

Crowd vs. Expert: What Can Relevance Judgment Rationales Teach Us About Assessor Disagreement?

Mucahid Kutlu^{*}, Tyler McDonnell[♦], Yasmine Barkallah^{*}, Tamer Elsayed^{*}, and Matthew Lease[♦]

^{*}Qatar University [♦]University of Texas at Austin
{mucahidkutlu,yasmine.barkallah,telsayed}@qu.edu.qa
{tmcdonnell,ml}@utexas.edu

ABSTRACT

While crowdsourcing offers a low-cost, scalable way to collect relevance judgments, lack of transparency with remote crowd work has limited understanding about the quality of collected judgments. In prior work, we showed a variety of benefits from asking crowd workers to provide *rationales* for each relevance judgment [21]. In this work, we scale up our rationale-based judging design to assess its reliability on the 2014 TREC Web Track, collecting roughly 25K crowd judgments for 5K document-topic pairs. We also study having crowd judges perform topic-focused judging, rather than across topics, finding this improves quality. Overall, we show that crowd judgments can be used to reliably rank IR systems for evaluation.

We further explore the potential of rationales to shed new light on reasons for judging disagreement between experts and crowd workers. Our qualitative and quantitative analysis distinguishes subjective vs. objective forms of disagreement, as well as the relative importance of each disagreement cause, and we present a new taxonomy for organizing the different types of disagreement we observe. We show that many crowd disagreements seem valid and plausible, with disagreement in many cases due to judging errors by the original TREC assessors. We also share our WEB-CROWD25K dataset, including: (1) crowd judgments with rationales, and (2) taxonomy category labels for each judging disagreement analyzed.

KEYWORDS

Crowdsourcing, Relevance Assessment, Evaluation, Disagreement

ACM Reference Format:

Mucahid Kutlu^{*}, Tyler McDonnell[♦], Yasmine Barkallah^{*}, Tamer Elsayed^{*}, and Matthew Lease[♦]. 2018. Crowd vs. Expert: What Can Relevance Judgment Rationales Teach Us About Assessor Disagreement?. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, July 8–12, 2018, Ann Arbor, MI, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3209978.3210033>

1 INTRODUCTION

Crowdsourcing platforms such as Amazon’s Mechanical Turk provide a low-cost and scalable way of collecting relevance judgments [2, 14]. While crowdsourcing is most often motivated by improved

scalability, it offers other potential benefits as well. Instead of relying on a single expert judgment for each document, a set of crowd judgments can be collected and aggregated to guard against human error (even trusted judges are fallible), or again human bias, by reflecting average opinion in what is an inherently subjective judging task. It may even be easier to find a crowd judge with relevant expertise than personnel available in one’s local area [7, 22].

While crowdsourced judgments have been used in developing several test collections [5, 16], crowd task designs require special attention to ensure the quality of the collected judgments. Therefore, understanding reasons for crowd disagreements with trusted assessors is important for designing better crowd tasks.

Despite many studies reporting high disagreement in relevance judging between trusted assessors [26, 27], disagreements with crowd workers are sometimes attributed to workers being lazy, stupid, or deceitful. While much prior work has sought to improve the quality of collected crowd data, relatively less work has sought to better understand and characterize the types of judging disagreement the crowd tends to exhibit. Moreover, most work has assumed crowd disagreement constitutes error rather than trying to distinguish valid disagreement from actual error.

Understanding the reasons behind judging disagreement is difficult without having insights into the judges’ thought-processes. Consequently, prior work studying relevance judgments of primary vs. secondary assessors has sometimes relied on research methods involving interaction with participant judges, such as think-aloud [1] and interviewing [26]. However, it can be challenging to apply these methods on the current crowdsourcing platforms.

Our earlier work [20, 21] proposed a *Rationale Task* (RT) design for collecting crowdsourced relevance judgments. In particular, simply asking judges to provide short excerpts from each document to explain their judgment for it was shown to yield a multitude of benefits. In this work, we investigate how we can further exploit these rationales to gain new insights into reasons for judging disagreement, especially with remote crowd work. Largely following our original RT design, we collect roughly 25K crowd judgments for 5K document-topic pairs sampled from the 2014 TREC Web Track [11]. As a refinement, we show that having judges focus on judging within a topic, rather than across topics, improves label quality. Overall, we find that crowd judgments are good enough for ranking information retrieval (IR) systems reliably.

Next, we conduct a qualitative analysis using rationales to understand the disagreements and present a novel taxonomy of types of disagreement. In our analysis, we *manually inspect* 1K crowd judgments for 200 documents (5 judgments per document) in which the aggregated crowd judgment differs from the original TREC judgment, and we assess the relative importance of each disagreement

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5657-2/18/07...\$15.00

<https://doi.org/10.1145/3209978.3210033>

cause. Our analysis distinguishes between valid disagreement due to subjective considerations (e.g., relevance thresholds) from consistently recognizable human error (e.g., clearly missed evidence of relevance). Among the 200 documents we inspected, we agreed with TREC assessors in only 51.5% of the cases, disagreeing otherwise due to perceived human error (in 21.5%), subjective considerations (20%), and other miscellaneous reasons (7%). Similarly, we agree with 51% of the 1K *crowd* judgments while disagreeing due to perceived human error (37% of cases), subjective considerations (6%), and other miscellaneous reasons (6%).

Contributions of our work are as follows:

- We show that topic-focused crowd judging improves quality vs. our earlier design judging across topics [21].
- We show that crowd judgments we collect are largely valid and plausible, and that they enable reliable ranking of participant IR systems in the 2014 TREC Web Track.
- We demonstrate the value of assessor rationales for helping to explain disagreements in relevance judging.
- We present a novel taxonomy over types of disagreement, qualitatively and quantitatively distinguish objective vs. subjective disagreements, and estimate the extent of human error in trusted judgments.
- We share our **WEBCROWD25K** dataset¹, including: (1) crowd judgments with rationales, and (2) taxonomy category labels for each judging disagreement analyzed. In separate work [13], we also describe and share (3) crowd judging behavioral data.

The remainder of this paper is organized as follows. We first present related work in Section 2. We then describe our prior Rationale Task design [21] in Section 3. Section 4 explains our crowd task design and collection, and evaluates the quality of the judgments. In Section 5, we discuss our qualitative analysis and outline our novel taxonomy of disagreement types, presenting the results of the analysis in Section 6. Finally, we conclude in Section 7.

2 RELATED WORK

2.1 Reasons behind Disagreements

Al-Harbi and Smucker [1] categorized the reasons for disagreement into four groups: difficulty in applying the search topic, difficulty in processing the document, secondary assessor mistakes, and primary assessor mistakes. Sormunen [26] found that disagreement between primary and secondary assessors become more likely with ambiguous topic descriptions. Through examining documents from the TREC 2009 Legal Track, Grossman and Cormak [15] showed that disagreement is mainly caused by human error. Chandar et al. [9] found that longer, less coherent and easily comprehended documents provoke more disagreements. Our analysis includes most of these disagreement reasons and adds subjective considerations and technical judging issues, including four types of crowd errors.

