

Evaluation in Information Retrieval

Edie Rasmussen
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA 15260
412/624-9459
erasmus@mail.sis.pitt.edu

ABSTRACT

Over its 40 year history information retrieval evaluation evolved with an emphasis on laboratory experimentation, the “Cranfield paradigm”, in response to early demands for experimental rigour. The current and highly successful TREC experiments follow this model. However the demand for results that are robust and scalable has caused an emphasis on system-centered evaluation at the expense of user-centered evaluation.

1. INTRODUCTION

In information retrieval, evaluation rules. Through what might be viewed as a series of historical accidents, evaluation has come to play a critical role in information retrieval research. It is expected that a researcher exploring some aspect of information retrieval will provide evidence of evaluation---not, after all, an unrealistic expectation. But the field has also developed a strong sense of what is an acceptable method of evaluation, as well as the kinds of performance improvements that are needed to show that a particular technique is significant. Few fields can claim to have devoted so much attention to the definition of experimental rigour. The view of the IR community on evaluation has imposed a set of standards which have been very effective in moving the field forward, but at a price---a loss of creativity in developing new evaluation techniques and an emphasis on laboratory evaluation which has focussed on the system rather than the user, or user plus system.

2. EARLY WORK

The Cranfield tests are undoubtedly the best known of the early information retrieval experiments, and those that were most significant in shaping the form which evaluation was to take for the next forty years, to the extent that we speak today of the “Cranfield paradigm” for information retrieval evaluation. The Cranfield tests date to the early days of computer-based text processing, and were conducted by Cyril Cleverdon and a group of researchers at the Cranfield College of Aeronautics, primarily as a test of indexing techniques. The first set of experiments, conducted in 1958-1962, tested four indexing systems, including the Uniterm system of coordinate indexing. The results, which suggested that two newer techniques, faceted classification and Uniterms, outperformed two traditional methods, UDC and alphabetical subject headings, seemed threatening to advocates of traditional methods, attracted widespread attention and were highly controversial. The controversy led to a critical examination of the methodology used, and the accusation that the results were an artifact of the research design. As a result of this

examination, Cleverdon devised a second set of experiments, with emphasis on rigour and a laboratory model. Cranfield II used 1400 documents and 279 queries. Research papers were used to instantiate queries, and the document collection was comprised of the pooled references. Relevance judgments were made by the question providers and augmented by students who screened the entire collection (Spark Jones, 1981a). Finally, recall and precision were the evaluation metrics used in the experiments.

Cranfield II thus led to the basic model for information retrieval experimentation in common use today: a document collection, a set of queries and associated relevance judgements, and measurement based on precision and recall. In terms of findings, the Cranfield experiments and a series of experiments on the SMART system (Salton, 1981) showed that the interest of the time in complex indexing systems was misguided, and that in general simpler indexing systems worked as well as more complex techniques. Spark Jones commented in 1981:

What, then is the Cranfield legacy? First, and most specifically, it has proved very difficult to undermine the major result of Cleverdon’s work, namely that indexing languages, including natural language, tend to perform much the same: the gross substantive result of the research remains true. Second, methodologically, Cranfield 2, whatever its particular defects, clearly indicated what experimental standards ought to be sought. Third, our whole view of information retrieval systems and how we should study them has been manifestly influenced, almost entirely for the good, by Cranfield. (Spark Jones, 1981a, p. 283).

3. CONSOLIDATING CRANFIELD

Summarizing IR research after 20 years of evaluation, Sparck Jones noted that there had been a general advance in the quality of retrieval experiments. The lessons learned included the need to use real data, to use enough data (both for queries and documents), to use multiple data collections, to use appropriate performance measures, to design tests carefully (e.g. by using Latin square designs for assigning tasks to people), and to evaluate findings adequately, for example by applying significance tests (Sparck Jones, 1981b). The need for statistical inference is reiterated by Tague[-Sutcliffe] (1981) in her invaluable guide, “The pragmatics of information retrieval experimentation”. Tague’s paper serves as a cookbook for the novice researcher; it addresses 12 questions on research design which, if answered well, constitute a successful experimental design for information retrieval. Ten years after issuing her first set of questions, Tague-Sutcliffe revisited her advice to researchers in the context of a broader idea of information

retrieval and what she termed a “paradigmatic shift ... in the research front, to user-center from system-centered models”. (Tague-Sutcliffe, 1992, p. 467).

