Measuring User Relevance Criteria

Falk Scholer Andrew Turpin Mingfang Wu

School of Computer Science & IT
RMIT University, GPO Box 2476V
Melbourne, Australia, 3001
{Falk.Scholer,Andrew.Turpin,Mingfang.Wu}@rmit.edu.au

Abstract

Recently, Scholer and Turpin [Proc. SIGIR 2008] proposed the use of techniques from the field of psychophysics to determine a relevance threshold *for a user. Using this threshold, they observed, one could match the relevance criteria of users to those of judges used to develop a test collection, hence selected users should have a better search experience with systems judged superior on that collection. In this paper we show that, when the level of relevance of documents is measured using a categorical scale such as TREC relevance levels, rather than a numerical or physical scale, then the psychophysical techniques for determining thresholds cannot be meaningfully applied in some cases. We demonstrate that the choice of mapping from the categorical scale to a numerical scale has a marked effect on the thresholds derived. Instead, we propose a simpler methodology for matching users to judges. Using the average split agreement approach, only 12 of our 40 student users can be considered aligned with the relevance criteria of TREC judges on three TREC topics.*

Keywords: *User study, relevance judgements, batch experimentation, TREC*

1 Introduction

Batch-based information retrieval experiments measure the performance of search systems by evaluating how documents are retrieved in response to a set of test queries. Central to this evaluation process is the notion of *relevance*: each document that is returned as an answer to a query is judged by a human as being either relevant to a search request, or not. Relevance is most commonly measured using a categorical scale, where different "levels" of relevance can be assigned to documents or other information resources. Based on these judgements, a variety of system performance metrics can be calculated.

Most large-scale experimental evaluations, such as those conducted using the TREC framework, use a binary relevance scale. Here, a document is relevant if it contains any information about the topic; otherwise, it is not relevant. The default TREC assumption can therefore be viewed as explicitly folding multiple possible levels of relevance into a binary criterion.

It is widely accepted in both information retrieval and information science that relevance has a personal dimension – a resource that is considered by one user to be relevant to a particular query or search request may not be considered to be relevant to the same request by another user [9, 13]. This difference in behaviour may be particularly prevalent for those documents that are only marginally relevant, containing only some limited information about a topic. For TREC-style system evaluation experiments to meaningfully transfer to a user population, it therefore seems plausible that the relevance criteria used by the members of that population needs to match the criteria applied by the TREC judges. This *relevance threshold mismatch* may be a key reason why many recent studies have failed to find consistent improvement in the performance of search system users when searching with a system that scores poorly under the TREC evaluation framework compared to when using a system that scores highly using the same evaluation approach [1, 2, 10, 16].

In this paper, we explore the relevance criteria of 40 users on TREC data. Firstly we calculate a *relevance threshold* (in "TREC relevance units") for each user using the definition of a threshold from the field of psychophysics. The sticking point with applying this well established methodology is that TREC relevance judgements are on a categorical scale, rather than a numerical scale. Hence we explore several mappings from the categorical scale to the numerical scale, showing that the choice of mapping has a marked affect on threshold values.

We also examine a simpler technique for determining user relevance criteria based on agreement scores between users and TREC judges. Of our 40 users, around 12 displayed a similar criteria to that of TREC

judges. There were also a large variation between agreement aggregated over topics, and also a large variation in agreement within topics at the individual document level.

Our results indicate that relevance thresholds vary significantly between individuals; that is, some searchers have a low tolerance for documents that are only of marginal use to an information need, while others consider these to be useful to the same search request. In addition to variation in thresholds across different users, we also find that relevance thresholds vary depending on the search topic, and on documents within the topic, relative to TREC judgements.

2 Background

Relevance

Relevance is a fundamental concept for the evaluation information retrieval systems. While many definitions — and indeed different types — of relevance have been proposed [12], most experimental evaluations in the IR field use a categorical scale [18]. In traditional *batch* experimental framework, a system is evaluated by running a set of search topics over a fixed collection of test documents. For each document that is returned in the answer list of a search system, a human judge determines whether the document should be considered to be relevant to the search topic, or not. In TREC (the ongoing series of Text REtrieval Conferences), judgements are usually made by paid information analysts, who are often also involved in the specification of the search requests (or "topics").

Under the batch approach, search systems are scored based on how well they are able to retrieve relevant documents; most metrics reflect some combination of how early in the answer list the relevant documents occur (precision), and the number of available relevant documents that are found (recall).

Although system performance metrics that make use of multiple-level relevance assessments have been proposed (for example, nDCG [11]), almost all widely-reported performance metrics are calculated based on binary relevance judgements. This includes metrics such as precision at 10 documents retrieved, R-precision, and mean average precision (MAP). Even in cases where multiple-level relevance judgements have been used, such as in the Terabyte track data that we use below, evaluation metrics are calculated after these levels have been folded into a binary scale.