2.2 Understanding Disagreements

Al-Harbi and Smucker [1] conducted a think-a-loud study to better understand secondary assessors' thought processes during relevance judging. Sormunen [26] interviewed secondary assessors to learn about reasons for disagreement. Wakeling et al. [28] recorded the behavioral data of assessors (i.e., mouse movements, screenshots and others) and then conducted interview in order to compare primary and secondary relevance judgments. While these methods are effective to understand the disagreement reasons, it is challenging to implement them with current crowdsourcing platforms.

Alonso and Mizzaro [2, 3] asked crowd workers to provide optional justifications for their judgments in free text format. In contrast, our RT design [21] requires rationales be provided in the form of document excerpts. While free text gives workers more flexibility, there is no easy quality check. On the other hand, requiring excerpts can be used to detect spammers, as shown in our analysis.

2.3 Assessing NIST Judgments

Scholer et al. [25] showed that NIST assessors judge similar documents differently, suggesting that their judgments can be inconsistent. Kazai et al. [18] found that there is a strong bias to Wikipedia pages (i.e., overrating them) in NIST judgments. Various studies also report that NIST assessors made mistakes in relevance judging [1-3, 8]. In our analysis, we also found that NIST assessors made mistakes but we also distinguish subjective vs. objective forms of our disagreement with NIST judgments.

2.4 Crowd Judgment Quality

Prior work has a mixture of findings about the quality of crowd judgments. Alonso and Mizzaro [2, 3] claim that crowd judgments can be a reliable alternative for relevance assessment. Blanco et al. [6] show that the ranking of IR systems do not change significantly when crowd judgments are replaced with NIST judgments. Kazai et al. [18] found that crowd workers perform as well as professional judges untrained in web search judging. On the other hand, there are also studies opposing using judgments of non-trained secondary assessors. Bailey et al. [4] show that judgments of primary or trained secondary assessors' judgments cannot be replaced by non-trained secondary assessors' judgments. Kinney et al. [19] report that judgments of non-expert assessors for domain-specific queries cause significant errors affecting system evaluation. Clough et al. [10] compare crowd judgments with expert judgments for domain-specific search tasks. They report that while crowd workers are able to rank systems correctly, they are less capable of differentiating levels of high accurate results than expert assessors. In our work, we show that crowd judgments do not yield significant changes in system ranking for general knowledge topics, establishing that they are reliable enough for practical use.

3 ORIGINAL RATIONALE TASK DESIGN

In this section, we describe the Rationale Task (RT) design from our prior work [21], which we adopt and modify in this work.

Instructions: We found in [21] that providing overly-specific instructions, examples, or corner-case notes ultimately left workers frustrated, both because of the length of instructions and because it made them feel more unsure about their final answer. As a result,

¹<http://qufaculty.qu.edu.qa/telsayed/datasets/webcrowd25k/>

we omitted all such specific instructions and provided a minimal task description, relying on a simple and intuitive judging scale to guide crowd workers (described below).

Judging Scale: Our RT design adopts a simple 4-point relevance scale: {Definitely Not Relevant, Probably Not Relevant, Probably Relevant, Definitely Relevant}. There are several proposed benefits of this scale: the relevance categories are evenly spaced across the spectrum of relevance, which makes conversion to binary judgments straightforward and offers flexibility to judges without over-complicating the scale; each of the relevance categories features colloquial language, rather than jargon common in the IR space; and the categories are symmetric with regard to adjectival descriptors (e.g., Probably Relevant vs. Probably Not Relevant).

Rationales: We asked workers in [21] to provide a rationale to support their labeling decision by selecting 2-3 sentences from documents. These excerpts enable new avenues of automated analysis and also provide expressive freedom for workers to argue in favor of the quality of their work and a new level of transparency which judgments alone do not provide.

Serving Tasks To Workers: In our earlier work [21], judging tasks were distributed to crowd workers uniformly at random from among the pool of all available judging tasks. As a result, workers might jump back and forth between different topics during judging.

4 MODIFIED DESIGN, DATA, & ANALYSIS

In this section, we present and evaluate our modified rationale task design at a much larger scale than in our earlier work [21]. We summarize the main differences in our experimental design vs. our prior study [21] as follows. Firstly, whereas our prior study used 700 NIST-judged documents from the 2009 Web Track, here we use 5K documents from the 2014 Web Track (Section 4.1). Secondly, we ask assessors to judge crawled webpages in the test collection, since this is what NIST does, rather than live versions of pages we used before. Thirdly, whereas our prior study judged a convenient, balanced, and largely random sample of documents, here we strategically sample documents to support ranking of participant IR systems. Fourthly, we employ the topic-focused judging order vs. judging across topics (Section 4.2). Fifthly, we decrease the price of relevance judging from \$0.10 to \$0.05 as a means of further exploring the stability of RT. Finally, in separate work [13], we describe, analyze, and share additional behavioral data we also collected during crowd judging.

In Section 4.3, we measure accuracy of crowd judges wrt. TREC assessors under varying aggregation methods. Section 4.4 reports the impact of crowd judging disagreement on the ranking of IR systems.

4.1 Test Collection & Document Sampling

We focus on the 2014 TREC WebTrack [11] (WT2014), which is the most recent TREC Web Track. WT2014 uses ClueWeb12² as document collection and contains 50 topics. The topics are a mixture of broad and specific queries, including navigational, informational and multi-faceted topics. For each topic, only a title and a one-sentence description were recorded. We focused on the ad-hoc search task, for which there are 14,432 relevance judgments. In comparison with our 4-point scale (Section 3), NIST’s six-point judging

scale slightly differs: 1) spam or junk; not useful for any reasonable purpose, 2) does not provide useful information, 3) provides some information, 4) provides substantial information, 5) dedicated to the topic, and 6) the home page of an entity named in the topic. To induce binary judgments, we map the first two categories to “non-relevant” and the remaining four to “relevant”. While this six-point scale may valuably support evaluation metrics using graded judgments, it may have side effects when collecting crowd judgments because of its complexity, as mentioned in Section 3. Additionally, we focus on binary disagreements (Section 5) and use *mean average precision* (MAP) in our evaluation (Section 4.4).

We sampled 100 documents to re-judge for each topic (i.e., $50 \times 100 = 5,000$ documents in total) using statAP [23] weighted sampling. According to the original TREC judgments, 45.4% of these documents are relevant (Table 3). Next, we collected 5 relevance judgments for each document using our NIST-Style Rationale Task on Mechanical Turk. In post-analysis, we found 9 documents having only 4 crowd judgments, so we removed them for consistency, leaving 24,955 judgments for 4,991 documents across 50 topics in our shared WEBCROWD25K dataset.

