AUTOMATIC INDEXING 1974

A State of the Art Review

KAREN SPARCK JONES

April 1974

Computer Laboratory
University of Cambridge
Corn Exchange Street
Cambridge CB2 3QG, England

This review was prepared under Grant No. SI/G/096 from the Office for Scientific and Technical Information, Department of Education and Science (now the British Library Research and Development Department).

The review was discussed at a Workshop held at Crawley, Sussex on April 29 and 30 1974. This was sponsored by the British Library Research and Development Division, and was organised by the University of Kent, Canterbury. A Report on the Workshop by Miss E. Wilson, Computing Laboratory, University of Kent, is in preparation.

C O N T E N T S

PREFACE

Terminology

Information retrieval	:	= document retrieval = reference retrieval = documentation (= retrieval).
Fact retrieval	:	(alias question answering) direct extraction of facts from an information store (or obtaining of answers therefrom)
Indexing	:	provision of descriptions of documents for retrieval purposes.
Extracting	:	provision of extracts of document texts for general purposes.
Abstracting	:	provision of summaries of documents for general purposes.
Automatic indexing	:	strictly, automatic provision of document descriptions for retrieval. But in this review extended to cover all the linguistic operations involved in analysing, describing and searching for documents, and in creating the index language required.
Analysis	:	of documents to select information required for indexing.
Description (1)	:	of documents by characterising selected information in an index language.
Searching	:	of file of document descriptions to find descriptions meeting a request specification.
Natural language	:	the language of documents, their titles and abstracts, and of requests.
Index language	:	the language used to describe documents and requests, which may be more or less artificial.
Document	:	specifically the actual articles in a library, but more generally the form of the document input to indexing, whether full text, title or abstract.
Surrogate	:	(alias representative) the input form of the document when this is not its full text.
Description (2)	:	the index language characterisation of a document.
Unit	:	in analysis or indexing, any word or sequence of words treated as a whole.
Word	:	a word in natural language.

Keyword	:	a word or word stem from the input natural language adopted as an index term.
Term	:	a word or word group in the index language used to describe documents; a simple rather than complex item.
Subject heading	:	also a word or word group in the index language; but normally a complex or compound item.
Descriptor	:	any unit in an index description, whether a keyword, term, subject heading or class label.
Precoordinate	:	as ordinarily used.
Postcoordinate	:	ditto.
Vocabulary	:	the set of descriptors of an indexing language.
Classification	:	any grouping of items, but ordinarily applied to descriptors.
Clustering	:	classification applied to documents.
Syntax	:	(alias syntagmatic) refers to relations between words or descriptors not constant in the language.
Semantics	:	(alias paradigmatic) refers to constant relations between words or descriptors; also simply to their own meaning.
Request	:	retrieval question.
Collection	:	in general, the set of documents in a retrieval system; but specifically, the set of documents and requests used for a retrieval experiment.
Recall	:	the ratio between relevant documents retrieved and all the relevant documents in a collection, for a request or set of requests.
Precision	:	the ratio between relevant documents retrieved and all the documents retrieved, for a request or request set.
Pullout	:	extent to which relevant documents are retrieved (in a general, not precisely measured way).
Selectivity	:	extent to which relevant documents only are retrieved (ditto).
Performance	:	of a retrieval system, measured in appropriate ways, for example by recall and precision.
Noticeable	:	interesting (as well as statistically significant) difference in performance.
Material	:	very interesting (ditto) difference in performance.

Data references

A particular convention will be adopted for specifying test collections. Thus an expression like "a 21x379 tropical foodstuffs collection" refers to a set of 21 requests and 379 documents dealing with tropical foodstuffs. These normally have associated relevance judgements, i.e. sets of documents defined as relevant to the requests. Items like "the 42x200 Cranfield collection" refer to well-known test collections. As collection specifications are in the form "mxn", the use of "m" or "n" refers to collections for which full details are not given, "m" representing an unknown number of requests and "n" of documents. The form "nK" is used where the number of documents is not given, but is manifestly large.

Literature references

To avoid overloading the text a slightly abbreviated form of reference has been adopted. "Snooks 1969,1970", for example, refers either to papers by the lone Snooks, or to papers of which the first listed author is Snooks, or to papers of which he is deemed the lead author because he is project director or most consistent author in a series of project papers. The relevant papers are all listed under Snooks in the bibliography. One or two sets of references use project or organisation names as leads.

Literature coverage

Sparck Jones 1973a attempted to cover much of the relevant literature for the period 1965-1970. This survey concentrates on the more recent period 1968-1973. It is based chiefly on papers published in the Journal of Documentation, the Journal of the ASIS, and Information Storage and Retrieval, but chapters like that dealing with question answering use material from other sources. I have also exploited other surveys, notably the relevant chapters of the Annual Review of Information Science and Technology (Cuadra 1966-). General acknowledgements must be made to Cleverdon 1966, Coyaud 1966, Lancaster 1968b, 1972a, Salton 1968a and Sharp 1965.

Organisation of the review

Section I defines automatic indexing for the purposes of the review and provides background on information retrieval and linguistics. Section II lists the linguistic components of an information retrieval system, from the point of view of syntax and semantics, and considers other system components which may be expected to influence indexing

performance. Sections III and IV deal with syntax and semantics respectively, describing syntactic and semantic approaches to document analysis, description and searching and to index languages. Section V examines 'indirect' indexing, represented by the use of citations and by document clustering. Section VI considers automatic indexing in operational mechanised retrieval systems, both standard and 'non-standard', the latter covering interactive retrieval and the use of printed indexes. Section VII makes a brief survey of work in the related areas of automatic abstracting and automatic fact retrieval. Section VIII summarises the main evaluation experiments involving automatic indexing, with special reference to the Smart project. Section IX attempts an overall conclusion and recommendations for the future.