Investigations of applying multiple-level relevance criteria in the TREC framework have suggested that the binary criterion may be overly simplistic [14]. The traditional criterion for TREC relevance states that a documents that makes any reference to the topic should be classed as relevant. This can therefore include a wide range of documents, from those that are only marginally relevant (containing no information beyond that which is already included in the topic description), to completely relevant documents (containing enough material to completely answer the information need). Analysis by Sormunen [14] showed that a large proportion of documents that are judged as relevant under the binary TREC criterion are in fact from the marginal category (50% of all relevant documents for 38 topics from TREC-7 and TREC-8).

In this paper we are particularly interested in matching users' relevance criteria to that of TREC judges, agreeing with Scholer and Turpin's conjecture that system comparisons made on TREC data can be carried over to user populations with such a similar relevance criteria [15]. As TREC documents are judged on a three point scale—"not relevant" (0), "relevant" (1), and "highly relevant" (2)—we quantify user relevance thresholds in these "TREC relevance units". The threshold for a user is defined as the TREC relevance category where the user will state that a document from that category is relevant 50% of the time. That is, there is a better than chance probability that the user will say a document is relevant if it is in that threshold relevance category, or a higher category.

Psychophysics

Psychophysics is the study of the relationships between stimuli and perception. The perceptual experience, which is intrinsically subjective, is measured through the use of a stimulus as a reference system. This allows a threshold — the intensity of a stimulus that is required for it to be consciously experienced — to be determined. Psychophysical methods have been applied in a wide range of domains, including sensory processes, memory, and learning [8].

Psychophysical thresholds can be measured using the method of *constant stimuli*. Here, for a particular stimulus that can occur at different levels of intensity, a fixed number different stimulus levels are repeatedly presented to a subject. After each presentation, the subject indicates whether the stimulus was detected or not. The order of presentation is random, but overall each level of the stimulus is presented an equal number of times [7]. After multiple presentations, the proportion of positive and negative responses at each stimulus level can be calculated. A *psychometric function* is then constructed by fitting an ogive curve through this data. The absolute threshold is that level of stimulus intensity for which the subject has a 50% chance of detecting the stimulus [8].

3 Methods

To investigate our hypothesis that user relevance thresholds can be measured, and vary between different searchers, we conducted a user study and applied

```
707  What evidence is there
     that aspirin may help
     prevent cancer?
770  What is the state of
     Kyrgyzstan-United States
     relations?
771  What deformities have been
     found in leopard frogs?
```

Figure 1. The three TREC topics judged by users in this study.

the psychophysical method of constant stimuli. To conduct such an experiment requires a group of subjects (users of a search system, whose thresholds are to be measured), and a stimulus.

40 users were recruited to participate in the experiment. The study was advertised using posters and online newsgroups, and participation was voluntary, with subjects being compensated for their time with a $50 gift voucher. All participants were students at RMIT University, and the study was conducted within the guidelines of the Human Research Ethics Committee of the university.

Topics and Documents

For the experimental stimulus, with which we aim to measure the subjects' perception of relevance, we used documents from the TREC GOV2 collection, a 426 Gb crawl of the `.gov` domain from 2004 [6]. This collection has 150 associated search topics, and a set of corresponding relevance judgements on a three point categorical scale: "not relevant" (0), "relevant" (1), and "highly relevant" (2).

The method of constant stimuli involves the repeated presentation of the stimulus at different levels, with an even number of presentations at each level. We therefore chose three TREC topics with a large corresponding number of judged documents at each level of the relevance scale, topics 707, 770 and 771, shown in Figure 1.

The thirty documents required for each topic — ten at each of the three relevance levels — were chosen by working down a sorted list of candidate documents taken by merging the top 50 documents from two runs with highest MAP scores from the Terabyte Track for 2004, 2005 and 2006. Only documents of type "text/html" were retained, with other types, documents smaller than 750 bytes or larger than 100,000 bytes, and all content within a `<script>` tag, being discarded.

The session began with three practice documents from topic 847; one of each TREC relevance category. These were discarded from the results logs before analysis began.

```
Imagine that you are writing a report
about the provided topic.  You will then
be presented with a series of documents,
one at a time.  Read each document, and
decide if it is relevant for the topic.
```

- `Relevant:  If the document contains any information that you would use for the report, then the entire document should be classed as relevant.  (This applies even if you have previously seen this information in another document.)`
- `Not relevant:  The document contains no information that you would use for the report.`

Figure 2. Judging task instructions.

Task

Users were asked to imagine that they are writing a report, based on an information need as specified in the description and narrative fields of a TREC topic. Documents were to be marked as relevant, or not relevant, for the topic in question. The precise instructions are shown in Figure 2.

The user study proceeded as follows. Users were presented with the description and narrative of a TREC topic, and asked to read through the information request. They were then presented with a list of documents, of different relevance levels, in turn. For each, they needed to decide if it was relevant (if the document contains any information that would be used in their simulated task), or not relevant (if the document contains no information that would be used in their simulated task), to the stated topic (that is, they made a binary decision).

For each of the chosen topics, users were presented with 10 documents at each of the three TREC relevance levels — 0, 1, and 2 — giving 30 judgements per topic in total. The topics, and documents, were presented to users following a balanced experimental design, to control for topic and document ordering effects.

