# I INTRODUCTION

## I . 1 The Problem

Automatic indexing was first taken seriously in the nineteen fifties. Luhn's automatic abstracts appeared at the ICSI in 1958. This survey attempts to answer the question: how far has automatic indexing got in the fifteen years between then and now? What is the state of the art now? Salton has repeatedly claimed that automatic indexing is competitive with manual. But remarks like Wellisch' in 1972, to take a random example, suggest that these claims have had little effect. Salton concluded a survey article in 1970 by saying:

> "A large number of automatic text analysis and indexing experiments have been examined. All the available evidence indicates that the presently known text analysis procedures are at least as effective as more conventional manual indexing methods. Furthermore, a simple indexing process based on the assignment of weighted terms to documents and search requests produces better retrieval results than a more sophisticated content analysis ... Such a simple automatic indexing procedure is easily implemented on present day computers, and there are no obvious technical reasons why manual document analysis methods should not be replaced by automatic ones." (Salton 1970a)

But Wellisch simply asserts that

> "the indexer's individual judgement is necessary for the choice of the right or most suitable term chosen from an array of possible candidate terms." (Wellisch 1972)

The literature as a whole suggests that while elementary automatic indexing based on titles, say, has been accepted for economic reasons, automatic indexing experiments of more sophisticated sorts have attracted less attention than they deserve. Manual thesauri continue to roll from the presses, while information scientists concentrate on secondary problems like costs, or go a-whoring after strange gods like psycho-sociological theories of information communication or ambitious formal models of retrieval systems.

The key questions are:

Has work on automatic indexing in fact produced any worthwhile results; and

How far have these been exploited?

### I.1.1 Linguistic components of information retrieval systems

I distinguish three stages in document (and request) processing in information retrieval systems:

1 the analysis of the input text, to identify its content;
2 the representation of its content in an index description;
3 the manipulation of the description in searching.

We also have to consider the formation of the index language exploited primarily in 2, but also in 3. These activities all ordinarily involve language. We thus have four linguistic components of a retrieval system:

1 document analysis
2 document description
3 language generation
4 document searching.

Two languages are involved: the natural language of the input texts, and the artificial language of the description texts. The latter may be more or less closely related to the former. (Any numerical or other codes used for index languages are irrelevant here.)

## I.1.2 Mechanisation

Mechanisation in retrieval systems may be a) clerical and b) substantive. Clerical mechanisation covers such activities as catalogue maintenance and loan recording. It is of no interest here. Automatic indexing implies substantive mechanisation.

In principle, automatic indexing or automatic information retrieval means automating all system components. A fully automatic system would be one in which documents (and requests) were analysed and described automatically, using an automatically generated index language, and retrieved by automatic searching. There should be no human intervention after the acceptance of raw documents and requests to be put to the system.

In fact, automatic information retrieval often refers simply to mechanised searching. Systems in which document files are machine held, while requests and documents are initially processed manually, are of relatively limited interest in the present context. Slightly greater interest attaches to systems in which a manually constructed indexing language is provided, and document analysis and indexing by assignment, as well as searching, is mechanised. The same holds for systems where documents are analysed manually but described automatically using a machine generated language (for example when manually extracted keywords are automatically grouped to provide thesaurus classes for description).

There are very few fully automatic operational or experimental systems involving significantly sophisticated analysis or description. Mixed approaches are more common. Mechanised searching systems involving little or no document and request processing, but simply title and request words, for instance, must also be considered. These are fully or largely mechanised systems, though ones of a very unambitious kind.

In this report I shall therefore be concerned either with substantially mechanised systems, even if they are simple, or with approaches attempting to mechanise more than searching, typically with some linguistic sophistication. The more ambitious approaches are usually still in the experimental stage. I shall not, however, distinguish research from operating systems. Equally, for most purposes, I shall not separate retrospective searching and selective dissemination of information.

## I.1.3 Indexing

In general, more sophistication in indexing means more structure. This may be semantic or syntactic (or, paradigmatic or syntagmatic). Semantic structure is primarily associated with the index language and syntactic with the form of description. (As is well known, a given conceptual relation may be expressed either semantically, usually if it is relatively static for a universe of discourse, or syntactically, typically if it is temporary.) The provision of structural and specifically syntactic information in a description however implies its recognition in the input. Syntactic structure is normally indicated for the sake of preciseness, while semantic structure allows greater flexibility in searching. In principle indexing systems may range from simple keyword ones without any use of structural information, to Syntol-type ones where both syntactic and semantic structure is exploited. However mixed systems allowing structure of one type but not the other, or different degrees of refinement, are often found, for example ones combining an elaborate semantic structure in the indexing lexicon with simple term coordination - which constitutes minimal or null syntax - in descriptions. Alternatively, systems may allow for sophisticated treatment at one stage, but not others, for example where input texts are parsed to identify nouns for adoption as keywords.

Ideally, the form of the document input to a retrieval system is its full text. In fact it is more likely to be a representative or surrogate like a title or abstract. The discrepancy between the form of the document input and the document itself, where these differ, is an underlying problem for retrieval, since the object of retrieval is documents themselves. In the same way that the indexing language in a system is logically distinct from the natural language of input documents, though the two may be very close, say when the former is a lexical subset of the latter, the description of a document is logically distinct from either the document itself or its representative, even if the latter is in fact simply taken over without modification as a description, as in title based systems. Typically retrieval applies at two removes from the actual document, or even three if we think of it as really being directed at a document's content or what it is about.

It is useful to distinguish various levels with respect to document analysis and description. We may refer to units of analysis and their components. A document or its representative may be treated as a single unit in analysis, say for the selection of words by statistical criteria. Alternatively individual sentences may be processed separately, say in parsing. Similarly a description may be a single unit, like a class specification in the UDC, or a set of units like a list of thesaurus terms. A unit of analysis or description is an item which is treated as essentially independent of others in processing, i.e. in analysis, description or searching.

Units of analysis or description may themselves be reducible to their components. For example, a representative title may be reduced to its component words; or if a description consists of a single subject heading it may not be treated as reducible, while if it consists, say, of a Syntol-type encodement, its individual member terms may be extracted from their relational matrix. It is possible to look at the same entity in two ways: for example a keyword list forming a description may be regarded either as a unit, with its component words standing in a mutually modifying relation to one another, or as a set of effectively independent units. The important question is how a document or its description is handled.

## I . 2 Evaluation

Automatic indexing can be evaluated at two levels:

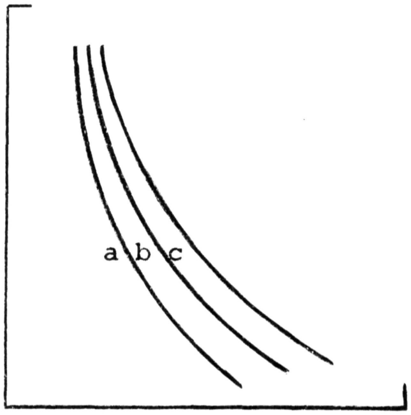
1. We may have external evaluation, in which automatic and manual indexing are compared. This may be macro-evaluation, in which the overall performance of whole systems is compared, or micro-evaluation, in which changes in one specific system component are considered. The first occurs where manual and automatic indexing of different types is involved, so that different methods of input analysis or search techniques normally follow. There are thus changes in more than one system component. This may be the case if manual indexing using a thesaurus is compared with automatic keyword extraction and classification, since although the intention of the two approaches may be the same, the details at several stages may differ. Micro-evaluation would apply, say, if manual and automatic word extraction were alternative analysis procedures for the same modes of description and searching. At the lowest level of detail it is almost impossible to confine comparisons strictly to one system component, but they may be sufficiently restricted for properly based evaluation.

2. Internal evaluation applies when different approaches to automatic indexing are compared. This again may be macro- or micro-evaluation. Macro-evaluation is illustrated by comparisons between automatic keyword and automatic document clustering, in which only the initial keyword lists for the documents are the same, but description and searching differ. Micro-evaluation is illustrated by different approaches to parsing input texts, for the same sort of description, or by different methods of grouping keywords to form an automatic thesaurus.

Retrieval system evaluation is not as good as it should be. There are obvious difficulties as soon as concepts like user happiness are invoked. But even allowing for genuine problems, there is still a depressing lack of rigour about strictly internal system evaluation, where performance is measured in terms of the system's ability to satisfy particular well-defined requirements which are independently assumed to be related to user happiness. The proportion of experimental reports failing to invoke any significance criteria is, for instance, still large.

The wide choice of measurement procedures in particular makes detailed comparisons between different experimental or project findings very difficult. This is not the place to enter into measurement controversies. It is sufficient to note that though recall and precision may be frowned on by purists, the fact that they are widely used makes cross-project comparisons possible; indeed cross-project comparisons are more or less compelled to refer to recall and precision. At the same time, such apparently trivial details as the choice of procedure for averaging over a set of requests often makes specific detailed comparisons impossible.

Since this survey is attempting a higher order evaluation, namely does automatic indexing work, the effort has to be made to judge reported results and to compare them. I shall adopt a fairly robust approach. In general, in the absence of significance tests, it is dangerous to assume that a performance difference of less than 5% is significant at some level. In any case, small differences, even if they are statistically significant, are not very interesting. In a broad way, I shall characterise performance differences, assumed statistically significant, as interesting if they are at least noticeable, i.e. of the order of 5-10% different, and as rather more interesting if they are material, i.e. more than 10%. These degrees of difference can be illustrated on the ubiquitous recall-precision graph, as follows:



b noticably better than a

c materially better than a

I shall ordinarily use recall and precision to refer only to the well known ratios. I shall use pullout and selectivity to refer in a general way to a system's retrieval of relevant and only relevant documents.

### I.2.1. Automatic vs manual indexing

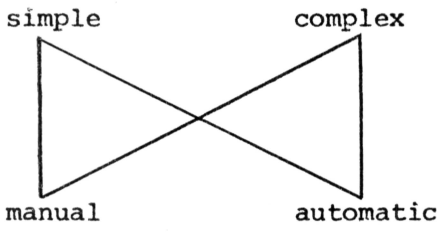
There are particular controversies concerning automatic indexing. The most important, hinted at in Salton's quotation is

simple automatic indexing vs elaborate manual indexing.

In its most extreme form, this would contrast simple keyword extraction and matching with manual relational indexing backed up by a thesaurus and complex search modification procedures. It is also exemplified by keyword extraction with some use of statistical associations for request expansion, on the one hand, and manual description involving the assignment of thesaurus headings and searching exploiting classificatory relations, on the other.

In practice it is difficult to achieve sophisticated input analysis and thesaurus construction automatically, but simple automatic indexing may be adopted because a sophisticated manual system is too expensive. However workers like Salton argue that really sophisticated manual systems do not pay their rent. It is certainly true that it is extremely difficult to achieve a generally high level of system performance by any means: gains in one area are normally only made at the cost of losses in another (the familiar R/P syndrome). This applies, as Cleverdon 1966 showed in the Cranfield 2 experiments, to wholly manual systems too, where simple procedures are typically as effective, or nearly as effective, as more complex ones.

The overall evaluation of automatic indexing must therefore rest on a series of comparisons, which can be read off the following diagram:



Using the symbol $\supset$ to mean at least noticeably, and preferably materially better than, we must ask

is complex manual indexing $\supset$ simple manual indexing
complex automatic indexing $\supset$ simple automatic indexing
simple automatic indexing $\supset$ simple manual indexing
complex automatic indexing $\supset$ complex manual indexing
simple automatic indexing $\supset$ complex manual indexing
complex automatic indexing $\supset$ simple manual indexing ?

### I.2.2 Simulation vs competition

In early work on automatic indexing, experiments in analysis and index language construction were deemed successful if they produced the same or similar results as manual indexing. Thus statistical extraction procedures which selected more or less the same words as human indexers were thought acceptable. Similarly, automatic thesaurus construction experiments were judged by their ability to produce groups of words with the same semantic base as manual thesauri, for example sets of synonyms. However retrieval evaluation experiments which showed that human indexers are not particularly consistent, and equally that they are not necessarily very effective, mean that automatic indexing should really be evaluated by its retrieval performance only. This is in any case more appropriate since automatic indexing cannot attempt detailed simulation of manual indexing, largely because manual indexing processes are not properly understood and the knowledge used by human indexers is not available. It must also be recognised that some automatic procedures, particularly statistical ones, are grossly unsuited to human simulation. Automatic and manual indexing must therefore be primarily compared through their retrieval performance.

### I.2.3 Black box vs human user

Fully automatic indexing could work perfectly well and not produce anything readily comprehensible to the human user between input document and request and output retrieved document specifications. For example, keywords and keyword classes may be identified by numbers. This is no criticism of automatic indexing as long as an automatic system with 'blind' searching is envisaged. But requirements for 'comprehensible' items like verbal document descriptions or classification displays present no difficulties of principle. Producing them is a purely clerical matter. Similarly, automatic indexing is compatible with user searching, for example by iterative methods. The fact that automatic systems may not be superficially convenient for human users should therefore not be held against them. However, it might be something of an enterprise to meet some human requirements by purely automatic means, for example a well-displayed traditional hierarchic classification with appropriate node labels. I shall not be concerned with requirements of this sort, though I shall consider some forms of user interface later under the heading of non-standard systems.

