Process and outcome: On the evaluation of IR systems in the age of interaction, GUIs and multimedia

Stephen Robertson

Microsoft Research Ltd

St George House, 1 Guildhall Street, Cambridge CB2 3NH, U.K.

ser@microsoft.com

Also City University London

The question I want to address in this talk has to do with the relation between traditional ideas of retrieval system evaluation and the present concern with highly interactive systems, maybe involving other media as well as or instead of text. For reasons that I hope to make clear, my major concern is with interaction rather than with the media themselves. Particularly, I want to consider the questions: Why has the notion of relevance been such a powerful device in the traditional model of evaluation? And: Why is it not so useful in the evaluation of interactive systems?

Black boxes, input, output and outcome

The Cranfield/TREC model of information retrieval system evaluation in some sense treats the system as a black box. That is, we package the system in its entirety, provide it with inputs (documents and requests), and examine and evaluate its outputs. Actually, we are not simply concerned with output, but rather with the *outcome* of the process – in some sense, what happens as a result of the output. Evaluation is not possible without some notion of outcome – it makes no sense to ask evaluation questions in relation to output alone.

There is in principle some indeterminism in the interpretation of outcome – in particular, it can be assessed close to the boundary of the black box (the immediate effects of the output), far away (the ultimate effects), or at any stage in between. I discuss below the particular interpretation of outcome assumed in the Cranfield/TREC tradition.

With knobs on: Using outcome to evaluate components

Given some evaluation of the output of the black box, we can open it up a little. This is not really a question of *looking* inside to see what is happening inside – more of poking around to see what will be the effect on output/outcome of changing something inside. Thus we may have knobs (e.g. tuning constants) which might be regarded simply as another form of input; or we may have packaged components (such as stemmers) which we can simply replace with variants.

So the general method is: evaluate the output/outcome from the black box before and after making changes, and infer the effect of the changed component on performance.

The indivisibility of outcome and process

One could argue that the distinction between outcome and process is artificial – every process is part of another process, the outputs of every process contribute to other processes, every process contains outputs of every sub-process. We tend to see retrieval as a self-contained activity on the part of the user, but in truth it must be part of some larger activity. This larger activity presumably informs the retrieval process and provides the context in which it make sense; but we may also presume that the interaction is two-way, in that retrieval affects the larger activity.

This would not matter if the effect derived entirely from the *outcome* of the retrieval process; in this case one would still, in principle, be able to treat retrieval as more-or-less self-contained. A problem arises, however, because it may be some other part of the retrieval *process* that interacts with the larger one. Suppose, for example, that you are writing a review article, and that as a result of searching (not the individual documents you retrieve, but the general picture you form during the course of the search), you change your mind about the nature or scope or approach of the article. Then evaluation of the outcome of searching (in terms of retrieved documents) no longer captures the full effect of the retrieval system.

Process as a way of life

We might take this argument further in the case of information retrieval, and argue that the whole point of retrieval is in the process rather than in any (definable, measurable) outcome. I believe that would be going too far, and would render evaluation practically impossible; however, I believe also that that view contains an element of truth. This would imply that any form of outcome evaluation that we use is likely to miss at least some of the point.

It can also be argued, by contrast, that any claim about the inadequacy of looking at definable/measurable outcomes is simply a reflection of the inadequacy of our definition of outcome. If the searching process has some effects that are not apparent when we look at the individual documents retrieved, this simply means that the outcomes of the search include other effects that take different forms. This is certainly a valid argument, though it does not help greatly, because it does not tell us how to define/measure these outcomes – merely that we need a much richer notion of output and outcome.

Relevance is an interpretation of outcome

It is worth now reconsidering the traditional Cranfield/TREC approach in the light of the above discussion.

I distinguished between output and outcome; in the traditional test, these are narrowly but distinctly different. The output of the system is a set of documents (ranked or otherwise), or perhaps some other representation of the document space, such as descriptions of clusters. Documents do not in themselves carry relevance values (though it is sometimes convenient to assume that they do); relevance values result from the interaction between the document and the user. Thus the *outcome* of the process is that the user now knows about certain documents which s/he regards as relevant to the need.

As indicated above, outcomes can in principle be assessed close to or far away from the boundary of the system; a wider ranging assessment might for example take the form of asking the question “Has the problem that the user started with been solved?”. Relevance is a very immediate outcome: it takes this point-of-assessment location to be as close as possible to the boundary of the system.

In this interpretation, we assume that retrieved documents are the only form of output. Furthermore, these documents are taken one-at-a-time – no interactions between documents are allowed for.

Why is it such a useful interpretation?