For this experiment, the TREC topics were framed in a task-based scenario: users were asked to imagine that they needed to write a report to fulfill the given information request. This simulated work task approach was used to ground the information need in a practical situation; previous work by Borlund has demonstrated that such simulated information needs can elicit searcher behaviour that is close to the behaviour that is exhibited when pursuing real information needs [5].

4 Results

The output of the user study was a set of responses from each subject: for a single topic, a subject had

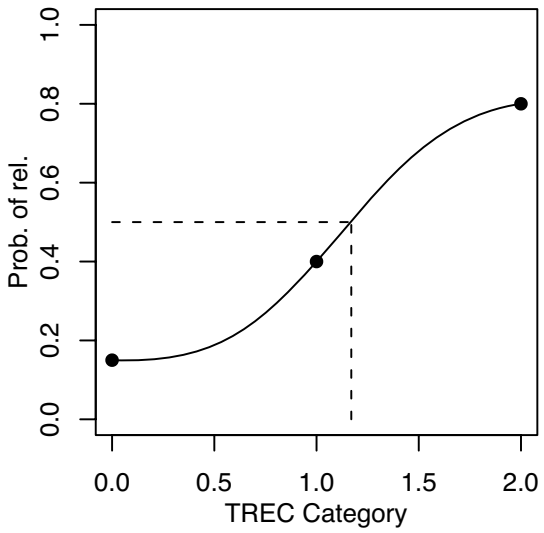

Figure 3. A psychometric function.

been presented with 10 documents at each TREC relevance level (the levels of the stimulus). Therefore, for each stimulus level, it is possible to calculate the proportion of times that the subject perceived the presented documents to be relevant, or not relevant.

Psychophysical Techniques

The subject's psychometric function is constructed by fitting a Weibull distribution through the data points. An example is given in Figure 3; here, the subject perceived TREC-0 stimuli to be relevant 15% of the time, while TREC-1 and TREC-2 stimuli were perceived to be relevant 40% and 80% of the time. To obtain the subject's relevance threshold, we take that point at which the subject would perceive a stimulus to be relevant 50% of the time. In the example, this is at a level of 1.2 on the relevance scale. TREC judges, who always select level 1 and level 2 documents as relevant, and level 0 documents as non-relevant (by definition), therefore have a threshold of 0.5.

Each subject's relevance threshold was measured on three topics. However, examining the time taken to make each judgment, there was an obvious fatigue effect over the (at least) two hour judging period. For 8 users the average time taken to judge the final 10 documents of the final topic was over 100 seconds less than the average time to judge the first 10 documents of the first topic; while for half the users the difference was greater than 60 seconds. Accordingly, we exclude the final topic for all users from further analysis.

After excluding the final topic, we have 27 users judging documents for topics 707 and 771, and 26 users for topic 770. To obtain a single relevance threshold for each subject, the average response rate across the two topics at each stimulus level was calculated, and then the Weibull curve fit and the 50% point extracted. For the initial fit, it was assumed that the categorical TREC relevance scale was in fact numerical: that is, a category 2 document was twice as relevant as a category 1 document. Likewise, the threshold values are assumed to be continuous on this numerical scale. We also excluded any users whose probability of saying relevant was not non-decreasing over the categories, as fitting an ogive curve to these values is troublesome. For example, a user with probability of saying relevant of 0.5, 0.25, and 0.75 for the TREC categories 0,1 and 2 was excluded. To properly fit a psychometric function to these users, we should collect more data using the method of constant stimuli to either confirm that their values are not non-decreasing (in which case the psychophysical methods should not be applied), or to correct an abnormally low reading due to random variations in the user's responses. Our analysis of the 31 users where curve fits were obtained (explained below) led us to conclude that the psychophysical technique is probably not suitable for this analysis, so we did not collect further data on these users.

Figure 4 shows the relevance threshold for each of the remaining 31 users, sorted in increasing order. A threshold can be interpreted as the TREC relevance level above which there is a greater than 50% probability that the user will save the document as relevant. For example, User 30 at the far right of the graph has a threshold of 2.0, and so 50% of the time User 30 will save a TREC 2 (rounding the threshold) document as relevant. At the other end of the graph, User 23 has a threshold of 0.0, and so will save any TREC document more than 50% of the time. User 32 has a threshold of 1.0, and so will save documents with a TREC relevance level of 0 less than 50% of the time, documents with a TREC relevance category of 1 about 50% of the time, and documents judged in TREC category 2 more than 50% of the time. This makes User 32 more TREC-like than Users 30 and 23.

Overall, one can see a large variation in behaviour of users. 5 users have thresholds below 0.5, indicating that they are "trigger happy" relative to TREC judges, and will save a document regardless of its TREC relevance category more than 50% of the time. 9 users have a threshold above 1.5, indicating that they are much more conservative than TREC judges, and will save any document at TREC relevance level 0 or 1 less than 50% of the time.