4.2 Topic-Focused Judging with Rationales

Our prior study [21] reported 92% agreement with NIST judgments in Web Track 2009 using the RT design. We adopt this design with a slight modification in the order of documents presented to the crowd workers, intended to better reflect the traditional judging approach employed by NIST. In the original RT design, a crowd worker might first judge a document for Topic X, followed by a document for Topic Y, followed by a Topic Z, before finally returning to Topic X. This varies significantly from the traditional TREC paradigm in which each topic is judged by a single (primary) assessor. To more closely mirror the TREC style of collecting relevance judgments, we propose a topic-focused, *NIST-Style Rationale Task Design* in which workers continue to judge documents from the same topic until it is exhausted, and only then move on to a fresh topic. This allows assessors to calibrate their internal topic definitions and relevance thresholds [24]. We will show that this yields higher judgment agreement than the original RT design.

To investigate the effect of topic-focused judging, we randomly selected 370 documents (out of the 4,991 sample above) across 27 topics and collected 5 relevance judgments per document (1850 judgments in total) using both the original RT design (random ordering) and the topic-focused judging described above. Overall, the topic-focused design produced 10.8% higher absolute accuracy (78.1% vs. 67.3% accuracy) over the original randomized ordering (aggregating crowd judgments by majority voting.). In analyzing the randomized ordering judgments, we identified several individual cases where workers were over-rating the relevance of documents, suggesting that they were unable to build an effective relevance threshold [24] while constantly oscillating between topics.

4.3 Accuracy of Crowd Judgments

We next discuss the quality of the relevance judgments we collected with respect to NIST judgments using the large sample of 4,991 documents. We used varying aggregation methods such as Majority Voting (MV) and Dawid-Skene (DS) [12]. We also considered a

²<http://www.lemurproject.org/clueweb12.php>

threshold filtering (TF) method we proposed earlier [21] which filters crowd judgments based on overlap between rationales and then applies a given aggregation method to remaining judgments.

Table 1: NIST-Style RT Agreement Results wrt. NIST.

Aggregation Method	Accuracy
Majority Voting (MV)	0.799
Dawid-Skene (DS)	0.798
Threshold Filtering & MV	0.779
Threshold Filtering & DS	0.749

Results in **Table 1** report simple accuracy of aggregated crowd judgments vs. TREC judgments. MV and DS appear effectively indistinguishable, while filtering performs slightly worse. Most notable, however, is that all of the accuracies are far lower than what reported on WT2009 in our prior study [21]: 0.92. One can imagine a variety of reasons for this; the start of Section 4 discusses a number of differences in experimental design and setup. We analyze the impact on ranking IR systems in Section 4.4, and further analyze disagreements in Section 5.

4.4 Effect on Ranking IR Systems

Next, we assess the impact of judging disagreement on ranking of IR systems participating in WT2014. Following typical TREC evaluation, we induce the ground-truth ranking of systems by using all (14,432) NIST judgments and evaluating systems by *mean average precision* (MAP). We refer to this ranking as *MAP-NIST*. In addition to this ground-truth ranking, we also rank the systems based only on the reduced set of 100 documents per topic sampled via statAP (Section 4.1), using either NIST judgments (StatAP-NIST) or crowd judgments (StatAP-Crowd). We calculate the correlation between these three rankings using Kendall’s τ and τ_{AP} [30], a variant of Kendall’s τ giving higher weight to swaps at higher ranks.

Table 2: Correlation of IR system rankings on WT2014.

Correlation Measures	MAP-NIST		STATAP-NIST	
	τ	τ_{AP}	τ	τ_{AP}
STATAP-NIST	0.905	0.876	1.0	1.0
STATAP-Crowd	0.937	0.921	0.947	0.939

Results are shown in **Table 2**. Remarkably, we see a higher ranking correlation score wrt. ground-truth ranking using crowd judgments than using NIST judgments (0.905 vs. 0.937 for τ and 0.876 vs. 0.921 for τ_{AP}). While this does not mean that crowd workers provide better judgments than NIST, it does indicate that disagreements between NIST and crowd judgments are not hurting IR system rankings according to either rank correlation metric.

In order to further explore this, we conduct an additional experiment. We simulate crowd error on the subset of statAP-sampled documents by randomly introducing errors on 20% of the documents (i.e., 1 - the 0.799 accuracy of MV-aggregated crowd judgments vs. NIST judgments). Next, we rank the systems using this judgment set and calculate Kendall’s τ and τ_{AP} wrt. using the real NIST judgments for the subset. We repeat this process 100 times

and calculate average τ and τ_{AP} across trials. Results of this simulation are lower ($\tau = 0.882$, with $\sigma = 0.032$, and $\tau_{AP} = 0.861$, with $\sigma = 0.039$) than when using the real crowd disagreements (Table 2). This suggests that crowd workers tend to disagree with NIST on documents that do not greatly impact the ranking (something Voorhees [27] reported earlier in analyzing disagreements among NIST judges).

Thus, despite the lower accuracy of crowd judgments seen in Section 4.3 vs. our prior study [21], we still see that using crowd judgments easily surpasses the traditionally established $\tau = 0.9$ threshold for reliable ranking of IR systems [27].

5 UNDERSTANDING DISAGREEMENT

In this section, we explain our qualitative analysis into reasons for disagreement using the rationales provided by the crowd judges. We first present the methodology of our analysis (Section 5.1), followed by the reasons for disagreement we identified (Section 5.2).

5.1 Methodology

To investigate the reasons for judging disagreements, we manually inspected a sample of topic-document pairs (See Section 6.1) in which the aggregated judgment of the crowd disagrees with the NIST judgment. Two authors of this paper judged the relevance of sampled documents independently and with no prior knowledge of other judgments by NIST or the crowd. Next, each of the two authors examined the respective NIST and crowd judgments and assigned one of four *stance* labels: 1) Strongly Agree with NIST; 2) Slightly Agree with NIST; 3) Slightly Disagree with NIST; or 4) Strongly Disagree with NIST. In “strong” cases, we perceive clear evidence in the document for our own judgment. With “slight” cases, we agree with one of the judgments but believe that the other is also reasonable, given a different interpretation of the topic or perception of relevance (e.g., relevance threshold). Finally, the two authors met in person to compare their labels and discuss their reasoning, ultimately arriving at a single, reconciled label for each case. Note that our stance labels (individual or reconciled) can be easily converted to graded relevance judgments.

Next, we sought to understand why there is disagreement by performing the following analysis for each document we judged:

- **Understanding disagreements with NIST:** For each case where we disagree with NIST, we carefully inspect the document and NIST judgment.
- **Understanding disagreements with the crowd:** We follow a similar process for each case in which we disagree with a crowd judgment. We consider each individual crowd judgments and *rationale*. Even when our judgment matches a crowd judgment, we still consult the rationale provided to verify it is reasonable.
- **Additional Categories:** Independent of our stance labels, we also annotated each document-topic pair to note if (1) the document is very long; (2) the document has low readability (e.g., poor web design or writing); or (3) expert knowledge appears necessary to judge the document-topic pair (e.g., knowledge of the American Revolution needed to judge a topic about the war). Note that these labels are not mutually exclusive; i.e., a document may have more than one.

