Is this document relevant? Errr it'll do

Mark Sanderson
University of Sheffield
m.sanderson@shef.ac.uk

Abstract *Evaluation of search engines is a critical topic in the field of information retrieval. Doing evaluation well allows researchers to quickly and efficiently understand if their new algorithms are a valuable contribution or if they need to go back to the drawing board. The modern methods used for evaluation developed by organizations such as TREC in the US have their origins in research that started in the early 1950s. Almost all of the core components of modern testing environments, known as test collections, were present in that early work. Potential problems with the design of these collections were described in a series of publications in the 1960s, but the criticisms were largely ignored. However, in the past decade a series of results were published showing potentially catastrophic problems with a test collection's "ability" to predict the way that users will work with searching systems. A number of research teams showed that users given a good system (as measured on a test collection) searched no more effectively than users given one that was bad.*

In this talk, I will briefly outline the history of search evaluation, before detailing the work finding problems with test collections. I will then describe some pioneering but relatively overlooked research that pointed out that the key problem for researchers isn't the question of how to measure searching systems accurately, the problem is how to accurately measure people.

Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009.