say, UDC, a faceted classification, alphabetical subject headings, and Uniterms. No matter how much variation there is among these languages with respect to their ability to define classes specifically, *if we equip each system with an identical entry vocabulary, they will all have the capability of achieving the same recall performance for any particular group of requests.* If, in the Cranfield investigation, identical entry vocabularies for the four systems had been built up, based on the original conceptual analysis of test documents, and if human variables in searching had been eliminated, the performance of the systems with respect to retrieval of known relevant documents would have been identical. For a particular collection of documents and of requests, any index can achieve the same recall performance as any other, providing they are both equipped with identical entry vocabularies, based on a common conceptual analysis. If the two systems should also have the same capability for uniquely defining classes, then both will also be capable of the same precision performance.

It would appear then that Dr. Richmond is erroneous in her contention that, with respect to the indexing of highly specialized subject matter, a tailor-made faceted classification or Uniterms can offer a "concentrated approach," whereas UDC and alphabetical subject headings can offer only a "dilute approach." It should not be assumed of the UDC that there is only one such beast. In fact there are as many UDC's as there are organizations using the scheme, since no two organizations use it in exactly the same way. Certainly no informed librarian would rely on the printed index to the schedules as a suitable entry vocabulary. Each library must develop his own entry vocabulary to reflect the way that documents are written in the subject fields of interest and, even more importantly, to reflect the way that requests are made by the library's user group. The richness of the entry vocabulary is a function of the exhaustivity of the indexing, and an individual library is able to control the recall powers of its version of the UDC on this basis. Similarly, the precision powers of the system can be controlled by the degree of specificity effected through synthesis of notational elements.

In retrospect, it can be seen that the early efforts of the Cranfield Project were imperfect. Cyril Cleverdon is the first to admit this. However, the comparative study of the four index languages was of great value in signposting the direction which further investigations should take. This, and subsequent work at Cranfield has done much to clarify thinking regarding the factors that affect importantly the operating efficiency of a document retrieval system.

■■