Is relevance hard work? Evaluating the effort of making relevant assessments

Robert Villa
Information Retrieval Group
Information School
University of Sheffield
r.villa@sheffield.ac.uk

Martin Halvey
Interactive and Trustworthy Technologies Group
School of Engineering and Built Environment
Glasgow Caledonian University, UK
Martin.halvey@gcu.ac.uk

ABSTRACT
The judging of relevance has been a subject of study in information retrieval for a long time, especially in the creation of relevance judgments for test collections. While the criteria by which assessors' judge relevance has been intensively studied, little work has investigated the process individual assessors go through to judge the relevance of a document. In this paper, we focus on the process by which relevance is judged, and in particular, the degree of effort a user must expend to judge relevance. By better understanding this effort in isolation, we may provide data which can be used to create better models of search. We present the results of an empirical evaluation of the effort users must exert to judge the relevance of document, investigating the effect of relevance level and document size. Results suggest that "relevant" documents require more effort to judge when compared to highly relevant and not relevant documents, and that effort increases as document size increases.

Categories and Subject Descriptors
H.3.3 Information Search and Retrieval

General Terms
Measurement, Experimentation, Human Factors

Keywords
User studies, user models.

1. INTRODUCTION
The judgment of relevance has been a heavily studied topic within the Information Retrieval (IR) field, with a considerable number of papers concerned with the definition and modeling of relevance [1,2,3]. Relevance judgment is important to both the search process itself [4], and in the creation of test collections [3,5]. With regard to the latter, there is a considerable body of work which has investigated the criteria assessors use to judge relevance for the creation of text collections [3,5].

Within the context of user evaluations, the judgment of relevance is often an inseparable part of a wider information seeking process [6]. On the other hand, when generating relevance assessments for test collections, the behavior of assessors is not normally considered as important, beyond the overall time taken to create a set of relevance judgments [5,7]. Given the importance of relevance assessment to the information seeking process, the relative lack of research studying assessors is perhaps surprising.

In this paper we address this current gap by considering the behavior and effort of relevance assessors as an important subject of study by itself. Learning more about the relevance judgment process has potential applications to a number of areas of continuing research in IR, and in particular, has potential application to user simulation and modeling. When simulating and modeling users, a range of simplifications must inevitably be used in modeling user search behavior, such as assuming that users will linearly look through a ranked list in order, from top to bottom [8]. By isolating and empirically investigating the judgment of relevance from the wider information seeking process, this study aims to provide insights which can be applied to such simulations, allowing the introduction of more realistic user behavior in a controlled manner. The approach taken is to apply a narrow, controlled, user study to one aspect of search, rather than considering the user search process as a complex undividable entity.

In this work, we investigate the following two research questions:

RQ1: Does the size of the document being judged affect the effort and accuracy of the judging process?

RQ2: Does the degree of relevance of a document to a topic affect the effort and accuracy of the judging process?

In both cases we are interested in two main responses: the effort required (including the users perceived effort), and the accuracy by which the relevance assessments can be made. In research question one the focus is on document length, while research question two considers the level of relevance. With regard to this latter question, our working hypothesis is that documents which are clearly either highly relevant or not relevant will require less effort to judge than relevant or partially relevant documents.

2. PREVIOUS WORK
Relevance has been a central focus of IR research from the inception of the field [1], with much research effort expended on defining and modeling relevance [1,2]. While research into relevance has been undertaken from a wide range of different perspectives [1,2] one important strand has been the generation of relevance judgments for use in developing sets of "qrels" for test collections [3,5]. The effort involved in the assessment process itself has not generally been a focus for this work, however, except when reflecting on the time and costs of generating relevance assessments for test collections as a whole [3], in order to minimize the test collection building effort. One exception is the work of Wang [7] which investigated the speed and perceived difficulty of relevance assessment in E-Discovery. Much past

