Mix and Match: Collaborative Expert-Crowd Judging for Building Test Collections Accurately and Affordably

Mucahid Kutlu
Qatar University
mucahidkutlu@qu.edu.qa

Tyler McDonnell
University of Texas at Austin
tmcdonnell@utexas.edu

Aashish Sheshadri
PayPal
aashish.sheshadri@gmail.com

Tamer Elsayed
Qatar University
telsayed@qu.edu.qa

Matthew Lease
University of Texas at Austin
ml@utexas.edu

ABSTRACT

Crowdsourcing offers an affordable and scalable means to collect relevance judgments for information retrieval test collections. However, crowd assessors may show higher variance in judgment quality than trusted assessors. In this paper, we investigate how to effectively utilize both groups of assessors in partnership. We study how agreement in judging is correlated with three factors: relevance category, document rankings, and topical variance. Based on this, we then propose two collaborative judging methods in which some document-topic pairs are assigned to in-house assessors for relevance judging while the rest are assessed by crowd workers. Results on two TREC collections show encouraging results when we distribute work intelligently between our two groups of assessors.

KEYWORDS

Information Retrieval, Relevance, Evaluation, Crowdsourcing

1 INTRODUCTION

Crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) provide a new avenue for scalable, low-cost collection of relevance judgments for constructing Information Retrieval (IR) test collections [3]. However, quality of crowd judgments is often more variable than with traditional use of in-house, known assessors. Consequently, researchers have explored a variety strategies for quality assurance with crowdsourcing, including effective *front-end* design [2], such as requesting rationales supporting labels [15] or relevance judgments [26], and *back-end* machine learning, such as statistical aggregation [19, 31, 36] and predicting label correctness [20]. Prior work has also considered how varying assessor expertise may impact judging for various IR judging tasks, such as systematic review [22, 27], e-discovery [16], and domain-specific search [12].

Building upon our prior work [27], we hypothesize that crowd judges may be better suited to judge some documents than others for relevance. If we could effectively distinguish such documents, then we could effectively route only those appropriate documents to the crowd for judging, while restricting our more limited and expensive trusted judges to remaining documents. Since our goal is to utilize trusted judges only for the documents that we believe the crowd is likely to label incorrectly, we investigate three broad factors which may correlate with agreement in judging: relevance category, document rankings, and topical variance. This builds on a long and storied history of research study on disagreement in relevance judging [1, 4, 7, 9, 11, 17, 21, 23, 24, 29, 33, 34].

Following this, we evaluate two collaborative judging approaches. The first *oracle* approach uses knowledge of disagreement for each topic to prioritize high disagreement topics for trusted judging. The second, practical approach focused on document importance rather than expected disagreement, using expensive trusted judges to judge those highly ranked documents which most greatly impact rank-based evaluation metrics. In particular, we use statAP method [28]'s weighting function to prioritize highly-ranked documents for trusted judging. We compare both approaches to a random document ordering baseline.

We report experiments using two NIST TREC[1] test collections for which both crowd and trusted TREC judgments are available. We show that collaborative judging offers a promising method to leverage the crowd in combination with trusted judges for accurate and affordable building of IR test collections.

2 DISAGREEMENT IN JUDGMENTS

To better understand on which topic-document pairs we might expect to see judging disagreement, we investigate three broad factors which may correlate with such disagreement: relevance category, document rankings, and topical variance.

2.1 Test Collections

We use two test collections to investigate judging disagreement between crowd vs. NIST assessors and to conduct rank correlation experiments using collaborative judging.

Million Query Track 2009 (MQ'09) [8]. The ClueWeb09 collection[2] and 100 MQ'09 topics were reused in the TREC Relevance Feedback (RF'09) Track [5]. Because RF'09 participating systems retrieved additional documents not judged for MQ'09, additional relevance judgments were collected for the track via MTurk. These judgments were also used for the subsequent TREC Crowdsourcing Track[3] and made freely available. Beyond judging new documents, 3,277 documents already judged by NIST were also re-judged as part of quality assurance during data collection. This yields 20,535 crowd judgments with which we can measure agreement with NIST. We also evaluate 35 system runs submitted to MQ'09 using these crowd judgments vs. NIST judgments, measuring rank correlation.

DESIRES 2018, August 2018, Bertinoro, Italy

[1] trec.nist.gov/ [2] lemurproject.org/clueweb09/ [3] sites.google.com/site/treccrowd/

Table 1: Confusion Matrices for Crowd (Cr) vs. NIST Judgments in MQ'09 and WT'14 Test Collections. 'R' represents relevant judgments and 'NR' represents not-relevant judgments. Bold indicates agreement.

