Time-Limits and Summaries for Faster Relevance Assessing

Shahin Rahbariasl
srahbari@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

Mark D. Smucker
mark.smucker@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

ABSTRACT

Relevance assessing is a critical part of test collection construction as well as applications such as high-recall retrieval that require large amounts of relevance feedback. In these applications, tens of thousands of relevance assessments are required and assessing costs are directly related to the speed at which assessments are made. We conducted a user study with 60 participants where we investigated the impact of time limits (15, 30, and 60 seconds) and document size (full length vs. short summaries) on relevance assessing. Participants were shown either full documents or document summaries that they had to judge within a 15, 30, or 60 seconds time constraint per document. We found that using a time limit as short as 15 seconds or judging document summaries in place of full documents could significantly speed judging without significantly affecting judging quality. Participants found judging document summaries with a 60 second time limit to be the easiest and best experience of the six conditions. While time limits may speed judging, the same speed benefits can be had with high quality document summaries while providing an improved judging experience for assessors.

KEYWORDS

Relevance Assessing; Time Limits; Document Summaries

ACM Reference Format:
Shahin Rahbariasl and Mark D. Smucker. 2019. Time-Limits and Summaries for Faster Relevance Assessing. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331270

1 INTRODUCTION

The idea of a speed accuracy tradeoff (SAT) is well known and embedded in people's common sense understanding [4]. For many tasks, we intuitively know that we can speed our work at the risk of having quality suffer. Likewise, we know that to obtain high quality work, we need to give ourselves time to do a job carefully.

When assessors are hired to collect relevance judgments for tasks such as test collection construction [9], speed and accuracy are chief concerns. The faster that assessors judge, the sooner we are done and usually the lower the cost. While faster judgments are better, we want to maintain accuracy or at least make an informed decision about the tradeoff between speed and accuracy.

Many factors influence the speed and accuracy of relevance judging with the chief ones being the assessors, the search topic or task, and the items being judged. For example, Smucker and Jethani [10] found that crowdsourced assessors judged newswire articles at an average rate of 15 seconds per document while supervised assessors in a laboratory setting took 27 seconds per document. In their experiment, they found lab workers to have somewhat better ability to discriminate relevant from non-relevant documents, but this difference was not statistically significant. For legal e-discovery tasks, Oard and Webber [8] report that a few minutes per document is typical.

In Heitz [4]'s review of the speed accuracy tradeoff, he notes that behavioral science has come to see the process as one where sensory information is accumulated in a sequential fashion and lower accuracy results from less accumulated information on which to make a decision. Assuming that we can extend this understanding of the speed accuracy tradeoff to relevance assessing, we can think of an assessor as scanning, skimming, and reading a document and at any moment, we can stop the sequential input and ask for a relevance judgment. The less time the assessor has, the less material has been consumed on which to make a judgment. Placing a time limit on judging is akin to limiting the assessor to a summary of the document, albeit a summary formed in their mind via their reading behavior up to the point at which time runs out.

In this paper, we investigate the effect of time limits and document summaries on relevance assessing speed and accuracy. We conducted a controlled laboratory study that used a factorial design and varied the time allowed (15, 30, and 60 seconds) and the document form (full length newswire or a summary). Our summaries were created based on a model of relevance for the topic, and thus our results for summaries represent the potential of summaries rather than the specific performance of a given summarization algorithm.

We found that providing more time does result in better discrimination between relevant and non-relevant documents, but most of this advantage was for full documents rather than summaries, i.e. time limits less than 60 seconds hurt the discrimination ability of assessors when judging full documents. Little difference was found in discrimination ability for short summaries at the various time limits. In addition, no statistically significant difference was found in accuracy of judging for the time limits or for full documents vs. document summaries.

