View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by RMIT Research Repository

Relevance Thresholds in User System Evaluations

Falk Scholer Andrew Turpin*

School of Computer Science & IT
RMIT University, GPO Box 2476V
Melbourne, Australia, 3001.
{aht,fscholer}@cs.rmit.edu.au

ABSTRACT

We introduce and explore the concept of an individual's *relevance threshold* as a way of reconciling differences in outcomes between batch and user experiments.

Categories and Subject Descriptors

H.4 [**Information Storage and Retrieval**]: Miscellaneous

General Terms

Performance, Design, Experimentation, Human Factors

Keywords

Search engines, information retrieval evaluation, user study

1. INTRODUCTION

Information retrieval (IR) experiments based on the Cranfield methodology measure system performance using a batch of queries and a test collection that has a subset of the documents judged as relevant or irrelevant by human judges for each query. The utility of a system is then computed using a metric that aggregates the relevance judgements for documents in ranked lists returned by the system for each query. For example, many papers report IR system comparisons using the TREC document collections, topics and judgements, using Mean Average Precision (MAP) as the metric [9].

An alternate way to evaluate systems is to take a group of human users and ask them to perform search tasks with the various systems, comparing outcome measures such as the time to complete a task, success or failure on a task, or subjective measures like user satisfaction. Previous studies [1, 2, 5, 6, 7, 8] have shown that attempting to transfer results from batch experiments into laboratory based user studies is difficult. That is, the systems rated as superior in batch experiments are unlikely to assist users in performing their tasks more quickly or more accurately than the systems that are rated poorly in the batch experiments.

There are many possible causes for this seeming mismatch between batch and user-based experimental outcomes. In this paper we introduce and test the idea of a mismatch in *relevance threshold* between the judges used to gather the batch data, and the users on which the systems are trialled.

*Supported in part by the Australian Research Council.

Copyright is held by the author/owner(s).
SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

System	P@1	Determination of relevance (4-pt scale)
U0	0	Both users 0.
U1	1	At least one user > 0.
V0	0	Neither user 2 or 3.
V1	1	Either user 2 or 3.
W0	0	No user 3, or one is 3 and the other < 2.
W1	1	Both users 3, or one 2 and one 3.

Table 1: Systems used (mapping of 4-pt relevance scale to P@1): Ux has a strict irrelevance criterion, Wx a strict relevance criterion, and Vx a mix of the two.

2. EXPERIMENTAL METHODOLOGY

Participants were recruited from our university, and experiments were carried out in accordance with the guidelines of the RMIT University Human Research Ethics Committee.

Using topics and documents from the TREC .GOV2 collection [3], relevance assessments were made for each document by two subjects on a four-point categorical (not numeric) scale using the following definitions. *Completely relevant(3)*: the document contains enough information to completely answer the information need, providing details on all aspects of the topic. *Highly relevant(2)*: the document contains answers to many aspects of the topic. *Marginally relevant(1)*: the document covers some aspects of the topic. *Not relevant(0)*: the document contains no information about the topic.

Using a similar framework to our previous studies [8], we constructed ranked lists using the known relevance levels of documents to achieve a given level of P@1. This resulted in eight search systems (sets of lists) as summarized in Table 1. Subjects were presented with information needs based on the TREC topics, and asked to find documents that help to resolve the need. We then measured the amount of time that a user needs to find a relevant document for an information need. For each document that a user viewed in a search results list, they could choose to save the document (indicating that it is relevant), or not save it (not relevant).

We attempted to measure the relevance threshold of individual users while they undertook the search task by examining the number of documents of each relevance level that each user read and then did, or did not, save. In this part of the experiment, we assumed that the true relevance level of a document was the ceiling of the average of the two user judgments that had been made on that document. Using techniques from psychophysics, we fit a Weibull psychometric function to the data of each user, and used the 50% point

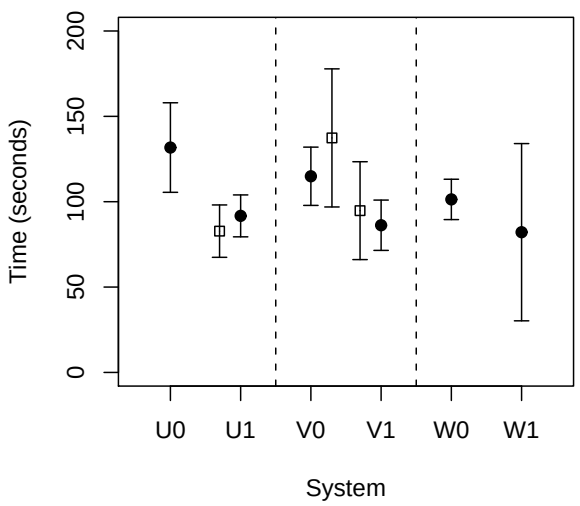

Figure 1: Mean time taken to save the first document using pairs of inferior and superior systems. Circles include all users; squares are a subset of users decided by relevance threshold (see text).

as the threshold from these curves [4].