In some cases, we identified more than one possible cause for disagreement: ambiguity in the topic description, and/or a crowd worker may have misunderstood the topic or have deemed the relevant content insufficient. In such cases, we determined the reason we believe to be most likely, yielding a *single best reason* for every case. We repeated this reasoning process over two full passes in order to further increase the consistency of our analysis. The entire process took around 30 hours.

Despite careful inspection, our method has certain limitations:

- (1) Our understanding of each topic is limited to what the original NIST assessor recorded, via topic definition and judgments. Some more nuanced understanding of an intended topic may have eluded us as secondary assessors.
- (2) While we carefully analyzed a small set of disagreements, we are certainly fallible and susceptible to human error.
- (3) To further understand the thought process of crowd workers, we rely on (our interpretation of) their provided rationales, which are limited to text excerpts from the documents. Rationales appear most useful to show why a document is relevant and less helpful for showing why a document is not relevant. In the latter case, we do our best to guess the reason for disagreement using all evidence we have at our disposal.

5.2 Reasons for Disagreement

In this section, we discuss in detail the reasons we observed for judging disagreements. We further induce a novel taxonomy over those reasons, presented in **Figure 1**.

5.2.1 Human Error. We sometimes observed seemingly unambiguous evidence for document relevance. For instance, in judging the document *clueweb12-0310wb-50-29927* for topic 273 (“Find Wilson’s Disease Association website”), the NIST assessor judged it as relevant, but it is neither the website nor has the URL for it. Therefore, in cases where we strongly disagree (with either NIST or crowd), we designate the reason for the disagreement as *human error*. Due to lack of any insight into NIST judgments beyond topic descriptions and judgments, we are unable to further understand this NIST judgment. However, inspecting crowd rationales led us to categorize 4 types of crowd errors:

(1) *Topic Misunderstanding.* In this category of mistakes, we believe the crowd worker made a good faith effort to judge the document but misunderstood the topic. Ideally, the rationale for a document judged relevant should indicate a part of the document making it relevant to a particular topic. Judging a document as relevant but providing a rationale that is not closely matched to the topic suggests that the crowd worker misunderstood the topic. For instance, in judging document *clueweb12-0204wb-61-01007* for topic 273 (“Find Wilson’s Disease Association website”), one crowd worker judged it “Definitely Relevant” with the following rationale:

Wilson’s disease is an inherited condition which causes copper to build up in the body. This excess copper tends to collect in the brain and liver but can also be found in the corneas (in the eyes) and the kidneys. If... not treated properly, it can cause very serious symptoms.

The chosen excerpt from the document explains Wilson’s disease, which indicates an over-interpretation of the topic definitions, such

that the crowd worker might have thought that this information about Wilson’s disease would be useful for a person who searches for WDA’s website. This is also similar to Wakeling et al. [28]’s finding: Secondary assessors can sometimes judge documents as relevant because of thinking that the information on that page can be also useful for a person who makes that search.

(2) *Missing Relevant Content.* Another reason for disagreement could be the lack of concentration or other unknown human error that causes missing a relevant content in a document. We used this label when crowd workers judged a document as “definitely not relevant” or “probably not relevant” with a rationale not relevant to the topic while there is a clear evidence in the document to be relevant.

(3) *Spammers.* Rationales are also useful in detecting workers who clearly do not follow task instructions, acting as “spammers”. We observed 4 types of behavior in this category:

- Providing rationales from other documents
- Providing off-topic rationales for documents judged relevant
- Reporting a page load error³ yet providing a rationale
- Judging a document with only textual content as relevant yet using our pre-defined rationale text for non-textual pages⁴

While we typically assume that crowd workers’ labels are correct when they match our own, when the label space is small (e.g., binary), we must also account for *accidental agreement*. We found rationales to be useful in identifying such cases. For example, a judge providing a rationale from the wrong document indicates that the judgment is likely spam, even if the label seems correct.

(4) *Relevant Rationale for Not-Relevant Judgment.* During our analysis, we also observed that some crowd workers judged a document as “Definitely Not Relevant” but provided a rationale which is a definitely relevant statement to the topic, i.e., a *conflicting* rationale. For instance, in judging the document *clueweb12-1611wb-41-22823* for topic 270 (i.e., “Find quotes from Sun Tzu”), a worker provided the following rationale:

sun tzu said, “the good fighters of old first put themselves beyond the possibility of defeat, and then waited for an opportunity of defeating the enemy.”

Though this rationale is a perfect example to judge the document as relevant, the worker judged the document as “Definitely Not Relevant”. Our earlier study [21] noted such behavior as well, which seems to be most likely due to clicking too quickly or misunderstanding the topic or the judging scale.

5.2.2 Ambiguous Topic Definition. While a primary assessor formulates a clear definition of the *information need* in their head, secondary assessors (including ourselves) are reliant upon the primary assessor’s written topic description to understand the information need. Overly terse, incomplete topic descriptions may introduce ambiguity in the information need and relevance judging.

In such cases, while we disagree with the NIST or crowd judgments, we think that his/her judgment is also reasonable based on

³In addition to our 4-point judging scale, workers were also provided a fifth option to indicate that a given web page did not load and so could not be judged.

⁴For cases in which a relevance judgment depends on non-textual content, we asked workers to enter a particular pre-defined text string.

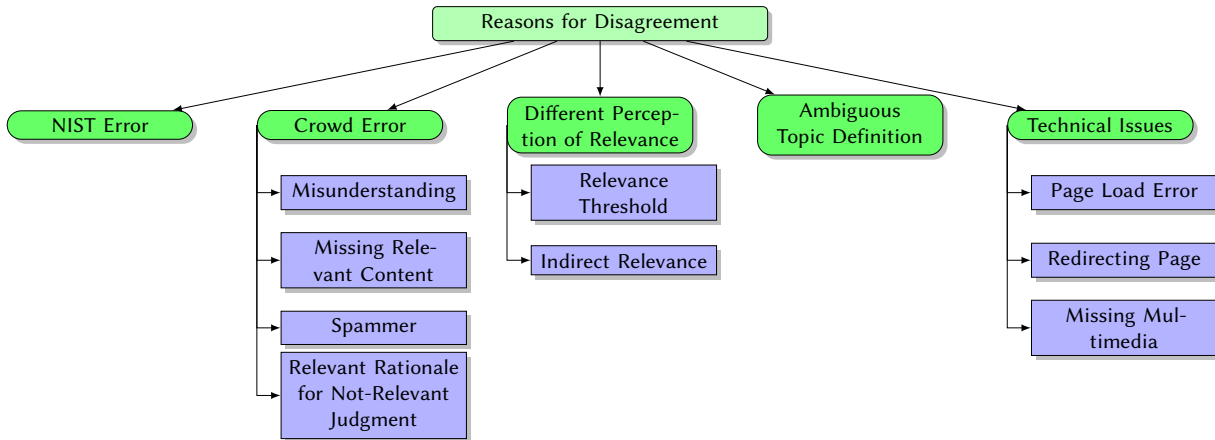


Figure 1: Reasons for Disagreement in Relevance Judging.

the topic description (given another interpretation of the topic). For example, in judging the documents for topic 261 (i.e., “What folk remedies are there for soothing a sore throat”), it seems that the NIST assessor did not consider home-remedies nor herbal-remedies as folk remedies and thus did not judge such documents as relevant.