	MQ'09					*WT'14*				
	Majority Voting **65%**		***Dawid-Skene*** **70%**			***Majority Voting*** **80%**		***Dawid-Skene*** **81%**		
	Cr-R	Cr-NR	Cr-R	Cr-NR	**Total**	Cr-R	Cr-NR	Cr-R	Cr-NR	**Total**
NIST-R	**44%**	10%	**41%**	13%	54%	**39%**	6%	**37%**	8%	45%
NIST-NR	25%	**21%**	17%	**29%**	46%	14%	**41%**	11%	**44%**	55%
Total	69%	31%	58%	42%	100%	53%	47%	48%	52%	100%

Web Track 2014 (WT'14) [13]. Recently, a new crowd-judgment collection has been released [18, 23]. 100 NIST-judged documents for each of 50 WT'14 topics were selected by statAP [28]'s sampling method. MTurk judgments for these documents were collected via [25, 26]'s rationale method. In total, 25,099 MTurk judgments for 5000 documents were collected (i.e., roughly 5 judgments per document). We evaluate 29 system runs submitted to WT'14 using these crowd judgments vs. NIST judgments, measuring rank correlation.

For both test collections, we reduce graded relevance judgments to being binary and report two different methods for aggregating them: majority voting (MV) and Dawid-Skene (DS) [14]. Whereas MV performs unweighted voting, DS performs weighted voting based on unsupervised individual reliability estimates. Agreement statistics in **Table 1** show the WT'14 crowd is 10-15% more accurate than the MQ'09 crowd, *wrt.* NIST judgments as the "gold standard" (80% vs. 65% in MV and 81% vs. 70% in DS). We also see that DS performs much better on MQ'09 (where it evidences more variability in crowd assessor reliability) than on WT'14, whose crowd demonstrates less variability.

2.2 Agreement vs. Relevance Category

Is there more disagreement on documents judged by NIST to be relevant or non-relevant? We might expect higher agreement for clear-cut cases of relevance/non-relevance, and higher disagreement for boundary cases [24, 34]. **Table 1** shows confusion matrices for both test collections and aggregation methods. For all settings, we observe that crowd judgments show higher agreement with NIST assessors on NIST-judged relevant documents than on non-relevant ones. Assuming DS aggregation, we see that crowd judgments agree with NIST on $\frac{41\%}{54\%} = 76\%$ (MQ'09) and $\frac{37\%}{45\%} = 82\%$ (WT'14) of NIST-judged relevant documents. On non-relevant documents, crowd judgments agree with NIST to a lesser degree: $\frac{29\%}{46\%} = 63\%$ (MQ'09) and $\frac{44\%}{55\%} = 80\%$ (WT'14). Higher agreement on NIST-judged relevant documents suggests that when in doubt, inexpert judges may be more liberal in judging documents as relevant [33].

2.3 Agreement vs. Document Rank

We next consider how disagreement is correlated with document rankings. Following the same logic discussed above regarding document relevance, we might expect highly relevant documents to retrieved at very high ranks, totally unrelated documents to be retrieved at low ranks, and borderline documents (leading to judging disagreement) retrieved at more middling ranks [24, 34]. To measure this, we compute the average rank of each judged document across all submitted runs for each track. We treat all documents retrieved at rank ≥ 1000 as having rank 1000, then bin documents into 10 groups based on their average ranks, using a fixed interval size of 100 ranks. Finally, we compute agreement statistics for each bin for NIST vs. crowd-workers.

Figure 1 shows the accuracy of the 10 groups over the two collections. In MQ'09, the accuracy for the first group (i.e., the average document rank ≤ 100) is higher than the accuracy when the average document rank is between 200-500 in both aggregation methods. Regarding the results for WT'14, we observe a more clear pattern: the accuracy is noticeably higher in the first group. The second group's accuracy is the lowest among other groups, and accuracy increases gradually as the average document rank increases.