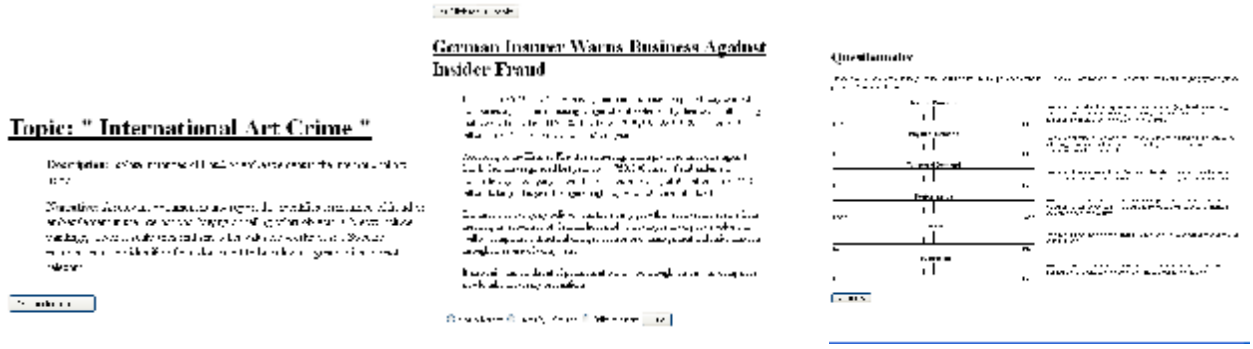

Figure 1: Screen shots of the experimental system showing the topic page (left), document page (center) and NASA TLX (right)

work has investigated the inter-assessor reliability of relevance assessments [2,5], and the criteria by which assessors judge documents [5]. Relevance judgments have also been studied as part of the overall information seeking process. Again, the focus has often been on the criteria by which users judge the relevance of a document to a task [4,9] but much of this work has also investigated how a user's conception of relevance changes over time, as the search process develops [4,6,9]. The effect of relevance level (i.e. multi-dimensional relevance judgments, partial relevance) has also been studied (e.g. [2]).

From the perspective of user studies, work has also investigated the effort involved in search as a whole. For example, in [10,11] the cognitive load in search was investigated, using a secondary task to indirectly measure the user's cognitive load on a main task. In [12] the NASA TLX [13] instrument was used to measure the task load of blind and non-blind users searching on a book search web site, finding that task load correlated with task time. Many other measures of task performance and effort have also been used in interactive studies, such as time and number of queries [14].

3. EXPERIMENTAL DESIGN

3.1 Design

In the experiment we are primarily interested in manipulating two independent variables: document size and relevance level. The study was designed as a relevance judgment task only, with users being presented with a search topic, and then a document. The task of the users was then to judge the relevance of the document to the topic. As users judged documents, the system would record the user's actions, and after each judgment the user's perception of task effort was gathered using a NASA TLX [13].

3.2 Data and topics

For the purposes of this study the AQUANT collection was chosen, along with the topics and relevance assessments (qrels) from the TREC HARD task from 2005 [15]. All 50 topics were used in the study (an example topic is shown in Figure 1). The relevance assessments available in the collection had three levels: "not relevant", "relevant" and "highly relevant", with all three relevance levels being used. Two independent variables were used: relevance, with three levels corresponding to the three TREC relevance levels, and document size, also with three levels (small, medium and large).

To classify AQUANT documents into the three sizes, word counts for all documents in the qrels were generated and sorted, and then split into three equally sized groups. The ten documents closest to the median length of each group were then extracted for each topic, and taken as representative small, median, and large documents. Not all topics had a full ten documents for each category. For the not relevant category, only documents judged as not relevant by TREC were used. The experiment used a randomized block design was used, where for each combination of document length and relevance level, a random topic and document was selected and presented to the participant.

3.3 Procedure

The study was implemented online, and was distributed to staff and students at Sheffield University, UK, as well as through social media channels. The webpages consisted of a short demographic questionnaire, a set of instructions, followed by nine topic and document combinations (Figure 1). The system first displayed the topic description to the participant, along with a button which could be used to display the document. At the bottom of the document page participants could then select the degree of relevance of the document (not relevant, relevant, and highly relevant), and then click to move on to the follow up NASA TLX. A button also allowed the participant to return to the topic description: they could move between topic and document as many times as they wished, but the view document button had to be clicked at least once.