Insisting that TREC Thresholds are Categorical

While interesting, the results presented in Figure 4 assumed that the TREC categorical scale could be interpreted numerically. This is a strong assumption, as TREC judges are not instructed to award a "2" to documents that are twice as relevant as documents that

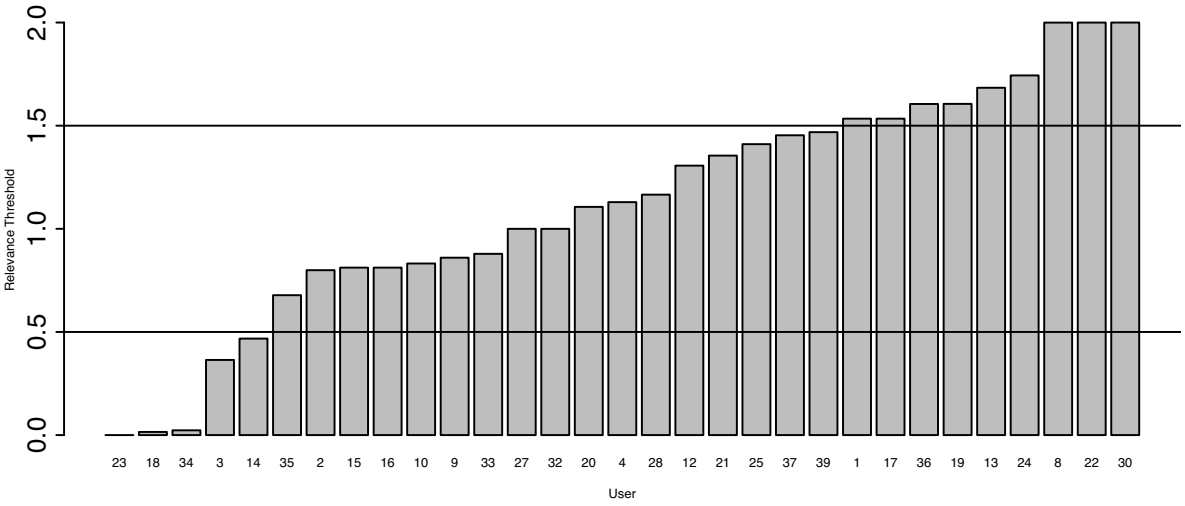

Figure 4. Thresholds of users for probability-of-saving, averaged over two topics, assuming TREC category 1 documents have numeric value 1, sorted in increasing order.

are judged "1". Accordingly, we varied the numerical weight of the TREC category one documents from 0.2 to 1.8 in steps of 0.2 and calculated thresholds as above. We kept TREC category 0 and 2 documents anchored at 0 and 2 on the numeric scale. We also round the numeric threshold obtained from the curve fit to the nearest category, to get a threshold of 0, 1 or 2.

Figure 5 shows four of the nine relevance thresholds we calculated for each user: 0.2, 0.4, 1.0 and 1.8. The five thresholds omitted for each user to avoid cluttering the figure all fell on points that are already plotted. As can be seen, 19 out of the 31 users have a consistent threshold. However, 12 users have one of two possible thresholds depending on the assumed numerical value of the level 1 category, with User 13 having all three possible values. The value of thresholds obtained with psychometric techniques, therefore, seems to depend heavily on the assumed underlying numerical relevance scale.

Agreement-based Approach

In effect we wish to categorise users into three categories: "generous", users who routinely say TREC category 0 documents are relevant (threshold < 0.5); "TREC-like", who do say TREC category 0 documents are irrelevant, but do say TREC category 1 and 2 documents are relevant ($0.5 \leq$ threshold < 1.5); and "parsimonious", who routinely say documents from TREC category 1 are not relevant (threshold ≥ 1.5). By examining the simple proportion of Category 0 documents saved and the number of Category 1 not saved, we can assign users to each class. From the top panel in Figure 6, we can see that Users 18, 34, 40, 23, 11 and 6 all say TREC category 0 documents are relevant for at least 50% of the 20 documents in this category that they judged. Similarly, in the bottom panel of Figure 6 we can see a group of 24 users that say that at least half of their 20 documents in category 1 are irrelevant. We note that two users, 6 and 50, appear in both groups. Using this approach to exclude generous and parsimonious users, we are left with 14 of our 40 users that seem to share, on average, the relevance criteria of TREC judges more than 50% of the time. We refer to this approach as the *split agreement* approach.

If we just compute simple agreement between each user and the TREC judges for the 60 documents judged by each user, only User 29 (43.3%) falls below 50% agreement with TREC judges. User 35 has the highest agreement score of 80%. The mean agreement over both topics for all users is 62% (standard deviation 7.8%). This is higher than in most previous studies: Voorhees [17] 32.8%, Sormunen [14] 39% and Turpin and Scholer [16] 45%, while on a par with a study by Al-Maskari at al. [3] which found 63% agreement on relevant documents.

