LOUIS KAPLAN

Information Retrieval From the Management Point of View

Several conclusions may now be drawn by management, based on results derived from several "laboratory" experiments in information retrieval. A major finding is that a controlled indexing language (controlled by an authority list of headings) will not provide more effective retrieval than will the uncontrolled type. Automatic indexing, using semantic and syntactic devices, does not improve upon the performance of a manual system. Increasing the number of subject entries per document (with or without computer) will increase the number of retrievals relevant to a question, but will at the same time disproportionately increase the number of nonrelevant references.

Introduction

A number of investigations conducted recently by documentalists have grave implications for those library administrators contemplating the development of a large-scale information system. In this paper some well-known experiments are discussed, and the results evaluated from a management point of view.

During the past few years a number of significant tests of information retrieval systems have been conducted, of which three are perhaps most important to librarians: the work by Cleverdon and his associates at the College of Aeronautics in Cranfield, England; by Saracevic, Rees, and others at the Center for Documentation and Communication Research at Case Western Reserve University; and by Salton and his co-workers in the Department of Computer Sciences at Cornell. These information scientists have indisputably advanced our understanding of information retrieval; on the other hand, their efforts to optimize retrieval have not met with undivided success. Furthermore, from the library management point of view, the depth of indexing employed, the construction of thesauri, and the sophisticated devices introduced seem terribly expensive. Nevertheless, it would be a mistake for librarians to ignore the implications of the work done by these information scientists.

Brief Description of the Tests Under Discussion

1. *The Cranfield tests.* The Cranfield tests emphasized the significance of language devices which influence recall and precision, such as roles, links, interfixing, partitioning; also studied was the influence of the number of coordinate terms in a search question and the depth of indexing.[1]

Three indexing languages were tested: single-terms, concepts, and a controlled language, all in the subject field of aerodynamics. With each language several recall devices were tested, and

Mr. Kaplan is Director of Libraries, The Memorial Library, University of Wisconsin, Madison.

for each of the languages several precision devices were used, including coordination.

2. *The Case Western Reserve tests.* Several indexing languages were tested by Saracevic and his team.[2] Those that need be referred to in this context are: (a) keywords assigned by indexers (that is, in the language of the text) and (b) a language based on the so-called "telegraphic abstracts" (a language employing a number of formal recall and precision devices).

The tests conducted at Case Western Reserve University emphasized the influence of the manipulation of search questions. Depth of indexing was tested by treating full texts, abstracts, and titles as independent variables. A third major variable was the indexing languages.

3. *SMART.* The SMART system (originally established at Harvard, now at Cornell) is described in a recent text by Gerard Salton and in a number of reports entitled *Information Retrieval System,* coming most recently from the Department of Computer Sciences at Cornell.[3] Unlike MEDLARS, where machine manipulation follows manual indexing, SMART indexing depends as well upon machine manipulation of the documents prior to the actual retrieval process. Each search question and each document is manipulated from the viewpoint of word and phrase frequency and from the viewpoint of establishing, by frequency studies, clusters of related documents.

In addition, dictionaries are provided to reduce the variety of words by compounding stems and suffixes; for example, one dictionary makes it possible to recognize the singular and the plural of a word as a single term, and words such as economize, economical, economies are also gathered up as a single term. Semantic relationships are established by means of a dictionary of synonyms, and the hierarchical relationships are established in a classified system. The syntactic relationship between phrases is controlled by phrase dictionaries, for example, library schools and schools of librarianship. The emphasis in SMART, then, is on the influence of these dictionaries on the document search and in the manipulation of the search questions. These dictionaries are studied independently and also with respect to their cumulative effect. Thus the SMART system identifies the single best dictionary, as well as those which in combination prove most efficient with respect to recall and precision.

Results of Tests

The inverse relationships of recall and precision. There is general agreement that there is usually an inverse relationship between recall and precision, that is, while recall can be raised to 100 percent, the cost in the number of nonrelevant documents retrieved is great. The nearer one approaches 100 percent recall, the greater proportionately is the drop in precision.