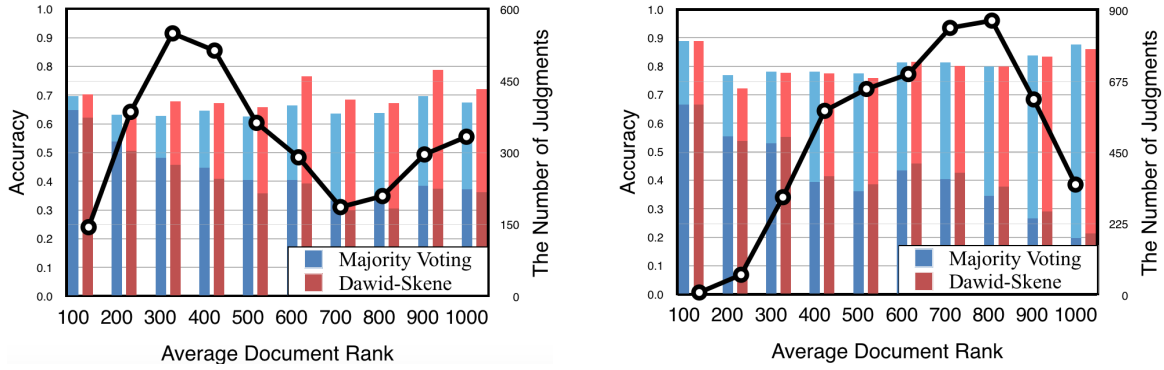
2.4 Agreement Across Topics

Because an individual with a given information need knows best what they are and are not looking for [11], NIST typically utilizes the same individual to both develop a topic (description) and perform judging for that topic. While written topic descriptions are useful, they are never complete, and so secondary assessors (be they NIST [34] or crowd [3]) have less information to go on when judging relevance of someone else's topic. This naturally leads to disagreement. While even NIST assessors are known to often disagree [34], crowd judging introduces further variability. For example, retired intelligence analysts working as NIST assessors may share a common geographic, cultural, and knowledge background, suggesting a consistent bias. Crowd workers may be far more diverse.

Figure 2 shows the distribution of judging agreement across topics for MQ'09 and WT'14. With MV aggregation, the standard deviation across topics is 0.13 (MQ'09) and 0.17 (WT'14). However, with DS aggregation, the standard deviation slightly tightens to 0.11 (MQ'09) and 0.15 (WT'14). Interestingly, the standard deviation is actually higher in WT'14, despite its crowd judgments having higher accuracy (See [23] for further discussion about disagreement variance across topics in WT'14). Regardless, we clearly do observe large variability in judging agreement across different topics in both collections, suggesting the importance of modeling topical factors in order to accurately predict assessor disagreement [10, 30, 35].

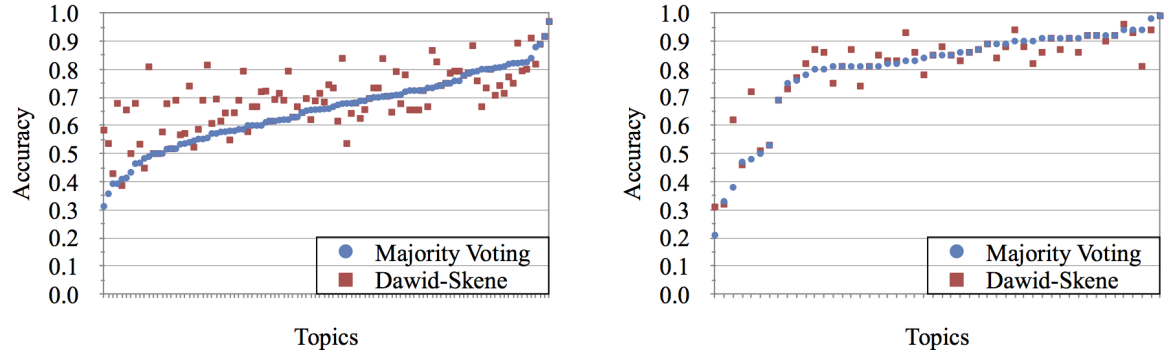
3 COLLABORATIVE JUDGING

Thus far we have seen: (1) greater agreement on NIST-judged relevant documents, potentially due to inexpert crowd judges being



(a) MQ'09 (b) WT'14

Figure 1: Accuracy of crowd judgments vs. average document rank. Bars show accuracy and are shaded: lower, darker regions represent the ratio of true positives, while higher, lighter regions represent the ratio of true negatives. The black line shows the number of judgments per bin.



(a) MQ'09 (b) WT'14

Figure 2: Distribution of crowd judging agreement with NIST across topics for MQ'09 and WT'14. Topics are ordered left-to-right by increasing Majority Voting aggregation accuracy.

liberal in judging relevance when in doubt [33]; (2) greater agreement on documents ranked very high or low [24, 34]; and (3) high variance in agreement across topics. