1

Introduction

Karen Sparck Jones

This book is about information retrieval experiment. Documentary information systems have changed in many ways in the last 20 years. The rapid growth of specialized literature has encouraged an intellectual development, post-coordination, and a technological development, the use of computers. Many questions about effective methods of document identification, and about efficient methods of document management, have naturally followed. These questions about the substantive and the economic aspects of retrieval systems have provoked a whole range of studies. Some of these may be described as investigations; others can properly be described as experiments. They are sometimes associated with generalizations or sometimes, more strongly, with theories or models.

These studies taken together have produced some moderately solid results, and have to some extent enlarged our understanding of the way retrieval systems work. For example it appears that good retrieval performance lies in the 40–60 per cent recall and precision area; and it seems that some probabilistic models offer valuable insights into system behaviour. But research progress has perhaps been less than might have been expected, and information system practice has in essentials been extremely conservative. The 1958 International Conference on Scientific Information, held in Washington, was widely felt to mark the beginning of a new era in information processing. The novel ideas and techniques to be developed were symbolized by the 'auto-abstracts' of conference papers produced by Luhn. By 1978 computers were firmly established in information work, but primarily, and almost entirely, for clerical operations: the crucial information processes of document and request characterization and matching are done by human beings along conventional lines.

There is no very good reason to suppose that the conventional methods are best, even in principle, let alone practice. There is in particular no good reason to believe that they are based on any thorough understanding of the nature of information systems. Information systems are manifestly complex. It is by no means certain that information science exists: but it is clear that information processes need scientific study. The requirement for, and role of, experiments in such a study is clear. A good deal of experimental and investigative work has been done in the last two decades; but while results

have been obtained and experience gained, a review of this research serves mainly to show how much more needs to be done. Individual experiments are still far too often methodologically inadequate, in some cases in obvious matters of design, in others in less obvious but nevertheless very important matters of scale.

Over the years, the character of experimental work in information retrieval has changed. Many early tests focused on the new information processing methods associated with automation: they were concerned either with the application of post-coordinate searching using manual indexing, or more ambitiously, with wholly automatic indexing techniques. However the main outcome of this work was the discovery that nobody really knew, in detail, how document retrieval systems behave or more importantly, why they behave in the way they do. Many test results were unexpected or ambiguous, and it became evident that this was due to weak experimental design, which was in turn a consequence of unexamined assumptions about retrieval systems. The need for a better standard of experiment, preferably informed by an explicit characterization of system properties, has been slowly recognized, and the quality of experiments has improved. Some research workers have in particular been able to capitalize on previous work through the use of common data, and also through the application of common techniques for system performance measurement. Thus although specific tests may have been done within the framework of particular operational systems and their assumptions, much of the research that has been done has been devoted to a painstaking analysis of typical system behaviour, over a wide range of factors.

These developments are in many ways illustrated by the objectives of the Cranfield 2 test, its results, and its subsequent influence on information retrieval research. The test was designed to compare the performance of different indexing languages, including simple natural language, in a laboratory environment. It was intended to improve on Cranfield 1 in experimental design, and was systematically conducted. The most striking result, the competitive performance of post-coordinate natural language terms, was not expected, but has been largely supported by subsequent experiments. The test also indicated that only medium levels of performance could be expected of retrieval systems. The inverse relationship between recall and precision was clearly displayed, and subsequently adopted as a generalization about retrieval systems, if not as a law.

The test was criticized for methodological inadequacies, for example in the way the test data was generated, and for being too small in scale. Further, while the test's concern with index languages and their application was clearly focused on the centre of retrieval systems, important factors, especially those involving users, were not studied. It has also been argued that recall and precision have been overvalued as measures and, more generally, that the bottom-up approach to the understanding of retrieval systems represented by Cranfield-type experiments may be unproductive or misleading, and that a top-down approach, guided by a theory or model, is preferable.

The impact of the Cranfield 2 test on later research has nevertheless been considerable. Specific projects have applied Cranfield procedures and used the Cranfield test data for comparative purposes. More broadly, much of the experimental research done since Cranfield 2 has been in the same tradition,

and addressed to similar topics, though with more emphasis on automatic natural language indexing techniques, and more interest in a wider range of system factors. Moreover, in other work motivated by the alternative approach, seeking to base experiments on theory and so explain rather than simply exhibit system behaviour, the influence of Cranfield is often visible. Thus while much experimental work is explicitly indebted to Cranfield 2, much more has been coloured by it. This book is rightly dedicated to Cyril Cleverdon.

The Cranfield test showed what hard work experimental information retrieval is: collecting data, conducting searches, computing performance figures all take substantial time and effort. There is work in the essentials of experiment: maintaining systematic procedures, imposing controls on variables; and there is work in the realities of research: varying approaches in the search for understanding and explanation, and rehashing past studies for consistency and solidity. Salton (Chapter 15) has noted that the Smart Project has involved thousands of tests, implying a large use of man and machine power; yet there are important aspects of retrieval system use and behaviour which have not been studied in these tests.

Since information is claimed to be of increasing importance, at least in twentieth century high technology cultures, while information systems are little understood, there is every justification for information system investigation and experiment. Existing information systems reflect long and extensive experience, but it does not follow that such systems cannot be improved, and that new technologies may not be exploited for wholly new types of system. It is arguable that our current understanding of information processing is like that of sixteenth century herbalists: it embodies some observation and insight, but lacks detailed analysis and supporting theory.