After each search task, we asked participants about the ease of judging and other factors, and judging summaries with a time limit of 60 seconds was most favored by participants. Indeed, the average time to judge a summary under the 60 second limit was only 13.4 seconds. In other words, with effectively no time limit, judging

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331270

summaries can be fast with little loss in quality. Time limits with full documents can be used to force the relevance assessor to make a judgment based on their limited examination of the document with little loss in accuracy, but this comes with the additional cost of time pressure and stress on the assessors.

2 RELATED WORK

In a series of four experiments, Maddalena et al. [6] investigated crowdsourced assessors' judging behavior and the effect of placing time limits on judging. In these experiments, the crowdworkers judged newswire articles from TREC 8. If workers were given a time limit, Maddalena et al. found that a shorter time of 15 seconds resulted in better assessor agreement with the official relevant judgments as compared to a longer 30 seconds time limit. Other experiments investigated judging without time limits, time limits from 3 to 30 seconds and the advantage of a fixed judging time vs. a time limit. Time limits less than 15 seconds resulted in lower judging quality. They also discovered that a fixed judging time of 25-30 seconds was best. A fixed judging time requires the assessor to submit the judgment at the end of the time period and likely encourages the assessor to keep reading a documment until the time is up. In contrast to our work, Maddalena et al. [6] focused solely on crowdsourced workers while we used university students in a supervised setting and also investigated summaries vs. full documents.

Wang and Soergel [12] performed a user study on different parameters in relevance assessment including agreement, speed, and accuracy. Moderate to a substantial agreement was reported between the two groups of students. There was no significant difference in the speed of the two groups of assessors but the speed varied between the individuals.

Similarly, Wang [11] studied the accuracy, agreement, speed and perceived difficulty in relevance judgments for e-discovery. Wang found speed and perceived difficulty to be correlated as well as perceived difficulty and accuracy, but Wang did not find a correlation between accuracy and speed. Regarding speed, only a small fraction of documents slowed down the assessors. In contrast to our work and Maddalena et al. [6], Wang [11] did not apply time limits and thus working faster or slower would have been at the discretion of the assessors. Assuming most assessors make a legitimate attempt to do a good job, it is reasonable to assume that assessors without time limits will modulate their time to achieve a uniform level of accuracy.

3 MATERIALS AND METHODS

We conducted a 2x3 factorial experiment. Our two factors were document type (full document or summary) and time limit (15, 30, and 60 seconds). In this section we describe the details of the experiment.

3.1 User Study

After receiving ethics clearance from our university's office of research ethics, we recruited participants, and after pilot testing, we had 60 people participate in the final study. Each participant started with a tutorial describing the experiment and practice using the interface to judge the relevance of 5 documents.

For the main task, participants judged 20 documents for each of six search topics that were different from the topic used in the tutorial phase. In total, participants judged 120 documents. The 2x3 factorial design resulted in 6 treatments. For a given topic, a participant received a single treatment. Of the 20 documents for each topic, half were relevant and half were non-relevant. Each participant saw a randomized order of the topics and documents. The six treatments were balanced across users and task order with a Latin square. Before each search task, participants answered a pre-task questionnaire that asked them about their knowledge about the topic and other matters. After each search task, participants answered a questionnaire about the judging experience. We paid each participant $20 for their participation.

3.2 User Interface

We designed the user interface to show the participants one document at time. Participants could see the title of the document and either the full document or a short document summary. The only actions allowed were to submit a judgment of relevant or non-relevant. Through the whole study, participants could see the search topic and its description for that task. Participants were also provided with information about the number of documents and the time left for judging each document.

The study involved three different time constraints for sets of 20 documents: 15, 30, and 60 seconds per document. When participants ran out of time, the document was hidden and participants had to submit their judgment to proceed. We recorded the overall time they spent on judging the document including the time after hiding the document.

3.3 Topics and Documents

In total, we used seven topics from the 2017 TREC Common Core track [1] and documents from the New York Times annotated corpus. Six topics were for the main task (topic ids = 310, 336, 362, 426, 427, 436), and one was for the tutorial phase (id=367). We combined the topic description and narrative into a single description for display to the participants.

