

# “Would You Trust Your IR System to Choose Your Date?” Re-thinking IR Evaluation in the 21st Century

Elaine G. Toms

Information School, University of Sheffield  
Regent Court, Portobello, Sheffield, UK  
e.toms@sheffield.ac.uk

**Abstract.** This talk examines interactive IR system evaluation from the holistic approach, including some of the pitfalls in existing approaches, and the issues involved in designing more effective processes and procedures.

**Keywords:** User-centred evaluation, interactive information retrieval.

## 1 Overview

In the quest for almost any concept or object, human nature is surprisingly and paradoxically predictable: people will either know it when they see it, or they know what it should be before they find it. Yet, the concept of the “right” document, photo or indeed any information object is in the mind of the seeker regardless of whether the seeker is looking for dinner, a new car, a date, a novel to read, or a solution to a work problem. To complicate the matter, that solution is also a moving target, and not necessarily a single reality out there to be uncovered. Once an object is found, whether it is pertinent, relevant, accurate or correct is a *human* judgment made at a particular moment in time for a particular purpose. An information retrieval (IR) system can only provide suggestions; its role is to support and facilitate, and not to replace the human decision-making process.

In the evaluation of interactive IR systems, this requires a directional shift from the typical TREC, INEX and even CLEF evaluation processes. At present, evaluation has moved from an emphasis on topical relevance, to an emphasis on measuring almost anything that can be quantified. This is more likely to be data extracted from transaction logs in an attempt to develop a more predictable personalized search with likely the same accuracy as forecasting the future by reading tealeaves in a cup!

We have failed to step back and assess the broader picture. What exactly are we evaluating and for what purpose? It could be said that we have focused far too long on the tool and not on what the tool will be used for. For example, if that IR system was choosing ones date, a restaurant or a medical treatment, would we use the same evaluation techniques in use today? This talk will consider evaluation from that holistic and contextual perspective. It will examine some of the pitfalls in existing approaches, and discuss the issues involved in designing more effective evaluation approaches for assessing interactive IR systems.