Proceedings, The Fourth AAAI Conference on
Human Computation and Crowdsourcing

Crowdsourcing Relevance Assessments: The Unexpected Benefits of Limiting the Time to Judge

Eddy Maddalena
University of Udine, Italy
eddy.maddalena@uniud.it

Marco Basaldella
University of Udine, Italy
basaldella.marco.1@spes.uniud.it

Dario De Nart
University of Udine, Italy
dario.denart@uniud.it

Dante Degl'Innocenti
University of Udine, Italy
dante.deglinnocenti@spes.uniud.it

Stefano Mizzaro
University of Udine, Italy
mizzaro@uniud.it

Gianluca Demartini
University of Sheffield, UK
g.demartini@sheffield.ac.uk

Abstract

Crowdsourcing has become an alternative approach to collect relevance judgments at scale thanks to the availability of crowdsourcing platforms and quality control techniques that allow to obtain reliable results. Previous work has used crowdsourcing to ask multiple crowd workers to judge the relevance of a document with respect to a query and studied how to best aggregate multiple judgments of the same topic-document pair.

This paper addresses an aspect that has been rather overlooked so far: we study how the time available to express a relevance judgment affects its quality. We also discuss the quality loss of making crowdsourced relevance judgments more efficient in terms of time taken to judge the relevance of a document.

We use standard test collections to run a battery of experiments on the crowdsourcing platform CrowdFlower, studying how much time crowd workers need to judge the relevance of a document and at what is the effect of reducing the available time to judge on the overall quality of the judgments. Our extensive experiments compare judgments obtained under different types of time constraints with judgments obtained when no time constraints were put on the task. We measure judgment quality by different metrics of agreement with editorial judgments. Experimental results show that it is possible to reduce the cost of crowdsourced evaluation collection creation by reducing the time available to perform the judgments with no loss in quality. Most importantly, we observed that the introduction of limits on the time available to perform the judgments improves the overall judgment quality. Top judgment quality is obtained with 25-30 seconds to judge a topic-document pair.

1 Introduction

Crowdsourcing is a recent approach that allows to reach large amounts of individuals on-line in order to complete short tasks that require human intelligence. In the context of Information Retrieval (IR) evaluation, crowdsourcing has been used to collect relevance judgments at scale. While crowdsourcing platforms enable the collection of large amounts of judgments from the crowd, the main challenge is quality assurance. Some crowd workers perform with lower quality for a number of different reasons including lack of clear instructions as well as malicious behaviors (Eickhoff and de Vries 2013). In order to deal with this aspect of crowdsourcing, quality control mechanisms as well as crowd answer aggregation techniques (i.e., collecting judgments for the same topic-document pair from different crowd workers and aggregating them together) have been proposed (Sheshadri and Lease 2013; Turpin et al. 2015; Venanzi et al. 2014; Hosseini et al. 2012).

As paid crowdsourcing (i.e., tasks are completed in exchange of a small monetary reward in platforms like Amazon MTurk) is commonly used to generate such relevance judgments, another challenge in using crowdsourcing for IR evaluation is the monetary cost attached to it. As each judgment is rewarded with a small monetary amount, scaling-up crowdsourced relevance judgments is difficult.

In this paper we look at how to make individual crowdsourced relevance judgments more time efficient and thus, more cost-effective. We specifically look at the time dimension involved in making relevance judgments by crowd workers and we study how much time workers need to make a judgment and what is the effect of reducing available judging time on the judgment quality. By identifying the optimal time needed to perform a high quality judgment, our proposed techniques allow to reduce the overall cost of generating an IR evaluation collection by means of crowdsourced relevance judgments, while maintaining an adequate quality.

We run extensive experiments by comparing different strategies to reduce the time available for a relevance judgment in crowdsourcing platforms. We experimentally show the trade-off of time/quality and which are the best choices, at the same cost, between asking for many quick judgments for the same topic-document pair as compared to fewer well-though judgments. We additionally compare different ways to enforce time constraints for judging tasks and their effect of judgment quality. Our main contributions are:

- A study on the time crowd workers take to make relevance judgments and what are the effects of training and worker properties as well as the impact of topics and documents on judgment time and quality.
- An analysis of how judgment quality degrades by reducing/increasing the time available to make a judgment in a

crowdsourcing setting.

- A comparison of alternative approaches to enforce time constrains on crowd judges (e.g., maximum time available vs. exact time to be spent on the task).
- The identification of the best way to spend a monetary budget in a crowdsourcing platform to collect relevance judgments looking at the trade-off between many quick judgments or fewer slow judgments.

This paper is structured as follows. In Section 2 we discuss related work in the areas of relevance judgment and crowdsourcing. In Section 3 we present the research questions addressed in this work. In Section 4 we look at how crowd workers behave when performing relevance judgments having no time constrains. In Section 5 we compare the effect of limits on the time crowd workers have to perform the task on judgment time and quality. In Section 6 we analyze the quality implication of forcing workers to spend an exact or a certain amount of time to complete the task. In Section 7 we present the results of the trade-off between collecting many quick judgments and few slow ones at the same crowdsourcing cost. In Section 8 we discuss the implication of our findings on IR evaluation collections based on crowdsourced relevance judgments. Finally, Section 9 concludes the paper summarizing our work.

2 Background

The concept of relevance and the process of relevance judgment have been extensively studied (Tang and Solomon 1998; Mizzaro 1997). Judging the relevance of a document is subjective and varies over time for the same assessor. It is however a key element of IR evaluation where the goal is to measure IR system effectiveness by measuring document relevance. Generating relevance judgments efficiently is a research question that has recently raised attention (Halvey, Villa, and Clough 2014). We first review relevant work looking on how much time relevance assessors need to make a judgment and what influences it. Next, we look at which crowdsourced relevance judgment quality techniques have been proposed so far and at how quick crowd judgments are a typical indication of low quality work.

2.1 Relevance Judgments and Time

Classic work (e.g., (Ahituv, Igbaria, and Sella 1998)) has shown that time pressure typically impairs performance of decision makers working with information. However, experience helps in dealing with time constraints. We claim that most crowd workers are used to optimize the time needed to complete tasks. Thus in this work we focus on cost/quality optimization for crowdsourced relevance judgments.