Unfortunately, though many retrieval experiments and investigations have been carried out in the last 20 years, much of the experience in the conduct of tests which has been gained is not very accessible. Published papers tend to be cleaned up accounts of objectives, general methods, and results. Project reports may be much fuller, but even here it is often extremely difficult to find out exactly what was done, or why it was done. Though an improvement in the quality of experiments is detectable, far too many of those reported are defective, and in many cases defective in recognized ways, for which remedies are available. As all but the most limited test is a major enterprise, it is a pity that so much effort should be wasted. The best that can be said about many reported studies is that even if they are individually dubious they may collectively point in the same direction. This is perhaps something, but it is not much. The object of this volume is to make available the experience in information retrieval testing of its contributors, in the hope that this will lead to more fruitful and useful testing in the future.

The book is designed to treat information retrieval experiment and investigation in a comprehensive way, relevant to both pure research and to operational practice. Pure research naturally leads to experiment, but system operators may also want to know how their system is working. Both research workers and practitioners may know what they want to find out, but it may not be at all obvious how it is to be found out, and in particular what the best specific testing methods are. This is clearly seen in evaluation tests. Most tests are intended to evaluate the performance of existing or proposed

retrieval systems or their individual components. A good deal has been written about evaluation: what aspects of system performance, under the general headings of effectiveness and efficiency, should be evaluated and what form the evaluation should take. Evaluation implies experiment, or at least observation. But what to do and how to do it are distinct. For example, there is a difference between an interest in studying a method of indexing, evaluating it via recall, and actually carrying out the study by having certain documents indexed by certain people and searched in a certain way to satisfy certain needs. The general assumption tends to be that if you know what you want to evaluate, with given evaluation criteria, the appropriate experiment is obvious. Experience shows that this is not the case, because the characteristics of retrieval systems are so difficult to determine and their implication for experiment so difficult to identify.

Much has nevertheless been learnt in the last 20 years about the conduct of retrieval system tests; and this book attempts to bring the varied experience of its contributors to bear on all aspects of retrieval experiment. Part 1 deals with general topics applicable to all experiments and investigation: methodological issues, the relation between testing and evaluation, the problems presented by inaccessible system factors like meaning, and the proper provision of data. Part 2 covers different types of test: real-life tests and laboratory tests, with a separate treatment of manual and automatic systems in each case; simulation tests, and gedanken experiments. Though the different types of test have much in common, they present distinct problems and demand different approaches in testing. In Part 3, specific retrieval tests are used to illustrate and amplify the points made in the previous sections. Chapter 12 considers the test history of the last 20 years through major and representative tests; Chapter 13 examines Cranfield 1 and 2; Chapter 14 analyses a specific experiment in detail, and Chapter 15 the long-term Smart Project. In the book as a whole, the emphasis is on a detailed treatment of the issues involved in information retrieval testing, and especially on the less obvious problems occurring in the design and conduct of tests. Throughout, examples are used to illustrate the points made.

It will be evident that the theme of the book is a large one, with many facets. The chapters moreover embody individual views of their specific topics. For this reason, it is appropriate to conclude this Introduction with the interpretations of the key terms used which underlie all the chapters. These terms have been used in these senses so far, but to lead into Part 1 they need an explicit rather than implicit characterization. However, since the characterization is filled out by the book as a whole, all that is required here is a summary indication of the use of the key terms informing the different chapters.

Thus as this book is about information retrieval experiment, we need to say what is meant by 'experiment', and how it is related to 'investigation'. An **experiment** is designed to answer the question 'What happens if you do X?', or 'What happens if you do X rather than Y?'. An **investigation** is designed to answer the question 'What happens in System S?', or 'What happens, in general?', or, more tentatively, 'What might be happening in System S or in general?'; most modestly, an investigation is designed to answer the question 'What data can we get which may tell us what might be happening?'

We can think of experiments as suggesting methods of investigation, and investigations as influencing the design of experiments. Experiment is typically hypothesis-guided, investigation hypothesis-generating. However we must allow that the hypothesis underlying an experiment may be implicit rather than explicit, and, further, may be of the weakest kind, namely 'This variable is important and therefore worth study'.

But this difference between experiment and investigation is not the significant one: the essential difference between experiment and investigation is in the application of control in experiment. One cannot test without control. The problem in experiments is then to control variables and not to suppress them.

Experiment implies, and investigation typically involves, measurement and specifically, since an information retrieval system has a purpose, measurement related to system merit in some sense. However investigations may be summed up in measurements of a descriptive rather than evaluative kind. Moreover both experiments and investigations may be concerned with individual system elements only, and with direct measurements of these, so performance measurement proper is indirect rather than direct.

Finally, although experiment and investigation may in principle refer equally to laboratory or operational system studies, in practice there have been few operational system experiments, and experiment and investigation tend to refer to laboratory and operational studies respectively.

In practice, this distinction between experiment and investigation may be difficult to apply to work that has been done in information retrieval, particularly in relation to operational systems. Thus while focusing on experiment, the book does include discussions of better-conducted and more systematic investigations. Equally, many of the points applicable to experiment are applicable to investigation. The word 'test' is therefore used as a global, neutral term to cover both experiment and investigation, and also as a stylistic variant to refer to either when the context makes the particular interpretation clear.