ASLIB CRANFIELD RESEARCH PROJECT

FACTORS DETERMINING THE PERFORMANCE OF INDEXING SYSTEMS VOLUME 2

Cyril Cleverdon and Michael Keen

An investigation supported by a grant to Aslib
by the National Science Foundation

Cranfield
1966

Published by

Cyril Cleverdon

Wharley End

Bedford

SUMMARY

The test results are presented for a number of different index languages using various devices which affect recall or precision. Within the environment of this test, it is shown that the best performance was obtained with the group of eight index languages which used single terms. The group of fifteen index languages which were based on concepts gave the worst performance, while a group of six index languages based on the Thesaurus of Engineering Terms of the Engineers Joint Council were intermediary. Of the single term index languages, the only method of improving performance was to group synonyms and word forms, and any broader groupings of terms depressed performance. The use of precision devices such as links gave no advantage as compared to the basic device of simple coordination.

All results have to be considered within the context of the experimental environment, but they can be said to substantiate or clarify many of the findings of Cranfield I. It is conclusively shown that an inverse relationship exists between recall and precision, whatever the variable may be that is being changed. The two factors which appear most likely to affect performance are the level of exhaustivity of indexing and the level of specificity of the terms in the index language. For any given operational situation, the optimum levels cannot be categorically stated in advance, but can only be determined by an evaluation of the system, the main consideration probably being the subject field.

It would be unusual if the characteristics of the subject field used for this test were such as to make it unique, so the high performance obtained with the single terms in natural language can be considered to be of some importance in regard to the use of natural language text as input to mechanised systems.

PREFACE

It was intended that this should be the final volume of the Report on Cranfield II. This may still be the case, but as the results were being prepared for publication, we were continually aware of the gaps that needed to be filled. The delay in the appearance of this volume is partly due to attempts to obtain some of the missing data, but a great deal still remains to be done. The detailed analysis of the reasons for failure to retrieve relevant documents or for the retrieval of non-relevant documents was an important part of Cranfield I, but so far in this project it has only been attempted in a superficial manner. It is most desirable that such analysis should be done, the more so because of the completely unexpected test results.

To the acknowledgements already made in Volume I, I would also wish to include Professor Salton and Professor Wilkins. It was a most happy chance that the National Science Foundation should fund two projects which, when they started, appeared to have quite different objectives, but which, as they progressed, were seen to be closely related. Cooperation with Professor Salton has enabled us to go much further than would otherwise have been the case, and I am most grateful for his continued assistance over the past two years and for providing data on his work which we have been able to present briefly in this report.

It was on the recommendation of a colleague that I read 'Social Deviance' by Professor L.T. Wilkins. Although dealing with a different subject, the author's opinions on what should be done in his particular field were entirely in line with my views on what is required in the field of documentation. For his agreement to allow me to include a series of short extracts from his book, I am most grateful to Professor Wilkins.

Cranfield
December 1966

Cyril Cleverdon

CONTENTS