(Yilmaz et al. 2014) looked at how effort taken to judge a document correlates with the utility of a document to an end-user with an information need rather than with document relevance. (Verma, Yilmaz, and Craswell 2016) claim that the effort to judge relevance should be included in IR system evaluation metrics together with the relevance of retrieved documents. (Halvey and Villa 2014) look at the effort needed to judge the relevance of images, and measure the effect of image features and topic properties (e.g., difficulty) on effort and quality of judgments. (Villa and Halvey 2013) look at the difficulty of judging document relevance showing how borderline relevant documents require more effort to judge and that document size has an impact on the judging difficulty as well. (Wang 2011) observes that relevance assessor speed increases when the perceived difficultly of the task is low and that the judging accuracy increases when perceived difficulty increases. Building on top of this work which showed what affects judging effort, we propose novel ways to make crowd judgments more efficient without renouncing at judgment quality.

2.2 Relevance Judgments and Crowdsourcing

Crowdsourcing has become a popular means to collect relevance judgments and to create IR evaluation collections (e.g., (Smucker, Kazai, and Lease 2013; Kazai et al. 2011; Dean-Hall et al. 2014)). Previous work has shown how the use of crowdsourcing to collect relevance judgments is reliable and that IR evaluation based on it is repeatable (Blanco et al. 2011). Moreover, crowdsourcing can be used to extend existing test collections with additional judgments over time (Tonon, Demartini, and Cudré-Mauroux 2015).

When relevance judgments are collected by means of crowdsourcing the classic approach is to ask for the same topic-document pair to more than one worker in the crowd and then to aggregate their answers to obtain a relevance label. Doing this, there is the need to use an aggregation technique the most popular one being 'majority vote' where the label selected the most is used as final one. To ensure better label quality, a number of more advanced aggregation techniques have been proposed. For example, in (Hosseini et al. 2012) authors shows that using an Expectation Maximization model to aggregate crowd judgments yields to more stable evaluations as compared to majority vote. More recently, it has been shown that (Venanzi et al. 2014) better results can be obtained considering a measure of worker similarity based on the type of errors they perform. Thus, by identifying worker communities better answer aggregation techniques can be defined.

To allow for standard and repeatable comparisons of crowd answer aggregation techniques, ad-hoc benchmarks have been proposed. For example, SQUARE (Sheshadri and Lease 2013) serves as a comparative test collection for aggregating relevance judgments collected from the crowd. A different benchmark based on simulated crowd answers (Hung et al. 2013) allows to study how to best set parameters of crowd aggregation models. In our work we do not focus on crowd answer aggregation techniques but we rather take the assumption that high-quality individual labels will lead to high-quality aggregated results. Instead, we focus on improving the efficiency of crowdsourced relevance judgments by reducing the time available to look at the document and make a relevance decision about it. This reduces the cost of creating the collection assuming we pay crowd workers for their time. In our work we investigate the feasibility of this approach in terms of individual and aggregated judgment quality.

(Anderton et al. 2013) look at mistakes by crowd workers performing relevance judgments, comparing to trained

TREC assessors. They show that very short time spent to make the judgment leads to worse quality judgments and that document length has little impact on the time spent to make a judgment. Our first experiments (Section 4) shows similar results over a different test collection for which we crowdsourced relevance judgments with no time constraints.

In the context of micro-task crowdsourcing, measures of work effort in completing tasks have been proposed (Cheng, Teevan, and Bernstein 2015). In that case the goal is to understand how much time is needed to complete a certain task type and decide for a proper reward accordingly. As compared to them, in our paper we focus on the task of judging relevance and look at the efficiency/quality trade-off. Moreover, as measure of work quality we use standard assessor agreement measures that well fit our research questions.

3 Research Questions

In order to optimize the cost of collecting crowdsourced relevance assessments, we assume a very basic model, where the monetary cost c of an assessment for a topic is simply the product of the time t taken to judge each document, the number of judgments per document j, the number of documents n, and the reward r assigned for a judgment:

$$c = t * j * n * r. \quad (1)$$

By considering the reward r and the pool n constant, the total cost is then affected by the time and the number of judgments per document. We do not consider other parameters like the time to read the topic/query, the time to express the judgment, the time to switch to a new document and/or topic, etc. as we expect them not to be different over workers and documents. We measure the quality of a judgment by its agreement with editorial judgments (we will see some specific agreement measures below).

On the basis of this model, we can frame the following four research questions:

RQ1. How much time t do crowd workers take to judge the relevance of a document if no time constrain is set?

RQ2. What is the minimum amount of time t we can ask crowd workers to take in judging the relevance of a document? How does the judgment quality decrease when less time is available to make a judgment?

RQ3. Which type of timeout is the most appropriate to foster effective judgments? An *exact-time* timeout, where the document is shown for a certain amount of time and the judgment cannot be expressed before, or a *maximum-time* timeout, where the judgment can be expressed also before the expiration?

RQ4. With a fixed budget c, what is the best trade-off between time available for a judgment t and number of judgments collected per document j? Is it better to ask for more judgments done quickly (higher j, lower t) or less judgments done with more time available (higher j, lower t)?

To answer those, we ran a battery of four experiments, described in detail in the following four sections.

4 E1: We Have All the Time in the World

4.1 Aims

The time spent by a relevance assessor to perform a judgment varies quite a lot according to existing literature. For example, (Villa and Halvey 2013) report an average of 100 seconds for documents in the AQUAINT collection; (Yilmaz et al. 2014) report that most expert judges take up to 140 seconds while crowd workers up to 90 seconds. The aim of this first experiment, that addresses RQ1, is to measure the range of time spent by individual workers in our setting and use the results to identify appropriate thresholds for timed judging tasks.

4.2 Experimental design

We used Sormunen's work (Sormunen 2002), which reassessed some TREC-7 and TREC-8 documents on a four-level relevance scale (`H` for highly relevant, `R` for Relevant, `M` for Marginally relevant, and `N` for Not relevant). We used five TREC-8 topics (403, 418, 420, 427, 448), selected based on the availability of multi-graded relevance judgments, of at least 2 documents per relevance level, and to avoid topics which are not anymore timely at present time. For each topic, we randomly selected eight documents having different lengths and different relevance levels. We selected two documents for each level, one long and one short. We define document length based on word count and uniformly sample documents sorted by length.