We carefully selected documents for the experiment to avoid biasing judging errors toward false positives or false negatives. For each topic, we first computed a document ordering using reciprocal rank fusion (RRF) [3] from the runs submitted by groups participating in the 2017 TREC Common Core track. Documents ranked highly by RRF are considered likely to be relevant by the majority of participating retrieval systems. Any highly ranked non-relevant document likely looks relevant on its surface and likely will lead to false positives by assessors. Likewise, a relevant document ranked lowly by RRF will likely generate false negatives by assessors. To get a good mix of documents, we divided each RRF list into an ordered list of NIST judged relevant and a list of non-relevant documents. We then split each list in half to produce four lists and then randomly sampled 5 documents from each list. In this fashion, we have balanced the types of errors that users will make when judging.

3.4 Document Summaries

For three out of the six main tasks, we showed the users paragraph-long summaries instead of full documents. We use a modification of the method of Zhang et al. [14] to create the summaries. This method trains a classifier to recognize relevant material and after dividing a document into paragraphs, selects the paragraph with the highest likelihood of being relevant as the summary. Our modification is that we trained the classifier based on the known relevant and non-relevant documents from the NIST judgments. As such, our document summaries represent an upper bound on the technique's potential for selecting the most relevant paragraph from a document.

3.5 Measures of Accuracy and Discrimination

We define the NIST relevance judgments to be the truth and measure performance against them. As such, a basic measure is accuracy:

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |FP| + |TN| + |FN|} \quad (1)$$

where $|TP|$ is the number of true positives (NIST says relevant and participant says relevant), and $|TN|$ is the number of true negatives (agree on non-relevant), $|FP|$ is the number of false positives, and $|FN|$ is the number of false negatives. Accuracy cannot tell us about the types of errors being made, and thus we also report the true positive rate (TPR) and false positive rate (FPR). The TPR is the fraction of relevant documents judged as relevant, and the FPR is the fraction of non-relevant documents judged as relevant. Relevance judging can be understood as a discrimination task, and as such, a measure of discrimination ability is d' [5]. Since true positive or false positive rates of 1 and 0 can lead to infinities in the computation of d', we employ a standard smoothing mechanism of adding a pseudo-document to the count of documents judged. Therefore, the estimated true positive rate (eTPR) is defined as: $eTPR = (|TP| + 0.5)/(|TP| + |FN| + 1)$, and estimated false positive rate (eFPR) is calculated as: $eFPR = (|FP| + 0.5)/(|FP| + |TN| + 1)$, and d' is computed as: $d' = z(eTPR) - z(eFPR)$, where the function z is the inverse of the normal distribution function.

To measure statistical significance of results, we used generalized linear mixed-effects models, as implemented in the lme4 [2] package in R. Both the participants and topics are random effects and the experiment factors (time limits and document type) are fixed effects. We measure the statistical significance of the factor on the dependent variable (e.g. accuracy) by building a complete model and then a model without the factor that we are testing. By comparing these two models with a likelihood ratio test, we obtain the reported p-value.

4 RESULTS AND DISCUSSION

Table 1 shows the average time participants took to judge a document, and Table 2 shows the fraction of judgments that exceeded the time limit. As the time limit increases, the average time to judge a document increases regardless of whether it is a full document or a summary. When the time limit is 60 seconds, participants rarely exceed the time limit and we see that full documents take on average 22.6 seconds to judge in comparison to only 13.4 seconds for a summary. When a time limit of 15 seconds is imposed, both full documents and summaries take only 9.8 and 9.1 seconds respectively. In effect, the 15 seconds time limit forces an assessor to consume material on par with a paragraph long summary. For similar judging tasks, Zhang [13] found average, unrestricted (no time limits) judging times of 22.7 seconds for summaries and 50.0 seconds for full documents.