3. JUDGE RELEVANCE THRESHOLDS

The aim of our experiment is to explore the effects of a mismatch in relevance threshold between judges and users by altering the relevance thresholds used in the batch experiments. In particular, we evaluate systems Ux, where category one (marginally relevant) documents are considered relevant (this is the default assumption in most TREC experiments); systems Vx, where category one documents are considered irrelevant; and systems Wx where only category three (completely relevant) documents are considered relevant. If users are not using the same relevance criteria as were used in the batch judgements, then we would expect differences in systems that appear in the batch experiments to not be reflected in the user experiments.

Differences in time to find relevant documents using different systems in shown by the circles in Figure 1. Using batch judgements where marginally relevant documents are considered irrelevant reduces the gap between the systems from a user perspective (circles are closer together in the Vx panel than in the Ux panel), but the difference is still significant (t-test, $p < 0.05$). When the batch judgments insist that only completely relevant documents are considered as relevant (systems Wx), then users do not notice a difference between the two systems ($p > 0.05$).

4. USER RELEVANCE THRESHOLDS

A user's relevance threshold should be less than one if their behaviour is to match that used in the batch experiments that assessed Systems U0 and U1 as the inferior and superior systems. That is, if a user read a category one document (marginally relevant), there should be a better than even (50%) chance that the user would save that document as relevant, since category one documents were considered relevant in the batch experiments. Our users have different relevance thresholds; if we were to exclude any users from the data that have a threshold lower than one, and reanalyze the time taken until the first document is saved, we would expect the system U0 to perform more poorly (time to save increases), and the time taken to save a document with system U1 to decrease. Similarly, for the batch experiments that evaluated system V1 as superior to system V0, it was assumed category one (marginally relevant) documents were irrelevant, and so users should have a threshold between 1 and 2 if they are to match the judges.

User	1	2	3	4	5	6	7	8	9	10	11
Thresh.	0.0	0.1	0.1	0.2	1.2	1.4	1.5	1.8	1.8	1.9	2.1

Table 2: Relevance thresholds for each user.

Table 2 shows the individual user relevance thresholds. The first four users all have a threshold below one; that is, there is a more than even chance that they would categorize a level one document (marginally relevant) as relevant. The remaining seven users, however, all have a threshold greater than one, indicating that there is less than a 50% chance that they would save a category one document.

If we exclude those seven users (5 to 11) who have a relevance threshold mismatch, and re-evaluate the time taken to save documents using systems U0 and U1, then we get the mean time shown by the square in the U1 section of Figure 1. It is clear that the mean time to save with U1 went down due to the exclusion of users with a threshold greater than one. Unfortunately, there was not enough data to conclude that mean time with U0 went up.

Re-evaluating Systems V0 and V1 using users with thresholds between 1.5 and 2.5, thus choosing the users whose relevance thresholds match the judges used in the batch experiment that says V1 is better than V0, we see that the gap between V0 and V1 widens (squares compared to circles in the Vx panel), as expected. Moreover, the difference between time is now statistically significant ($p < 0.05$). Thus, when relevance thresholds match, batch differences are more clearly reflected in the user experience.

There were other sources of mismatch that were explored in this study, but space prohibits their discussion in this abstract.

Acknowledgments

We thank Justin Zobel and Steve Garcia for valuable discussions.

5. REFERENCES

[1] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. *SIGIR'07*, p773–774.

[2] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? *SIGIR'05*, p433–440.

[3] C. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2004 terabyte track. Gaithersburg, MD, 2005.

[4] G.A Gescheider. *Psychophysics: method, theory and application*. Lawrence Erlbaum Ass., New Jersey, 1985.

[5] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kraemer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? *SIGIR'00*, p17–24.

[6] D. Kelly, X. Fu, and C. Shah. Effects of rank and precision of search results on users' evaluations of system performance. TR-2007-02, U. of North Carolina, 2007.

[7] A. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. *SIGIR'01*, p225–231.

[8] A. Turpin and F. Scholer. User performance versus precision measures... *SIGIR'06*, p11–18.

[9] E. M. Voorhees and D. K. Harman. *TREC : experiment and evaluation in information retrieval*. MIT Press, 2005.