In our first experiment E1, each worker (we recruited highest quality workers as provided by CrowdFlower) was shown the TREC topic (title, description, and narrative) and, after an initial test question that verified that the topic had been understood, had the task of judging the relevance of the eight documents, shown in a permuted order such that any document appeared exactly 5 times in each of the eight positions. A worker could use as much time as he/she wanted on each document before going to the next one, and he/she was allowed to perform other units, but only on different topics (i.e., workers could not re-judge the same topic, to avoid a learning bias). We collected 5 judgments for each document in each position, so for each topic 5*8=40 workers were needed in this first experiment. Finally, we ran the experiment twice, for both India and U.S. based workers independently, for a total of 3200 judgments.

4.3 Results

Judgment Time Figure 1 shows the task execution time distribution. As expected, in both runs, many workers took little time to complete the judgment with a tail of very long execution times. Figure 2 shows the distributions of time spent by individual workers over each topic and document position. We can observe that for some topics (e.g., 420) the first document to be judged takes more time because of learning effects. On average there is little delay for the first document to be judged as compared to the others. Around 97% of the times recorded are in the range between 2 and 100 seconds, around 80% are below 35 seconds and around 60% spent 5–35 seconds (see also Table 1).

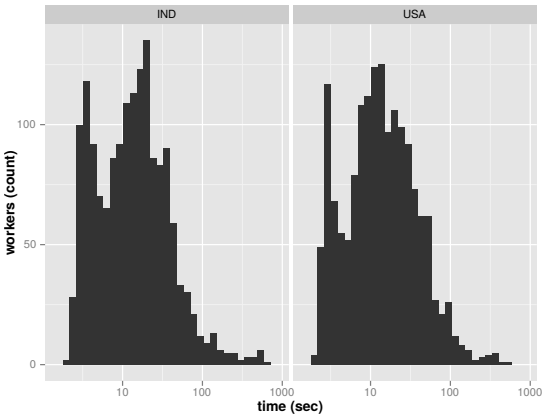


Figure 1: E1: Time required by workers to judge a document (log scale).

Table 1: E1: Distributions of time spent for US- and India-based workers.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
US	1.98	6.62	13.00	24.20	27.20	580.00
IN	1.90	5.46	13.30	25.40	25.70	634.00

We found no correlation between judging time and topic or document length (measured as number of chars or words, or with the ari readability index (Senter and Smith 1967)). This is consistent with the findings of (Anderton et al. 2013) who looked at these dimensions within the TREC 2012 Crowdsourcing track collection. We also did not observe correlation of time spent with relevance level, nor with agreement rates with Sormunen and TREC judgments.

Judgment Quality We now study the quality of relevance judgments obtained from the crowd with no time constrains imposed on the task. We measure quality as the level of agreement (measured with Cohen's Kappa) against Sormunen's 4-level judgments. We also measure agreement against binary TREC judgments, in three ways: by considering TREC relevant as H and nonrelevant as N, and by thresholding the 4 levels into 2 levels (both N-MRH and NM-RH). As additional measure of quality we also considered the average distance of the category selected by the worker to Sormunen's category (i.e., accuracy) which showed analogous results. The same metrics have been previously used to measure crowd judgment quality (Nowak and Rüger 2010; Kazai 2011). We compute such quality metrics both using the crowd judgments considered individually as well as with judgments aggregated together following the standard majority vote approach.

Figure 3 shows how quality changes over execution time for individual workers. The horizontal axis shows the deciles of time spent (deciles values are shown in the table below the plot). The vertical axis reports the quality, measured both with Sormunen distance and Cohen's Kappa. Note that for our dataset (documents with uniform distribution of relevance) a random assessor would obtain an average distance of 1.75 (dashed horizontal line). The two quality measures consistently show that the highest quality values are obtained in the central part of the curve, corresponding to around 5–50 seconds, and especially 5–25 seconds (values in bold in the table). This is consistent with Figure 2 and confirms that the time interval 5–30 seconds covers most worker activities. In the following experiments we will focus on such time interval to identify the execution time that leads to highest quality judgments.

Table 2: E1: Agreement, measured as Cohen's Kappa, between Workers and Sormunen (W-S) and Workers and TREC (W-T) over different document timeouts and positions, for both U.S. and India based workers, both individual (I) and aggregated (A). W-T Kappa is computed with three different weights: (i) default; (ii) `NM` into 0 and `RH` into 1; and (iii) `N` into 0 and `MRH` into 1. For comparison, S-T Kappa is 0.59, 0.55, and 0.77 with the three weights, respectively.

Position		p1	p2	p3	p4	p5	p6	p7	p8	AVG
U.S.										
W-S	I	.26	.34	.36	.46	.33	.49	.44	.43	.39
	A	.41	.36	.47	.65	.46	.63	.54	.47	.50
W-T	I	.20	.26	.27	.29	.24	.29	.32	.24	.26
	A	.30	.31	.35	.41	.34	.35	.42	.21	.34
W-T, `NM-RH`	I	.16	.22	.22	.21	.18	.25	.26	.22	.21
	A	.25	.27	.34	.34	.37	.34	.34	.15	.30
W-T, `N-MRH`	I	.28	.33	.41	.51	.34	.47	.48	.37	.40
	A	.52	.43	.52	.69	.38	.57	.78	.38	**.53**
IN										
W-S	I	.32	.39	.42	.22	.24	.30	.36	.31	.32
	A	.38	.57	.60	.44	.38	.34	.67	.38	.47
W-T	I	.21	.25	.29	.20	.14	.29	.31	.26	.24
	A	.26	.41	.36	.33	.25	.37	.56	.34	.36
W-T, `NM-RH`	I	.19	.27	.26	.17	.14	.28	.28	.26	.23
	A	.20	.51	.33	.30	.25	.43	.68	.27	.37
W-T, `N-MRH`	I	.30	.33	.39	.26	.24	.32	.44	.37	.33
	A	.36	.55	.59	.59	.43	.36	.71	.55	**.52**

We also look at judgment quality variations over the document order presented to workers. Table 2 shows that, as expected, Kappa values of judgments aggregated by majority vote are higher than those of individual judgments. We also note that the first two judgments are typically of lower quality, most likely because of training effects.