Automatic indexing. Using SMART methods Salton came to the conclusion that, "Fully automatic text analysis and search systems do not appear to produce a retrieval performance which is inferior to that obtained by conventional systems using manual document indexing and manual search formulations."

Precision and recall devices. Precision devices, except for coordination, proved of little value. Of the various recall devices, the use of synonyms proved significant, while the hierarchical (classified) proved less effective than had been supposed. At Case Western the use of role indicators proved to be significant only when the full text was available to the indexers; with abstracts, role indicators and other retrieval devices were not superior. At Cranfield, the controlled language performance was not improved by manipulating it hierarchically.

At Cranfield a surprising outcome was the realization that the uncontrolled single term natural language of the text was little improved by most recall or precision devices. At Cornell, it was found that the cumulative effect of all the dictionaries was more effective than any lesser combination.

In summary, in any system a significant recall device is the dictionary of synonyms, but the hierarchical element is not of major significance. Coordination is a powerful retrieval procedure.

Controlled languages. At Cranfield, a rank order of thirty-three indexing languages and devices was published, indicating their power of recall. The top seven languages were all uncontrolled. The best controlled language ranked tenth; its recall ratio was 61 percent compared to 65 percent for the best of the uncontrolled languages. The statistical difference between them is regarded as significant.

Some Observations from the Managerial Point of View

Cost factors. Information scientists have not seriously attacked the question of the cost of the various indexing languages.[4] It would appear, given the emphasis placed on the indexing languages at Cranfield and the search strategy at Western Reserve, that a number of those engaged in the testing were probably well acquainted with the subject matter of the tests. Despite this, Saracevic reported that the single greatest and most important variable was the quality of the indexing. A study of MEDLAR failures shows that with respect to recall, 72 percent of the failures can be attributed to faulty indexing or to faulty search strategy, while with respect to precision the number attributable to these two factors was 45 percent From these bits of evidence the relative insignificance of the indexing system and language, compared to the indexing itself, and the imaginativeness of the search strategy, rises to haunt us. Furthermore, realizing that automatic indexing is not now superior to manual indexing, and guessing at the cost of this kind of indexing, the prospects are anything but bright.

Depth of indexing. Also significant is the considerable depth of indexing employed in these tests, depth considerably greater than is provided by conventional subject catalogers. At Western Reserve, the number of indexing terms extracted from the full text ranged from thirty-six to forty, while twenty-three to thirty were taken from the abstracts.

The significance of the depth of indexing can be seen in the statistics supplied by Cranfield in tests run on the single-term, natural language indexing language: with fourteen index terms, the recall ratio was 62.8; twenty-two terms produced a ratio of 63.5; and thirty-three terms produced a ratio of 65. However, there is a law of diminishing returns with respect to the depth of indexing. When an average of sixty terms were taken from abstracts, the recall ratio dropped to 60.9.

Automatic indexing. Turning to automatic indexing, of considerable significance from the managerial point of view is the fact that the intellectual effort required is considerable and of great significance with respect to the results. In the absence of a good dictionary of synonyms, the results can be disappointing, while the time required to compose a dictionary is an imposing consideration, as Salton has noted.

On the average, using all the devices available, SMART performs as follows:

Recall Ratio	*Precision Ratio*
10	85-95
50	60-80
100	30-45

As Salton himself has admitted, these are not satisfactory levels of performance.

Coordinate indexing. The first Cranfield study (1962) tested four indexing systems, of which one was a coordinate system, best known as Uniterms. As summarized by Cleverdon, "It achieved the best overall figures in the test, it presented no difficulties for the technical searchers . . . and was notably successful with short indexing times. It appears to have as good a relevance figure as any other system."

Nevertheless, the Cranfield group refuses to concede any natural advantage to Uniterms (a "post-coordinate" system) over the others tested (the "precoordinate" types). The capability of retrieving any combination of terms is a feature of a post-coordinate system, yet "the results of the investigation show that this advantage, though it existed, was not large." Also: "the difference between the two types of system is therefore shown to be not a fundamental difference but merely one of cost or convenience, and it has not been proven as yet on which side the advantage lies."[5]

It should be made clear in this connection that the Uniterm index system tested at Cranfield was devoid of various precision devices which are a feature of other coordinate indexing systems (such as the metallurgical index at Case Western Reserve). In the presence of such precision devices, the recall ratio found at Cranfield presumably would have been lowered.

