Richer Theories, Richer Experiments

Stephen Robertson
Microsoft Research, Cambridge

The Cranfield approach to evaluation and that of its successors, including TREC, is oriented towards system effectiveness. The experimental paradigm is that we have a number of alternative systems, and the research question under investigation is: 'Which system is best'. If we take seriously the notion that we are engaged in developing a *science* of search, then Cranfield would seem to fit with the idea of a scientific experiment, specifically a laboratory experiment, designed to test out ideas and to help in the development of models or theories. In fact, Cranfield would seem to give us the only notion that we have of a laboratory experiment in search. However, an analysis of the role of empirical knowledge in general and laboratory experiment in particular, in relation to models or theories, reveals some limitations of the Cranfield approach. Despite the huge advances in this experimental paradigm since Cranfield itself, due in large measure to TREC, I believe we are only scratching the surface of what experiments can tell us.

In the scientific approach, we would be looking for models or theories to explain and interpret the phenomena we see around us. In the case of information retrieval, we have some notion of what phenomena are of interest to us: people writing documents; other people (users) needing information in order to solve some problem or accomplish some task; these users undertaking search or information-seeking tasks; and the various mechanisms which might help them do this, by delivering or pointing at documents, or even by answering questions using information extracted from documents. Finally, we have a notion of success or failure, or perhaps degrees of success, in this process. This notion of success or failure we have taken to be central, exactly because we are trying (as engineers) to construct new and better mechanisms with a view to helping the users.

Again, in the scientific approach, we would be looking to the models or theories to tell us things about the phenomena that we did not know or understand before. We can see this as a process of *prediction* – a model might say, in effect, 'if you do this [which we had not done before], or look at the phenomena in this way [which ditto], then this is what you will observe.' In the IR case, because of our engineering emphasis on constructing mechanisms which work well, we have seen the function of models as telling us how to make them work better. Typically this is all we ask of a model in IR. We regard this as the only test we need to make of a model, that it gives us good retrieval effectiveness. Thus the function of experiment is (only) to tell us how well we are doing.

This feels like a major limitation. To be sure, the predictions about how to do things well are going to be the main *useful* predictions and applications of such models – although we might also ask if the same models are capable of making other useful predictions. But in any case, testing a model should not be restricted to testing its useful predictions. Less useful or even completely useless predictions may well tell us as much about the model and how to improve it as the useful predictions.

Furthermore, this seems to be one source of the (partial) standoff between the laboratory experimental tradition in IR and the user-oriented, often observational work on information seeking. While the user-oriented world may acknowledge the notions of success and failure (albeit with a somewhat broader notion of these qualities), there are many other aspects of information seeking processes, often orthogonal to the success/failure axis, that are of interest. In particular, user behaviours come to mind. In my view, one way to advance the field of IR would be to seek a much richer range of theories and models, and a correspondingly richer range of experimental and observational studies, with the primary aim of validating, or refuting, or deciding between, the models. I think we are in fact moving in this direction, but slowly.

I believe that what we need now is not so much better systems (though they are always welcome) as better understanding of the phenomena.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.