5 E2: Faster! Faster! Sorry, Too Late

5.1 Aims

Based on the analysis of the time taken by crowd workers to judge the relevance of a document (see Section 4), we now study the effect on judgment quality of reducing the time available to crowd workers to look at the document to be judged. To understand which is the minimum amount of time required to perform relevance judgments by crowd workers (RQ2) we designed the following experiment (E2).

5.2 Experimental Design

We display a document to crowd workers for a predefined amount of time and ask for a best effort relevance judgment. Given the results from E1 (Section 4), we set the following *timeouts* (i.e., time after which the document disappears and

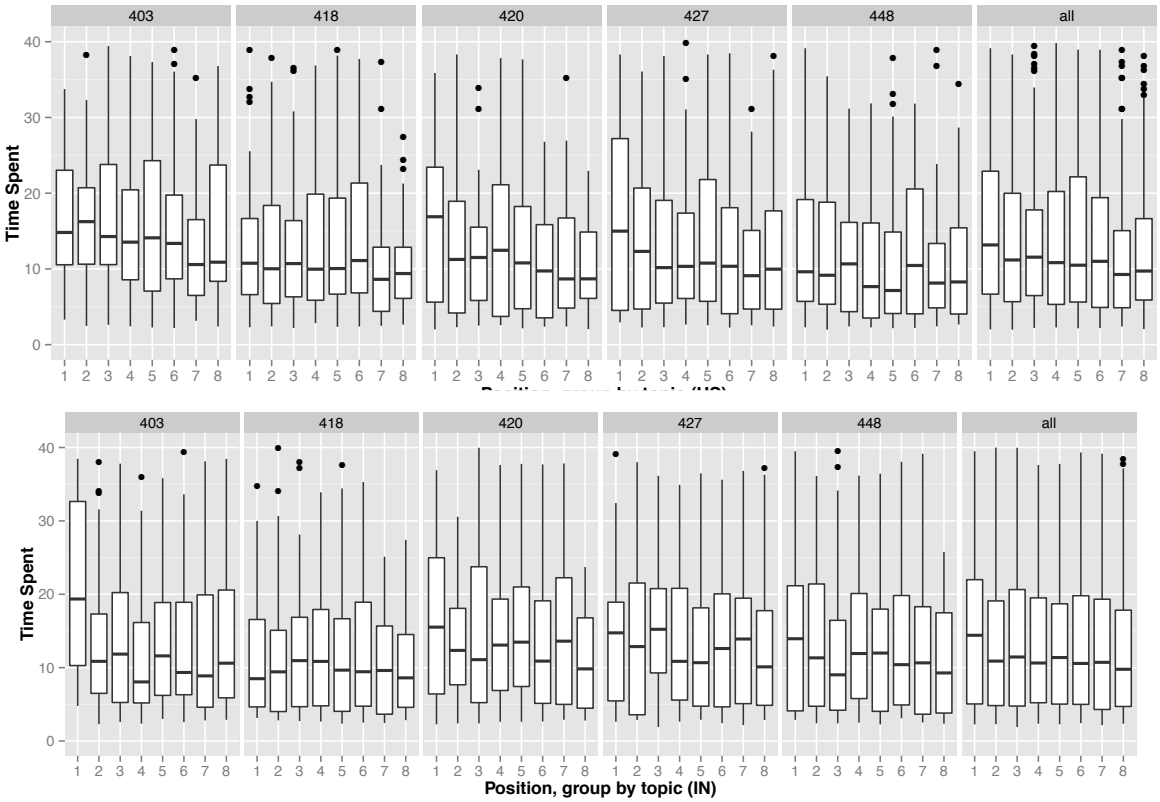
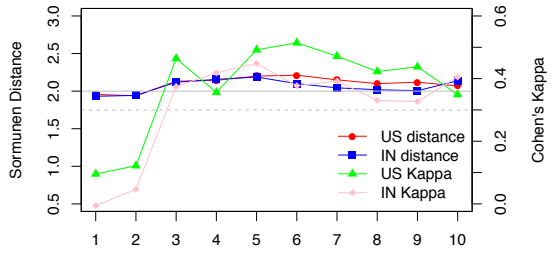


Figure 2: E1: Time spent by individual workers to judge document relevance broken down over topics and judging order for US based workers (above) and India based workers (below).



	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
U.S.	2.0	3.2	**5.1**	7.6	10	13	17	**23**	32	51	580
IN	1.9	3.4	**4.5**	7.0	9.9	13	17	**22**	31	46	630

Figure 3: E1: Variation of quality over time spent by individual workers (binned by deciles). Quality is measured both as distance from Sormunen label and Cohen's Kappa. The two quality measures have Kendall's correlations of 0.82 (U.S.) and 0.87 (IN).

a judgment has to be made): 3, 7, 15, 30 seconds as we observed best quality to be around the time interval $5-30$ seconds.[1] In this experiment, workers are allowed to complete the judgment before the timeout (i.e., we set the *maximum time* to judge) and can proceed to the next document. We use five TREC-8 topics (405, 408, 415, 416, 421). We selected 6 highly relevant, 6 relevant, 6 marginally relevant, and 6 not relevant documents per topic. We ask each worker to judge the relevance of 8 documents in total where the first two documents (a long and a short one) are displayed for 30 seconds, the following two documents for 15 seconds, other two for 7 seconds and, finally, two documents for 3 seconds. We set decreasing timeout values in order to let workers get prepared for shorter document visualization times and to learn how to grasp relevance signal in limited time. Any other timeout order would penalize short timeouts even further. Note also that even if some learning effect is present during the judgment of the first document (see Figure 2), having a first timeout at 30 seconds allows enough time to perform their judgment for at least 80% of workers (Figure 3). We also point out that the topics (and documents) used in E2 are different from E1, for two reasons: we needed topics having 6H, 6R, 6M, 6N documents for each topic, and we wanted to avoid possible biases from workers participating in both E1 and E2. Figure 4 summarizes the experimental design of E2.

[1]We can interpret judgments completed in less than 5 seconds as spam and tasks completed in more than 30 seconds as done by multi-tasker workers.