	Time Limit (seconds)				
Doc. Type	**15**	**30**	**60**	**Mean**	***p*-value**
Full doc.	9.8	15.2	22.6	15.9	$p \ll 0.001$
Summary	9.1	11.5	13.4	11.3	
Mean	9.5	13.4	18.0	13.6	
***p*-value**	$p \ll 0.001$				

Table 1: Average time in seconds to judge a document.

	Time Limit (seconds)			
Doc. Type	**15**	**30**	**60**	**Mean**
Full doc.	0.25	0.14	0.04	0.14
Summary	0.17	0.04	0.00	0.07
Mean	0.21	0.09	0.02	0.11

Table 2: Fraction of judgments that exceeded time limit.

	Time Limit (seconds)				
Doc. Type	**15**	**30**	**60**	**Mean**	***p*-value**
Full doc.	0.70	0.71	0.73	0.71	$p = 0.18$
Summary	0.73	0.72	0.74	0.73	
Mean	0.71	0.71	0.74	0.72	
***p*-value**	$p = 0.09$				

Table 3: Average accuracy.

Table 3 shows the average accuracy for the 60 participants and the six experimental treatments. As explained in Section 3.5, we report the statistical significance of each of the experimental factors' effect on accuracy, and as Table 3 shows, neither the time limits nor the document type had a statistically significant effect on accuracy. In contrast to accuracy, Table 5 shows that time has a statistically significant effect on the d' measure of discrimination ability. As with accuracy, discrimination ability with summaries is not different from full documents at a statistically significant level.

Tables 6 and 7 show the true positive and false positive rates for each of the six treatments. The higher true positive rate for summaries along with the effectively same false positive rate as for judging full documents shows that summaries have the potential to help assessors identify relevant material while not simply raising the false positive rate. With full documents, as the participants had more time, they were able to consume more material and suppress false positives. Thus, while there is little advantage to judging full documents in terms of accuracy, the evidence shows that full documents result in more conservative judgments.

Finally, Table 4 shows the results for a selection of the post-task questionnaire's questions. In comparison to full documents, participants found judging summaries with a 60 second time limit, which is akin to having no time limit for such small amounts of text,

Question	Treatment (time limit, doc. type)						
	(15, F)	(30, F)	(60, F)	(15, S)	(30, S)	(60, S)	All
Difficulty (1=very difficult, 5=very easy)	2.5	3.1	2.8	3.2	3.4	3.6	3.1
Experience (1=very unenjoyable, 5=very enjoyable)	3.0	3.0	2.9	3.2	3.5	3.4	3.2
Mood (1=very bored, 5=very engaged)	3.3	3.3	3.2	3.6	3.3	3.2	3.3
Concentration (1=very hard, 5=very easy)	2.9	3.0	3.1	3.4	3.5	3.7	3.3
Confidence (1=very uncertain, 5=very confident)	3.2	3.2	3.4	3.3	3.6	3.8	3.4
Time Pressure (1=very stressed, 5=very relaxed)	2.1	2.4	2.8	2.3	2.9	3.0	2.6
Accuracy (1=very inaccurate, 5=very accurate)	3.2	3.4	3.4	3.4	3.7	3.9	3.5

Table 4: Post task questionnaire. Participants rated their feelings and self-perceived performance. A value of 3 is "neutral".

	Time Limit (seconds)				
Doc. Type	**15**	**30**	**60**	**Mean**	***p*-value**
Full doc.	1.07	1.18	1.31	1.19	$p = 0.13$
Summary	1.27	1.22	1.36	1.28	
Mean	1.17	1.20	1.34	1.23	
***p*-value**	$p = 0.047$				

Table 5: Average ability to discriminate (d').

	Time Limit (seconds)				
Doc. Type	**15**	**30**	**60**	**Mean**	***p*-value**
Full doc.	0.64	0.64	0.64	0.64	$p = 0.03$
Summary	0.70	0.66	0.67	0.68	
Mean	0.67	0.65	0.65	0.66	
***p*-value**	$p = 0.63$				

Table 6: Estimated true positive rates.

