Retrieval and relevance:
on the evaluation of IR systems

The ISI Lazerow Lecture,

University of California, Los Angeles

Stephen E. Robertson

Centre for Interactive Systems Research
Department of Information Science
City University
Northampton Square
London EC1V 0HB
UK

November 11th, 1993

Abstract

An overview of the history of the testing and evaluation of information retrieval systems, from the late fifties and Cranfield to the present day and TREC, is presented. Some themes are highlighted, particularly the idea of a test collection and more recent work on interactive systems. This legacy contrasts with the situation in system evaluation in other areas (*e.g.* expert systems or interface design).

The definition of the "system" presents problems in IR experimentation. In particular, some aspects of the user's mental models and/or cognitive processes should be included in the system, if the task is taken to be helping the user to resolve her/his ASK. This point combines with the dominance of interactive systems to reinforce the polarity between laboratory and operational experiments, and consequently the difficulty of designing good IR experiments.

The concept of relevance and its uses in information retrieval is discussed in this context. Finally, the experimental environment based on the Okapi system at the City University is described, and some results are presented.

Ladies and Gentlemen,

I am honoured to have been invited to deliver the 1993 ISI Lazerow lecture, and it gives me very great pleasure to be talking to you today.

1 Introduction

Information retrieval systems have been around for at least two-and-a-half thousand years; mechanized systems for around 60; computer-based for around 35. We have also been evaluating IR systems for about 35 years. We could argue about which of the two started first, but the history of evaluation certainly started independently of the history of computer-based systems; the first systems to be evaluated were manual ones.

This situation contrasts sharply with that relating to most other classes of computer-based systems, particularly those which show some similarity in difficulty of evaluation to IR systems. For example, neither the evaluation of expert systems, nor that of user interfaces, has any remotely comparable legacy.

Such historical baggage can be both an advantage and a burden. Consider, for example, the situation of a doctoral student whose project involves the design and implementation of a new or modified retrieval method or technique. She or he must undertake some evaluation of the method, and this entails extensive assessment of the evaluation literature and the design of an experiment or experiments according to a large base of state-of-the-art. Certainly some students of my acquaintance would prefer a less exacting or constraining history!

2 Some history

Conflicting philosophies

The first substantial comparative test of information retrieval systems was the first Cranfield test, between 1958 and 1962.[1] This was designed to com-

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

pare approaches based on rival philosophies of IR: each of the four systems under test was taken as representing one of these philosophies.

In the event, the differences between the four systems were small (the test therefore failed to satisfy the proponents of *any* system!). Furthermore, it became evident that system performance was strongly affected by the details of implementation: the faceted classification scheme initially performed the worst of the four systems; when they made a change in the method of generating the indexes, facet moved from worst to best. In the (comparatively few) tests of rival philosophies made since then, this general effect has been observed many times.

Techniques

The second Cranfield experiment[2] therefore concentrated on the detailed mechanisms of index language construction, and thereby set the tone for most subsequent testing. At around the same time, the big Medlars evaluation[3] tested only one system, but concentrated on a detailed failure analysis of the results, ascribing failures to retrieve relevant documents (or, conversely the retrieval of non-relevant documents) to a variety of causes such as index language design, specific index language elements, indexing rules, specific indexing decisions, search formulation rules *etc.*

Empirical exploration, combinatorial experiments

A concentration on techniques, in a laboratory experiment where everything is repeatable, suggests an exploration of the possibilities; any technique that can be thought of can be tried, irrespective of whether it has any basis in theory. Furthermore, techniques that can be used in combination must be tried in combination (particularly if there is no theory of interaction); this

[2] C.W. Cleverdon, J. Mills & E.M. Keen, *Factors determining the performance of indexing systems.* (2 vols.) Aslib Cranfield Research Project. Cranfield, U.K.: College of Aeronautics, 1966. For a discussion of both Cranfield projects, see also K. Sparck Jones, The Cranfield tests. In: K. Sparck Jones (ed.), *Information retrieval experiment.* London: Butterworths, 1981, pp 256–284.

[3] F.W. Lancaster, *Evaluation of the Medlars demand search service.* Bethesda, Md: National Library of Medicine, 1968.

leads to combinatorial experiments (trying out all combinations of a set of variable system elements).