5.2.3 Different Perception of Relevance. Assessors may disagree also due to the inherently subjective nature of relevance judging. Many such factors impact relevance judgments, such as novelty of the document [29], reliability of the source [31] and others. We may have disagreed with crowd workers or NIST assessors due to such perception. We identified two types of cases:

(1) *Relevance Threshold.* The amount of relevant content in a document has a large impact on our relevance judgment, as expected. We see this impact more obvious by checking the rationales. For instance, in judging the document clueweb12-1601wb-52-10051 for topic 290 (i.e., “How do you identify a Norway Spruce?”), one of the crowd workers provided the following rationale:

The Norway Spruce (Picea abies) is a large evergreen coniferous tree, growing to 35-55 m tall and with a trunk diameter of up to 1-1.5 m. The shoots are orange-brown and hairless.

While showing that the document provides some relevant information for the topic, the crowd worker still judged it as “Probably Not Relevant”, indicating that although s/he found some relevant content, s/he did not find that the relevant material was sufficient to satisfy the information need. Interestingly, three crowd workers also provided rationales covering these sentences but judged differently (i.e., two as “Definitely Relevant,” and one as “Probably Relevant”), further suggesting different thresholds for relevance.

We also found cases in which we disagreed with NIST likely due to different relevance thresholds. For example, for topic 278 (i.e., “What are the lyrics to the theme song for ‘Mister Rogers’ Neighborhood”?), the NIST assessor judged document clueweb12-1006wb-74-08027 as “Not Relevant”, likely due to it only containing the first four lines of the lyrics, rather than the full lyrics. We believe this should be sufficient to judge the document as relevant.

(2) *Indirect Relevance.* According to track judging instructions [11], a page should be judged as “Not Relevant” if it does not contain a relevant content but only provides a link or mentions a resource name (e.g., a book or a course given by a university) directly related to the topic of interest. However, we found that NIST assessors and crowd workers consider the links or resource names as relevant information in many cases we inspected.

For example, the NIST assessor judged document clueweb12-0700wb-28-32790 as relevant to topic 267 (i.e., “What are the lyrics to the song Feliz Navidad”) though the page does not contain the lyrics, only a link named “View Feliz Navidad Lyrics”. Also in judging the document clueweb12-1509wb-34-18722 for topic 276 (i.e., “How has African American music influenced history, including cultural history”), we observed that the NIST assessor judged the document as relevant even though it only lists courses offered by a university. While some courses seem relevant to the topic, there is no relevant information for the topic in the page. Similarly, we also observed that some crowd workers give the link as their rationale, suggesting that they consider links in their judging process.

5.2.4 Technical Issues. Correctly rendering crawled web pages is challenging due to the complex structure of web pages (e.g., containing multimedia files from the host server and also from other web addresses). Therefore, relevance judging of crawled web pages imposes many technical challenges which may have an impact on the relevance judgments. We identified the following issues:

(1) *Page Loading Error.* Crowd workers explicitly indicated some pages did not load to be judged.

(2) *Missing Multimedia.* Rendering web pages using crawled pages is challenging because many pages use varying multimedia files from other web addresses (e.g., YouTube) which may not be available at the time of collecting relevance judgments. Even if the multimedia files are captured during the crawling, it can be challenging to modify the web page source to get the images to display correctly. Missing multimedia becomes a big problem especially with topics such as topic 258 (i.e., “Find pictures of a hip roof”).

(3) *Redirecting*. In our analysis, we noticed that some pages are redirecting to other pages, causing judging of different documents than we believe were viewed by NIST assessors. This problem was potentially more severe in our earlier study using live pages [21].

6 RESULTS & DISCUSSION

In this section, we present and discuss the results of our analysis. We first explain how we sampled the documents to be manually inspected (Section 6.1) and then show our own judgments for the sampled documents (Section 6.2). Finally, we present the distribution of disagreement reasons (Section 6.3).

6.1 Sampling Documents to be Inspected

Table 3 shows the distribution of documents at varying *agreement levels* (AL) (i.e., the percentage of crowd workers agreeing with the NIST assessor in judging a particular document when graded judgments are collapsed to binary judgments). To select documents to be analyzed, we opt for stratified sampling. In our sampling method, we consider 3 different agreement levels of crowd workers that cause disagreement with NIST judgments (i.e., 0%, 20%, 40%) and two possible NIST-judged binary relevance judgments (i.e., relevant and not relevant) separately, yielding 6 (=3x2) different combinations. We sample 25 documents from each case. In order to have a better representation of the disagreements, we also randomly sample 50 more documents from the remaining disagreement cases, resulting in 200 documents in total. The distribution of the 6 cases in our sample is given in **Table 4**.

Table 3: Distribution of Documents at Varying Agreement Levels between NIST and Crowd Workers. The shaded rows represent disagreement between crowd and NIST based on majority voting.

Agreement Level	NIST-Judged Not Relevant	NIST-Judged Relevant	Total
0%	1.4%	0.6%	2%
20%	4.7%	1.6%	6.3%
40%	8%	3.7%	11.7%
60%	11.7%	7.8%	19.6%
80%	14%	10.9%	24.9
100%	14.8%	20.6%	35.5%
Total	54.6%	45.4%	100%

Table 4: Document Distribution of the Manually Inspected Sample.

Agreement Level	NIST-Judged Not Relevant	NIST-Judged Relevant	Total
0%	28	25	53
20%	39	29	68
40%	48	31	79
Total	115	85	200

6.2 Our Own Relevance Judgments

In this section, we discuss our own relevance judgments for the sampled documents. We discuss the results in two different ways: 1)

results within the stratified sample of documents (Section 6.1); and 2) results projected to each of the six cases NIST and aggregated crowd judgment differ, based on the frequency of each case. See the first three rows of Table 3 to do this projection.

The summary of our judgments is given in **Table 5**. In 23% and 25.5% of the cases, we slightly and strongly *disagree* with NIST, respectively. This means that we disagree with NIST assessors in 48.5% (=23% + 25.5%) of the cases we inspected, projected to 47% of all disagreement cases. This suggests that the quality of the crowd judgments is much higher than we earlier calculated (Section 4.3).

Considering only NIST-judged relevant documents, as expected, we agree with NIST in many more cases than we disagree: 51 (=7+10+12+4+6+12) vs. 34 (=7+5+3+7+8+4). However, we see the opposite pattern over the NIST-judged non-relevant documents: 52 (=10+9+21+1+6+5) vs. 63 (=8+11+12+9+13+10). This suggests that our judgments are even more liberal than NIST assessors [26].

We also observe that as fewer crowd workers agree with NIST, our own agreement with NIST similarly decreases, as expected. Specifically, among documents with crowd AL of 40% (i.e., both NIST-judged relevant and non-relevant documents), we disagree with NIST assessors in 36.7% (= $\frac{12+10+3+4}{48+31}$) of the cases. Our disagreement with NIST increases to 54.4% (= $\frac{11+13+5+8}{39+29}$) and 58.5% (= $\frac{8+9+7+7}{28+25}$) when the crowd AL is 20% and 0%, respectively.

6.3 Distribution of Disagreement Reasons

In this section, we first discuss the results of our qualitative analysis in the all sampled documents (Section 6.3.1). Next, we focus on only hard-to-process documents (Section 6.3.2) and the problematic topics where there is a high disagreement between NIST and crowd workers (Section 6.3.3).

6.3.1 Distribution across the whole sample. In our analysis, we assessed 200 NIST judgments and 1000 (= 200x5) crowd judgments. The distribution of each category of agreement/disagreement is shown in **Figure 2**.

There are several observations we can make from these results. Firstly, our agreement ratios with NIST and crowd workers are similar (51.5% vs. 51.2%). However, among the agreements with crowd judgments, 27.1% (= 1% + 13.6% + 12.5%) of them appear to be *accidental agreement* (Section 5.2).

Secondly, the human error ratio of NIST assessors is lower than human error ratio of crowd workers, as expected (21.5% vs. 36.7%). Among crowd errors, missing relevant content and misunderstanding are the main problems, contributing to about 80% of all crowd errors. Judgments from spammers are just 16.3% of the crowd errors, that is, 6% (= 36.7% x 16.3%) of all cases we inspected. We also noticed that among spam judgments, 78% of them use rationales from other documents. Therefore, most of such spam can be easily detected by checking whether each rationale actually exists in the document being judged. We automatically identified that 2,878 crowd judgments in the entire collection (i.e., 11.5% of all crowd judgments) use rationales from other documents.

Misunderstanding and conflicting judgments might be resolved by providing better topic descriptions and task instructions, suggesting that 42.8% (= 39.5% + 3.3%) of crowd errors (i.e., 15.7% (= 42.8% x 36.7%) of the sample and 18.6% (projected) of the all disagreements) could be eliminated by the “perfect” task design. However, missing

Table 5: Our Relevance Judgments. The ratio of each case with respect to the total sample size is given in parentheses.

NIST Binary Judgment Crowd Agreement Level	Not Relevant			Relevant			Total
	0%	20%	40%	0%	20%	40%	
Strongly agree with NIST	10 (5%)	9 (4.5%)	21 (10.5%)	7 (3.5%)	10 (5%)	12 (6%)	69 (34.5%)
Slightly agree with NIST	1 (0.5%)	6 (3%)	5 (2.5%)	4 (2%)	6 (3%)	12 (6%)	34 (17%)
Slightly disagree with NIST	8 (4%)	11 (5.5%)	12 (6%)	7 (3.5%)	5 (2.5%)	3 (1.5%)	46 (23%)
Strongly disagree with NIST	9 (4.5%)	13 (6.5%)	10 (5%)	7 (2.5%)	8 (4%)	4 (2%)	51 (25.5%)
Total	28 (14%)	39 (19.5%)	48 (24%)	25 (12.5%)	29 (14.5%)	31 (15.5%)	200 (100%)

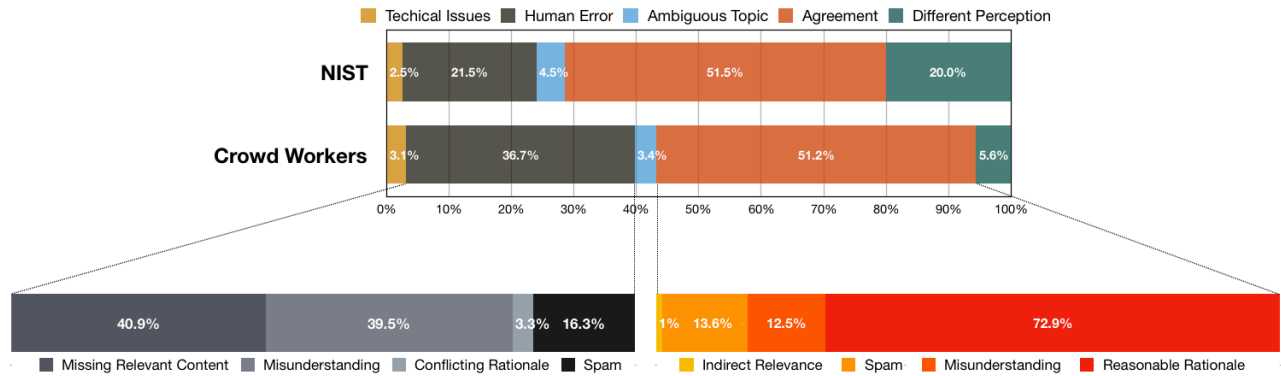


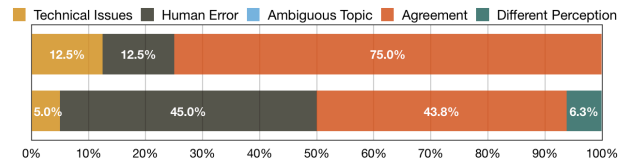
Figure 2: Distribution of Agreement & Disagreement Reasons.

relevant content, which contributes to 15% of the sample and 12% (projected) of the all disagreements, may be a harder problem to fix. On the other hand, simply paying more could incentivize higher quality work given a more complex task [17].

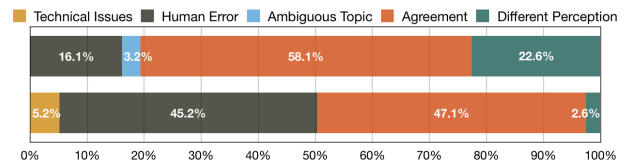
Thirdly, the ratio of different perceptions of relevance for NIST judgments is much higher than its ratio for crowd workers (20% vs. 5.6%). Relevance threshold constitutes to 65% and 80% of those cases for NIST and crowd workers, respectively.

Interestingly, we also found that even expert NIST assessors did not always follow track judging instructions [11] wrt. indirect relevance, sometimes judging a document as relevant even when it only pointed to a resource (a web page or a course/book name) that is potentially relevant (7% of all cases we inspected).