The argument has been made that a Uniterm system will break down if used with a large collection of documents.[6] Cleverdon disputes this, though neither disputant can argue from experience. Still another theoretical argument against the Uniterm system is that it might prove less effective with social science and humanistic materials than with materials in the natural sciences.

Computer manipulation of a manual-based system. Such a system is MEDLARS; it is not an automatic system in the sense of the SMART system. In the MEDLARS system, other than the machine search itself, the indexing operations are performed manually. The MEDLARS system, on the average, provides the user with about 60 percent of the relevant documents in the collection, but of the total documents retrieved, about 50 percent will not be relevant.

It is widely believed that computer manipulation when applied to a controlled indexing language will greatly improve its efficiency. This is not true; even if more subject terms per document are posted, the overall efficiency of a controlled indexing language will not be significantly improved by computer manipulation, assuming that improvement of the recall factor alone is not enough.

This raises a perplexing question. Are all our users equally allergic to an increase in the number of nonrelevant documents, given an increase in the number of relevant ones? For example, in this regard are historians to be equated with chemists? With economists?

Another perplexing question is this: librarians suspect that scholars do not use the subject catalog extensively, and most often use it with respect to subjects outside their own speciality. Is this mainly because the subject catalog is inadequate, or because their more urgent retrieval needs lie in nonmonographic documents not now indexed in our subject catalogs?

Search strategy. Also of importance is the amount of manipulation of questions (commonly termed search strategy) that took place in these experiments. In university libraries few questions are manipulated to the extent that took place in the tests under discussion. In the Cranfield tests and at Case Western the manipulation of the questions was extensive. At Case Western each question was searched in four different ways, namely: (1) the searchable terms found in the question itself; (2) to (1) is added terms taken from a thesaurus; (3) to (1) is added terms taken from encyclopedias and sources other than the thesaurus; (4) a combination of (2) and (3).

The considerable influence of these four manipulations can be seen in the number of relevant and nonrelevant documents retrieved:

	(1)	*(2)*	*(3)*	*(4)*
Relevant	106	130	180	192
Nonrelevant	124	197	509	598
Recall Ratio	.43	.52	.72	.77
Precision Ratio	.54	.48	.34	.33

At Cornell, various semantic and syntactic procedures are applied both to the questions and to the documents; to put it otherwise, the heart of the SMART system is the correlation coefficients by which terms in the question are matched with terms from the documents.

Except in libraries serving a small group of users, in a manual setting this kind of question manipulation will not be possible unless highly skilled librarians in considerable numbers are employed. In the automatic system the manipulation of the questions is mandatory.

Whether the costs of sophisticated information systems can be justified in either the manual or the automatic mode remains to be seen. At the moment we have no idea what costs would be incurred by systems such as the SMART system in the setting of a large library with a large number of scholars engaged in research. As for making the system available to undergraduates, this involves an entirely different order of cost magnitude.

References

1. Cyril W. Cleverdon, *Factors Determining the Performance of Indexing Systems* (Cranfield, England: College of Aeronautics, 1966), 2v.
2. Tefko Saracevic, *An Inquiry into Testing of Information Retrieval Systems* (Cleveland, Ohio: Center for Documentation and Communication Research, Case Western Reserve Univ., 1968).
3. Gerald Salton, *Automatic Information Organization and Retrieval* (New York: McGraw-Hill, 1968).
4. Frank B. Rogers, "Costs of Operating an Information Retrieval Service," *Drexel Library Quarterly* 4:271–78 (Oct. 1968).
5. Cyril W. Cleverdon, *Report on the Testing and Analysis of . . . Indexing Systems* (Cranfield, England: College of Aeronautics, 1962), pp.101–02.
6. Arthur D. Little, Inc., *Centralization and Documentation. Final Report to the National Science Foundation.* 2d ed. (Cambridge, Mass.: 1964).
7. Cleverdon, *Report.*