Effect of Topic

Thus far we have only examined the behaviour of users averaged over both topics that they judged. Figure 7 shows the same data as Figure 6, but with the three topics shown separately. Recall that only the first two topics per user are included. Again, using the 50% cutoff criteria, the far right of both panels of the figure confirms the users that are either in the generous or parsimonious class from Figure 6. What is new in

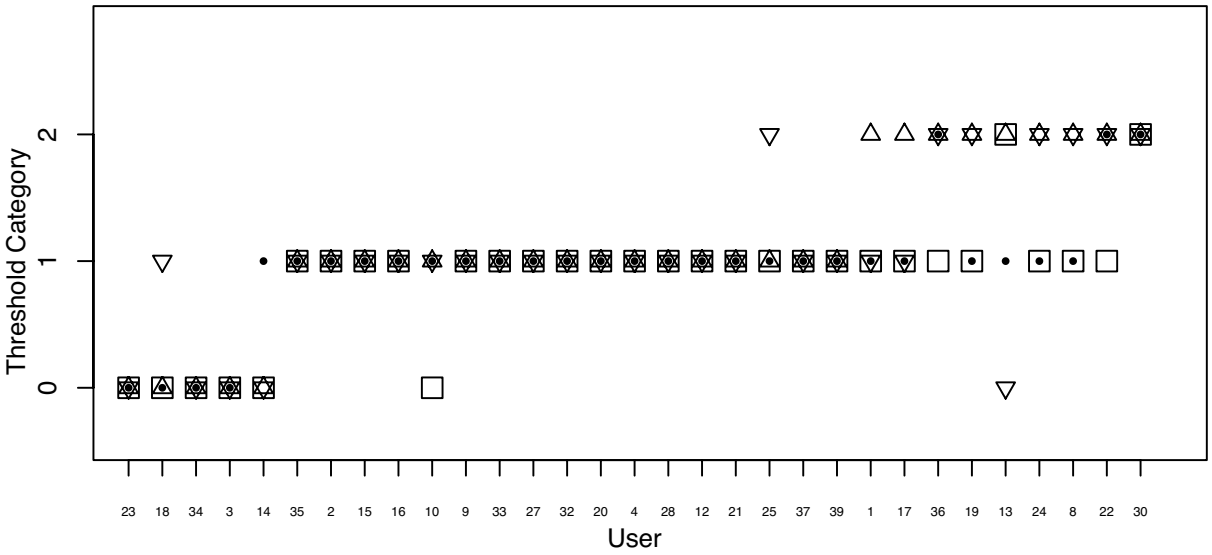

Figure 5. Thresholds of users for probability-of-saving averaged over two topics and rounded to the nearest TREC category, assuming TREC category 1 documents have numeric value 0.2 (solid circle), 0.4 (square), 1.0 (triangle up), and 1.8 (triangle down) sorted as in Figure 4.

TREC	User	Topic		
Category	Judgement	707	770	771
0	relevant	26%	24%	43%
1	not relevant	75%	42%	40%

Table 1. The proportion of "generous" (row 1) and "parsimonious" judgements (row 2) for each topic over all users.

Figure 7, however, is that there are several users that disagreed with the TREC categories by more than 50% on only one of their two topics, and hence the average remained below 50% in Figure 6. In particular, in addition to the 24 users in the bottom panel, there are 8 users who treated TREC category 1 documents as irrelevant more than 50% of the time for at least one topic. Likewise, in the top panel there are 10 more users that would be characterised as generous if topics are considered separately, than if the average over topics was used alone.

Examining the graphs in Figure 7, it also seems apparent that most generous users occur for Topic 771, and most parsimonious users occur with Topic 707. Table 1 shows the overall percentages of generous and parsimonious users for each topic. Pairwise t-tests support the intuition from Figure 7: Topic 771 contributes more generous users ($p < 0.005$) and Topic 707 contributes more parsimonious users ($p < 0.003$) that the other topics.

Drilling further into each topic, Figure 8 shows the proportion of users that judged a document into the opposite category from the TREC judgement for each document in each topic. Recall that each bar represents the proportion out of (at least) 26 users that judged that document. This demonstrates the reason many users are parsimonious on Topic 707: six of the ten documents in TREC category 1 for that topic are judged as irrelevant by nearly all users, and 8 of the 10 documents are judged irrelevant by over half the users. We have examined the documents, and observed that for at least 5 of them, the TREC judgements do not agree with the constraints given in the description and narrative of the topic (for example, the document contains only links to information, or does not specify a particular organisation).

Summary

In summary, we have examined the judgements of 40 users relative to TREC judgements for 30 documents on 3 topics. By using different criteria for quantifying the agreement between users and TREC judges, we get differing numbers of users who are "TREC-like". Table 2 summarises the four ways we segregated users. To be TREC-like, documents in categories 1 and 2 are considered relevant, and documents in TREC category 0 are irrelevant. Generous users said at least 50% of documents in TREC category 0 are relevant. Parsimonious users said that at least 50% of documents in TREC category 1 are irrelevant.

