Computerized indexing need not be impossible

Colin Bell and Kevin P. Jones

Within less than 20 years vast new areas of the solar system have been brought closer: the Moon has been visited by man and Mars, Venus and even some of the further planets have been brought into sharper focus than could ever have been achieved through astronomy. This new era of exploration has been founded on skills in rocketry, but none of this could have been attained without enormous advances in computation. All would have seemed impossible far less than a lifetime ago. Therefore, the time is ill-chosen to state that something may be impossible (as Mr Neil Fisk[1] implies in his letter about computerized indexing), even though it may be difficult to achieve at present and be of doubtful economic value.

Some processes involved in indexing are entirely mechanistic and closely related to sets of rules. Examples of these are alphabetization and the permutation of index entries, both of which have been successfully computerized. Other processes are far less clearly understood. Nevertheless, some, or even all, of these processes may also turn out to be mechanistic, and therefore capable of being computerized, once the full structure of the indexing operation has been elucidated. It is also possible that a few operations are not mechanistic. It may still be possible to substitute mechanized processes for these, however, as computer routines are far more consistent. A properly programmed computer cannot forget to do something.

The major difficulty facing the potential mechanizer is that what constitutes the indexing process is poorly understood. Even the assessment of 'good' indexing practice is largely subjective, although Keen[2] has been attempting to establish objective criteria at the College of Librarianship, Wales. Therefore, before a case can be established as to whether indexing is or is not capable of being computerized, it is necessary to postulate the processes which may be involved. Some of what we believe to be the major activities are outlined below:

i. assessment of what the text is 'all-about'
ii. rejection of unsuitable words for indexing
iii. identification and extraction of suitable words for indexing
iv. identification of singular root forms
v. assessing the relative value of the extracted words as potential index entries
vi. recording semantic relationships between potential index entries
vii. recording syntactic relationships
viii. conversion of words selected into standardized morphological forms: e.g. plurals where appropriate
ix. creation of cross references
x. imposition of economic constraints
xi. alphabetization and permutation.

There are many programs available to perform the last named and this will not be given further consideration.

The designer of machine systems can either attempt to mimic human actions closely or can apply purely mechanistic alternatives. In some cases these alternatives may be improvements on human processes: computers are much better at counting, for instance. In others the computer may produce less aesthetically satisfying but serviceable results: the extremely ugly KWIC (Keyword-in-Context) indexes are capable of retrieving relevant material.

In an earlier exploration of the indexing process by one of the present authors[3] it was argued that simple repetition of text words is one of the major selection criteria employed by human indexers. In recent experiments conducted by members of the Aslib Informatics Group[4] it was found that inter-indexer performance was most consistent when words occurred frequently in the texts selected. On the contrary, passages were assessed to be difficult to index when word repetition was low.

The process of computerization involves strict discipline and a close attention to detail. This can be illustrated by reference to the previous paragraph in which it was assumed that the reader would understand what was implied by 'words in texts'. It is relatively trivial to program a computer to recognize *words*: they are strings of alphabetical characters separated by spaces or punctuation. Compound words (*fountain pen*, for instance) can be regarded as exceptions in that they may need to be treated as a unit rather than as two distinct words. This particular problem will be given further consideration subsequently. Another assumption was made in the previous paragraph, namely that the reader would realize that the 'words' would not be closed-system, function words (articles, pronouns, prepositions, etc.), but would be open-system words capable of being defined by dictionaries rather than by grammars. Nevertheless, it is simple to make the computer disregard the closed-system words[5] as these form a finite collection which is but rarely extended. Further, it is relatively simple to allow for exceptions, mainly proper names like *Helen of Troy*, where the whole phrase needs to be selected. In the indexing of monographs and specialized serials it is essential to be able to reject words which merely relate what that particular publication is all about—*rock* may not be a useful index entry if it occurs on every page of a geology textbook. It would be relatively simple to create programs to establish what the text is all about on a statistical basis.

It seems probable that the human indexer pays greater regard to words in chapter headings, sub-headings and words printed in unusual type faces such as bold and italics. This is not surprising, as the author is in effect telling the reader 'here is something new', 'this is especially important', or 'briefly, this is what it is about'. Fortunately the computer can be programmed to recognize unusual type styles where texts have been prepared for computer type-setting. This is an increasingly common occurrence and is usually achieved by employing control characters which may either be unusual characters or combinations of characters: for instance, $$ might be used to indicate that italic type was specified for subsequent text. The human indexer may place greater weight on words in the opening and closing sentences of paragraphs and similarly in opening and closing paragraphs. Even words consisting of unusual letter combinations (the *fj* in *fjord*, for example) are more likely to be selected. All of these can be identified mechanically, but the weighting given to them will have to be established pragmatically.