After making a relevance judgment a NASA TLX questionnaire would be displayed. Only part 1 of the questionnaire was utilized, which is composed of six semantic differentials (mental demand, physical demand, temporal demand, performance, effort and frustration, all rated between 0 and 100). After completing this questionnaire the next topic would be displayed, and this process would continue for each of the nine topic and document combinations. No payment was made for participation. The study webpage was designed to control the size of page which would be viewed by the participant, as far as possible. On starting the study a new browser window would be opened with a specified width and height (1000x800 pixels), and a simple page design was used to ensure consistency between browsers. Not all web browsers allow a window to be fixed, although if the browser window was resized it would result in a logged event. A range of other events were also tracked, such as page scrolling.

4. RESULTS

In total 49 participants completed the survey: 27 females and 22 males with an average age of 29. The participants were

multinational and all indicated that they had advanced proficiency in English. In total the participants judged 409 unique documents across all 50 topics.

As much of the data analysed showed significant differences for Levene's Test, non-parametric statistical tests were used. Friedman tests were used, with pairwise comparisons made using Wilcoxon sign ranked tests (adjusted alpha = 0.0167). To determine the level of effort involved in making document judgments, both behavioral and subjective data was recorded. Behavioral data included time to make a judgment, and number of topic view clicks (i.e. number of times a user reviewed a topic). Subjective data was recorded with the NASA TLX.

4.1 Performance

Performance was compared using accuracy of judgment, true positive rate (TPR) and false positive rate (FPR), with the relevance assessments from the TREC HARD track used as the gold standard. The overall performance of users is comparable to users in a similar experimental set up by Smucker et al. [16] in terms of TPR and FPR indicating that our participants are representative. The accuracy of judgments was not influenced by document size ($c^2(2)=0.545$ p=0.761). The accuracy of judgment was influenced by document relevance level ($c^2(2)=11.091$ p=0.004). With significant differences between non-relevant and relevant documents (Z=-2.474 p=0.013) and highly relevant and relevant documents (Z=-2.889 p=0.004).

Table 1: Number of true negatives (TN), true positives (TP), false negatives (FN), false positives (FP), and accuracy of judgment by both document size and relevance. True positive rate and false positive rate by document size are also shown.

	TN	TP	FN	FP	Acc.	TPR	FPR
All	119	221	72	28	0.7727	0.7543	0.1905
Sml	42	72	25	7	0.7808	0.7423	0.1429
Med	37	73	25	12	0.7483	0.7449	0.2449
Lrg	40	76	22	9	0.7891	0.7755	0.1837
Not	119	--	--	28	0.8095	--	--
Rel	--	99	47	--	0.6781	--	--
High	--	122	25	--	0.8299	--	--

4.2 Effort

4.2.1 Time and topic views

Document size had a significant effect on time to make a relevance judgment ($c^2(2)=73.658$ p<0.001). With pairwise comparisons showing differences between small and large (Z=-8.030 p<0.001) and medium and large (Z=-6.439 p<0.000). Document size did not influence topic views.

Table 2: Mean time (SD) to judge document and mean topic view clicks (SD), by document size and document relevance

	Time secs (SD)	Topic View (SD)
Sml	63.28 (49.8)	0.40 (0.59)
Med	86.97 (78.6)	0.40 (0.58)
Lrg	145.59 (24.8)	0.33 (0.54)
Not	100.65 (20.8)	0.25 (0.48)
Rel	95.40 (85.2)	0.47 (0.61)
High	100.73 (15.3)	0.40 (0.40)

Document relevance had a significant effect on time to make a relevance judgment ($c^2(2)=7.575$ p=0.023). No significant differences were found in the pairwise comparisons at the adjusted alpha. Document relevance had an influence on topic views ($c^2(2)=12.444$ p=0.002). There were significant differences between not-relevant and relevant (Z=-3.641 p<0.000) and not-relevant and highly relevant (Z=-2.868 p=0.004).