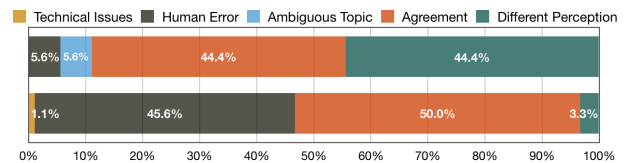
6.3.2 Distribution across Hard-to-Process Documents. As noted earlier, during our inspection, we labeled the documents that are hard to process using three different labels. **Figure 3** shows the distribution of reasons for the documents with these labels. In all three cases, the human error ratio for crowd workers increases compared to its ratio among all sampled documents (36.7% vs. 45%, 45.2%, and 45.6%). On the other hand, human error ratio for NIST is lower than its ratio on all sampled documents (21.5% vs. 12.5%, 16.1%, and 5.6%) suggesting that NIST assessors are not affected by the difficulties in processing the documents. Among crowd errors on document-topic pairs requiring expert knowledge, we found that 55% of the disagreement reasons is misunderstanding. On the other hand, we found that 69% and 83% of the crowd errors are due to missing relevant content for long documents and documents with low readability, respectively.



(a) Document-Topic Pairs needing Expert Knowledge (16). Top Bar: NIST, Lower Bar: Crowd.

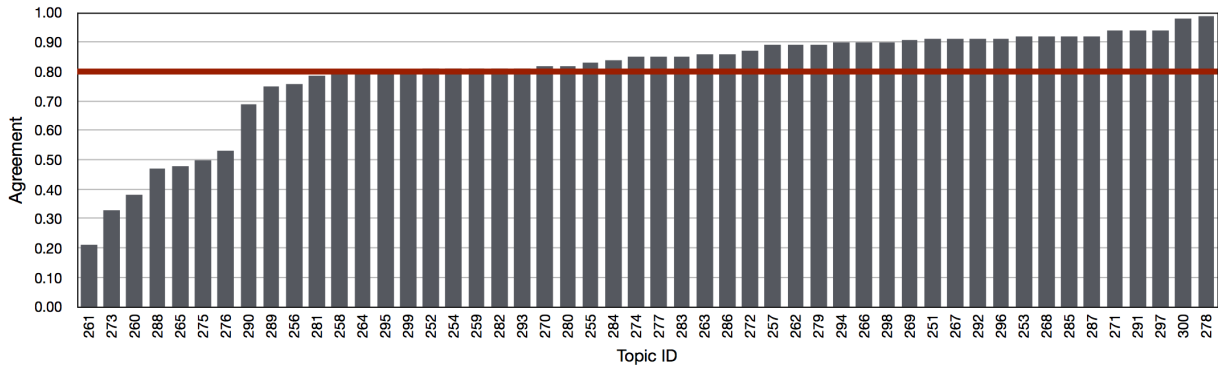


(b) Long Documents (21 cases). Top Bar: NIST, Lower Bar: Crowd.



(c) Documents having Low Readability (31). Top Bar: NIST, Lower Bar: Crowd.

Figure 3: Distribution of Agreement & Disagreement Reasons for Hard-To-Process Documents. The number of documents for each case is given in parentheses.

Figure 4: Agreement between NIST and Crowd Workers. The horizontal line represents the average agreement among all topics.**Table 6:** Topics with the Lowest Agreement between NIST and Crowd Workers.

Topic ID	Topic Description	Sample Size	Our Judgment	
			Agree w/ NIST	Disagree w/ NIST
261	What folk remedies are there for soothing a sore throat	9	1	8
273	Find Wilson (Wilson’s) Disease Association web site	16	12	4
260	Find a list of the major battles of the American Revolution	8	7	1
288	Find quotes from Fidel Castro	13	9	4
265	What were the ten worst tornadoes in the USA	7	5	2

6.3.3 Topic Specific Disagreement Reasons. In our analysis, we noticed very low agreement between NIST and crowd workers for a few particular topics. **Figure 4** shows the agreement between NIST and aggregated crowd judgment for each topic. We investigate 5 topics where the agreement is lower than 50% to better understand such topic-specific problems. Descriptions for these topics and our stance labels (collapsed to binary) are shown in **Table 6**.

Topic 261. We realized that the NIST assessor did not think any remedies named as home, herbal or even grandma’s remedy (in the document with id clueweb12-1911wb-01-10721) as a folk remedy. We disagree with NIST (i.e., agree with the aggregated crowd judgment) in 8 cases out of 9 in our inspected sample, where 7 of them are due to topic ambiguity. We also determined that, for this topic, 19 crowd judgments, out of 45 ($= 9 \times 5$), in our inspected sample and 184 judgments (among $100 \times 5 = 500$) in the whole collection are actually spam.

Topic 273. We mostly agree with NIST assessors (in 12 cases out of 16). In 46 crowd judgments, out of 80 ($= 16 \times 5$), we noticed that the crowd workers misunderstood the topic and usually provided a text that describes Wilson’s disease.

Topic 260. We determined that 14 judgments in our inspected sample and 194 judgments in the whole collection for this topic are actually spam. Misunderstanding is the second most important disagreement reason we found (25% of the judgments we inspected). We observed that the crowd workers were very liberal in judging documents as relevant if the document is somehow related to American Revolution even though it does not mention any battle name. This appears consistent with past work [19] finding that laymen often fall back on simple query term matching in assessing relevance for topics which exceed their level of topical expertise.

Topic 288. In our inspected sample of this topic, 40% of our disagreements with crowd workers appear to be due to workers missing relevant content. We labeled 5 out of 13 documents as either "too long" or "low readability" and found that 60% of the disagreement reasons for these 5 documents are missing relevant content, suggesting these factors may have negatively impacted judging. We also found that 25% of the judgments suffered from technical issues (20% page load error and 5% redirecting).

Topic 265. The main problem appears to be the high ratio of spam (51% of the crowd judgments we inspected and 46% of all crowd judgments over the entire collection for this topic).

We automatically detected spammers who provide rationales that do not exist in the documents to be judged and noticed that automatically-detected spam ratios are particularly high for a few specific topics, and that the agreement ratios for these topics were generally lower than others. A likely explanation is that our topic-focused judging design (Section 4.2) had the unintended side-effect of also concentrating spam within a topic instead of distributing it evenly across topics.

7 CONCLUSION & FUTURE WORK

While crowdsourcing offers a low-cost, scalable way to collect relevance judgments [2, 14], lack of transparency with remote workers has limited understanding about the quality of collected judgments. In prior work [20, 21], we investigated the value of asking crowd workers to provide *rationales* explaining each relevance judgment. In this work, we scaled up this rationale-based judging design to assess its reliability in practice to support a real TREC track evaluation: the 2014 WebTrack [11]. We investigated having crowd

judges focus on judging within a topic, rather than across topics and showed this improved the quality of collected judgments. Overall, we showed that we were able to reliably rank IR systems using crowd judgments.

