THE TESTING OF INDEX LANGUAGE DEVICES

CYRIL W. CLEVERDON
Director, Aslib Cranfield Research Project
and
J. MILLS
Deputy Director, Aslib Cranfield Research Project

One-day conference, London, 5th February 1963

INTRODUCTION

THE evaluation of information retrieval systems has recently become an important matter. In the past, however, most reports or proposals on this type of work appear largely to have ignored the efficiency of operation of the central core of an IR system, namely those operations concerned in the compilation and use of the index. The only aspects to receive consideration are the physical form of the index and the design of thesauri or classifications. The former activity has been slanted towards the use of computers and has tended to assume that this type of equipment will, *ipso facto,* give an improved performance but has made no attempt to justify cost factors which may be one hundred times that of conventional techniques. Work on thesauri and classifications, where it has been practical in nature, appears to consist of compiling lists of terms which go out of favour as quickly as any list of subject headings in the past; the more popular theoretical approach is the setting up of models or the use of increasingly abstruse and complex algebras. From the results and conclusions of the experimental work at Cranfield, it would seem that many of these investigations are comparatively trivial.

In this paper we set out the fundamental operations involved in compiling and using an index, show how the various factors can influence the operating efficiency, and consider the methods to be used in the present Aslib Cranfield investigation.

DEFINITIONS

As the analysis of indexing has become more detailed, there has been an increasing requirement for the more precise definition of the various operations. We have endeavoured to use terms in their conventional meanings wherever possible, but it has frequently been necessary to modify, or to find new terms.

An *information retrieval system* is the complete organization for obtaining, storing and making available information. This could be a definition of a

EXHIBIT 2037
Facebook, Inc. et al.
v.

conventional library, but an IR system would be expected to exploit the information in a positive manner and to have extra facilities such as people on the staff capable of evaluating information before it is passed to the inquirer. It would also be expected to have a *subject index* to the items in the store, the index being the physical equipment which permits of the retrieval of references in the searches. The index may be in the form of a card catalogue, a printed list, a set of peek-a-boo cards, a computer, or any other convenient equipment. The arrangement within the index will depend upon the *index language** and this may be a straightforward alphabetical arrangement of terms, or a classified arrangement of terms, or any variation of these methods. The index language may be used in a *pre-co-ordinate* or *post-co-ordinate* manner. The former implies that the co-ordination of separate concepts is done at the time of indexing and the entries in the subject index will show this co-ordination. The latter implies that the co-ordination of concepts is done at the time of searching, so the entries in the subject index will refer only to single elements.

The *vocabulary* of the index language is the complete collection of sought terms in the natural language, including all necessary synonyms, that are used in the set of documents and are therefore required for entry points to the index language. An *index term,* on the other hand, is an actual term or heading used in the index language, and may be a word or words, as with alphabetical subject indexes, uniterm indexes or zatocoding, or may be notational elements, such as a group of numbers in the Universal Decimal Classification, or may be non-meaningful groups of letters, as in the Western Reserve University Metallurgical Index.

Concept indexing is the intellectual process of deciding which are the concepts in a particular document that are of sufficient importance to be included in the subject index. Conventionally, this involves a 'Yes' or 'No' assessment, for a concept either is, or is not, considered worthy of inclusion in the subject index. It is possible for the indexer to indicate the relative importance of different concepts in a document by *weighted indexing,* which involves the assignment to each concept of a weighting number.

The *exhaustivity* of the concept indexing is a comparative term; at a high level it implies that an entry is made for every possible concept in a document. At a low level it implies that a selection has been made and a smaller number of concepts have been used. *Specificity* is also a comparative term. A concept can be translated into an indexing language in such a way that the index term is co-extensive with a concept. This is a high level of specificity and implies that the index term covers the concept but nothing else besides the concept. Alternatively, the translation can be to a less specific (often called 'broader') index term which includes the concept being indexed as well as other concepts.

Syntactic indexing implies the use of headings which display the relationship between the various elements, as distinct from those which merely show the existence of several attributes relevant to the subject indexed.

* In recent papers we have been using the term 'descriptor language' following on the usage of Mr B.C. Vickery. However, Mr Calvin Mooers has pointed out that the word 'descriptor', although now somewhat debased in common usage, originally had a precise meaning. We have agreed to restrict our use of the word to this precise meaning and have therefore decided upon the term 'index language'.

A *search programme* is the formalization of the search request and it can show the same characteristics as outlined above for indexing, i.e. it entails a statement of the concepts, which can be at varying levels of exhaustivity, and can be translated into indexing terms of varying specificity.

The *operating efficiency* of an index language will depend upon its performance as regards recall and relevance. *Recall ratio* equals $\frac{100R}{C}$, where C equals the total number of documents in the collection which have an agreed standard of relevance to a given question, while R equals the number of those relevant documents retrieved in a single search. On the other hand, *relevance ratio* equals $\frac{100R}{L}$, where L equals the total number of documents retrieved in a single search. Operating efficiency is affected by the exhaustivity and specificity of the indexing, as well as the search programme, and by varying any of these factors, one will obtain a performance curve which plots recall ratio against relevance ratio. *Economic efficiency* deals with the performance of the complete index.

A *set of documents* is any collection of documents which are, or will be, used as the basis of a single subject index. The set can be large or small, restricted to an organization's research papers, or be a heterogeneous collection of journal articles, research reports, patents, etc., in many different languages, but homogeneous in that they will be used as the raw material of a single index. A *set of questions* is a collection of questions to be put to a single subject index, either at present, or at any time in the future when the index is still intended to be operating.

THE PREPARATION OF AN INDEX

Assuming there is agreement concerning the set of documents to be indexed, the following operations have to be carried out in compiling and using an index:

1. Assess the subject matter of each document in relation to the requirements of the users, and decide which subjects should be included in the index. This is concept indexing, and is at present, and for the foreseeable future, an intellectual operation. With a pre-co-ordinate index, it is also necessary to decide on the appropriate combinations of concepts and, if the index language is to show relationships, the syntax.
2. Translate the subject concepts into the index language. This is a clerical task, except in those cases where a new term has to be added to the vocabulary of the index language.
3. Place the indexing decisions into the index, which may involve preparing and filing catalogue cards, punching holes in a card or cards, or making marks on tape. This again is a clerical process.
4. Make a concept analysis of the question and decide on the priority of alternative search programmes. As with concept indexing, this is an intellectual process.
5. Translate the search concepts into the index language, a purely clerical task.
6. Operate the physical retrieval mechanism of the index.

The Aslib Cranfield Project has been primarily concerned with operations 1, 2, 4 and 5, and it has only been due to the necessity of having the index in the physical form that we have been involved with 3 and 6. It is certain that these two latter points play no part in deciding on the operating efficiency, except in so far as that one technique might be more, or less, prone to clerical errors than another. They can, however, significantly affect the economic efficiency of an index.

Involved in these operations is the variable of the index language. Whichever type of index language is used, it is certain that all the stages 1 to 6 have to be carried out. More important is it to note that the only two operations which have a true intellectual content are completely divorced from any consideration of the index language.* The basic concept analysis of the document and the basic concept analysis of the question, with the auxiliary decision of which concepts should be included in the index or the search programme, will be the same irrespective of which index language is used. It is probably the case that many indexers tend to think in the terms of the index language and their concept indexing decisions may be thereby influenced, but fundamentally it is true that concept indexing is a separate process which should not be affected by the index language.

INDEX LANGUAGES

The common basic requirement of all index languages is a complete vocabulary of all the sought terms, including all necessary synonyms, that are used in the indexing of a set of documents. This may be likened to an uncontrolled set of uniterms, and must be the basic structure for all index languages; and, whatever ultimate form an index language may take, it can only operate at maximum efficiency by having such a vocabulary. To this basic structure can be added a number of devices which are intended to improve the recall ratio or the relevance ratio. These devices (see Vickery[1]) can be listed as follows:

A. *Devices which, when introduced into an uncontrolled vocabulary of simple terms, tend to broaden the class definition and so increase recall*

1. Confounding of true synonyms.
2. Confounding of near synonyms; usually terms in the same hierarchy.
3. Confounding of different word forms; usually terms from different categories.
4. Fixed vocabulary; usually takes the form of generic terms, but may use 'metonymy' for example, representing a number of attributes by the thing possessing them.
5. Generic terms.
6. Drawing terms from categories and, within these, facets; this controls the generic level of terms, and to a certain degree controls synonyms.

* It is, of course, true that the compilation and maintenance of the index language can fairly be said to be an intellectual task. Its use, however, within the context of an indexing operation is a separate matter which requires only clerical operations.

7. Representing terms by analytical definitions (semantic factors), in which inter-relations are conveyed by relational affixes or modulants; the generic level will usually be more specific than when control is by categories.
8. Hierarchical linkage of generic and specific terms, and, possibly, of co-ordinate terms.
9. Multiple hierarchical linkage, i.e. linking each term to a number of different generic heads.
   It should be noted that devices 8 and 9 are not usually (as the others are) methods of class definition determining the structure or constituents of individual subject descriptions; they are ancillary devices (manifested as systematic sequence, or classified arrangement, as a thesaurus, as a network of *see also* references, etc.) indicating the existence of classes wider than these individual descriptions.
10. Bibliographical coupling, and citation indexes; these, also, are ancillary devices which indicate the existence of wider classes, the latter reflecting the use made of the documents and a probability of relevance arising from this.

B. *Devices which tend to narrow the class definition and so increase relevance*

11. Correlation of terms: although implicit in some form in all practical indexing, this device is not inevitable; i.e. the use of a single term to define a class may retrieve quickly and economically, if the term is sufficiently rare in the context of the system.
12. Weighting, i.e. attempts to express the particular relevance of each concept used in indexing a document to the whole document. It may take two forms:
    i. An attempt to assess subjectively the relative 'information content' of each term within the context of the system;
    ii. An objective measure, based on statistical counting of the word frequencies, etc.
13. Indicating connections between terms (interlocking):
    a. Without explicit expression of particular relations (interfixing); this may take at least three forms:
        i. Partitioning of the document; e.g. if the same document deals with the Conductivity of titanium and the Hardness of copper at a particular temperature, partitioning makes it clear that the document has at least two separate 'themes'.
        ii. Interfixing within a theme (or 'information item'); e.g. Lead (1) Coating (1) Copper (2) Pipes (2) makes it clear that the subject is the Lead coating of copper pipes and not the Copper coating of lead pipes.
        iii. If terms are recorded physically in a linear sequence, a citation order (a regulated sequence in which terms from different categories are cited), will convey relations.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | **sales@docketalarm.com** | **1-866-77-FASTCASE**

fastcase®
Smarter legal research.