This approach is best exemplified by the early (and continuing) experiments on the SMART retrieval system,[4] and also by the early experiments by Sparck Jones on clustering and term weighting.[5]

Test collections and batch systems

The SMART and Sparck Jones experiments are also characterised by the repeated use of test collections of documents, requests and relevance judgements, and by the view of information retrieval as a batch process. In some sense these two ideas go together (and with combinatorial experiments); in order to do such experiments, it is necessary to have stable sets of documents and requests, and to have already collected all the relevance judgements that will be needed. Interaction with the user is not really compatible with this approach.

The test collections used for those experiments and for most subsequent ones (for example the Cranfield collection) have generally been build for specific experiments, and most certainly not for the range of experiments to which they have been subjected. In the UK in the seventies, there was a movement to design and construct a bigger and better test collection, specifically for re-use in a wide range of experiments: the so-called 'Ideal' test collection[6] (even then there were quote marks around the word 'ideal'!).

A considerable amount of work was done on the 'ideal' test collection design, but when it came to the crunch, we failed to find the combination of will and resource to build it. This was a great disappointment at the time, although it was clear that the concept of test collection experiments had some major limitations (concerning which more later).

[4] see for example G. Salton, The Smart environment for retrieval system evaluation. In: K. Sparck Jones (ed.), 1981, *op. cit.*, pp 316–329.

[5] see *e.g.* K. Sparck Jones & R.G. Bates, *Research on automatic indexing, 1974–1976.* Cambridge: Computing Laboratory, University of Cambridge, 1977.

[6] K. Sparck Jones & R.G. Bates, *Report on a design study for the 'ideal' information retrieval test collection.* Cambridge: Computing Laboratory, University of Cambridge, 1977.

The Book

The first twenty years or so of IR experimentation were very well summarised and encapsulated by the 1981 book *Information Retrieval Experiment,* edited by Karen Sparck Jones.[7] This remains the only substantial source devoted to the process of experimentation, and to the variety of experimental methods and techniques that may be used in different circumstances.

I contributed a paper to that volume.[8] Ten years later, in 1991 (the date is important), Micheline Hancock-Beaulieu and I wrote a paper on evaluation[9] in which we argued that there had been substantial changes in the scope and methods of IR system evaluation, with the emphasis moving away from laboratory experiments on test collections and towards more user-oriented studies.

Widening the boundaries?

Experiments involving user interaction with a system were not of course new. Just two examples may illustrate this point: the experiments on the THOMAS system by Bob Oddy[10] and the MONSTRAT experiment by Nick Belkin and others.[11] But it seemed clear that the emphasis was shifting and would continue to shift. A return to substantial test collection experiments seemed highly unlikely to us. . . until. . .

Or not?

In late 1991, after we had completed our paper, TREC was announced. For those of you who are unfamiliar with TREC, it is a mammoth competitive text retrieval experiment.[12] Research groups which participate agree

[7] K. Sparck Jones (ed.), 1981, *op. cit.*

[8] S.E. Robertson, The methodology of information retrieval experiment. In: K. Sparck Jones (ed.), 1981, *op. cit.*, pp 9–31.

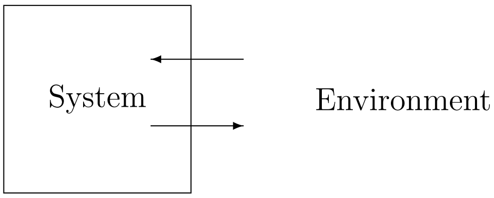
[9] S.E. Robertson & M.M. Hancock-Beaulieu, On the evaluation of IR systems. *Information Processing and Management* **28**, 457–466, 1992.

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

[11] N.J. Belkin, R.D. Hennings & T. Seeger, Simulation of a distributed expert-based information provision mechanism. *Information Technology* **3**, 122–141, 1984.

[12] D.K. Harman (ed.), *The First Text REtrieval Conference (TREC-1).* NIST Special Publication 500-207. Gaithersburg MD: National Institute of Standards and Technology,

Figure 1: An open system



to mount a substantial collection of full-text documents (provided by the TREC organisers) on their own retrieval systems, and run a series of topics (again centrally provided) as searches against the database. Results are submitted to the organisers for relevance evaluation, and the pooled relevance judgements are available for subsequent runs.

TREC is, in fact, the 'ideal' test collection reincarnate. My research group at City is taking part, and we find the experience fascinating and also extremely valuable. So I do not intend to denigrate TREC when I say that it is a real throw-back to an earlier era of evaluation. Nevertheless, this is the case.

Three weeks ago, I made a similar remark in a seminar at Rutgers University. David Lewis, who was in the audience, told me that at an early stage in the discussion of possible large-scale retrieval projects involving ARPA, he (David) had sent them a copy of the 'ideal' test collection report. I had not until that moment realised how strong the historical connection was. But it reinforces my point.