In terms of performance, the Cranfield studies and those that followed showed retrieval results described by Sparck Jones as “pretty middling”. Even in the best cases it was difficult to improve performance above a broad 50% precision –50% recall level (Sparck Jones, 1981b). Of the many techniques employed, term weighting and relevance feedback were the methods that seemed most consistently promising in terms of performance improvements.

With the evaluation methodologies developed through the Cranfield studies firmly established, information retrieval research continued to develop, exploring new models and new parameters within those models. Sparck Jones (1974) had set a standard for the field in her analysis of what constituted a reasonable performance improvement. In her view, performance differences of less than 5% should be disregarded; and differences, even if statistically significant, were termed “noticeable” if the difference was 5-10%, and “material” if greater than 10%. This somewhat arbitrary pronouncement had the effect of discouraging a flood of papers reporting trivial performance improvements and kept the focus on substantive improvements.

Through the 1970’s and 1980’s the “laboratory” for information retrieval research expanded, with new test collections, query sets and relevance judgements. The collections grew steadily larger, though still falling far short of those found in operational systems. Among the two largest were the NPL (or Vaswani) collection with 11,429 documents, and the UKCIS collection, with 27,361. Documents in the laboratory were still, however, document surrogates rather than full text documents: titles and abstracts, or in some cases merely titles.

Among the many small-scale, document surrogate, laboratory experiments being performed, the STAIRS study by Blair and Maron (1985) stood out as a rare example of a study of operational retrieval in a full-text system. Like the earlier Cranfield study, the STAIRS study generated controversy because the results were unpalatable to many. The study used lawyers and paralegals involved in a complex lawsuit to search for documents within the full text of the collection of legal documents surrounding the case. The searchers provided their own queries and relevance judgments, and were encouraged to search until they found enough useful documents to defend the case (which they estimated as at least 75% of the relevant documents and 100% of the vital documents). However, the mean for recall observed in the study was 20%, a figure well below the “middling” results common in laboratory experiments. As Blair (1996) later pointed out, it seemed that system growth (resulting in part from the “information landfill” approach arising from the low cost of information storage) was making effective, high recall retrieval harder to attain. In fact, performance was worsening faster than it could be improved.

4. EVALUATION AND INNOVATION

The STAIRS study aside, careful attention to evaluation had resulted in what seemed to be irrefutable evidence that information retrieval researchers, applying techniques such as the

vector space and probabilistic models, relevance feedback, and query expansion, could build systems which significantly outperformed the standard Boolean model. However there was a growing concern in the field over the schism between research and practice. As Sparck Jones commented in 1981:

It has taken quite a long time for the test results of the 1960s to filter through into practice, and we must therefore expect as slow responses to the experiments which have been carried out in the 1970s, ..., or which should be carried out in the 1980s. (Sparck Jones, 1981b, p. 250).

The major vendors of information retrieval systems such as Dialog and BRS continued to use a basic Boolean model. Developers of off-the-shelf software for information retrieval rarely implemented ranked output systems, and when they did, as in the case of Personal Library Software (PLS), met with a mixed reaction. (PLS, derived from the SIRE laboratory system, was the first off-the-shelf software based on non-Boolean IR models, and was viewed with some suspicion by searchers trained on Boolean models.) The frustration of researchers in the field with their inability to affect the status quo is shown clearly in a special issue of *Information Processing and Management* in 1988 on “Trends in Research on Information Retrieval: the Potential for Improvements in Conventional Boolean Retrieval Systems”. In the issue the editor, Radecki (1988) and the contributors summarized what information retrieval research had accomplished--what it had to offer to the field. As a counterpoint, Smit and Kochen (1988) reported on a survey of database vendors who were asked about the “impediments to innovation” in their systems, and found that there was a lack of knowledge about retrieval enhancements reported in the literature, and that many of those who were aware of them did not know how to implement them.

A few years later, Ledwith (1992) provided the viewpoint of a developer for a large commercial online system (STN). He considered the difficulty of extrapolating the results of IR experiments to the searching of large scientific databases, and voiced three concerns: the size and composition of the test collections and scalability to files 750 times larger than the largest test collection; the general nature of the test queries which would produce intractable retrieval in real-world collections; and whether the improvement in performance (if attained) would be worth the cost of implementation.

Thirty years of IR experimentation in the Cranfield model, then, had resulted in a limited success. There was confidence within the IR community that real performance improvements could be achieved; yet the transfer of the technology to the commercial section was extremely limited. The most compelling reasons put forward for this failure to implement the findings were first, lack of information on the part of those who might implement them, and second, skepticism about scalability among those who understood the innovations being proposed.

5. TREC-KING FORWARD

The most dramatic response to criticism of the real-world relevance of small-scale Cranfield-style experimentation was the initiation in 1992 of the Text Retrieval Conference (TREC), hosted by the National Institute of Standards and Technology

(NIST). TREC, open to a steadily growing group of researchers world-wide, provided the infrastructure for large-scale IR evaluation, and in doing so, aimed to encourage communication and technology transfer among the industrial, academic and government sectors. TREC also had as a goal the improvement and dissemination of evaluation techniques, and one of its strengths has been the use of standard routines to analyze the results of query runs, eliminating the variance in analysis of data that exists in the IR literature in general.

For TREC participants, NIST provided gigabytes of data and a set of realistic queries (posed as topics with a narrative statement plus an indication of the characteristics of a relevant document), and obtained relevance assessments based on pooling and manually examining the top-ranked documents from runs across multiple systems. Results were submitted to NIST and analyzed there, and an annual meeting of each year's participants ensured communication across sectors and stimulated technology transfer. This approach was particularly successful in ensuring early adoption of successful innovations by other TREC participants, so that the first several years in particular were marked by overall increases in performance.

TREC has improved on the Cranfield era laboratory environment in many ways. The test collection is very much larger than any used in prior work, and it is heterogeneous, composed of documents from many sources in order to eliminate the bias that comes from using a single-source test collection. While the relevance assessments are based on the pooling method rather than comprehensive, an analysis has been conducted to determine that the impact of unjudged relevant documents does not bias the results (Zobel, 1998). Standardized analysis of data from the participants by NIST through a set of "trec-eval" routines ensures consistency of interpretation, in contrast to prior work in which it was difficult to compare results across studies since a variety of techniques were used to produce recall/precision graphs. The continuity provided by TREC over its ten-year history has also allowed cumulative data, time and initiative for some deeper analysis; for example, Voorhees in a recent study addressed the question "How many queries is enough?", that is, the effect of topic set size on retrieval experiment error (Voorhees, 2002). It is surprising that the question had not been addressed before this, and a measure of the value of the TREC data and infrastructure that it has been addressed now.

The retrospective retrieval task, regarded historically as the fundamental task of information retrieval over 40 years of experimentation, was one of two initial TREC tasks (referred to as "ad hoc retrieval"). This task was discontinued after 1999, when it was noted that while performance had improved dramatically since the inception of TREC, it had leveled off and it was felt that the community was not learning enough from the runs for substantive further improvements (Voorhees & Harman, 1999). The second task, though less studied prior to TREC, was also regarded as a classic: the SDI (selective dissemination of information), or (as it was called in TREC) "filtering" task, which was based on the idea of optimizing a query to extract information from a dynamic data stream.

A significant success of the TREC environment has been its ability to broaden its definition of information retrieval tasks with

the addition of new tracks, begun in TREC-3. Bringing together a community of information retrieval researchers and providing a hospitable environment for new ideas inevitably led to the identification of new tasks, and the formation of sub-communities to study them. Some of these tracks were short-lived (e.g. the "confusion" track and the "high precision" track) while others had a longer life span, the "interactive" and "cross-language" track being among the longest-lived of the current tracks. Recent introductions are the "question-answering" track, the "Web" track, and the "video" track, the latter requiring development of a video test bed. (Voorhees, 2001). With a stable environment from year to year, TREC provides an unprecedented opportunity to take the time to define the task and experiment with evaluation techniques and metrics in a collegial way. Perhaps the best example of this is the Interactive Track, which over its nine-year span has struggled with the problems of user-centered evaluation, and is currently working on developing a metrics-based comparison of interactive systems.

Measured by many variables, TREC has been successful. The conference series has had an enormous impact on information retrieval research. It has met its goals of fostering communication and technology transfer, providing a scaled-up test bed for research, and promoting standardization in IR evaluation. It has provided a body of data to answer questions about experimental design. A perhaps-unanticipated bonus has been its encouragement of discussion about a wider range of IR tasks. There is no doubt that the IR community owes a great deal to NIST and the TREC team for providing the necessary infrastructure and venue.

Criticism of TREC has for the most part focused on minor, perhaps unavoidable outcomes. It has been suggested that the TREC cycle, with its tight deadlines for submission of data, discourages any significant introspection about individual results, particularly any form of failure analysis. The emphasis on query averages results in a failure to examine or explain system differences for individual queries, focusing instead on "robust" evaluation measures such as the precision-recall curve. Dissemination of results beyond the immediate TREC community has also been identified as a problem, since the tight annual schedule sometimes discourages publication of TREC results in the journal or conference literature.

Precision and recall have remained the evaluation measures of choice, and both have been sought in equal measure. (An early exception to this was van Rijsbergen's E-measure, which allowed the researcher to set a parameter rewarding either precision or recall to fine tune system performance.) Some research has recognized that these may not be the optimum measures, with the value of recall being questioned in many common situations where the searcher does not really need everything on a topic. This has been particularly evident in the Web retrieval environment, where the massive size and of the database and the redundancy in the information it contains makes recall not only unnecessary, but also undesirable.

The major criticism of TREC, however, has been that it is a "child of the Cranfield paradigm". As in the Cranfield model, two simplifying assumptions are made: one, that relevance is binary and static, and two, that for the purpose of evaluating the IR

system, the user and the user's interaction with the system can be ignored. (An exception is found in TREC's Interactive Track, but even here researchers are struggling to find the appropriate method and metrics.) The impact of these two assumptions has been the subject of considerable study.

6. THE ISSUE OF RELEVANCE

Relevance is the pillar on which information retrieval evaluation stands. A fairly narrow view of relevance was developed in support of laboratory experimentation. For this purpose, relevance was considered to be topical relevance, a subject relationship between document and query. It was viewed as a judgment that could be made by an individual external to the search process, and one which was binary: a document was relevant, or it was not. This view of relevance was invoked not because it was a valid mirror of reality, but as an assumption to simplify the experimentation.

While this simplified view of relevance was adopted as the basis for experimentation from the earliest studies, researchers continued to explore its nature. The literature on relevance is extensive and has been summarized by Schamber (1994) and Mizzaro (1997). Relevance is recognized as one of the central concepts in information retrieval, and one that is not well understood. It is seen as a relationship between any one of a document, surrogate, or information, and a problem, information need, request, or query (Mizzaro, 1997). Researchers such as Schamber (1994) and Barry (1994) have looked at the criteria on which users base their relevance judgments, finding them to be subjective, dynamic, and multi-dimensional. More recently, researchers have examined the meaning of relevance and users' judgements in new retrieval environments, for instance in image retrieval (Choi and Rasmussen, 2002) and the World Wide Web (Rieh & Belkin, 2000). It seems clear that as new tasks are identified, a first consideration must be to establish what relevance means in the context of that task, before comprehensive evaluation methods can be established. This is particularly true if the goal is to support a user-centered view of evaluation, where relevance judgements that are binary, static, and query-based may be unattainable.

7. THE ISSUE OF INTERACTION

According to Harter and Hert, "The omission of the user from the traditional IR model, whether it is made explicit or not, stems directly from the user's absence from the Cranfield instrument." (1997, p. 14). They identify several validity and reliability issues arising from the Cranfield model: its implication of a batch approach to IR, the real-life searching and stopping behaviour of real users, human factors issues (for instance in information display), oversimplification of the concept of information need, problems with recall as a measure and with relevance viewed as a binary and static variable. There is also an implicit assumption that the queries, users and documents are representative of some larger population.

In discussing "the dilemma of measurement in information retrieval research", Ellis (1996) argues that the historical view of information retrieval as a discipline made tractable by quantification has in fact restricted its development both theoretically and methodologically. He identifies three qualitative research approaches: cognitive, behavioural, and affective which

may give a richer picture of the complexity of the retrieval interaction than the rigidity of the Cranfield approach.

Other researchers have tried to incorporate the role of interaction within the Cranfield paradigm. This approach is exemplified by the Interactive Track in TREC, which over a period of nine years has worked on developing a standardized and quantifiable approach. Participants in the track address methodological issues, and tend to explore very different aspects of the retrieval task, so that unlike other tracks there is very little comparability across systems.

8. TASKS AND QUERIES

For researchers in related fields considering the adoption of the Cranfield/TREC paradigms, task analysis is another priority. Most information retrieval evaluation, certainly the early evaluation, was based on the idea of a single retrieval task: a retrospective search, that is, a search through a static collection of documents for all those documents which matched a query, or met a certain information need. The measures of success were recall and precision, measuring the ability of the system to deliver all the relative documents and only the relative documents. Subsequent work, notably the TREC experiments, has recognized other tasks, such as filtering, in which items are selected from a dynamic stream of data, high precision, in which obtaining a few highly relevant items is the goal, and question-answering, in which a snippet of text containing the answer to a specific question is sought.

9. LESSONS FROM INFORMATION RETRIEVAL EVALUATION

There are many good models to be found in information retrieval evaluation. The benefits to the community of sharing well featured, tunable search engines (such as SMART and InQuery) and other components are well demonstrated. The development of shared large-scale test collections in the Cranfield mode greatly furthered research in the field. The ready availability of systems and data has enabled researchers to devote their energies to what is new and innovative, rather than basic system building. Early agreement on norms for evaluation also simplified research design, decreased researcher bias, and facilitated comparisons. The TREC style of community-based research has been shown to "hustle research communities very effectively, leading to significant and rapid improvement in task performance" (Sparck Jones, 1995, p. 313).

On the other hand, information retrieval evaluation has been fairly narrowly focussed in terms of tasks. It lacks a taxonomy of tasks and suites for evaluation. It also lacks a taxonomy of queries that would permit analysis at the query type rather than the system level. While much more is known about the nature of relevance than in the early Cranfield days, an effective means of integrating a more complex view of relevance into the evaluation process has not been developed. Researchers are still working to develop appropriate measures for interactive evaluation, including standard instruments for users. Because evaluation is based on query averages for precision and recall, there is little emphasis on the flexibility to match user needs and outcomes. From the user's perspective, process and strategy may be as important as outcomes, and mechanisms to study process and measures to evaluate it are still lacking.

While the resources for IR evaluation, in terms of systems, tools, and data environments, are readily shared, this takes place largely on an ad hoc and inter-personal level. One of the goals identified by an NSF workshop on toolkits for information retrieval was the development of a more formal infrastructure such as a clearinghouse to facilitate shared research environments (Korfhage & Rasmussen, 1998). More formal sharing of IR tools would help to broaden the retrieval community and promote a high evaluation standard.

Researchers in information retrieval have devoted a significant amount of time and energy to developing good, standardized evaluation techniques, perhaps inevitably compromising on complexity in the process, particularly with respect to user-centered evaluation. For a new field in the early stages of reaching consensus on evaluation techniques and metrics, the time may be right to explore system-centered and user-centered evaluation in parallel.