To investigate the potential of rationales to provide new insight into judging disagreements between expert and crowd assessors, we then analyzed 200 disagreements between TREC and crowd judges. We found that rationales for judging relevant do provide useful insights into crowd workers' thought process and can be used to better understand disagreement reasons. However, negative rationales (for judging a document non-relevant) were usually not helpful for disagreement analysis. In total, we disagreed with NIST assessors in 48.5% of the cases we inspected, finding that many crowd disagreements appear valid and plausible. We presented a novel taxonomy over reasons for disagreement, and we share our WEBCROWD25K dataset, including: (1) crowd judgments with rationales, and (2) taxonomy category labels for each judging disagreement analyzed. In a separate work [13], we also describe, analyze, and share (3) behavioral data collected during crowd judging.

Overall, we believe that forming a rationale is critical to forming a coherent relevance judgment, whether or not judging instructions explicitly require it. Our earlier results [21] showed that requiring annotators to provide rationales incurs almost no additional time, suggesting that annotators might be already doing so implicitly. While we have investigated collecting judgment rationales for crowd work, as we have earlier argued [20, 21], we believe that asking (traditional) expert judges to also provide rationales could provide a myriad of benefits, enriching both the quality and value of collected relevance judgments.

ACKNOWLEDGMENTS

We thank the many talented crowd contributors and NIST relevance assessors who provided the data for our study, and the reviewers for their valuable feedback. This work was made possible by NPRP grant# NPRP 7-1313-1-245 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

REFERENCES

- [1] Aiman L Al-Harbi and Mark D Smucker. 2014. A qualitative exploration of secondary assessor relevance judging behavior. In *Proceedings of the 5th Information Interaction in Context Symposium*. ACM, 195–204.
- [2] Omar Alonso and Stefano Mizzaro. 2009. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, Vol. 15. 16.
- [3] Omar Alonso and Stefano Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *Information Processing & Management* 48, 6 (2012), 1053–1066.
- [4] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P de Vries, and Emine Yilmaz. 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 667–674.
- [5] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A test collection with query variability. In *Proceedings of the 39th ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 725–728.
- [6] Roi Blanco, Harry Halpin, Daniel M Herzig, Peter Mika, Jeffrey Pound, Henry S Thompson, and Thanh Tran Duc. 2011. Repeatable and reliable search system evaluation using crowdsourcing. In *Proceedings of the 34th ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 923–932.
- [7] Muhammed Fatih Bulut, Yavuz Selim Yilmaz, and Murat Demirbas. 2011. Crowdsourcing location-based queries. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on*. IEEE, 513–518.
- [8] Ben Carterette and Ian Soboroff. 2010. The effect of assessor error on IR system evaluation. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 539–546.
- [9] Praveen Chandar, William Webber, and Ben Carterette. 2013. Document features predicting assessor disagreement. In *Proceedings of the 36th ACM SIGIR conference on Research and development in information retrieval*. ACM, 745–748.
- [10] Paul Clough, Mark Sanderson, Jiayu Tang, Tim Gollins, and Amy Warner. 2013. Examining the limits of crowdsourcing for relevance assessment. *IEEE Internet Computing* 17, 4 (2013), 32–38.
- [11] Kevyn Collins-Thompson, Craig Macdonald, Paul Bennett, Fernando Diaz, and Ellen M Voorhees. 2015. TREC 2014 Web Track Overview. In *Proceedings of the Twenty-Third NIST Text REtrieval Conference (TREC)*.
- [12] Alexander Dawid and Allan Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics* (1979), 20–28.
- [13] Tanya Goyal, Tyler McDonnell, Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. 2018. Your Behavior Signals Your Reliability: Modeling Crowd Behavioral Traces to Ensure Quality Relevance Annotations. In *6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*. 10 pages.
- [14] Catherine Grady and Matthew Lease. 2010. Crowdsourcing Document Relevance Assessment with Mechanical Turk. In *Proc. of the NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 172–179.
- [15] Maura R Grossman and Gordon V Cormak. 2012. Inconsistent responsiveness determination in document review: Difference of opinion or human error. *Pace L. Rev.* 32 (2012), 267.
- [16] Maram Hasanain, Reem Suwaileh, Tamer Elsayed, Mucahid Kutlu, and Hind Almerkhi. 2017. EveTAR: building a large-scale multi-task test collection over Arabic tweets. *Information Retrieval Journal* (21 Dec 2017).
- [17] Chien-Ju Ho, Aleksandr Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 419–429.
- [18] Gabriella Kazai, Nick Craswell, Emine Yilmaz, and Seyed MM Tahaghoghi. 2012. An analysis of systematic judging errors in information retrieval. In *Proceedings of the 21st ACM conference on Information and knowledge management*. 105–114.
- [19] Kenneth A. Kinney, Scott B. Huffman, and Juting Zhai. 2008. How Evaluator Domain Expertise Affects Search Result Relevance Judgments. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*. ACM, New York, NY, USA, 591–598. <https://doi.org/10.1145/1458082.1458160>
- [20] Tyler McDonnell, Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. 2017. The Many Benefits of Annotator Rationales for Relevance Judgments. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*. AAAI, 4909–4913.
- [21] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why is That Relevant? Collecting Annotator Rationales for Relevance Judgments. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*. AAAI, 139–148. *Best Paper Award*.
- [22] Jean-Francois Paiement, James G Shanahan, and Remi Zajac. 2010. Crowdsourcing local search relevance. In *Proceedings of CrowdConf*.
- [23] V Pavlu and J Aslam. 2007. *A practical sampling strategy for efficient retrieval evaluation*. Technical Report. Technical report, Northeastern University.
- [24] Falk Scholer, Diane Kelly, Wan-Ching Wu, Hanseul S Lee, and William Webber. 2013. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th ACM SIGIR conference on Research and development in information retrieval*. 623–632.
- [25] Falk Scholer, Andrew Turpin, and Mark Sanderson. 2011. Quantifying test collection quality based on the consistency of relevance judgements. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 1063–1072.
- [26] Eero Sormunen. 2002. Liberal relevance criteria of TREC-: Counting on negligible documents?. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 324–330.
- [27] Ellen M Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & mgmt.* 36, 5 (2000), 697–716.
- [28] Simon Wakeling, Martin Halvey, Robert Villa, and Laura Hasler. 2016. A comparison of primary and secondary relevance judgements for real-life topics. In *Proc. of the ACM on Conf. on Human Information Interaction and Retrieval*. 173–182.
- [29] Yunjie Xu and Zhiwei Chen. 2006. Relevance Judgment: What Do Information Users Consider Beyond Topicality? *Journal of the American Society for Information Science and Technology (JASIS&T)* 57, 7 (May 2006), 961–973.
- [30] Emine Yilmaz, Javed A Aslam, and Stephen Robertson. 2008. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual ACM SIGIR conference on Research & development in information retrieval*. 587–594.
- [31] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdzka. 2014. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proc. of the 37th ACM SIGIR conference on R&D in information retrieval*. 435–444.