3 Systems, users and boundaries

I would like to explain what I mean by boundaries in this context.

We may start by thinking about an open system in the sense in which the term is used in general systems theory. The diagram in Figure 1 might have been taken from the first page of a book on GST.

1993. The report on TREC 2 is forthcoming; TREC 3 will take place in 1994.

Figure 2: The 2-layer model

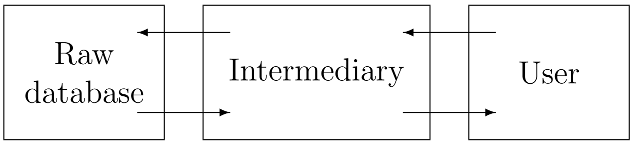
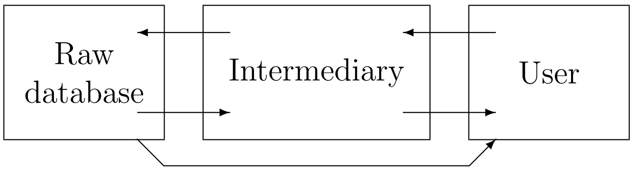


Figure 3: Modified 2-layer model



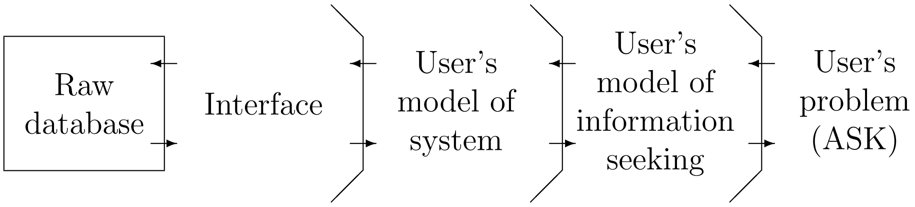
A simple interpretation of Figure 1 in the context of IR is to think of the IR system (in the usual sense of that term: the database/retrieval engine/user interface) as the "system" in the figure, and the user as part of the environment, feeding in a query and receiving documents/items/records/references in response. In a recent paper with Efthimis Efthimiadis,[13] we found it useful to add a layer to this model, as in Figure 2.

In Figure 2, the intermediary may be human or machine; either way, we wanted to consider the various kinds of information that might flow along the arrows in different kinds of interaction. A further complication is that, if the user is present at the search session with the human intermediary, she or he would normally be able to see the system (*i.e.* database) output directly, as well as having interaction with the intermediary, as in Figure 3.

Figures 2 and 3 address only the system end of the process. At the user end, it may be appropriate to consider various layers in the following sense. The user starts with some problem that she/he perceives as requiring infor-

[13] E.N. Efthimiadis & S.E. Robertson, Feedback and interaction in information retrieval. In: C. Oppenheim, C. Citroen & J.-M. Griffiths (eds) *Perspectives in information management 1.* London: Butterworths, 1989, pp 257–272.

Figure 4: A multi-layer model



mation, which we may describe as an ASK (anomalous state of knowledge) in Belkin's terms.[14] However, any interaction with an IR system will be mediated through various other models and processes in the user's mind. In Figure 4, I identify two such layers: the user's model of the information seeking process in general, and the user's model of the particular system.

From the point of view of the resolution of the ASK, the "system" should clearly include those aspects of the user's cognitive activity which fall within the other two layers. However, this idea introduces immediate and obvious problems: specifically, we cannot observe the information events represented by the arrows which cross the boundary of this expanded "system".

4 Some evaluation issues

I would now like to discuss some of the difficult issues in IR system evaluation.

Laboratory *versus* operational conditions

The question as to whether to conduct IR experiments under laboratory conditions (*in vitro*), or under operational or live-system conditions (*in vivo*), is a fundamental one. It is not, of course, a pure dichotomy: there are methods of evaluation that combine features of both. But the conflict is real, both in the obvious sense of realism *versus* controllability, and because there are several associated questions.

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

Batch *v.* interactive experiments

Batch search systems lend themselves to laboratory evaluation. In a batch search experiment, the inputs and the outputs are clearly identifiable, and the assessment of output against input can be made outside the searching process. Interactive system experiments are much less susceptible to laboratory methods.

Repeated *v.* one-off searches

In a laboratory environment, searches may be repeated (under the same or different conditions) any number of times. Under operational conditions, various kinds of learning effect may prevent the repeated use of the same query/information need.

Frozen requests *v.* instances of use

Laboratory conditions require "requests" which are fixed as written statements or in some other form. Instances of use of a system, representing instances of recognition of ASKs by users, can be the subject of operational experiments only.

As indicated, none of these is a pure dichotomy, nor is the correspondence between them absolute: nevertheless, we can discern two distinct IR evaluation paradigms, represented respectively by the combination of the left-hand, or the right-hand, sides of these pairs. The idea of a test collection fits with the left-hand-side only.

Diagnosis

In open system terms, the starting point of Cranfield 1 was to regard the system as a black box (completely described by a label such as "faceted classification"), and to look only at the inputs and outputs. I have already pointed out that it was then seen as necessary to look at techniques and methods in detail. This might be regarded as a diagnostic (as opposed to black-box) approach: in general, we need to discover how and why things work or do not work, rather than *that* they work or not.

Testing with a diagnostic aim might take two broad forms. True diagnostic testing is best exemplified by detailed failure analysis in the style of the

Medlars experiment; the alternative (not true diagnostic testing, but providing some kinds of diagnostic information) is to do combinatorial experiments (trying every combination of a set of variables), in the style of SMART. The two are not in opposition: they are likely to provide complementary diagnostic information. However, they do not go together very well in the same experiment: combinatorial experiments require large numbers of runs, while failure analysis on just one run is likely to be extremely time-consuming (of human time, that is).

System boundary

I have already indicated the problem with identifying an appropriate system boundary for the purpose of evaluation or experimentation. In practical terms, there are great difficulties with identifying a boundary inside the mind of the user (for example as indicated by Figure 4); the experimenter must devise means of inferring the appropriate cognitive events from indirect evidence—in other words, one must of necessity take a diagnostic approach.

Some consequences

The importance, already noted, of methods and techniques and details of implementation, as opposed to broad philosophical or theoretical principles, emphasises the necessity for diagnostic and/or combinatorial experiments; this has been recognised for some time. However, the importance of considering parts of the "system" that are in the mind of the user stresses the need for diagnostic (not combinatorial) experiments, and for operational conditions. Again, the importance of interaction in IR stresses the need for operational conditions and non-repeated searches. There is thus considerable pressure to move away from test collections and laboratory experiments.

Nevertheless, the tremendous twin advantages of laboratory experiments remain: those of scale and repeatability. These are not to be lost lightly; diagnostic tests in operational conditions are expensive and time consuming, for much more limited returns. This strong polarity is a major dilemma for anyone contemplating experimentation in information retrieval.

5 The problem with relevance

Given the above discussion, I would like to indicate what I see as the major problem with the use of relevance in IR system evaluation. Because I remain firmly of the opinion that relevance is, despite many problems, a most valuable concept in IR, I will try to counterbalance this discussion of problems with some assessment of its value.

First, what the problem is not. I do *not* see any problem with the subjectivity of relevance. This merely reflects the subjective nature of the ASK and of the information-seeking process; I would be highly suspicious of any claim to be able to evaluate IR systems on a wholely objective basis. I do *not* see any problem in the range of different factors which might contribute to a user's assessment of a document; I regard relevance as a portmanteau concept which can include a variety of factors, not solely of a topical or subject nature.

What the problem *is* has three parts to it:

Relevance evaluates only one kind of system response...

That is, of all the kinds of information that can flow along any of the left to right arrows in any of figures 1, 2, 3, or 4, relevance only addresses the information identifying a specific item.

...and at only one level of aggregation...

That is, it considers only whole items (usually documents). It is often suggested that either parts of documents, or sets of documents, might be assessed, but this very seldom happens in experiments, and in any case would introduce a great many substantial problems associated with abandoning the assumption that information comes conveniently packaged in discrete units.

...and at only one time.

That is, the information-seeking task is seen as being bounded by some time cut-off (say shortly after the search), and relevance judgements are taken at that point. The idea that a user may change his or her mind about relevance is seen as a problem, while in reality it may indicate a change in the user's ASK (which is, of course, precisely what the information system is trying to help the user to achieve).

The uses of relevance

Nevertheless, relevance has been and remains a concept of deep and lasting significance in information retrieval:

> *As the basis for evaluation measures such as recall and precision.*

For all the above comments, measures such as recall and precision are powerful tools in the armory of the retrieval system experimenter.

> *As a partial basis for diagnostic studies.*

The importance of diagnosis has already been stressed. One form of diagnosis, as pioneered in the original Medlars experiment, is to examine and seek explanations for failures in the sense defined by relevance (*i.e.* relevant items not retrieved or non-relevant items retrieved). Such analysis is not easy, and is certainly not the only form of diagnostic work open to an experimenter, but it is potentially powerful.