10. REFERENCES

- [1] Barry, C.L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, 45: 149-159.
- [2] Blair, D.C. (1996). STAIRS Redux: thoughts on the STAIRS Evaluation, ten years after. *Journal of the American Society for Information Science*, 47(1): 4-22.
- [3] Blair, D.C. & Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communications of the ACM*, 20: 289-299.
- [4] Choi, Y. & Rasmussen, E. (2001). Users' relevance criteria in image retrieval in American history. *Information Processing & Management* 38: 695-726.
- [5] Ellis, D. (1996). The dilemma of measurement in information retrieval research. *Journal of the American Society for Information Science* 47(1); 23-36.
- [6] Harter, S. & Hert, C.A. (1997). Evaluation of information retrieval systems: approaches, issues and methods. *Annual Review of Information Science and Technology* 32: 3- 94.
- [7] Korfhage, R. & Rasmussen, E. (1998). *NSF Invitational Workshop on Information Retrieval Tools*. <http://www.interact.nsf.gov/cise/conferences.nsf/wkshpdescription/idmirtools?OpenDocument>
- [8] Ledwith, R. (1992). On the difficulties of applying the results of information retrieval research to aid in the searching of large scientific databases. *Information Processing & Management* 28(4): 451-455.
- [9] Mizzaro, S. (1997). Relevance: the whole history. *Journal of the American Society for Information Science* 48(9): 810-832.
- [10] Radecki, T. (1988). Trends in research on information retrieval---the potential for improvements in conventional Boolean retrieval systems. *Information Processing & Management* 24: 219-227.
- [11] Rieh, S.Y. & Belkin, N.J. (2000). Interaction on the Web: scholars' judgment of information quality and cognitive authority. In: *ASIS 2000: Proceedings of the 63rd ASIS Annual Meeting, Chicago, IL*. (N.K. Roderer and D.H. Kraft, eds.). Medford, NJ: Information Today. Pp. 25-38.
- [12] Salton, G. (1981). The Smart environment for retrieval system evaluation---advantages and problem areas. In: *Information Retrieval Experiment*. (K. Sparck Jones, ed.). London: Butterworths. pp. 316-329.
- [13] Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology* 29: 3-48.
- [14] Smit, P.H. & Kochen, M. (1988). Information impediments to innovation of on-line database vendors. *Information Processing & Management* 24(3): 229-241.
- [15] Sparck Jones, K. (1974). Automatic indexing. *Journal of Documentation* 30(4): 393-432.
- [16] Sparck Jones, K. (1981a). The Cranfield tests. In: *Information Retrieval Experiment*. (K. Sparck Jones, ed.). London: Butterworths. pp. 256-284.
- [17] Sparck Jones, K. (1981b). Retrieval system tests 1958-1978, In: *Information Retrieval Experiment*. (K. Sparck Jones, ed.). London: Butterworths. pp. 213-255.
- [18] Sparck Jones, K. (1995). Reflections on TREC. *Information Processing & Management* 31(3): 291-314.
- [19] Tague[-Sutcliffe], J. (1981). The pragmatics of information retrieval experimentation. In: *Information Retrieval Experiment*. (K. Sparck Jones, ed.). London: Butterworths. pp. 59-102
- [20] Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, 28(4): 467-490.
- [21] Voorhees, E. (2002). The effect of topic set size on retrieval experiment error. In: *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland* (??, et al, eds.). ACM: ?? pp. 316-323.
- [22] Voorhees, E. & Harman, D. (1999). Overview of the eighth Text REtrieval Conference (TREC-8). In: *The Eighth Text REtrieval Conference (TREC-1999)* (E.M Voorhees and D.K. Harman, Eds.) NIST Special Publication 500-246. Available at: http://trec.nist.gov/pubs/trec8/papers/overview_8.pdf
- [23] Voorhees, E. & Harman, D. (2001). Overview of TREC 2001. In: *The Tenth Text REtrieval Conference (TREC-2001)* (E.M Voorhees and D.K. Harman, Eds.) NIST Special Publication 500-250. Available at: http://trec.nist.gov/pubs/trec10/papers/overview_10.pdf
- [24] Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia* (W.B. Croft, et al, eds.). ACM: New York. Pp. 307-314