Human memory is sufficiently flexible to permit comparisons to be made between a word and its derivations without conscious effort. In the machine, inter-word comparison is facilitated by the automatic conversion of all words into their singular, root forms. This can be achieved. MORPHS[6] (Minicomputer Operated Retrieval—Partially Heuristic—System) has been automatically establishing root forms for the past three years and operates at over 90% accuracy.

Having established a set of significant words it is important to identify the relationships between them and also to words which may *previously* have been considered to be insignificant. The semantic relationships can be identified through a series of frames, some of which are illustrated below (the letters are used to represent words):

i.	A is/are B	(synonym/hierarchical relationship)
ii.	A termed B	(synonym/definition)
iii.	A defined as B	(definition/hierarchical relationship)
iv.	A or B	(synonym)
v.	A classes B, C, D	(hierarchical relationship)
vi.	A (B)	(synonym)
vii.	A containing B	(inclusive relationship)

Examples of some of these are listed below (the category is indicated in parentheses after the example):

a *blow-hole* or *gloup* (iv)
visual display units (V.D.U's) (vi)
generalized transformations fall into two classes: *embedding* and *conjoining rules* (vii).

The list of frames is far from complete and a successful algorithm would depend upon elucidating a relatively comprehensive set. From the above it is evident that the frames rarely fall into mutually exclusive categories, but this may not be a major hindrance, and will be of little consequence in the operation of extending the set of indexable words. The hierarchical and synonymic relationships can also be used for forming cross-references. In the case of synonyms the author's preferred terminology should be readily identified on a statistical basis. The statistical relationship between words can be established also by clustering (Sparck Jones[7]). Clustering is a computational technique for classification which has no exact manual equivalent. Nevertheless, it appears to approximate closely to the almost intuitive ability which humans have for grouping words and which finds ultimate expression in Roget's *Thesaurus*. Unfortunately it is difficult to illustrate this multi-dimensional technique in non-mathematical terms, but it is possible to appreciate its potential when it is used to produce contours of sets of similarities between words, as in the work by Susan Jones[8].

The frames and clustering techniques afford methods for establishing cross references and the latter have been used to pre-coordinate index entries. It is also possible to relate some words through syntactic relationships; prepositional relationships are particularly important. One other criterion, and possibly the one which most closely matches human indexing activity, is word proximity: co-occurrence within a sentence or paragraph. An extreme example of this approach is found in compound words which are discussed subsequently. Fortunately, this last would be simple to establish computationally.

It is envisaged that the main working file would consist of a concordance of all significant words (i.e. non-closed-system function words) in their singular root form. The roots would be accompanied by all derivations, possibly including plurals, detected in the text together with information about semantic and syntactic relationships. Programs for concordance construction have existed for approximately 25 years. Obviously the concordance envisaged is more complex, but is probably less so than the files associated with many information retrieval systems. It would need to be large, but very large files are handled for information retrieval and storage costs are tending to fall.

A subsidiary file would consist of closed-system words, and lists of plural endings, prefixes and suffixes would be needed. These would be comparatively small (that is between 500 and 1,000 words) and it should not be too difficult to incorporate a small dictionary of common homographs (*seal* for instance). The tendency towards decreasing storage costs coupled with increasing speed of access should enable some dictionary, encyclopaedia and thesaurus facilities to be built in or possibly provided on-line. The human indexer has access to these tools and it may be expecting rather much for the machine to function without them. On-line access to dictionaries and other reference 'books' should follow the widening availability of services like VIEWDATA—it is envisaged that PRESTEL will include an encyclopaedia.

To present a moderately coherent picture it has been necessary to skip rather quickly over some fairly difficult problems. Thus the selection of words from text may present some difficulties. Compound words (like *fountain pen*) will have to be identified either by statistical association (*fountains* and *pens* are unlikely to co-occur within the *same* text except as *fountain pens*) or by employing a more complex process which divides the text where closed-system words or punctuation occur and then subjects the ensuing collections of words (pseudo-phrases) to a series of tests to determine whether they are compound words or merely phrases. Algorithms[9] have recently been developed which may be of assistance. Proper names and other special vocabularies (e.g. medical, chemical and botanical names) could be identified computationally—programs exist which can detect chemical nomenclature[10], for instance.