> *As the basis for probabilistic models in IR.*

The entire field of probabilistic models in information retrieval would not have come into existence without the concept of relevance. I believe that the probabilistic approach has given the field some valuable insights as well as useful methods (but that is of course another seminar!).

> *As the basis for relevance feedback.*

Relevance feedback could obviously not have happened without the idea of relevance. I believe that this technique has proved to be a most valuable device, both within systems based on probabilistic models and elsewhere.

6 Okapi: some experiments and results

Finally, I would like to turn to a facility which we have built up at City University to enable us to evaluate a range of techniques and methods within one experimental system, and in the context of live user information-seeking behaviour.

The system is Okapi, originally built by Stephen Walker and colleagues at the Polytechnic of Central London.[15] The starting point for Okapi was that a user should be able to walk in off the street and use it. It was originally designed as an online public access library catalogue, at a time when at least some users could not have been expected to have any experience of computer systems of any kind. It is, however, a general purpose text-retrieval system, and is being used with abstracts and full-text databases as well as library catalogue records.

The major design features of Okapi are: free-form natural language queries; searching mainly on word stems (though with a small dictionary of phrases and synonyms); weighting of search terms and ranking of document output; relevance feedback with query expansion. The relevance feedback feature requires that the user should answer a relevance question whenever she/he views a full record (*"Is this the kind of thing you want? (y/n)"*). It sometimes annoys the users to have to answer it, but it is extremely useful from an evaluation point of view.

Evaluation facility

Stephen Walker and the Okapi project moved to City a few years ago, with the aim of developing a live-use evaluation facility. Okapi is available on the campus network (and indeed over Internet), with various databases (currently the City University library catalogue; Bath University catalogue with some of the records enhanced by the addition of contents-page and other descriptive information; and a section of Inspec). Users have to register to use it (there is a terminal in the library, which it is possible to use anonymously, in the style of most library catalogues, but users from elsewhere on the network require a user id). In registering, they agree that we may observe their use of the system by examining their logs, and perhaps invite them to take part in other experiments, *e.g.* by completing questionnaires or doing additional searches in the Centre, *etc.*

[15] Okapi and the experiments using it have been described in a number of papers and reports: see *e.g.* M. Hancock-Beaulieu & S. Walker, An evaluation of automatic query expansion in an online library catalogue. *Journal of Documentation* **48**, 406–421, 1992.

Some observations

The following are a few of the points to emerge from a series of studies of uses and users, both quantitative and qualitative.

> *Users commonly repeat searches, either with minor variations or identically.*

That is, a user may log on one week and undertake a search, and then log on again the next week and start with a very similar, or even identical, initial query.

This surprised me when we first observed it. However, if we consider a user such as a doctoral student, involved in a long-term project, it is perhaps not so surprising. It seems that the user may use the initial query as a way to locate her/himself at a known place in the database, and then make use of the interactive facilities for subsequent navigation.

> *Relevance feedback is used moderately frequently.*

In particular, it is used in about one-third of the searches in which it is available (it only becomes available as an option when the user has marked one or two items as relevant). This is not a huge amount of use, but is substantial, particularly when compared with the use of "advanced" facilities on other online catalogues. It is clearly accepted by many users as a useful and valuable method of interaction. Further, in those searches in which it is used it is responsible for the retrieval of about one-third of the items the user marks as relevant.

> *Users would like to use relevance judgements experimentally or constructively.*

When we first introduce relevance feedback, I imagined that users would make their relevance judgements in a relatively naive fashion, without thinking too much about the consequences. I could not have been further from the truth: many users would like to experiment with their relevance judgements as a way of controlling the direction of navigation. This fits very well with the first observation above: relevance feedback is seen as a method of navigation, rather than as a method of homing in on an ideal set.

7 Final comments

None of these observations could have been made in a laboratory environment. Furthermore, they reinforce the points made earlier about evaluation of interactive retrieval and the importance of the user's mental model. The first observation, in particular, suggests that we should be looking at users over a longer time period than just one session on a system; the resolution of the ASK should clearly be measured in relation to the problem situation (again in Belkin's terms), which in these cases extends over weeks or longer.

In the Okapi project we are endeavouring to devise tools to help in live-use, interactive-system evaluation. This is not an easy task, and such evaluation must be taken as complementing (rather than replacing) the more traditional laboratory experiments. However, I believe the task to be well worthwhile.