The first row of the table uses simple agreement between all of a user's judgements and the TREC judgements. Using the 50% cut-off, 39 of the 40 users behave TREC-like using this criteria. The second row

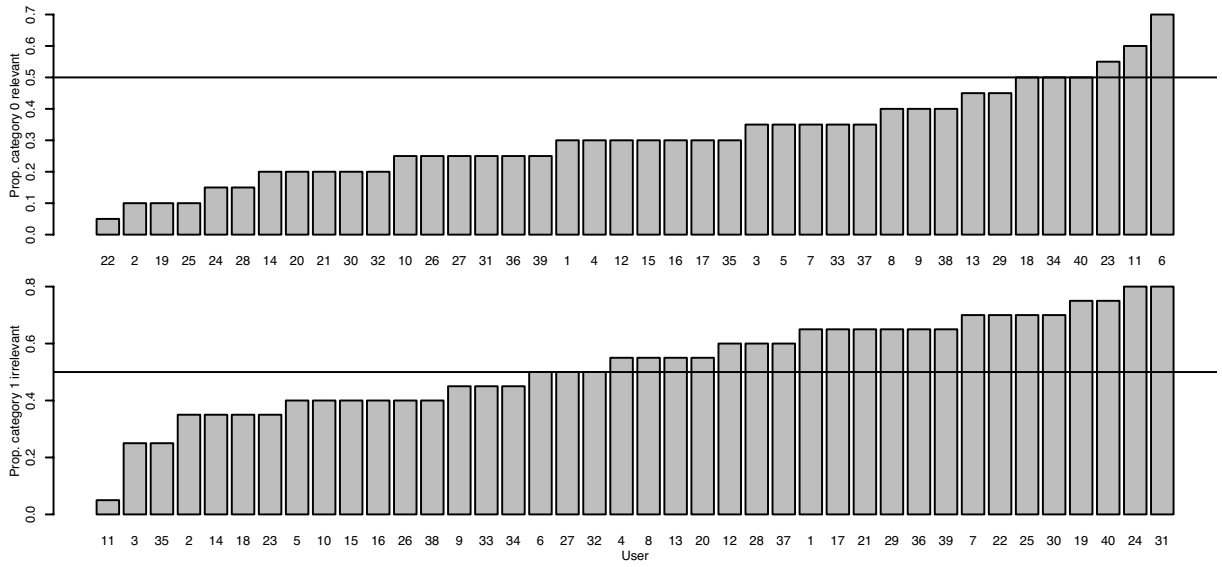

Figure 6. Proportion of documents of TREC category 0 that each user says are relevant (top panel); and the proportion documents of TREC category 1 that each user says are irrelevant (bottom panel).

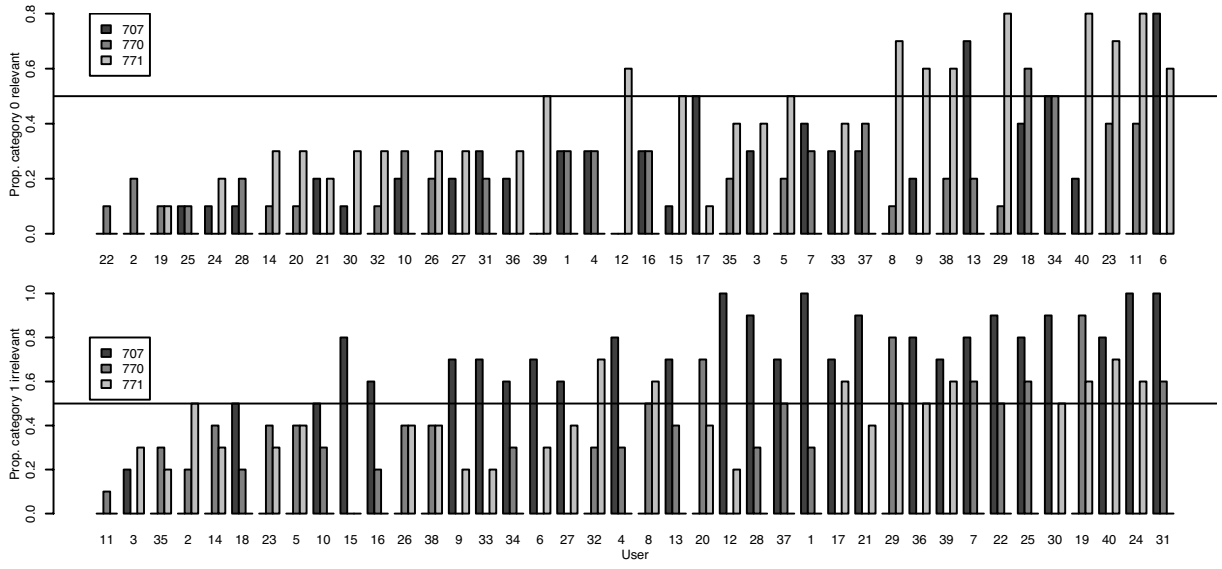

Figure 7. Proportion of documents of TREC category 0 that user says are relevant for each topic (top panel); and the proportion documents of TREC category 1 that each user says are irrelevant for each topic (bottom panel).