4.2.2 Subjective Effort

Each of the 6 NASA TLX semantic differentials was compared across document size and document relevance level. The general trend for most of the categories is that demand increases as size of document increases, the exception being perceived performance where the values decrease as document size increases. For mental demand the differences were found to be significant ($c^2(2)=21.669$ p<0.001. *Post hoc* tests showed differences between small and large documents (Z=-4.270 p<0.001). For physical demand the differences were found to be significant ($c^2(2)=29.903$ p<0.001). *Post hoc* tests showed differences between small and large documents (Z=-5.370 p<0.001) as well as medium and large documents (Z=-4.440 p<0.001). For temporal demand the differences were found to be significant ($c^2(2)=35.804$ p<0.001). *Post hoc* showed differences between small and medium documents (Z=-3.804 p<0.001), small and large documents (Z=-5.698 p<0.001) and medium and large documents (Z=-3.476 p=0.002). Differences in effort were found to be significant ($c^2(2)=13.386$ p=0.001). *Post hoc* tests showed differences between small and large documents (Z=-3.732 p<0.001) and medium and large documents (Z=-2.567 p=0.010). Differences in frustration were also found to be significant ($c^2(2)=18.922$ p<0.001). *Post hoc* tests showed differences between small and medium documents (Z=-3.488 p<0.001) and small and large documents (Z=-4.449 p<0.001). There was also a significant difference in terms of perceived performance ($c^2(2)=8.646$ p=0.013). *Post hoc* tests showed differences between small and medium documents (Z=-2.476 p=0.013) and small and large documents (Z=-2.773 p=0.006).

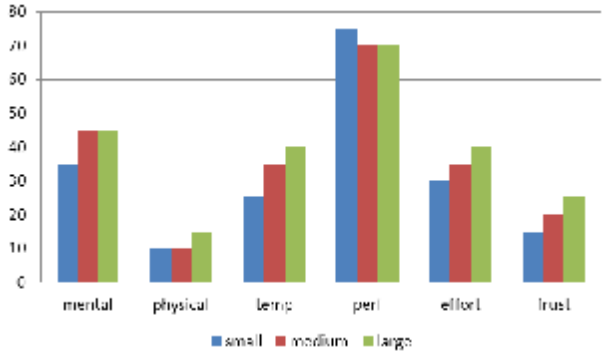

Figure 2: Median subjective ratings for each of the 6 semantic differentials by document size (y-axis: user rating 0-100)

When the results for document relevance were analysed the general trend is that the relevant documents required the highest workload to judge. For mental demand the differences were found to be significant ($c^2(2)=11.499$ p=0.003). *Post hoc* tests showed differences between non-relevant and relevant documents (Z=-3.445 p=0.001) and highly relevant and relevant documents (Z=-2.550 p=0.011). For physical demand the differences were found to be significant ($c^2(2)=7.154$ p=0.028). *Post hoc* showed

differences between not relevant and relevant documents (Z=-2.483 p=0.013). Differences in effort were found to be significant ($c^2(2)$=12.725 p=0.002). *Post hoc* tests showed differences between not relevant and relevant documents (Z=-3.198 p=0.001).

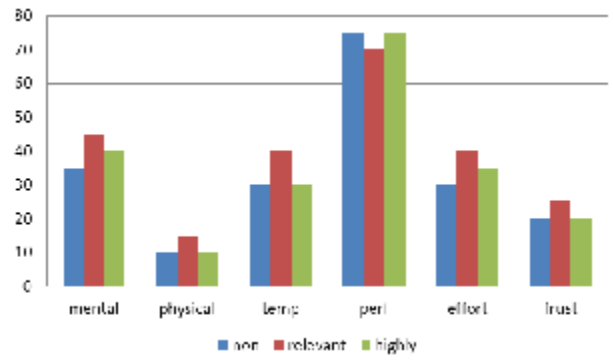

Figure 3: Median subjective ratings for each of the 6 semantic differentials by document relevance (y-axis: user rating 0-100)

5. DISCUSSION

Considering RQ1, which looked at the relationship between document length and both effort and accuracy, it can be seen from Table 1 (5th column) that accuracy is not affected by document size. However, looking at Table 2, it can be seen that document size does have a significant effect on how long participants took to judge a document: as might be expected, longer documents took longer to judge (Table 2, 1st column). Looking next at subjective effort, the general trend is for effort to increase as document size increases (Figure 2) with the exception of perceived performance, which shows the reverse. This suggests that participants did perceive the judging of longer documents as requiring more effort.

