Final version published as:
Journal of Information Science, Vol. 34, No. 4, 439-456 (2008)
DOI: 10.1177/0165551507086989
© 2008 Chartered Institute of Library and Information Professionals
Available here

On the history of evaluation in IR

Stephen Robertson[1]

Microsoft Research

Abstract

This paper is a personal take on the history of evaluation experiments in information retrieval. It describes some of the early experiments that were formative in our understanding, and goes on to discuss the current dominance of TREC (the Text REtrieval Conference) and to assess its impact.

Keywords: information retrieval, evaluation, experimentation, relevance, Cranfield, TREC.

1. Introduction

The foundation of the Institute of Information Scientists in the UK in 1958 coincides closely with the beginning of the notion of experimental evaluation of information retrieval systems. Although there had been some earlier attempts, we usually mark the start of the tradition as the Cranfield experiments, which ran from 1958 to 1966. Information retrieval is commonly regarded as a core component of information science, and systematic empirical evaluation of IR systems probably represents the strongest claim that information science can make to being a science in any traditional sense. There is a nice irony here: the founder of the empirical tradition in IR, the Cranfield librarian Cyril Cleverdon, was not at all a supporter of the Institute. But more of this anon.

As for the present, and despite the concerns of the founders of the Institute, academic information science is now quite closely associated with the former library schools, many of which have adopted titles which include the word 'information'. However, a lot of current work in IR, theoretical and experimental, takes place elsewhere, mainly in computer science departments, though several other academic domains are represented. It probably comes as a considerable surprise to a current PhD student, working on (say) a machine learning optimisation technique applied to search engine ranking, that he or she is in thrall to an experimental tradition founded by a librarian, working with card indexes, a half-century ago.

Thus the history that is the subject of this paper is not too readily defined in terms of institutional or academic boundaries – or national ones. Despite this, it can be seen as a remarkably coherent development of a set of principles and methods. Like all academic subjects it generates argument and disagreement and heated disputation, but there remains a relatively stable common core, which has, despite its limitations (I will argue), served us well over the last 50 years. Furthermore, while its present international status developed out of a US dominance for a large part of that period, the strength of the UK contribution has been remarkable.

[1] *Correspondence to: Stephen Robertson, Microsoft Research 7JJ Thomson Avenue Cambridge CB3 0FB UK; e-Mail: ser@microsoft.com*

In this paper, I will be surveying the history of this experimental tradition, both from the point of view of the ideas involved and also from that of some of the people and groups who contributed, and the environments in which they worked. In these latter respects, the paper will have some focus on the UK, and on groups and projects in which I have been involved myself in one way or another. I make no apology for this personal focus; the paper is as much 'history as I remember it' as formal documented history.

1.1. A note on sources

I will be citing original material throughout this paper, but the single best source for the first half of the period covered (up to 1980) is the book called *Information Retrieval Experiment*, edited by Karen Spärck Jones [1]. The project that has dominated the last fifteen years of experimental work in the field is TREC, the Text REtrieval Conference; this too has been the subject of a recent book [2] which makes a great introduction to a huge volume of work.

2. Cranfield

In the 1950s, many of the ideas (there were indeed many) about how to do information retrieval could be traced back to library classification schemes and their embodiment in card catalogues. Printed indexes, which now survive only as back-of-the-book indexes, were also common; there were also pre-computer forms of mechanization, specifically a whole range of different forms of punched cards used in different ways. But the library classification model was the dominant one. Under this model a document had to be classified/indexed by a human being, and the result of this process was a short description or representation of the document in a more-or-less formal indexing language. The particular form of the indexing language; the kind of analysis that went into constructing it in the first place, and then applying it to a document; the amount of detail you could represent in this way; the specificity of the representation and its divisibility; all of these and more were subjects of fierce argument.

2.1. Arguments

Some of this argument revolved around anecdote. That is, researchers would try to come up with examples to understand the differences between methods, or to demonstrate why one particular system would work and others would not. Two such examples were 'venetian blinds' (as distinct from 'blind Venetians'[2]) and 'lead coatings on copper pipes'[3].

But the core of the argument was generally not empirical, but philosophical. Library classification schemes tend to carry with them entire philosophical world-views, concerning the nature of human knowledge, and to some extent of its representation in documents. But the nature of *language* as such was somewhat separate and peripheral – in some sense the object of a formal classification or indexing system is to avoid all the vagaries and pitfalls of natural language. Of course one has to describe and define the concepts or categories of the scheme in natural language, but the function of this description might be regarded as pedagogic – to help the librarian or user towards an understanding of what the concept or category *really* is, and to see underneath the surface of language. In constructing such a scheme, one might appeal to literary warrant, but that would not absolve one of the responsibility of understanding the concept.

[2] The 39th Doge of Venice at the turn of the thirteenth century, Enrico Dandolo, was blind.

[3] One is somewhat reminded of Noam Chomsky's famous example, dating from the same period, of a sentence that is grammatically correct but apparently meaningless: 'colourless green ideas sleep furiously'.

In the context of these arguments, empiricism (let alone a formal scientific experiment) was a radical notion. There was resistance to a strictly functional view of such schemes, quite apart from the difficulties of first formulating the functionality and then operationalising an experimental framework.

2.2. The beginning of experimentation

Some ideas were being floated in the literature in the 1950s (stimulated by the Royal Society Scientific Information Conference in 1948 [3]), and one or two small experiments were reported in the UK, US and Netherlands. Interest was further developed by the International Conference on Scientific Information in Washington DC in 1958 [4]). But by this time Cyril Cleverdon, Librarian at the then Cranfield College of Aeronautics, had got the bit between his teeth, and started (with funding from the US National Science Foundation) the first of two Cranfield projects, eventually published in 1962 [5]. This project tackled the philosophical divisions in the field head-on, by subjecting four indexing schemes – exemplars of opposing views of how information should be organised – to a direct experimental competition.

Each scheme was to be operated by experts in that scheme. They would construct the scheme itself, index the documents, design the search strategies and undertake the searching. Thus we might regard the necessary human expertise as part of the 'system' in a broad sense. The four schemes were: the Universal Decimal Classification (a hierarchical library classification); an Alphabetical Subject Catalogue (subject headings expressed as phrases); a Faceted Classification Scheme (allowing the construction of complex categories by combination of elements from different facets); and the Uniterm System of Co-ordinate Indexing (terms relatively freely assigned and combined).

Both the methods used and the results obtained provoked much debate and led to the formulation of a second Cranfield project. Methods will be discussed further below, but one of the results is worth noting. On the primary measure of effectiveness used, the four competing systems did not show huge differences; however, the faceted classification scheme came out worst of the four. An analysis of failures then identified an issue with the chosen card-index representation of the faceted classification (a form of so-called 'chain indexing'), and an alternative representation was tried – this boosted the scheme to best of the four. This was somewhat surprising to the proponents of the various schemes: it meant that at least to some degree, the determinants of effectiveness might not be the major principles on which the schemes were based, but the details of implementation.

2.3. Cranfield 2

In terms of both methodology and content, the transition from Cranfield 1 to Cranfield 2 was a great leap forward [6].

For content, the focus was still on 'indexing languages': artificial languages constructed to allow the representation of documents and requests in some partially-formalised way. But rather than treating such a language as a black box defined by some overriding general principle, an attempt was made to disentangle the detailed processes of building a language into small steps and to evaluate the steps. The broad-brush conclusion, that the best thing to do was to search on combinations of words, leaving the natural language words almost untouched, was quite shocking to many people at the time, although it would be much less surprising now, given the predominance of word-based search engines.[4]

But the more significant achievement of Cranfield 2 was to define our notion of the methodology of IR experimentation. The basic ideas of collecting documents and queries were inherited from Cranfield 1, but the biggest change concerned the notion of a good answer. The method in Cranfield 1 was to use a 'source document' – that is, to start from a known document and formulate a question to which that document was a suitable answer. (More specifically, the author of the document was asked to formulate the question which prompted the work to

[4] We may note again that even this word-based searching experiment at Cranfield 2 was done without the aid of computers in any form.

be done in the first place.) Then the criterion for the search system was whether or not it retrieved this source document.

This method was explicitly intended to avoid judgements of relevance. Some earlier work had attempted to obtain relevance judgements by agreement among a group of judges, but found it very difficult. But the source document method was severely criticised, for three main reasons:

1. The queries might be regarded as unrealistic;
2. Retrieval of the source document is not a good test;
3. The resulting measure evaluates recall only, not precision.[5]

The response in Cranfield 2 was to continue to use the method as a way of generating queries, but to deal with items 2 and 3 as follows. Source documents were removed from the test collection, and relevance judgements were made by judges. No attempt was made to get agreement between judges; the judgements for each query were made by a single judge.

Documents to be judged for each query were selected by a variety of methods including manual searching and the use of a form of citation-based indexing. The aim was completeness – to discover all (or nearly all) the relevant documents in the collection. (The extent to which this aim was achieved was the subject of much argument.)

2.4. Cyril and Jason

As an aside from the main theme of this paper, it may be of interest to reflect on the characteristics of some of the disputations involved. In this regard, two people stand out.

Cyril Cleverdon I have already named; the other is Jason Farradane, one of the founders of the Institute of Information Scientists, as well as of the Information Science Department at City University. Farradane had been a technical information officer in the food industry, and before that a chemist, and his views of the field of information science were strongly influenced by this hard-science background. Both men had strongly-held opinions, and both expressed themselves forcibly – and they could not stand each other. Give one of them a platform at a conference or meeting, and the other would be in the audience, just itching to jump up and explain why the speaker had got it all wrong. Farradane conducted his own rather smaller evaluation of his own rather idiosyncratic method of indexing in the middle 1960s, and believed that Cleverdon's application of scientific method to the construction of index languages and to experimental design for the test was fundamentally flawed. Their arguments were fierce and unrelenting, sometimes well beyond the boundaries of civility.[6]

This particular animosity was perhaps extreme in its mixture of the personal and the professional, but it was by no means the only argument engendered by the project. In the US, Don Swanson [7] was almost equally trenchant in his criticisms of the methods and results of Cranfield, albeit in the form of a paper published five years later, rather than a person-to-person public confrontation. Many other authors have contributed arguments and criticisms to the debate, and Cranfield-type methods tend to generate strong reactions both for and against. A later example (now in relation to TREC) can be found in a paper by Blair and the subsequent responses [8].

Nevertheless, the methods pioneered at Cranfield survived and prospered over the next forty years, most directly through the agency of TREC, the Text REtrieval Conference. They did so despite their very real

[5] Recall: proportion of the relevant documents in the collection that are retrieved (ability of the system to find the relevant documents). Precision: proportion of the retrieved documents that are relevant (ability of the system to weed out non-relevant documents). In the present state of the art, we have a large menagerie of measures in common use; most of them derive directly from, or are inspired in some way by, these primitive notions of recall and precision, albeit adapted to measure ranked output rather than set retrieval (ability of the system to rank relevant documents highly).

[6] Personal anecdote: as a masters student of Farradane's in 1967-8, and publishing my first paper, based on my master's dissertation on evaluation measures, in 1969, I incurred by association some of Cleverdon's wrath. It was some years before my relations with Cleverdon recovered from that poor start.

limitations and distortions. There are indeed many things wrong with them, but (I will argue further below) they have also yielded real and valuable results.

3. Some experiments

In this section, I will describe a small selection of experiments that took place in the quarter-century or so following Cranfield 2. This is selective both in respect of the set of experiments and in respect of the details concerning each one. But both the selection and the sequence are chosen to bring out the development of some of the critical ideas in the field.

3.1. SMART

The most significant early series of experiments in computer-based retrieval was that on the SMART[7] system from the very early 1960s [9]. The project, led by Gerard Salton, started at Harvard but was based for most of its life at Cornell University in the US, and continued until the 1990s. Many of the ideas that are currently taken for granted in the web search engines were pioneered there; in particular, the use of purely automatic methods based on the text of the documents, the notion of a scoring function (to measure the extent to which each document matches the query), and the consequent ranking of documents or references to documents for display to the user. It is worth observing that the scoring-and-ranking idea, built into SMART from the very beginning and taken up by many other researchers, did not even begin to appear in any commercial system until the late 1980s, and really only took off with the web search engines in the middle 1990s.

The model used in the SMART system is normally described as the vector space model. It can be conceived as a vector space where the axes are defined by terms (typically words), and each document and each query is represented by a vector of weights – a point in the vector space. Document similarity, or the similarity between a document and a request, is seen as (the reverse of) a distance measure in the space. The scoring function most commonly used in SMART was not based on a Euclidean distance measure, but on cosine correlation, which itself is based on angles between vectors.

The vector components, the weights, can be simple binary values (1='term present', 0='term absent'), but can be more complex, based for example on statistical information. The SMART team used the term frequency in the document (TF), and then combined it with the inverse document frequency (IDF) devised by Spärck Jones (see section 3.3). The idea of relevance feedback was also pioneered in SMART, by Rocchio [10]: using relevance judgements by the searcher to improve the ranking for the current search, or to enhance the indexing of documents for the benefit of subsequent searches.

The early SMART evaluation experiments, from the middle 1960s on, were conducted on a variety of small test collections (perhaps tens or hundreds of documents), either built in-house or re-used from other experiments. In particular, they made use of the Cranfield collection when it became available in machine-readable form (again, see section 3.3 for further discussion of this issue). By the early 1980s, they were using and in some cases constructing somewhat larger test collections. But it is worth noting here, in this age of plentiful and cheap computing power on every desktop, that in 1973 a single computer run on the 1400-document Cranfield collection took 11.2 minutes of processor time and cost $86.22.[8]

[7] It is hard to find the origin of some of the names of systems discussed here. A dictionary of information science defines SMART as System for the Mechanical Analysis and Retrieval of Text. However, in one of the very early reports from Cornell by one of the researchers there, it was said to mean 'Salton's Magic Automatic Retriever of Texts' (almost exactly this form also appears in Wikipedia). I can find no reference by Salton himself to any expansion.

[8] Professors were charged for computer time for their research projects, and had to include such expenses in grant applications. On the above figure, 20 runs might easily cost more than a professor's salary for a month. I am indebted to Donna Harman for this datum.

3.2. Medlars

The Medlars[9] Demand Search Service was one of the early operational computer-based retrieval systems – the predecessor of Medline and PubMed today, covering the medical research literature. Most of the early computer retrieval systems were devoted to the scientific and technical literature, riding on the back of the computerisation of the production of abstracts journals. At the time searching on Medlars was based solely on human-assigned indexing terms, from a controlled indexing language (MeSH[10]). Queries used Boolean logic; output was an unranked set of references. Queries were typically formulated by expert searchers, on the basis of an interaction with the user, a face-to-face reference interview or correspondence by mail. Readers much younger than myself might like to note that this was well before the days of network access, let alone online operation, and the computer was located in the US, at the National Library of Medicine (NLM) outside Washington DC. Searches were run in batches over night, and a printout of results was posted back to the user.

The experiment, conducted by Lancaster [11], was completed in 1968. It was aimed at evaluating the index language itself and the methods and procedures used to index documents and formulate searches. Users were invited to participate at the time of initial contact with NLM; thus the queries were real ones representing real information needs, and the users made relevance judgements in relation to those needs. The study was unusual in including a very detailed failure analysis: an attempt was made to attribute each failure in search (either type: in identifying good documents or in rejecting bad ones) to one or more of a variety of system causes, for example the structure of the index language, indexing policy or practice, the language or logic of query formulation.

One particular result of the Medlars experiment is of some interest. As mentioned, the interaction between the user (medical researcher) and the intermediary (expert searcher) might take place by correspondence or by face-to-face interview – although all searches took place at NLM, intermediaries were located at significant centres around the world, and users could visit them. One experimental question was: how much does it help to have a face-to-face interview? The answer was surprising: it hinders! That is, effectiveness on the searches arising from face-to-face interviews was somewhat worse than on those arising from correspondence. The explanation put forward by the researchers was that a user who writes a letter has to think carefully about his/her information need and how to define and describe it, without being constrained by the language of the system. A user who walks into the office of an intermediary has probably not gone through this process, and the intermediary is liable to go straight into the system language without spending enough time on understanding the information need. Training of intermediaries was changed as a result of this discovery.

3.3. Karen Spärck Jones

For about twenty years from the early 1960s, Karen Spärck Jones (at the Computer Laboratory, Cambridge, UK) conducted a series of computer-based experiments into term clustering and term weighting.

Unlike Cleverdon (or indeed most of the other experimenters working at that time), she did not build her own test collection. That made her particularly receptive to the idea of re-using collections built by other researchers; she was one of the first to perceive the possibilities and difficulties of this mode of operation. As soon as she could after the completion of Cranfield 2, she obtained that collection (now in machine readable form), and her first series of experiments [12] was based entirely on this collection. During this period she invented the form of term weighting based on the number of documents in which the term occurs (inverse document frequency weighting, IDF) [13, 14]. The combination of IDF with within-document term frequency TF, by the SMART team, was to dominate thinking on document-ranking systems for many years, and to have a profound influence on the next generation of term-weighting and document-ranking algorithms.

However, when she subsequently repeated the experiments on further collections, while IDF proved itself, some of the clustering results she had obtained on Cranfield were not confirmed. This encouraged her to think

[9] The Medical Literature Analysis and Retrieval Service.

[10] Medical Subject Headings.

more seriously about the design and construction of test collections. From the middle 1970s until the early 80s, she led a significant effort to clarify and explain the basic paradigm and to improve both the methods and the materials of experimentation in information retrieval. The effort involved many other people, in the UK and elsewhere (including for example Keith van Rijsbergen [15], who has lived and worked in many countries, but spent the bulk of his working life in the UK). The effort had two main outcomes: a proposal and a book.

The second of these, the book *Information Retrieval Experiment* [1], appeared in 1981. It was a collection of papers by a dozen other authors working in the field, with two contributed by Spärck Jones herself. But she also ensured a rare level of coherence in the whole enterprise, by careful planning of the whole, by writing introductory and connecting material, and by suggestion and comment to the other authors. This book was for many years the sole coherent source on how to plan and execute an information retrieval experiment, arguably until the publication of the book on TREC in 2005 [2].

The other outcome of Spärck Jones's work in that period was a design for a new and better test collection. Each of the collections existing at that time had been designed and created for a specific experimental comparison, but typically these same collections were then re-used for many other experiments, for which they were clearly not ideal. So the proposal for the 'ideal' test collection was born. The quotation marks were deliberate, reflecting not so much what might be achievable as an aspiration. The study included some careful preparatory work on a number of important details, such as the pooling technique for obtaining relevance judgements in a relatively large collection [15, 16]. A costing was also made.

But the proposal then hit a wall. The financial support available for basic IR research in the UK in the late seventies was deemed insufficient to fund the project. The 'ideal' test collection idea went onto a back shelf, and sat there for more than a decade, until the TREC project began in the US in the early 1990s (section 4 below). Spärck Jones herself diversified her research into areas relating to natural language processing. However, she returned to very active involvement in experimental IR with the development of TREC.

3.4. Keen

Michael Keen was a member of the Cranfield 2 team, who then went to the US to join the SMART team for a period, and returned to the UK (the College of Librarianship Wales at Aberystwyth) to conduct his own experiments throughout the 1970s. While in the US he made a significant analysis of the various measures of retrieval effectiveness that had evolved over the course of many experiments from Cranfield on [17]. The issue of the choice of measures and their analysis was then and remains now a common concern of researchers in the field – a theme which generates a significant and probably increasing number of new papers every year.

Keen then undertook an evaluation of index languages in the information science domain [18], followed by a study of the searching of printed subject indexes [19]. The first provided some further evidence (confirming the Cranfield 2 conclusion) that straight English words make for a good and effective indexing language, and that it is seriously hard to do better.

Somewhat unlike Cranfield 2 but perhaps more in tune with Cranfield 1, both of Keen's studies, particularly the second on printed subject indexes, are characterised by serious attempts to address questions relating to human searching behaviour. That is, they tended to regard the searcher as being as much part of the 'system' as any set of rules or algorithms. This required them to allow searchers to take the kind of on-the-fly, intelligent decisions that searchers typically take, rather than trying to reduce every aspect of search to the application of predefined rules. In keeping with the usual practice of the times, the model was of 'delegated search' rather than 'end-user' search – in other words, the searcher was assumed to be a professional information scientist acting on behalf of the real user or person needing the information. However, despite this delegated-search idea, these experiments look forward to the more user-oriented search studies of recent years, as much as to the more mechanistic and algorithmic approaches that tend to dominate TREC and SIGIR.

3.5. Belkin and Oddy

This user-orientation also characterizes (although in quite a different fashion) the smaller project of Robert Oddy [20], and his subsequent work with Belkin.

Oddy, doing a PhD in Newcastle UK in the early 1970s, in the Computer Science department, was interested in designing highly interactive systems, and in the difficulties of so-called 'end-user' search (the person with the information need searching on his or her own behalf)[11]. These include the difficulty of constructing a good search strategy, which in the days of Boolean search was quite a technical skill; but also the conceptual difficulty of describing an imprecisely-understood need for information. Oddy designed a prototype system, Thomas[12], which was intended to maintain some kind of model of the user's interest, inferred from the user's actions and responses, which would minimize the user's effort in reaching some information goal, and particularly avoid the search strategy issue. Although he was not able to get real users to use the system, his experiments involved a form of simulation of real-user interaction, using real queries and real relevance judgements obtained from experiments similar to the Medlars experiment described above. The primary objective, which the experiments suggested had been achieved, was to allow the user to obtain similar results to the more conventional system but with less effort.

Nick Belkin, also doing his PhD in the UK in the 1970s[13], had come up with the ASK model, addressing exactly this idea of an imprecisely-understood need [21]: that which stimulates a user to start seeking information is characterised as an Anomalous State of Knowledge. Belkin went to City University, and teamed up with Oddy for a design study [22] of an IR system based on the ASK model. This study did not include any search evaluation tests of the sort described above, but attempted instead to validate and explore certain aspects of the user model, by means in part of small-scale user experiments. Belkin has continued, initially at City and then at Rutgers University after his return in 1985 to the US, to explore the areas of user interaction, user cognitive processes and the task context of information seeking. Some of this work has taken the form of more-or-less conventional retrieval system tests, on the Cranfield-TREC model and in particular within TREC, but the emphasis has always been on the user side, with the challenge of integrating into the experimental methodology a real understanding of user behaviour. The interaction between the user orientation and TREC is explored further below.

3.6. Okapi

The final two projects I would like to discuss in this section came a little later. The Okapi[14] project began in the early 1980s. At that time, there was much interest in OPACs, online public access catalogues, in libraries. One could argue that this was the first sign that information retrieval might become something of interest to the man-in-the-street (or, in the traditional UK phrase, the 'man on the Clapham omnibus'[15]). The web search engine was still well over a decade away, but the notion that someone walking into a library (say, a student in a college library, or even a user of a public library – not a search expert) might reasonably want to do a search on a computer-based system was clearly in the air.

Okapi started life as just such a system, intended to provide access to a library catalogue, primarily in the form of subject searches. It was developed initially at the then Polytechnic of Central London, by Stephen Walker

[11] Technological note: Oddy was working in an environment which gave him access to visual display units, as opposed to the printing terminals that I was using at the same time. However, the interaction mechanisms were strictly keyboard input/ command-line/scrolling character/line display – nothing remotely like windows or mice yet.

[12] Thomas is not an acronym. One source of inspiration for it was Thomas the Tank Engine, a character in a children's book (my thanks to Bob Oddy for this information, which has not been published before). Not to be confused with a later system called THOMAS from CIIR, based on Inquery, for the Library of Congress.

[13] At the same time and place as I did mine, under the supervision of Bertie Brookes at University College London.

[14] 'If this has to be an acronym it stands for "Online keyword access to public information"' [23].

[15] The full phrase, due to the journalist Walter Bagehot in the nineteenth century, is 'the bald-headed man at the back of the Clapham omnibus'. The Clapham in question is a suburb of south London; since my home is in south London, I do use London buses, and I am bald-headed, he might mean me!

and colleagues [23], then moved with Walker in 1989 to join my team at City University[16]. Here it was developed into a general-purpose text search engine, used both with library catalogue data and the somewhat older domain of scientific abstracts. Although it was an experimental system, its *raison d'être* was to provide a real, live service to users, whose behaviour could be studied, or who could be enlisted to take part in further experiments.

Okapi implemented a simple text search based on a free-language query, typed into a box – in a form that would be instantly recognisable today but was then quite unfamiliar. The search mechanisms were based on IR methods developed experimentally (but with no real users) in the 1970s, including weighting of terms and ranking of output. The interface mechanisms were relatively primitive, though some advance on the command-lines of the 1970s.[17]

The service to real user groups allowed a range of different kinds of experiment [24, 25, 26] with user-based evaluation. These included extensive experiments in relevance feedback with automatic query expansion. In general, Okapi users were asked to provide relevance judgements as a matter of course; these could be used both for evaluation and for improving the current search (relevance feedback). On the other hand, it was not normally possible in such an environment to make any kind of attempt to discover all the relevant documents in the collection.

3.7. Croft

Bruce Croft, originally from Australia, and then taking a PhD under Keith van Rijsbergen in Cambridge in the late 1970s, moved on to the US and joined the University of Massachusetts, where he has been ever since. The group built around him at UMass, the Center for Intelligent Information Retrieval (CIIR), has become the strongest IR research group in the US and indeed in the world, supplanting the SMART group at Cornell in this respect.

One characteristic of CIIR's work is the prime status of experiment. A succession of experimental systems and toolkits (I3R, Inquery, Lemur, Indri ...) has provided the basis for this work; but the main determining factor is the attitude to theory and experimentation. CIIR has generated a wide range of exciting ideas and models, but these ideas and models do not count for much (in the view of the group) until they have been subjected to rigorous experimentation. Furthermore, experiments are required to have good baselines – to have a chance of surviving, an idea has to be shown not only to be good, but to be better than previous ideas.

CIIR has tended not to get involved in test collection creation, but to use test collections and materials built elsewhere. A recent book on the work of CIIR [27] contains just one chapter specifically on experimental methodology, devoted to a form of exploratory data analysis. But almost every chapter reports experiments, many using TREC data. However, as we shall see below, the dominant model is now changing – many test collections are built cooperatively over a number of research groups, often involving individuals from the groups taking part as request formulators and relevance judges. In this respect, CIIR now contributes extensively as one of the collaborating groups.

3.8. Postscript to three decades

After the apparent failure of the 'ideal' test collection project, and being closely involved in the Okapi projects, Micheline Beaulieu and I wrote a paper in 1991 [28] in which we argued that we seemed to be moving away from the Cranfield test collection paradigm, towards much more user-oriented work. This turned out to be a bad call, both for the field as a whole and for ourselves. In between the writing and the publication of this paper in

[16] Where I had moved in the late 1970s.

[17] One version of the Okapi interface was based on a vt100 terminal protocol. This had a screen of 24 lines of 80 characters. Each character position in the grid was addressable, and the available characters included a limited number of graphics characters which could be used to construct very limited visual display devices, such as rectangular boxes. Menus could be displayed, and menu items chosen by hitting a single numeric or alphabetic key.

1992, TREC was announced. TREC was the 'ideal' test collection writ large, underwritten by US funds, and planned on a grand scale. While we did not abandon our commitment to live-user experiments, TREC was too good an opportunity to miss, and the Okapi team joined the TREC effort with enthusiasm.

Nevertheless, the fault-line had been there from the beginning and remains there to this day. On the one hand, we can do experiments in a laboratory, characterised by control and artifice. The control enables us to set up formal experimental comparisons and to expect scientifically reliable answers, confirmed by statistical significance tests whose primary requirement is simply enough data; but the artifice requires us to abstract from the real world, to eliminate whatever messiness it might introduce as noise into our experiments. On the other hand, we can seek external validity and attempt to observe real world events in their natural setting, which involves waiting for them to happen and minimising any controls and any observer effects – and therefore get potentially rich but messy and noisy results, probably both unreliable and hard to interpret.

This is not a strict dichotomy, but is very much an opposition. The technology has a strong influence – sometimes technical developments make real-world observation easier, but sometimes they require us to invent new forms of control or put phenomena beyond our reach. In the days of library buildings and physically located catalogues, we could and did stand beside the catalogue or at the entrance to the library and ask users questions (rich data but potential for observer interference). Nowadays we can log all their activities on a previously unimaginable scale (much more data, no observer interference) but it is seriously hard to ask a user *why* they did something which we observed.

In the end, any experimental design is a compromise, a matter of balancing those aspects which the experimenter feels should be realistic with those controls which are felt to be necessary for a good result. Furthermore, the field advances not by deciding on a single best compromise, but through different researchers taking different decisions, and the resulting dialectic.

4. TREC

The initial organiser of TREC, and main architect of its success, is Donna Harman; for the last few years it has been run by Ellen Voorhees [2]. Among the many people who contributed to both its design and its success was Karen Spärck Jones, main author of the 'ideal' test collection proposal a decade and a half earlier.

At the time of writing, TREC is in its sixteenth year. It is an annual competition / collaboration / bake-off / get-together between research groups interested in different aspects of information retrieval. Every year, a set of tasks is defined, broadly information retrieval/search tasks. They may be defined by any or all of the following: the nature of the material to be searched, the type of user, the type of search request, the task context in which the user is operating, the timescale of the information need, the form(s) of interaction allowed, etc. Specific tasks typically persist over several TRECs, but may eventually be replaced by others. At the beginning there were just two tasks, to which all participants contributed; now there are separate tracks each with its own handful of tasks, and participants typically choose one or a small number of tracks.

The entire process is masterminded by the US National Institute of Standards and Technology (NIST), but tracks are largely organised by the participant research groups. The usual process goes something like this. Track co-ordinators and participants agree their tasks and their raw data; this last might consist of (a) a collection of documents, obtained from some external source; (b) a collection of requests or topics, which may also be obtained externally or may be created internally; and (c) a set of relevance judgements. The creation of the topics and/or the relevance assessments may be done by people employed by NIST for this purpose; the main group of assessors is a set of retired news analysts, formerly employed by one of the security agencies.

The documents and requests are distributed to participants, and each participant indexes the documents and runs the requests through some experimental search system. Some set of results is submitted to NIST, and results for each request from all participants are then pooled in some way for relevance assessment. Everyone gets together for a conference in November each year, and discovers (normally only on arrival) how well they have done in the competition.

This broad-brush sketch is intended only to provide an overview; individual tracks and tasks often deviate from this model. A few of the tasks are described further below; but interested readers are referred to the book cited earlier [2] or to the annual reports at [29].

4.1. Ad-hoc retrieval

The user model invoked here is what has now become the most obvious one for search: user has an information need, sits down in front of a system and conducts a search against an existing collection of documents, over a limited time period. This is known in TREC jargon as an 'ad-hoc' search – an earlier more-or-less equivalent term was 'retrospective searching'. System produces a ranked list of items, which the user consults in rank order. Users may judge documents good or bad, but in principle there may be any number of good documents in the collection. (The one significant change in this model from the Cranfield view is that there is now an assumption that each system will rank its results list.) It is often asserted that there is also an assumption that requests are topical or subject-based (documents *about* X); indeed the TREC jargon, which is to call requests 'topics', encourages that view[18]. However, although most of the requests used in TREC (all of those in early TRECs) are indeed topical, there is no necessary requirement of the model that this should be so, or that they should be *purely* topical. In some sense the nature of the requests is determined by the relevance judgements; if the judgements depend on other criteria than pure topicality, then that is the nature of the task.

However, it is fundamental to the model that the judge or assessor should indeed be able to make a judgement on each document, actually a binary one for most of the TREC ad-hoc tasks, and should be able to make the judgement irrespective of the order of presentation of the documents. This last precludes, for example, embedding a criterion of novelty in the usual ad-hoc task relevance judgements (although one track did investigate novelty by making judgements of novelty separated from the relevance judgements).

4.1.1 Methodology

There are many methodological issues here, but a major one concerns the set of documents to be judged for relevance. The ideal since Cranfield has been completeness – discovering all (or in practice most) of the documents in the collection that might be judged relevant. The practice of employing people to create the topics in the first place and then to make judgements allows a significant amount of effort to go into this phase, but certainly does not allow a complete scan of a reasonable size collection by each judge. Therefore relevance judgements have to be selective. One method used extensively in TREC is the pooling method –given the outputs of a range of different systems, judging a pool of the top 100 ranked items from each system is likely to give a reasonable variety of relevant documents. Furthermore, there is some evidence that under some conditions this is likely to result in the discovery of most of the relevant documents in the collection.

What are the required conditions? Well, the evidence suggests that we need to start with a good range of different kinds of systems – preferably, in particular, including some manual systems involving human-designed search strategies and (preferably again) some degree of interaction in the search. Second, we need reasonably deep pools (preferably 100+ from each system, not 10). Third, the collections themselves cannot be too big.

In all these respects, the pooling method is currently under suspicion. Given the increasing range of tasks at TREC, the number and variety of participants in each task has declined. Although some tracks involve an explicit manual or interactive task, such tasks are hard for participants to undertake, and many of the TREC tracks do not attempt such a task. Finally, the scale of the document collections used for TREC has increased hugely since the beginning. One target for the last few years has been to reach or at least approach web scale. TREC is not there yet, but has been taking large steps in that direction.

[18] A typical TREC 'topic' has a short title (which might be used directly as a query), and additional information about the supposed information need under the headings 'description' and 'narrative'. The narrative contains explicit rules for judging the relevance of documents. An example title and description are: Hydroponics—Document will discuss the science of growing plants in water or some substance other than soil.

A further issue is that in some of the tracks (and also in many of the more recent TREC-like initiatives) either or both of the request formulations and relevance judgements are made by the track participants rather than by judges employed by NIST. This is essentially volunteer effort, and effectively precludes the judging of thousands of documents per query.

Thus some effort in the last few years has been devoted to alternatives to the pooling method. Various methods based on sampling are currently being tried. A major motivation for the completeness target has always been the re-use of test collections by other researchers after relevance judgements have been obtained. A challenge for the sampling approaches is to maintain re-usability.

4.1.2 Ranking algorithms

At the heart of a search engine in the modern sense is a scoring-and-ranking algorithm. This may be used for various tasks, but most directly for ad-hoc retrieval, and therefore these algorithms became a major focus for the TREC ad-hoc task.

At the start of TREC, the best-known and most well-established ranking algorithms were those associated with SMART and the vector-space model – essentially cosine correlation, either with simple TF*IDF term weighting or with one of a small number of variants developed as part of SMART. There were several other algorithms in existence, including the Robertson/Karen Spärck Jones relevance weight (RSJ [30]: IDF, no TF, but with a relevance feedback component) and some other approaches implemented in Inquery (based on an inference model of retrieval [31]). But most researchers would treat one of the SMART variants as baseline: that which they would like to improve on.

The early years of TREC proved revolutionary in this respect. Using a further development of an old probabilistic model, we in the Okapi team developed a much extended version of RSJ weighting, now commonly known as (Okapi) BM25: this makes effective use of TF and also document length [32]. Its success at TREC helped stimulate a whole host of other developments in ranking. A new form of model, known as Language Modelling [33], appeared in the late 1990s, and has also been influential. There is absolutely no doubt that today's ranking algorithms are *far* better than those of 1990.

Nowadays, BM25 has something of the status that the vector space model had in 1992. That is, many researchers use BM25 as baseline: that which they try to beat. It is also the case that BM25 has made it into several commercial search systems. However, modern commercial ranking algorithms tend to be much more complex, leveraging different kinds of information. They do not treat documents as undifferentiated blobs of text (which is what both the vector space model and BM25 do), but extract different kinds of evidence which need to be combined in an optimal way for good ranking. Typically in such an environment, an algorithm like BM25 will provide part of the evidence, but will be combined with other clues for the final ranking.

4.2. Feedback

A second strand of experiments that was present from the beginning of TREC is those associated with feedback, specifically with per-request feedback based on relevance judgements. Relevance feedback (RF) is the process of getting the system to learn some characteristics of relevant documents, over and above what can be inferred directly from the request itself. The idea has been around since Rocchio's work on the SMART system in the 1960s [10], and as indicated above, was also implemented in the Okapi system.

It's not so easy to design a good evaluation of an RF system. The SMART researchers devised various methods simulating an initial search, user examination of a few top documents, followed by a new search. A tricky issue concerns the treatment, while evaluating the new search, of the documents already 'seen' by the user. Okapi evaluation with real users concluded that given a simple interface, (a) after having made some relevance judgements (as required anyway by the system) some users would make use of an RF facility;[19] (b) some of these

[19] ...although the usage of the RF facility declined as we moved into the window display era and interfaces became more complex.

users would then mark as relevant items that they would have been unlikely to find in the original search. Such data provides circumstantial evidence that RF can be beneficial, but does not provide a good basis for comparative evaluation of different methods.

The first few rounds of TREC had a task called 'routing'; the model was as follows. We assume we have an existing collection of documents, and for each request we already have (that is, the system has access to) relevance judgements, more-or-less complete, on this existing collection. Now we want to search a new collection – we need to formulate a query based on the original request and the relevance judgements from the old collection. This particular formulation of RF task is relatively clean from an experimental point of view but very unrealistic.

Subsequently, a task called 'adaptive filtering' was developed.[20] Here the model is that documents arrive in the system in a stream; for each request, a decision has to be taken on each incoming document, concerning whether the requester should be notified about it or not. If yes, the requester is then assumed to provide a relevance judgement on it; s/he is also assumed to have provided two or three examples of relevant documents on initially formulating the request. This design, while still artificial in many ways, is clearly more defensible than the routing model.

In many of these experiments, the RF notion has proved extremely powerful. In general, documents judged relevant (or not) by the user or requester provide extremely rich information about the (hidden) user information need, above and beyond what is provided by the stated request. In fact the idea has extended into theories and models; the notion that documents may be judged for relevance to the need becomes not just a mechanism for evaluating systems, but a basic concept in design.[21]

4.2.1 Machine learning

Feedback-related tasks, or similar ones like text categorisation based on human-assigned examples, have also been instrumental in introducing a new community into IR. The machine learning (ML) community thrives on learning from examples, and although much of the early work on routing was based on home-grown methods from the IR world (Rocchio, Okapi...), it became increasingly common to see methods and ideas brought in from the ML community. Nowadays, such ideas and methods are pervasive in IR, not just in the RF context. In particular, modern ad-hoc scoring-and-ranking algorithms often depend on ML methods for optimisation. This is no longer a question of learning about a particular user's underlying information need, but about learning at some level of abstraction, about what characteristics or features of request and document combined are good predictors of a user relevance judgement.

4.3. Tackling the Web

When TREC was announced at the tail end of 1991 the World Wide Web scarcely existed, though the problems of information discovery on the internet were already being recognised. The initial technical challenge at TREC-1 was to index and search a text collection of the order of two gigabytes in size[22] – and this was indeed a seriously hard task for some of the participating groups. But beside today's web search engines (which claim to index 20 billion *pages* and upwards), 2Gb pales into insignificance. Nor is size the only characteristic of the Web which differentiates it from other collections of documents for search purposes. Its heterogeneity, its extremely variable quality, the presence of web spam are all major features. In addition, there are features which (as we have

[20] Old hands will recognise in the idea of filtering an earlier form of search system known as 'current awareness' or 'selective dissemination of information'. The latter phrase was usually abbreviated to SDI – long before Ronald Reagan purloined the acronym for his star wars project.

[21] The first theory to incorporate relevance explicitly was a probabilistic indexing model due to Maron and Kuhns in the US in 1960 [34], around the time that the relevance idea was being operationalised for experimentation but more-or-less independently. Subsequently Rocchio's method [10] and, in the UK, the Robertson/Spärck Jones model [30] tied the two together more firmly.

[22] Actually two collections of approximately 1Gb each.

discovered over the past decade) are positively useful for search purposes, the most obvious one being the linkage between pages.

TREC has tried to tackle some of the issues of web search – both the technical problems of dealing with the sheer size of data and the search effectiveness issues which are the main theme of this paper (and which may also be related to size). A succession of tracks has pushed up the size of experimental collections, starting with the Very Large Collection, through the Web Track, and to the Terabyte Track. This last was based on a crawl of the .gov domain in the US, and resulted in a collection which was actually somewhat less than half a terabyte, but nevertheless much larger than previous TREC collections. At the same time, the issue of query types has been tackled. It is now understood that web search engines receive a variety of types of query. The commonly-cited classification is informational / navigational / transactional; however, at TREC the following types have been used, in addition to traditional topical queries: 'topic distillation' (find a good overview page to browse from); 'home page' (find a home page for e.g. a person, company, product); 'named page' (find a specific page such as a form for a particular purpose).

The amount and variety of evidence that can be used to help web search engines rank effectively, beyond the text of the page, is surprisingly wide. Google has made famous PageRank, which is a query-independent measure of how good a page is, based on linkage; but all web search engines make extensive use of other kinds of evidence. Anchor text (taking snippets of text from a referring page to describe the referred-to page) is probably the strongest single piece of evidence to help home page queries. Usage data, based for example on click-through logs, also appears to be of major importance. All of this poses challenges to the TREC environment – it might be seriously hard for a public project like TREC to get hold of the necessary data to do experiments. At the time of writing, TREC has clearly demonstrated the benefits of anchor text, but has yet to tackle click-through, or web spam (there has, however, been a track devoted to spam email detection).

4.4. Interactive experiments at TREC

Typically, a TREC task involves each group trying out a small number of variants of its own system, with a view to addressing a research question or questions of particular interest to that group. In addition, there is the cross-group comparison, which makes up the competitive element of TREC.

Probably for most participants and observers, the competitive element dominates: the scope for serious internal experimentation to be covered by a set of submissions for a TREC task is somewhat limited, and is probably better done offline, on some existing set of TREC test material.

However, using human searchers in interactive searching experiments within the TREC framework introduces serious problems into the competitive aspects of TREC. The primary issue is that human searchers vary vastly.

In statistical terms, a typical TREC set of results has two main sources of statistical variability, between systems and between requests, and also an interaction effect between the two. The variability between requests is well-known to be large (actually larger than the variability between systems); this in itself is problematic, though it can be dealt with by having enough requests. But if we now include human searchers in the equation, they introduce their own variability, and also probably two more interaction effects, any or all of which may be large. Even if it were feasible to control such variability by numbers, it is very likely that different groups would have access to different types of searcher, so it would be very hard to impose controls sufficient to allow cross-group comparisons.

This lesson was learnt, somewhat painfully, over three or four successive years of the TREC interactive track. In part, therefore, the emphasis has been on developing methodologies which allow a group to set up a relatively controlled internal experiment, and hope to get statistically valid and reliable results on its own research questions. In particular, this means a design which allows the teasing out of the various main and interaction statistical effects mentioned above. Also required was a rich set of additional data-gathering tools, including detailed records or logs of the search processes followed by the searchers, both automatic logs and methods such as think-aloud recordings.

The development of these methods and tools has been impressive, but it is probably true that we have only scratched the surface of what we might learn from them.

Of the research groups most active in the TREC interactive track, we have come across two already. Nick Belkin's group at Rutgers took part in the track in all nine years that it ran, and the Okapi group (from various institutional homes in the UK) in the first seven. Two other groups, one from the US and one from Australia, also took part most years.

4.5. Final remarks on TREC

The extraordinary success of TREC, over a decade-and-a-half, has transformed both the state of the art of information retrieval in general and that of IR experimentation in particular. Even if TREC were to stop today, it would have had the following effects:

- It has stimulated a series of substantial advances in information retrieval techniques, particularly for example in ranking algorithms.
- This stimulation has fed back into the theories and models that underlie the techniques. The most significant advances are those that required the re-thinking of old theories or the development of new ones.
- One mechanism for this stimulation has been the element of explicit, open competition in TREC itself. Although it has been engineered to avoid claims of 'winning', and retains a very strong collaborative atmosphere, relative success at TREC has nevertheless carried considerable kudos, and has also encouraged the rapid spread of good ideas.
- Another mechanism has been the development and provision of test material, of a quality, quantity and variety quite unlike anything that went before. The proportion of research papers in the field that make some use of TREC-derived data is huge.
- In addition to test material, TREC has also greatly encouraged the development of good methods of experimentation. The standard of rigour of experimental methodology has been vastly improved.
- TREC has stimulated a number of imitations. I do not use the word in a pejorative sense; on the contrary, these projects (NTCIR, CLEF, INEX etc.) are themselves producing important results as TREC has done.

These achievements are huge and extraordinary, and deserve to be shouted from the rooftops. Without in anyway belittling them, we need also to be aware of the negative aspects of TREC. I will mention three.

The first concerns its competitive nature. This has in itself been immensely stimulating, in exactly the ways described above; but it also engenders a focus on results (based on effectiveness measures like recall and precision) which sometimes gets in the way of other things. It is very unusual now to see the kind of detailed failure analysis that characterised the early Medlars experiment. Similarly, theories or models tend to be the subject of experimental investigation *only* in terms of the effectiveness of the resulting system. Seen as an application of the usual scientific method, of challenging theories by trying to derive falsifiable consequences, which may then be tested experimentally, this is extremely limited.

The second aspect concerns TREC's laboratory nature. It is a laboratory experiment, and the materials and methods it has generated are materials and methods for laboratory experiments. Any laboratory experiment is an abstraction, based on a set of choices: choices to represent certain aspects of the real world (directly or indirectly) and to ignore others. Choices are made deliberately for the end in view – to isolate certain variables in order to be able to understand them. But choices are also made perforce – because certain aspects of the real world are highly resistant to abstraction. This factor introduces inevitable biases in what is studied: some groups of variables are more amenable than others to abstraction into a laboratory setting. From this point of view, the most important grouping of variables in the IR field is of those that directly concern users and those that do not. On the whole, user variables are resistant to abstraction.

The standard way to deal with this issue in laboratory experiments, inherited from Cranfield, is to reduce the user variables to requests and relevance judgements. This is, we have seen, an extraordinarily powerful abstraction; but it does not allow us to answer all the research questions we might reasonably ask. Parts of TREC, particularly the interactive tracks, have attempted and to some extent succeeded in pushing outside this limitation; but the bias remains. It is very much easier for (say) a PhD student in the field to work on mathematical models and ranking algorithms, using the TREC material in the usual way and never questioning the validity of relevance judgements, than to venture into the jungle of real users with real anomalous states of knowledge.

Other aspects of the real world make their way into TREC, but with more or less difficulty. While TREC was originally designed to allow experiments to scale up to collections of realistic size, at the same time real-world collections have themselves got bigger and more complex. In respect of the Web in particular, as indicated above, TREC has so far addressed only a few of the issues.

The third limitation I would like to address relates to the second, but is distinct. In the process of abstracting from the real world, we define artificial and restricted goals for our systems. The primary system goal addressed in TREC and in most such experiments is the retrieval of items of information. From the point of view of a user engaged in a larger task, or from the point of view of an organisation or institution or community trying to improve communication among its members, the retrieval of items of information must at best be a subgoal. Our understanding of the validity of this as a subgoal, and how it relates to the achievement of wider goals, is limited, and deserves more analysis – theoretical, observational, and experimental.

Experimentation in IR is a large domain. TREC occupies a big part of it, but by no means all.

5. Some current concerns

The various TREC 'imitations' use TREC-like methods to conduct further experiments and to generate new test materials [35]. In part this involves applying essentially the same methods as developed and used for TREC, but to different materials. For example, CLEF (the Cross-Language Evaluation Forum, based in Europe) covers retrieval in multiple mainly European languages, mixed-language collections, and queries and documents in different languages. NTCIR (NII Test Collection for IR Systems, based in Japan) has included material in Japanese, English, Chinese and Korean, and also has some patent data and scientific abstracts. However, some of the tasks force the development of new methods. For example, INEX (Initiative for Evaluation of XML, based in Europe) addresses the question of retrieval from collections of structured documents, where the appropriate unit of retrieval might, depending on the request, be a section or subsection of an original document. Since the traditional Cranfield/TREC method involves treating documents as indivisible units, both for retrieval and for relevance judgement, the INEX tasks require significantly different methods.

Moving a little away from search tasks, DUC (the Document Understanding Conference, US-based) addresses various questions around summarising documents (single documents or sets), which requires very different kinds of evaluation. Within TREC itself, there has been for some time a Question-Answering track: the aim is to generate a specific answer to a question, to be extracted or inferred from a collection of documents, rather than to retrieve (references to) documents which might contain such an answer. In such ways the experimental approach is being extended to a wider range of information-related functions and tasks.

As discussed above, there are interesting new concerns even within the Cranfield/TREC paradigm, for example the discovery of relevant documents in the collection. This is even more critical for many of the TREC imitations, which are often trying to accomplish TREC-like results with much less in the way of resources. In some cases (including some TREC tracks), either or both of request generation and relevance judgements are done by the participating research groups (that is, by the researchers themselves) rather than by assessors employed for the purpose. Thus there is very considerable interest in methods which promise to reduce the amount of effort needed to obtain a useful set of results.

A recurring theme in current work is the issue of measures. There are many things that cause researchers to worry about measures, including for example the case of retrieving parts of structured documents, or the use of

graded rather than binary relevance judgements, or methods such as sampling for choosing which documents to judge. In this last case, one of the concerns is with measures that are robust under incomplete judgements. More broadly, methods of analysis of results, including for example statistical significance analysis, are the subject of much current work.

6. Finally

At an earlier stage in the history of IR experimentation, one might have been tempted to conclude that the basic methodological work was already done – that we might settle down into a common, agreed way of doing experiments. This is far from the case. Although some of the ideas have remained remarkably stable, the field of IR experimentation is as exciting now, and is changing as fast, as it was at the time of my own initial immersion, in the days of Cyril Cleverdon and Jason Farradane.

7. Acknowledgements

I am grateful to the following for extremely useful comments on drafts of this paper: Alan Gilchrist, Donna Harman, Ellen Voorhees, Chris Buckley, Bob Oddy, and an anonymous referee. The historical inaccuracies and misinterpretations, however, are all mine.

8. References

[1] K. Spärck Jones (Editor), *Information retrieval experiment,* Butterworths, London, U.K., 1981.

[2] E.M. Voorhees and D.K. Harman (Editors), *TREC: Experiments and evaluation in information retrieval*, MIT Press, Cambridge MA, 2005.

[3] *The Royal Society Scientific Information Conference, 21 June–2 July 1948: Report and papers submitted*, Royal Society, London, 1948.

[4] *Proceedings of the International Conference on Scientific Information, Washington, DC*. National Academy of Sciences, Washington, DC, 1958.

[5] C.W. Cleverdon, *Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems,* College of Aeronautics, Cranfield, U.K., 1962. Aslib Cranfield Research Project.

[6] C.W. Cleverdon, J. Mills, and E.M. Keen, *Factors determining the performance of indexing systems*, College of Aeronautics, Cranfield, U.K., 1966. (2 vols.) Aslib Cranfield Research Project.

[7] D.R. Swanson, Some unexplained aspects of the Cranfield tests of index language performance. *Library Quarterly*, 41:223–228, 1971.

[8] D.C. Blair, Some thoughts on the reported results of TREC, *Information Processing and Management*, 38:445–451, 2002. Letters to the editor, *Information Processing and Management*, 39:153–159, 2003.

[9] G. Salton (Editor), *The SMART retrieval system: Experiments in automatic document processing*. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[10] J.J. Rocchio, Relevance feedback in information retrieval. In Salton [9], pages 313–323.

[11] F .W. Lancaster, MEDLARS: Report on the evaluation of its operating efficiency, *American Documentation*, 20:119–148, 1969.

[12] K. Spärck Jones, *Automatic keyword classification for information retrieval,* Butterworths, London, U.K., 1971.

[13] K. Spärck Jones, A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, 28:11–21, 1972.

[14] S.E. Robertson, Understanding Inverse Document Frequency: On theoretical arguments for IDF, *Journal of Documentation*, 60:503–520, 2004. Available at http://www.soi.city.ac.uk/~ser/idf.html (Accessed 17th August 2007)

[15] K. Spärck Jones and C. J. van Rijsbergen, *Report on the need for and provision of an 'ideal' information retrieval test collection,* Computing Laboratory, University of Cambridge, Cambridge, U.K., 1975.

[16] K. Spärck Jones and R.G. Bates, *Report on a design study for the 'ideal' information retrieval test collection,* Computing Laboratory, University of Cambridge, Cambridge, U.K., 1977.

[17] E.M. Keen, Evaluation parameters, In Salton [9], pages 74–111.

[18] E.M. Keen, The Aberystwyth index languages test, *Journal of Documentation*, 29:1–35, 1973.

[19] E.M. Keen, *Evaluation of printed subject indexes by laboratory investigation,* BLR&DD, London, U.K., 1978. Final report to the British Library Research and Development Department BLR&D Report 5454.

[20] R.N. Oddy, Information retrieval through man-machine dialogue, *Journal of Documentation*, 33:1–14, 1977.

[21] N. J. Belkin, Anomalous states of knowledge as a basis for information retrieval, *Canadian Journal of Information Science*, 5:133–143, 1980.

[22] N.J. Belkin, R.N. Oddy, and H.M. Brooks, ASK for information retrieval. Part I: Background and theory; Part II: Results of a design study, *Journal of Documentation*, 38:61–71 and 145–164, 1982.

[23] N. Mitev, G. Venner, and S. Walker, *Designing an online public access catalogue; Okapi, a catalogue on a local area network,* The British Library, London, U.K., 1985. Library and Information Research Report no. 39.

[24] S. Walker, The Okapi online catalogue research projects, In C Hildreth, (Editor), *The online catalogue: developments and directions*, pages 84–106. Library Association, London, 1989.

[25] S.E. Robertson, Overview of the Okapi projects [introduction to special issue of Journal of Documentation], *Journal of Documentation*, 53:3–7, 1997.

[26] M. Hancock-Beaulieu, Experiments on interfaces to support query expansion, *Journal of Documentation*, 53:8–19, 1997.

[27] W.B. Croft (Editor), *Advances in information retrieval: recent research from the Center for Intelligent Information Retrieval*. Kluwer, Boston MA, 2000.

[28] S.E. Robertson and M. Hancock-Beaulieu, On the evaluation of IR systems, *Information Processing and Management*, 28:457–466, 1992.

[29] NIST, National Institute of Standards and Technology, Text REtrieval Conference, http: //trec.nist.gov/ (Accessed 13 August 2007).

[30] S.E. Robertson and K. Spärck Jones, Relevance weighting of search terms, *Journal of the American Society for Information Science*, 27:129–146, 1976, Available at http://www.soi.city.ac.uk/~ser/papers/RSJ76.pdf (Accessed 17th August 2007)

[31] H.R. Turtle and W.B. Croft, Evaluation of an inference-network based retrieval model, *ACM Transactions on Information Systems*, 9:187–222, 1991.

[32] K. Spärck Jones, S. Walker, and S. E. Robertson, A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, 36:779–808(Part 1) and 809–840 (Part 2), 2000, http://www.soi.city.ac.uk/~ser/blockbuster.html (Accessed 17th August 2007)

[33] W.B. Croft and J. Lafferty (Editors), *Language Modelling for Information Retrieval,* Kluwer, 2003.

[34] M.E. Maron and J.L. Kuhns, On relevance, probabilistic indexing and information retrieval, *Journal of the ACM*, 7:216–244, 1960.

[35] Websites for various initiatives: CLEF: http://clef.iei.pi.cnr.it/; NTCIR: http://research.nii.ac.jp/ntcir/; INEX: http://inex.is.informatik.uni-duisburg.de/; DUC: http://duc.nist.gov/; TREC QA: http://trec.nist.gov/data/qamain.html (All accessed 6th September 2007).