		Category			
Method	Data	Generous	TREC-like	Parsimonious	Total
Simple agreement	Average 2 topics	1 (3%)	39 (97%)	–	40
Thresholds	Average 2 topics	5 (16%)	17 (55%)	9 (29%)	31
Split agreement	Average 2 topics	6 (15%)	12 (30%)	24 (60%)	40
Split agreement	Either topic	16 (40%)	4 (10%)	32 (80%)	40

Table 2. Characterisation of our 40 users into three categories using three different methods.

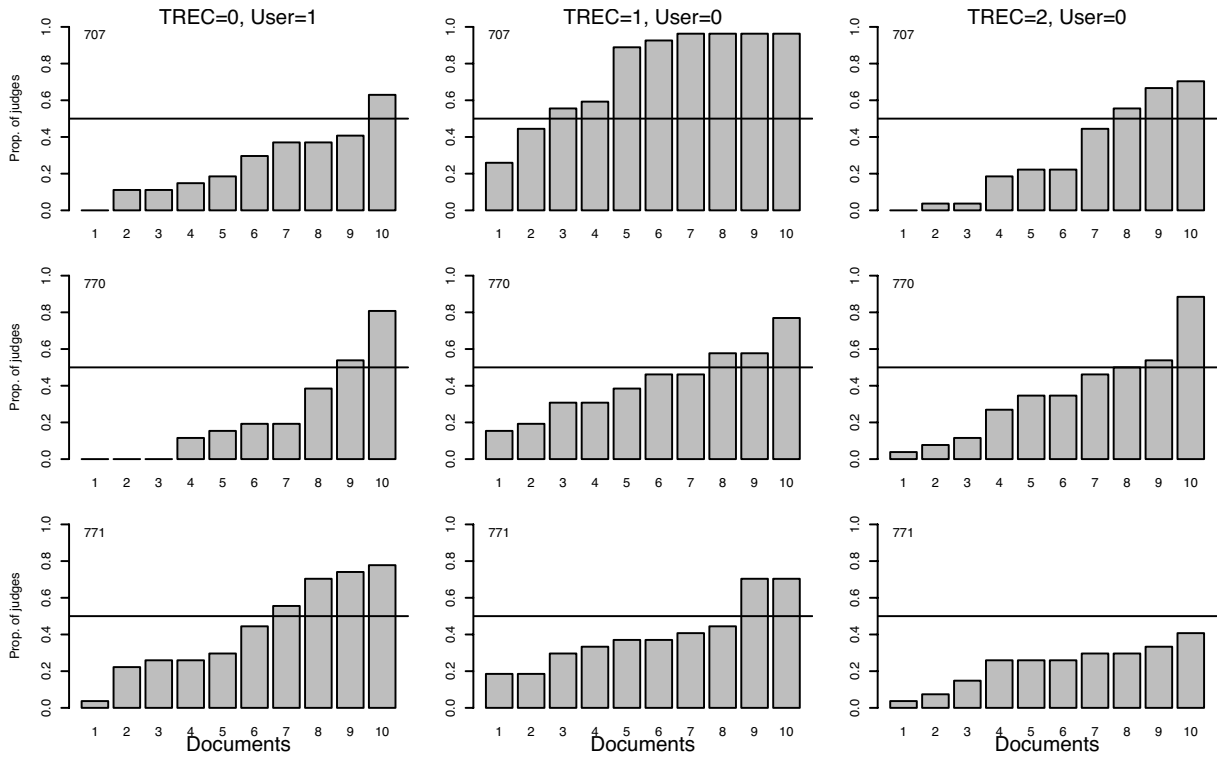

Figure 8. Proportion of users that judged documents in opposition to the TREC category assigned for each of the 10 documents for each topic.

of the table uses the technique from psychophysics to derive user's relevance thresholds, and records only 17 users as TREC-like. The final two rows use the split agreement strategy; firstly averaged over the two topics each user judged to give 12 TREC-like users; and secondly using a stricter criteria where the user is classed as generous or parsimonious if their judgements on either of their topics can be classed in such a way. Using this stricter criteria, only 4 of the 40 users are TREC-like on both of their topics. The percentages of the final two rows do not sum to 100, because a small number of users are both generous and parsimonious.

5 Discussion

TREC batch experiments assume that category 1 and 2 documents are relevant, and that category 0 are irrelevant. From our experiments, clearly there are users who do not use these relevance criteria, despite being instructed to judge documents in a similar fashion to TREC judges. Using strict alignment criteria, only ten percent (4/40) of our student users judged documents in a similar manner to TREC judges on two TREC topics. Perhaps it is unsurprising, therefore, that recent studies have found a mismatch between batch experiments on TREC data, and performance on users on the same data [1, 2, 10, 16].

We have asked the users in this study to perform a search task on the same data set that was used in this study of relevance judgements, and it will be interesting to see if the more TREC-like users perform better with systems ranked highly with batch experiments, while generous and parsimonious users do not.