Considering RQ2, first considering accuracy (Table 1), there were significant differences between relevant documents and both not relevant and highly relevant documents. For this latter case, Table 1 shows accuracy for relevant documents decreasing to 67.8%, from 80.1% and 83.0% for the not relevant and highly relevant cases. While a significant overall effect was found between time and document relevance level, no significant pairwise comparisons were found. Perhaps surprisingly, on average participants judged relevant documents quicker than not relevant and highly relevant, although these pairwise differences are not significant. Topic view clicks were higher for relevant documents when compared to not relevant and highly relevant, suggesting that participants tended to switch between the topic and document more when judging relevant documents.

Lastly, looking at the subjective effort (Section 4.4.2), results are more complex. Looking at document relevance (Figure 3), the results suggest that it is the relevant documents which require most effort to judge (significant differences were found for mental demand, physical demand, and effort). As can be seen in Figure 3, a similar non-significant trend can be seen for temporal demand and frustration. Interestingly, for *performance* this trend is reversed: the trend is for users to be less secure in their performance for relevant documents.

6. CONCLUSIONS AND FUTURE WORK

From the results presented in this paper, we can make the following two conclusions: (1) document length does affect the effort required to judge a document, but does not affect the accuracy; and (2) the degree of relevance of a document does affect both accuracy and effort: the trend is for accuracy to decrease for relevant documents, and perceived effort to increase.

Implications: simulations and evaluation metrics should take account of both document size and relevance level (where possible) when simulating users. While length does not appear to affect accuracy, it does affect effort, and simulations should take account of this. Similarly, the effort required to judge the relevance of a document varies based on its degree of relevance. In future work we aim to consider how the results of this study can be integrated into simulations of the search process.

7. REFERENCES

[1] Mizzaro. S., 1997. Relevance: The whole history. *Journal of the American Society for Info. Science*, 48(9), 810-832.

[2] Spink, A., Greisdorf, H., and Bateman, J. 1998. From highly relevant to not relevant: examining different regions of relevance. *Inf. Process. Manage.* 34, 5 (Sept 1998), 599-621.

[3] Carterette, B., Allan, J. and Sitaraman, R. 2006. Minimal test collections for retrieval evaluation. *SIGIR 2006*, 268–275.

[4] Tang R., Solomon P. 1998. Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior, *Inf. Process. Manage*, 34 (2-3), 237-256.

[5] Sormunen, E. 2002. Liberal relevance criteria of TREC: counting on negligible documents? *SIGIR 2002*, 324-330.

[6] Taylor, A. 2011. User relevance criteria choices and the information search process. *IP&M*. 48, 136-153.

[7] Wang, J. 2011. Accuracy, agreement, speed, and perceived difficulty of users' relevance judgments for e-discovery. SIGIR 2011 Information Retrieval for E-Discovery (SIRE) Workshop, Beijing, China, July 28, 2011.

[8] Carterette, B. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. *ACM SIGIR '11*. ACM, New York, NY, USA, 903-912.

[9] Vakkari, P. 2000. Relevance and contributing information types of searched documents in task performance. *SIGIR 2000*, 2-9.

[10] Jacek Gwizdka. 2010. Distribution of cognitive load in Web search. J. Am. Soc. Inf. Sci. Technol. 61, 11 (Nov 10), 2167-2187.

[11] Gwizdka, J. (2009). Assessing Cognitive Load on Web Search Tasks. *The Ergonomics Open Journal*. Bentham Open Access.

[12] Iizuka, J., Okamoto, A., Horiuchi, Y., Ichikawa, A. 2009. Considerations of Efficiency and Mental Stress of Search Tasks on Websites by Blind Persons. UAHCI '09, 693-700.

[13] Hart, S.G., Staveland, L.E. 1988. Development of a NASA-TLX (Task load index): Results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, 139–183.

[14] Louise T. Su. 1992. Evaluation measures for interactive information retrieval. *IP&M* 28, 4 (March 1992), 503-516.

[15] Allan, J. 2005. HARD Track Overview, TREC 2005

[16] Smucker M.D., Jethani, C.P. 2011. Measuring assessor accuracy: a comparison of nist assessors and user study participants. *SIGIR 2011*, 1231-1232.