5.3 Results

Figure 5 shows how judgment time varies given a set timeout. We can observe that both in the case of 3 and 7 second timeouts workers judge relevance *after* the document display time is expired (red horizontal line). We call the time

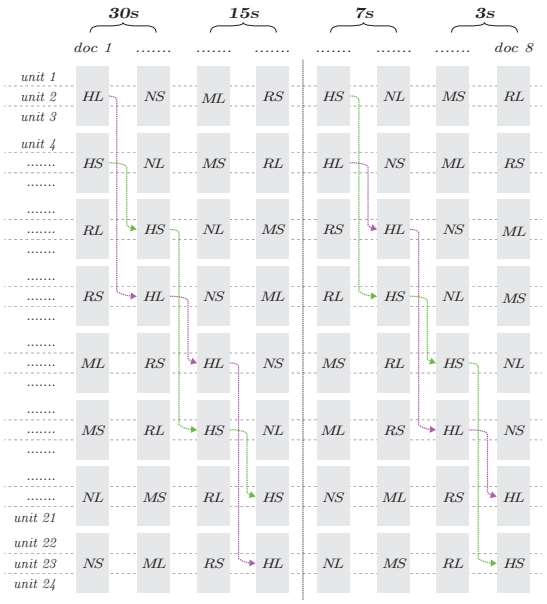


Figure 4: E2: Experimental design. Each worker is presented with 2 documents (one short and one long) over 4 different timeout values (30, 15, 7, and 3 seconds). Documents presented to the worker are different (i.e., 8 different documents in each row). Documents appearing in each of the 8 positions are different (i.e., 24 different documents in each column). The 8 presented documents are grouped in four pairs, each with a different relevance value (based on the NMRH scale of Sormunen). In each row, the choice of the documents that are presented to the worker is random inside each of 3 documents blocks (thus respecting the constraints).

between the timeout and the judgment *latency*. In the case of 15 and 30 second timeout, the judgment happens, in the median case, before the timeout.

Table 3 shows the Cohen's Kappa agreement values of aggregated judgments over 5 crowd workers with Sormunen judgments on a four level relevance scale and with TREC binary judgments. We can observe that in most cases 3 and 7 seconds are not enough to make a relevance judgment as agreement values are consistently lower than for other timeout values and the judgments are expressed well after the deadline (see the two rightmost panels in Figure 5). We can also see that US workers tend to have higher quality judgments than workers based in India.

When comparing agreement rates at 30 and 15 seconds we can see better results at 15 seconds. This can be explained by the learning bias in position 1-2 as confirmed by the E1 results (see Table 2) and of the following experiment presented in Section 7. It has also to be noted that the quality difference between 15 and 30 seconds is not statistically significant (t-test $p = 0.39$). Compared to TREC assessments, after transforming aggregated crowd judgments into binary ones by means of a threshold, we can see that agreement of crowd workers is lower than that of Sormunen assessments.

Table 3: E2: Cohen's Kappa values between Workers and Sormunen (W-S) and Workers and TREC (W-T) over different document timeouts and positions, for both U.S. and India based workers (see also Table 2).

Timeout(sec)	30		15		7		3	
Position	p1	p2	p3	p4	p5	p6	p7	p8
U.S.								
W-S	.43	.44	.52	**.53**	.51	.51	.35	.43
	.43		**.52**		.51		.39	
W-T	**.37**	**.37**	.34	.31	.34	.32	.21	.32
	.37		.32		.33		.27	
W-T, NM-RH	.28	.34	**.35**	.17	.30	.32	.22	.32
	.31		.26		**.31**		.27	
W-T, N-MRH	.37	.37	.30	.42	**.44**	.30	.23	.36
	.37		.36		**.37**		.29	
IN								
W-S	.39	.42	**.44**	.42	.35	.38	.36	.34
	.40		**.43**		.36		.35	
W-T	**.31**	.25	.28	.28	.26	.28	.29	.19
	.28		**.28**		.27		.24	
W-T, NM-RH	**.31**	.27	.29	.27	.24	.24	**.31**	.19
	.29		.28		.24		.25	
W-T, N-MRH	.28	.28	.25	.34	.30	**.41**	.30	.25
	.28		.30		**.36**		.28	

6 E3: Selecting the Best Timeout

When comparing judgment quality obtained with crowdsourcing (measured by Cohen's Kappa agreement with the original relevance assessments) we can observe that using some sort of timeout leads to better quality judgments as compared to judgments obtained with unlimited amount of time (Section 4). E2 results in Table 3 show that best judgment quality is obtained with 30s for TREC judgments (Kappa = .37) and with 15s for Sormunen judgments (Kappa = .52). Average Kappa values with TREC and Sormunen was 0.34 and 0.50 respectively in E1 (Table 2).

6.1 Aims

E2 results are not directly comparable with E1 results for two reasons: the documents used in E1 are different from those used in E2, and there could be a learning effect in E2 which could make the comparison biased by document positions and different time slots used. To allow for a direct comparison with E1 and to understand which timeout value leads to better worker performance, we run two modified versions of E2 with 15 and 30 seconds available to workers.

6.2 Experimental Design

The experimental design of E3 is very similar to that of E2 but with two important differences: the documents used are the same used in E1 and the time available to each worker for viewing the document is the same for all the 8 documents (fixed to 15 or 30 seconds). Like in E2, workers are free to judge a document before or after its disappearance thus not using all the time made available to them (*maximum-timeout*). We use the same quality checks as for previous experiments (i.e., topic understanding question and high-quality workers from the platform).

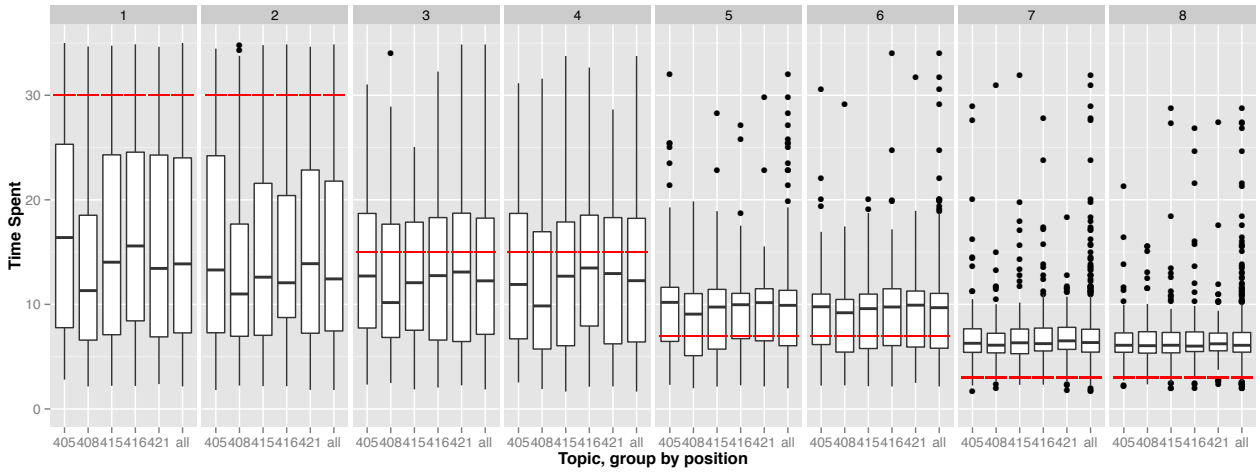


Figure 5: E2: Time spent by individual US based workers, breakdown over topics and document positions. The red lines show the timeouts.

Table 4: E3: Average of Cohen's Kappa values measuring the agreement between Workers and Sormunen (W-S) and Workers and TREC (W-T) over 8 documents for both U.S. and India based workers.

Country	U.S.		India		Average	
Timeout(sec)	30	15	30	15	30	15
W-S	0.47	0.49	0.48	**0.50**	0.48	0.50
W-T	0.33	**0.38**	0.32	0.35	0.33	0.37
W-T, NM-RH	0.31	0.37	0.31	**0.38**	0.31	0.38
W-T, N-MRH	0.47	**0.64**	0.49	0.50	0.48	0.57

6.3 Results

Table 4 shows the agreement between workers and the assessment of TREC and Sormunen obtained in E3. For both TREC and Sormunen comparison, the table shows that workers performances are always better for 15 than for 30 seconds and the difference is even statistically significant (t-test, $p < 0.01$) for TREC values.

Comparing E1 and E3 quality levels, we observe that, when looking at Sormunen judgments, in E1 the agreement between workers and gold standard is .47 for Indians and .5 for Americans while looking at TREC judgments, in E1 American and Indian workers agree with the gold standard respectively at .28 and .24 (see Table 2). In E3 (Table 4), the average values for 15 seconds (rightmost column) are always higher. We can thus conclude that the introduction of timeouts in crowdsourced relevance judgment tasks is also beneficial in terms of judgment quality. This does not yet completely answer RQ3; we will come back to this issue at the end of the next section.

7 E4: Many Fast&Furious, a Few Laid-Back?

7.1 Aims

Given the results so far, we now want to understand, given a fixed budget, what is the most effective way to collect crowdsourced relevance judgments with time-bound tasks. We study the trade-off between collecting many judgments with a very short timeout as compared to very few judgments with long time available to complete them (RQ4). With the assumption that the cost is computed, according to our model in Equation (1), by the actual workforce time spent on tasks (i.e., we pay workers for the time they spend on our tasks), we aim at finding the best timeout t and number of assignments values j for a fixed monetary budget c.

Table 5: E4: Different timeouts (rounded) and number of judgments, with the same monetary cost of a total of 150 seconds.

Timeslot(sec)	6	7.9	10	13.7	16.7	21.5	25	30	37.5	50
Assignments	25	19	15	11	9	7	6	5	4	3

7.2 Experimental Design

In order to compare the judgment quality over different trade-offs, we crowdsourced a number of different timeout/assignment combinations with the same total monetary budget for a number of topic-document pairs. Table 5 shows the ten time/judgments combinations (with a constant cost c according to Equation (1)).

To also address RQ3, as compared to timeouts used in E2 and E3 (i.e., *maximum-time*), in E4 we instead use the *exact-time* alternative. That is, we set a time after which the document disappears and workers have to make a relevance judgment, but we do not allow workers to judge and proceed before the given time, if they wish to do so.

7.3 Results

Figure 6(a) shows the judgment quality of different timeouts using 3 assignments in each case. We can observe that the quality increases as more time is available to the judge (thus, at a higher cost), but after 30s there is a sort of "plateau" with no noticeable increment of quality (whereas the cost c in-

Table 6: E4: Intraclass Correlation Coefficient of the workers judgments over the different topics, calculated for the assignments defined in Table 5.

Topic	403	418	420	427	448
Median ICC	.41	.39	.28	.21	.27

Table 7: E4: Intraclass Correlation Coefficient over different timeouts, calculated for the assignments defined in Table 5.

Timeslot(sec)	6	7.9	10	13.7	16.7	21.5	25	30	37.5	50
ICC	.28	.29	.31	.25	.30	.32	.35	.30	.35	.41

creases noticeably according to our model in Equation (1)); indeed, the quality seems to decrease after a "peak" at 30s. In Figure 6(b) the cost is kept constant (the number of judgments for each timeout is the one in Table 5), and we use all the available judgments at each timeout. Here the peak at around 25-30s is even more clear. Thus, since the cost c is constant for each timeout level, the highest quality is obtained with $t \in [25, 30]$s. Comparing workers based in US and India, we can see from Figure 6(b) that a budget gives significantly (t-test $p < 0.01$) better quality judgments when spent giving worker based in India 25s timeouts and to workers based in US 30s timeouts. Figure 6(c) shows the effect of the topic. The 25-30s range results in the highest quality across (most of the) topics; note that our collection includes both easy and difficult topics, with clearly lower quality judgments for topics 427 and 420.

This lower quality may be a sign of ambiguity of the documents contained in that topic, which seems confirmed by the calculation of the Intraclass Correlation Coefficient (ICC) of workers' judgments (Table 6). Topic 427 is, in fact, the topic which displays the lower agreement between workers themselves. Manually looking at the two most difficult topics we notice that they both contain the most technical concepts (427: 'UV damage, eyes' and 420: 'carbon monoxide poisoning') and have strict criteria of relevance specified in the narrative field. This indicates that time-constrained judgments may be more appropriate for general topics as compared to more technical/scientific ones which are probably more difficult for the average crowd worker to assess.

Table 7 shows ICC values over different timeout levels confirming that workers agree most when having at least 25s available for the judgment task.

As a last result, we go back to RQ3 as promised at the end of Section 6.3. By comparing the W-S Kappa values in Table 8 (or in Figure 6(b)) with the Kappa values in Tables 2 and 4 we observe that:

- We have already seen in Section 6.3 that maximum-time timeouts (used in E2 and E3) seem effective, since there is some increase in Kappa values from Tables 2 to Table 4.
- However, exact-time timeouts, used in E4, do cause a clearly higher increase: the Kappas obtained in E4 (Table 8) are clearly higher than those of E1 and E3, both at the 30s timeout, and at 13.7s and 16.7s (the closest values to 15s).

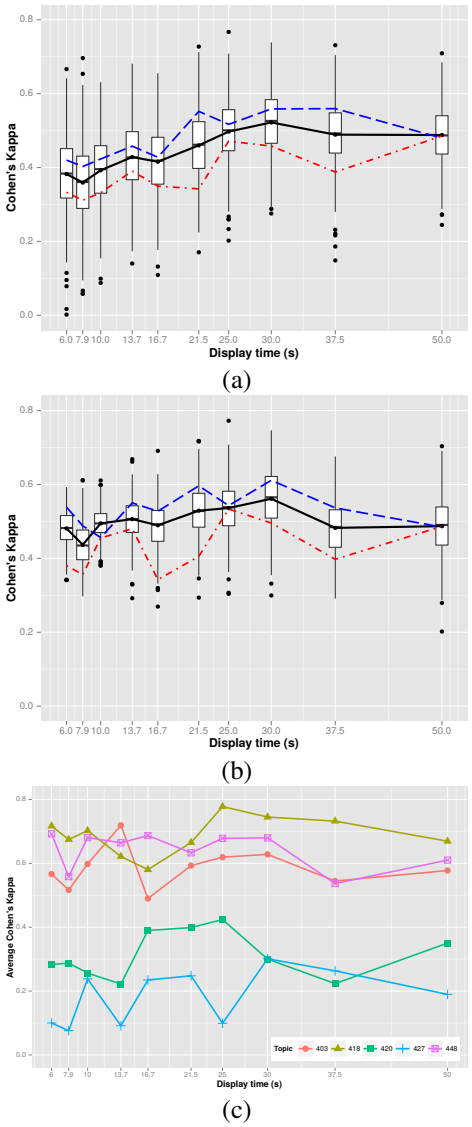


Figure 6: E4: Judgment quality (measured as Cohen's Kappa) over the 10 display time conditions with the same (three) number of workers (a), with the same cost (b), and breaking down on the five topics (c). In (a) and (b), the lines connect the average Kappa values: the solid line is for all workers, the dashed for the US workers only, and the dashed and dotted line for the IN workers only.

8 Discussion

In this section we summarize our main findings on how time constrains affect relevance judgments in a crowdsourcing setting across all the different experiments we performed.

8.1 Findings

We have observed in both E1 and E2 that there is a learning effect at the beginning of judgment tasks which implies that the first couple of judgments a worker completes will be of lower quality as compared to the following ones. From

Table 8: E4: Cohen's Kappa values over the ten timeslots, for U.S. and India based workers.

Timeslot(sec)	6	7.9	10	13.7	16.7	21.5	25	30	37.5	50
U.S										
W-S	.54	.49	.46	.55	.53	.60	.54	**.61**	.54	.48
W-T	.37	.30	.26	.41	.37	.42	.39	.35	**.44**	.39
W-T, NM-RH	.34	.31	.34	.37	**.43**	.40	.40	.34	.40	.34
W-T, N-MRH	.61	.47	.30	**.74**	.43	.69	.55	.50	.67	.55
IN										
W-S	.38	.36	.46	.48	.34	.40	**.53**	.50	.40	.49
W-T	.32	.18	.33	.32	.23	.29	.36	**.47**	.35	.34
W-T, NM-RH	.29	.21	.37	.37	.21	.31	.44	**.52**	.37	.34
W-T, N-MRH	.53	.10	.43	.43	.39	.44	.47	**.65**	.54	.46

E1 and E2 we have observed crowdsourced relevance judgments taking more than 30s tend to show lower quality. Indeed, best quality in E4 is obtained making workers judge relevance within a [25,30] seconds interval. We have noted that topic difficulty (measured by judgment quality as well as by ICC) does not impact the best timeout to be used (Figure 6(a) and Table 6 for E4). We measured judgment quality using both Cohen's Kappa as well as the distance between crowd and editorial judgments (both by TREC and Sormunen). Results obtained with both metrics concur. We also generally observed that workers based in US provided, in our experiments, better quality judgments as compared to those based in India (see Tables 3 and 8).

Answering RQ3 we observed that, given a certain timeout value, the best option is to have the worker to view the document for all the allocated time (i.e., not allowing him/her to proceed to the next document earlier) and to limit that time (E3 vs. E4).

Comparing the average Cohen's Kappa values between crowdsourced relevance judgments and judgments collected by Sormunen (Sormunen 2002) we can observe that the quality obtained in our first experiment E1 (Section 4) where workers could take any amount of time to complete the judgment task as it is traditionally done in crowdsourced relevance judgments, is 0.5 (Table 2) on average for US workers as compared to the average Kappa of 0.61 obtained with a timeout of 30 seconds for US workers in E4 (Table 8).

An interesting and recurrent aspect emerged in all the experiments is about Cohen's Kappa values between workers and TREC (W-T): very often we observed higher quality work judgments when thresholding them as N-MRH instead of NM-RH. This shows that workers are similar to TREC assessors in separating strictly not relevant documents to those they judged being somehow relevant.

8.2 Limits of This Study

Our work looks at optimizing relevance judgment time in crowdsourcing settings. While workers in the crowd are different (Kazai, Kamps, and Milic-Frayling 2011) and show different behaviors (Kazai, Kamps, and Milic-Frayling 2013), our approach does not take individual differences into account. We rather aim at finding a general strategy that works well over all workers. As future work, we will investigate the effectiveness of personalized approaches to time-bound relevance judgment by evaluating adaptive timeouts over different worker types and expertise levels.

In our work we used majority vote as a technique to aggregate crowd judgments. While more sophisticated techniques to aggregate crowd labels exist (e.g., (Hosseini et al. 2012; Venanzi et al. 2014; Sheshadri and Lease 2013)) in this work we do not focus on obtaining highest label quality but rather on observing the effect on quality degradation due to given time constrains in completing the task. For the same reason, we only include basic quality checks (e.g., topic understanding questions, use of high quality workers provided by the platform, etc.) when collecting data from the crowd. Thus our quality measurements indicate lower bounds and can easily be improved by combining other techniques to improve quality still reducing the cost of creating IR evaluation collections. Previous research has shown that even if assessor agreement levels are lower in crowdsourcing as compared to editorial judgments, IR evaluation results are still reliable and experiments are repeatable if we consider IR system ranking correlation levels (Blanco et al. 2011).

9 Conclusions

In this paper we have addressed the problem of limiting the time available for crowdsourced relevance judgment tasks. This is an important problem as controlling the judgment time allows for faster data collection (i.e., avoids the common *starvation* effect when batches of tasks do not finish as some workers take very long time to complete) as well as limits the cost of evaluation collection creation if workers are paid for the time they spent executing judgments.

We performed extensive experiments using standard test collections to evaluate crowdsourced judgment quality as compared to editorial judgments in a number of controlled experimental settings to understand the effect of limited time on quality. Results clearly show that limiting the time to perform a relevance judgment brings benefits both in terms of cost (and this was expected) as well as of quality (and this was unexpected). We observed that the best timeout value to be used lies in the interval of $25 - 30$ seconds and does not depend on topic, document, or crowd. Our findings are key for those researchers using crowdsourcing for the creation of large-scale IR evaluation collections as they can better control the creation cost still obtaining high quality annotations thanks to our proposed techniques.

Acknowledgments We would like to thank all the crowd contributors who participated to this study. This work was partially supported by the *UK EPSRC grant number EP/N011589/1*.

References

Ahituv, N.; Igbaria, M.; and Sella, A. 1998. The effects of time pressure and completeness of information on decision making. *J. Manage. Inf. Syst.* 15(2):153–172.

Anderton, J.; Bashir, M.; Pavlu, V.; and Aslam, J. A. 2013. An analysis of crowd workers mistakes for specific and complex relevance assessment task. In *CIKM '13*, 1873–1876.

Blanco, R.; Halpin, H.; Herzig, D. M.; Mika, P.; Pound, J.; Thompson, H. S.; and Tran Duc, T. 2011. Repeatable and reliable search system evaluation using crowdsourcing. In *SIGIR '11*, 923–932.

Cheng, J.; Teevan, J.; and Bernstein, M. S. 2015. Measuring crowdsourcing effort with error-time curves. In *CHI '15*, 1365–1374. ACM.

Dean-Hall, A.; Clarke, C. L. A.; Kamps, J.; Thomas, P.; and Voorhees, E. M. 2014. Overview of the TREC 2014 contextual suggestion track. In *Proceedings of TREC 2014*.

Eickhoff, C., and de Vries, A. P. 2013. Increasing cheat robustness of crowdsourcing tasks. *Inf. Retr.* 16(2):121–137.

Halvey, M., and Villa, R. 2014. Evaluating the effort involved in relevance assessments for images. In *SIGIR 2014*, 887–890.

Halvey, M.; Villa, R.; and Clough, P. D. 2014. SIGIR 2014: Workshop on Gathering Efficient Assessments of Relevance (GEAR). *SIGIR Forum* 49(1):16–19.

Hosseini, M.; Cox, I. J.; Milic-Frayling, N.; Kazai, G.; and Vinay, V. 2012. On aggregating labels from multiple crowd workers to infer relevance of documents. In *ECIR '12*, 182–194.

Hung, N. Q. V.; Tam, N. T.; Tran, L. N.; and Aberer, K. 2013. An evaluation of aggregation techniques in crowdsourcing. In *Web Information Systems Engineering–WISE 2013*. Springer. 1–15.

Kazai, G.; Kamps, J.; Koolen, M.; and Milic-Frayling, N. 2011. Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking. In *SIGIR '11*, 205–214.

Kazai, G.; Kamps, J.; and Milic-Frayling, N. 2011. Worker types and personality traits in crowdsourcing relevance labels. In *CIKM '11*, 1941–1944.

Kazai, G.; Kamps, J.; and Milic-Frayling, N. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Inf. Retr.* 16(2):138–178.

Kazai, G. 2011. In search of quality in crowdsourcing for search engine evaluation. In *ECIR'11*, 165–176.

Mizzaro, S. 1997. Relevance: The whole history. *JASIS* 48(9):810–832.

Nowak, S., and Rüger, S. 2010. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR '10, 557–566.

Senter, R., and Smith, E. 1967. Automated readability index. Technical report, DTIC Document.

Sheshadri, A., and Lease, M. 2013. SQUARE: A benchmark for research on computing crowd consensus. In *Proceedings of HCOMP 2013*.

Smucker, M. D.; Kazai, G.; and Lease, M. 2013. Overview of the TREC 2013 crowdsourcing track. In *Proceedings of TREC 2013*.

Sormunen, E. 2002. Liberal relevance criteria of TREC: Counting on negligible documents? In *SIGIR 2002*, 324–330.

Tang, R., and Solomon, P. 1998. Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior. *Information Processing & Management* 34(2-3):237 – 256.

Tonon, A.; Demartini, G.; and Cudré-Mauroux, P. 2015. Pooling-based continuous evaluation of information retrieval systems. *Inf. Retr. Journal* 18(5):445–472.

Turpin, A.; Scholer, F.; Mizzaro, S.; and Maddalena, E. 2015. The benefits of magnitude estimation relevance assessments for information retrieval evaluation. In *SIGIR 2015*, 565–574. ACM.

Venanzi, M.; Guiver, J.; Kazai, G.; Kohli, P.; and Shokouhi, M. 2014. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of WWW '14*, 155–164. ACM.

Verma, M.; Yilmaz, E.; and Craswell, N. 2016. On obtaining effort based judgements for information retrieval. In *Proceedings of WSDM '16*. New York, NY, USA: ACM.

Villa, R., and Halvey, M. 2013. Is relevance hard work?: Evaluating the effort of making relevant assessments. In *SIGIR 2013*, 765–768.

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*.

Yilmaz, E.; Verma, M.; Craswell, N.; Radlinski, F.; and Bailey, P. 2014. Relevance and effort: An analysis of document utility. In *CIKM 2014*, 91–100.