Obviously, completing the task of programming would be difficult, but it is, we believe, most unlikely to be impossible. We are not sufficiently arrogant to propose that such a system would work in an error-free manner. Indeed it is essential to accept that such systems must produce errors in the course of operation as programming for perfection is literally impossible in anything involving ordinary text. Errors are accepted in human systems (studies associated with information retrieval systems suggest that inter-indexer consistency rarely exceeds 40%[11]). If errors are tolerated in human systems then it is unrealistic to expect perfection in mechanized ones.

It is hoped that the above sketch will be sufficient to demonstrate that computerized indexing should be possible eventually. To postulate that it will always be impossible is tantamount to claiming that further effort in areas like artificial intelligence should cease to be funded in spite of such major advances as Winograd's program[12] and the very extensive progress which has been made in mechanized information retrieval.

The authors are, respectively, Head of Physical Testing Group and of Information Group of the Malaysian Rubber Producers' Research Association.

References

1. Fisk, N. *The Indexer,* Vol. 11, No. 2, October 1978.
2. Keen, E. M. *Eurim II*; ed. W. E. Batten. London: Aslib, 1977, 174-177.
3. Jones, K. P. *J. Docum.* 32, 1976, 118-123.
4. Aslib Informatics Group. Practical Indexing Workshop. 30th October 1978.
5. Winter, E. *in Informatics 3*; ed. K. P. Jones and V. Horsnell. London: Aslib, 1978.
6. Bell, C. L. M. *and* Jones, K. P. *Program* 10, 1976, 14-27.
7. Sparck Jones, K. *Automatic Keyword classification for information retrieval.* London: Butterworths, 1971.
8. Jones, S. *in Informatics 2*; ed. V. Horsnell. London: Aslib, 1975.
9. Bell, C. and Jones, K. P. [In preparation.]
10. Heym, D. R., Siegel, H., Steensland, M. C. *and* Vo, H. V. *J. Chem. Inf. Comp. Sci.* 16, 1976, 171-176.
11. Keen, E. M. and Digger, J. A. *Report of an information science index languages test.* Aberystwyth: College of Librarianship, Wales, 1972.
12. Winograd, T. *Understanding natural language.* Edinburgh: University Press, 1972.

Letters to the Editor

More about computer-assisted indexing

Note by Professor P. A. Samet, President, The British Computer Society

Following my letter to *The Indexer* (Vol. 11, No. 2, October 1978, p.114) I wrote to a colleague, Donald Knuth, at Stanford University about his experience of computer-assisted production of an index, as he remarks in his series of books *The Art of Computer Programming* that he had used a computer to help him. I was, in fact, intrigued by his wording, for at the end of the first two volumes inaccuracies are excused by saying that the index had been 'prepared' with the help of a computer, whereas in the third volume the word used was 'sorted'. His reply appears below:

Today I am preparing an index to a 200-page book and it is my firm conviction that no *automatic* tool is good enough (or will ever be). A good index needs to be thought about and *understood*—every entry looked up in the text again, and reflected on. . . . I spent six weeks per index in vols. 1, 2 and 3. However, computers can certainly be aids in the clerical parts of this task. I used a simple sorting programme for the initial data reduction. Today I am doing it differently: using a very versatile CRT display text-editor and a few macros to do the whole thing. It's like having a super-intelligent typewriter, making a special-purpose system unnecessary. Anyway, that's my opinion.

Yours sincerely,
Professor D. E. KNUTH,
Stanford University.

The structure of sub-headings

Am I right in thinking that traditional sub-heading procedure in book indexing flows, in terms of subject importance, from 'greater' to 'lesser'? It is, that is to say, linear in construction.

In many cases this seems satisfactory, but I have never felt it to be entirely so for scientific works, because science is not, in this sense, linear; with its constant inter-relationships, it is more of a lattice.

My interest was therefore aroused recently when I came across this idea cogently illustrated in a book, *Molecular aspects of gene expression in plants,* (Ed.) J. A. Bryant, Academic Press, 1976/77.

The index was excellent: to illustrate my point, it contained under the heading 'Ribonucleic acid' 9 sub-headings and 8 *see also* references. Looking these up one found that they in their turn had one sub-heading in three cases but 9, 11, 18, 31, and 39 sub-headings respectively in the remainder. The overall result was an admirably integrated format which I do not believe the linear method would have achieved.

Yours sincerely,
CHARLES LISTER.

Indexes are not Frothy

A.R. reports the following passage from recent correspondence with a well-known publisher. He is not sure whether the story has a moral.

A.R.: '. . . I hope the book is to have an index.'

Publisher: '. . . I appreciate your crusading zeal in the cause of indexes, but feel that in this instance one is not only unnecessary but might actually detract from the somewhat frothy tone of the book.'