## I . 3 Information retrieval : historical and general background

### I.3.1 Use of computers

This may be divided into three phases.

a) before 1960

As noted, the use of computers for information retrieval was first taken seriously in the nineteen fifties. By 1960 it was recognised that they could take over much of the low level drudgery of information systems. It was also believed that they could undertake some of the intellectual work involved in document processing and retrieval. Thus it was accepted that they might be used for searching, and it was hoped that they could be used for analysis and description. In searching they could carry out not merely straightforward hit or miss scans of a file, but more complex Boolean searches, and could exploit relations indicated in a thesaurus, say. In analysis and description true linguistic analysis was thought possible (this was the time of hope for machine translation), and alternative statistical techniques for picking up semantic information on a distributional basis had been suggested.

b) 1960-1970

Computers were adopted for clerical work wherever this seemed economical; the Marc project (Avram 1968) is symbolic. Their use for searching was well established and increasingly sophisticated. Operational systems with extremely large machine-held files, of which Medlars (Austin 1968) is an example, appeared. A natural consequence of these developments was the international circulation of standard data tapes. Serious work on automatic techniques for analysing and describing documents and for generating index language was carried out during the decade, but the use of computers for these purposes was not really widely accepted.

In general during the decade workers in information retrieval benefitted from the enormous improvement in computer resources and knowhow. A large amount of effort was expended in the early years in achieving economies in space or time which are now less necessary. In fact it looks now as if the problem is not machine capacity but the solution of fundamental linguistic and communication problems. Developments like on-line computing, which became established during the decade, have made many clerical tasks easier, and have stimulated interactive search systems. But it must be emphasised that these do not necessarily reflect progress in an understanding of the real requirements for effective automated retrieval, merely attractive substitutes for it.

c) after 1970

The trends of the decade 1960-70 have mostly continued. In particular large international semi-automatic processing and retrieval systems are well established, and on-line search systems proliferate. However some of the more ambitious aims of the nineteen sixties have been set aside. This is partly because well-founded, economic program packages for automatic indexing of a relatively sophisticated kind are simply not available, and partly because it is suspected that current not overly sophisticated approaches are not likely to be particularly effective. Allowance must also be made for the sheer inertia of the major operational juggernauts. A shift of emphasis is also detectable, in that more attention is paid to the user and less to the core system. An increasing concern with costs is not surprisingly evident. Neither of these need undermine work on automatic indexing, and user studies could be of value, but in practice they distract.

### I.3.2 Experiments and methodology

In the nineteen sixties the first serious retrieval experiments were carried out, presumably in response to the need for the cheapest and most effective ways of managing the increasing mass of documents produced. They were primarily concerned with manual indexing (except for automatic searching), and covered either test collections, as in Cleverdon 1966, or operating services, as in Lancaster 1968a. The two main conclusions drawn from these investigations were that retrieval performance is typically only middling, and that performance differences for alternative methods of description and searching are typically fairly small. Some of the relevant work will be discussed later, when automatic indexing is evaluated.

An important consequence of the tests was a rather better understanding of the complexity of retrieval systems, and appreciation of the many components involved, as well as a recognition of the need for a proper experimental methodology. Though much research is still defective, there has been a general improvement in experimental standards; tests are more likely to be conducted in a relatively controlled manner, with numerical rather than intuitive characterisations of performance. As noted, there is a good deal of controversy about measurement, and there is still a lack of worthwhile formal models of retrieval systems (as opposed to box diagrams or Californian formulae), but these are fairly active areas of work which contribute something to the administration of experiments.

## I . 4 Related areas

There are two obviously related areas of work. One is linguistics; the other is information analysis and retrieval in a wider sense. Some types of work under the second heading are not particularly relevant to automatic indexing and document retrieval: for example so-called content analysis, and data retrieval from rigidly controlled files. But research on automatic extracting and abstracting on the one hand, and on automatic question answering (sometimes called fact retrieval) on the other, are of interest. These are considered in Section VII.

Since automatic indexing is a linguistic activity, some comments on the state of the art in linguistics and its relevance to retrieval are appropriate here.