The power of the Cranfield paradigm lies precisely in this judicious and useful choice of the point of measurement of outcome. It could not be any closer to the boundary of the system and remain a form of evaluation; any further away, and it would introduce many variables and questions which would make it largely unusable, at least on the scale it has been used on. As it stands, it is, quite simply, what has made laboratory experimentation in IR possible.

Furthermore, as well as serving for black-box evaluation, it *also* provides some basis for diagnostic work. The kind of detailed failure analysis that we occasionally do in laboratory experiments, although not easy, is at least possible and can be informative with document-by-document relevance judgements.

Why is its application to interaction limited?

The problem is that there is simply no equivalent point-of-assessment location in an interactive system. There seem to be two possibilities:

Outcome measured during interaction

Here we take the presentation of a document to the user in the course of the interactive session as the output. Then a judgement of relevance is one possible outcome, but there are others (e.g. this document is not relevant but it gives me some information on the basis of which I can improve my search formulation). Also, there are other outputs (e.g. suggested query expansion terms, or summary information concerning the search), which may have similar outcome.

In this interpretation, the restriction of assessed output to the relevance of the retrieved documents is simply untenable. (It is also the case that input is problematic, since it must include all user responses to any system action.)

Outcome measured at the end of the session

Here we look for output only at the termination of the session. The user leaves the session with certain knowledge about documents that s/he did not have at the start.

There may again be other kinds of information that the user gathers during the search, e.g. concerning the existence of certain kinds of documents, or the structure of the subject, as suggested above. These non-document outputs with persistent outcomes would not be captured by the usual relevance assessment. In principle, this nevertheless seems more defensible than 7.1, or at least closer to the assumptions of the Cranfield tradition. But in this case it follows, for example, that precision is not a good measure, because rejection of non-relevant documents would normally occur *during* the search process – they are no longer part of the output. The measure to set against recall is the time/effort of the search process, but this immediately makes it very situation-specific, and less suited to

a laboratory experiment. There are also problems with understanding input, which must now catch the user at the point of approaching the system, whether or not they have yet verbalised their request.

In either case, the diagnostic potential of relevance judgements is greatly reduced over the black-box version

The tyranny of evaluation

One problem with our current view of information retrieval system research is that we assume that *experimentation* = *evaluation*. We construct theories and models, make predictions from them about what would be a good way to do information retrieval, and test these predictions through experiments. In principle, this sounds like good science, but for the fact that we only ever look for one kind of prediction: those that relate to performance. Good science would require us to look for *any* testable predictions that can be made from our models. Such approaches are occasionally visible in IR research, but mostly we remain blinkered by the view that the only interesting predictions are those that involve performance improvements. Once again, the black box view has influenced us strongly here: given that we decide in principle only to look at output, and given that output and outcome are so closely tied in this model, it hardly seems like much of a restriction. But once we abandon that view (as the investigation of interactive systems forces us to do), it becomes a much greater constraint.

The investigation of process suggests that we should be interested in the *behaviour* of systems, without any necessary, immediately implied assignment of *value* to different behaviours.

Multimedia

I have largely concentrated on the evaluation problems arising from interactive systems. There is no necessary connection between multimedia systems and interaction; it is certainly possible to conceive of circumstances or environments or tasks involving non-text or multimedia information retrieval but no strong interactive requirement (just as a for instance, matching a proposed new trademark against existing ones strikes me as being fundamentally a not-very-interactive task). But it is also the case that many of the tasks one can imagine involving different media are likely also to involve a high degree of interaction.

I believe that the major problems of evaluation arise from this fact, and are present in the evaluation of interactive text retrieval systems. I do not believe that other- or multi-media retrieval as such is particularly different from text retrieval from an evaluation point of view; the problems we really have to get to grips with are those of interaction.

Conclusions

In this paper, I have presented (yet another!) interpretation of relevance. This is nothing to do with the nature of relevance or how we should define or measure it; it relates only to the role it plays in evaluation.

To the extent that we believe that evaluation of IR systems has to do with the outcomes of searching, and to the extent that we are concerned with not-very-interactive systems, relevance is the obvious and almost ideal outcome characteristic for evaluation purposes. But for highly interactive systems (whether text-only or other- or multi-media) this is no longer the case. Nor is there anything obvious to take its place. My main conclusion, which I have no doubt any of you could have told me before I started, is that the evaluation of interactive systems is orders of magnitude more difficult than the evaluation of non-interactive systems.

Nevertheless, my hope is that maybe the analysis has some value. Somehow we need to get to grips with the question: what are the outcomes of searching, in the different contexts in which this activity takes place? And at what point(s) can we capture these outcomes?

Furthermore, I believe that we also need to follow in parallel a quite different path: to put aside for the moment questions of value, and seek instead to understand process. The processes that we need to understand include both those internal to our systems, and all the various external ones which interact with them.

I don't, I am afraid, have any recipes.