Given that the TREC judges are different for different topics, it may be unreasonable to expect a single user to have the same relevance criteria as TREC judges on a range of topics. The results from this study therefore suggests which TREC topics are suitable for use in user studies where the population is drawn from undergraduate computing students. From Table 1 it seems clear that Topic 770 encourages user judgements that are more TREC-like than the other two topics used in this study. In addition to some problem cases where the TREC judgements do not seem to match the specifics given in the description and narrative fields of the topic, we plan to further investigate other possible causes of the lower agreement for Topic 707; perhaps the documents for topic 707 are less readable (either in layout or content), or our users have less previous knowledge of the topic.

Also, as observed by Bailey et al. [4], TREC judges have proposed the topic that they are judging, and have some knowledge of the collection and the types of doc-

uments they are looking for as relevant. Our users, on the other hand, have no knowledge of the GOV2 collection we used, and most likely no knowledge of the topics. The TREC judges are "gold" judges, and our users are "bronze" judges, in the terminology of Bailey et al.

Perhaps even more disconcerting is the variety of agreement that users have with TREC judges for documents in the same TREC category for the same topic. While category one documents for Topic 770 were uniformly voted irrelevant by our users, for all other categories and topics, there were documents for which all of our users agreed with the TREC judge, and others in the same category and topic for which our users generally disagreed.

Throughout this paper we have assumed that getting more than 50% agreement between TREC and user relevance judgements for a topic/document is desirable. This was chosen because our users were asked to do binary relevance assessments, and 50% would be the result of agreement expected if random judgements were made. If this threshold is raised to a more rigorous level, so that users must agree with TREC judges by much more than chance, then naturally the number of TREC-like users will fall. Examining the bottom panel of Figure 7 shows that if the criteria is raised to 80% agreement, then all but one user would be parsimonious, and all except three greedy!

Finally, we observe that while the psychophysics based approach [15] is attractive because of its well established rigor, there is a problem in its application when using a categorical (TREC) scale and trying to map this to a numerical scale: the choice of mapping level can have an effect on the final threshold values. In future work we plan to investigate whether this limitation can be overcome by using different approaches for measuring the relevance level of documents, for example through user-based judgements on a continuous scale. Secondly, from our data it was only possible to fit sensible psychometric functions to 31 of the 40 users. This may be due partly to problems with interpretations of the TREC topic specifications. Also, using different scales and measures to represent the relevance levels of documents may help to reduce the incidence of cases where the psychometric functions are abnormal.

When using a categorical TREC scale for relevance measurement, we would encourage the split agreement methodology used in this paper as a way of quantifying whether users are TREC-like, or not. In particular, the approach where the cutoff point is averaged over two topics leads to a minimal number of "unusual" users who are not TREC-like in both the parsimonious and generous dimensions.

References

[1] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 773–774, Amsterdam, Netherlands, 2007.

[2] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 433–440, Salvador, Brazil, 2005.

[3] A. Al-Maskari, M. Sanderson, and P. Clough. Relevance Judgments between TREC and Non-TREC Assessors In Proceedings of Thirty-First ACM SIGIR International Conference on Research and Development in Information Retrieval, pages 683 – 684, Singapore, 2008.

[4] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. de Vries and E. Yilmaz. Relevance Assessment: Are Judges Exchangeable and Does it Matter. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 667–674, Singapore, 2008.

[5] P. Borlund. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56(1):71–90, 2000.

[6] S. Büttcher, C. Clarke, and I. Soboroff. The TREC 2006 terabyte track. In *The Fifteenth Text REtrieval Conference (TREC 2006)*, Gaithersburg, MD, 2007. National Institute of Standards and Technology.

[7] W. H. Ehrenstein and A. Ehrenstein. Psychophysical methods. In U. Windhorst and H. Johansson, editors, *Modern techniques in neuroscience research*, pages 1211–1241. Springer, 1999.

[8] G. Gescheider. *Psychophysics: The Fundamentals*. Lawrence Erlbaum Associates, 3rd edition, 1997.

[9] S. P. Harter Variations in relevance assessments and the measurement of retrieva l effectiveness. *Journal of the American Society for Information Science*, 47(1):37-49, 1996.

[10] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kraemer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 17–24, Athens, Greece, 2000.

[11] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[12] T. Saracevic. Rlevance: A review of the literature and a framework for thinking on the notion in information science: Part II: Nature and manifestations of relevance. *Journal of the American Society for Information Science and Techology*, 58(13), pages 2126–2144, 2007.

[13] T. Saracevic. Effects of inconsistent relevance judgements on information retrieval test results: A historical perspective. *Library Trends*, 56(4), pages 763–783, 2008.

[14] E. Sormunen. Liberal relevance criteria of TREC – counting on negligible documents? In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 324–330, Tampere, Finland, 2002.

[15] F. Scholer and A. Turpin. Relevance thresholds in system evaluations. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 693-4, Singapore, 2008.

[16] A. Turpin and F. Scholer. User performance versus precision measures for simple web search tasks. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 11–18, Seattle, WA, 2006.

[17] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. Journal of Information Processing and Management, 36(5), pages 691–716, 2000.

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