### I.4.1 Linguistics

There is no need to go into recent developments in linguistics in detail. Lyons 1968 presents a thorough overview, and specific attempts to relate linguistics and information retrieval have been made by Montgomery 1972, Coyaud 1972, Gardin 1973 and Sparck Jones 1973a.

Documentalists have two linguistic concerns: the treatment of natural language in document and request analysis, and of artificial indexing languages in document and request descriptions. Linguists should in principle have something to offer on the first, and perhaps, through their interest in linguistic universals, something on the second. But the survey in Sparck Jones 1973a shows that there has been almost no contact between linguistics and information retrieval, largely because current linguistic theory does not have anything obvious to contribute to the solution of documentalists' problems. There are three reasons for this. One is the linguists' lack of interest in writing systematic grammars as opposed to discussing the problems of how to do it properly. The second is the bias towards abstract models of linguistic competence rather than towards procedures reflecting linguistic performance; thus there is more concern with the correct representation of sentence structure than with the means of discovering it. The third is the continuing absence of a comprehensive treatment of semantics, which is of course of prime interest to documentalists. Linguists have not had much to offer on the related topic of gross discourse, as opposed to within sentence, structure either. It is, however, arguable whether the characterisation of such language using processes as providing a summary of a text, which is the essential objective of indexing, is properly a concern of theoretical linguistics as it is ordinarily regarded.

The following is therefore simply intended to provide a frame of reference, for example for the discussion of approaches to syntactic analysis for document retrieval.

Theoretical linguistics in the last fifteen years has been dominated by Chomsky. Chomsky 1965 is representative, and Kimball 1973 is a useful synthesis of his treatment of different types of grammatical, or syntactic, theory. Chomsky's approach to the provision of linguistic models, and his specific model, transformational grammar, have been widely accepted. Even where substantial modifications of transformational grammar have been proposed, as for example by the generative semanticists, the broad framework remains. More generally, most people working with language recognise the basic Chomskian distinction between surface and deep structure: text sentences exhibit superficial structures which are merely particular manifestations of underlying logical structures. Thus the sentences "Algernon proposed to Cecily" and "Cecily was proposed to by Algernon" have different surface structures but the same deep structure. Different views of the precise form of deep structures, and of the exact relation between deep and surface structures, have been put forward, but there is general agreement on the need to recognise two such structural levels.

Transformational grammar is most conveniently presented as an alternative to phrase structure grammar. A phrase structure grammar simply characterises a sentence as a nested bracketing (alternatively represented as a tree). It cannot therefore deal effectively with logically associated but physically separated units: for example in "Gregory brought the logs in", "brought" and "in" are related but do not adjoin and so cannot be grouped together. It may also be the case that the structural representations of logically closely related sentences, like an active and corresponding passive, are quite different. A transformational grammar characterises a text sentence as the result of the application of a series of transformations to an initial phrase structure. This is the form of the deep representation of a sentence, and the transformational rules of the grammar modify it in different ways, for example by moving or deleting components.

Chomsky's view is that deep structures are purely syntactic, and are processed independently to provide a semantic interpretation for a sentence. Other linguists seek to characterise deep structures semantically. In any case, very little has been done to provide the full specification of the semantic side of the grammar which anyone attempting automatic text analysis, like documentalists, would need. The rather crude proposals of Katz and Fodor 1963, seeking to exploit collocations between general conceptual classifiers as a vehicle for meaning identification and characterisation, can be taken as an illustration. More recently attention has focussed on one type of semantic sentence property, associated with logical relations between sentences like that of presupposition which holds, for example, between "It surprised John that Fred left" and "Fred left". The relevance of this development for documentation is not clear, though it has implications for question answering. Semantic approaches like Katz and Fodor's tend to be adopted in practice, if only in an ad hoc way, by anyone attempting reasonably comprehensive text analysis procedures.

The problems of identifying the deep structures underlying given sentences automatically are considerable, particularly if parsing in strict accord with Chomsky's model is envisaged, since this involves inverting an essentially one way process so as, for example, to replace (unknown) deleted items. The type of strategy which has to be adopted is discussed by Petrick 1973. Other parsing techniques with the general aim of identifying deep structures, but not necessarily Chomskian ones, have been studied by workers on question answering and will be considered later. In practice a good deal of mileage may be got out of relatively simple automatic phrase structure analysis. As a general syntactic model phrase structure is demonstrably inadequate, though large parsers on this basis were constructed (see Kuno 1965, 1966). But their manifest low level convenience means that they tend to be implemented, particularly where only restricted text analysis is required.