	Time Limit (seconds)				
Doc. Type	**15**	**30**	**60**	**Mean**	***p*-value**
Full doc.	0.28	0.26	0.22	0.25	$p = 0.29$
Summary	0.28	0.27	0.23	0.26	
Mean	0.28	0.26	0.23	0.26	
***p*-value**	$p < 0.001$				

Table 7: Estimated false positive rates.

to be easier, more enjoyable, and less stressful. At the same time, participants were more confident and believed they were more accurate in their judging of summaries under the 60 second limit.

5 CONCLUSION

We conducted a user study and investigated the impact of time limits and document size on relevance assessing. We found no difference in the quality of judgments with summaries in comparison to full documents. As noted in Section 3.4, our summaries represent an upper bound on performance as we selected the paragraph most likely to be relevant given a classifier trained using known relevance judgments. Even so, these results are in line with past research that has shown little difference in relevance judging accuracy between summaries and full documents [7].

Imposing a time limit can speed judging of both full documents and summaries without a loss of accuracy, which shows how little material assessors need to consume from typical newswire documents for making judgments. An aggressive time limit of 15 seconds produced judgments at a rate of 9.5 seconds per document, but such time limits increased the stress on the participants and reduced their enjoyment of the task.

Summaries appear to be a better solution to speeding relevance judging than time limits. With a generous time limit of 60 seconds, our assessors averaged 13.4 seconds per document and this treatment resulted in the best experience as measured by our post-task questionnaire.

ACKNOWLEDGMENTS

Thanks to Nimesh Ghelani and Adam Roegiest for their technical contributions. Thanks to Gordon Cormack and Maura Grossman for their feedback. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (Grants CRDPJ 468812-14 and RGPIN-2014-03642).

REFERENCES

[1] Allan, J., E. Kanoulas, D. Li, C. V. Gysel, D. Harman, and E. Voorhees (2017). Trec 2017 common core track overview. In *TREC*.
[2] Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.
[3] Cormack, G. V., C. L. Clarke, and S. Buettcher (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *SIGIR*, pp. 758–759.
[4] Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience 8*, 150.
[5] Macmillan, N. and C. Creelman (2005). *Detection theory: a user's guide*.
[6] Maddalena, E., M. Basaldella, D. De Nart, D. Degl'Innocenti, S. Mizzaro, and G. Demartini (2016). Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge. In *Fourth AAAI Conf. on Human Comp. and Crowdsourcing*.
[7] Mani, I., G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim (2002). Summac: a text summarization evaluation. *Natural Language Engineering 8*(1), 43–68.
[8] Oard, D. W. and W. Webber (2013). Information retrieval for e-discovery. *Foundations and Trends in Information Retrieval 7*(2-3), 99–237.
[9] Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval 4*(4), 247–375.
[10] Smucker, M. D. and C. P. Jethani (2011). The crowd vs. the lab: A comparison of crowd-sourced and university laboratory participant behavior. In *Proceedings of the SIGIR 2011 Workshop on crowdsourcing for information retrieval*.
[11] Wang, J. (2011). Accuracy, agreement, speed, and perceived difficulty of users' relevance judgments for e-discovery. In *Proceedings of SIGIR Information Retrieval for E-Discovery Workshop*, Volume 1.
[12] Wang, J. and D. Soergel (2010). A user study of relevance judgments for e-discovery. *Proc. of the Assoc. for Information Sci. and Tech. 47*(1), 1–10.
[13] Zhang, H. (2019). *Increasing the Efficiency of High-Recall Information Retrieval*. PhD thesis, University of Waterloo.
[14] Zhang, H., M. Abualsaud, N. Ghelani, A. Ghosh, M. D. Smucker, G. V. Cormack, and M. R. Grossman (2017). UWaterlooMDS at the TREC 2017 common core track. In *TREC*.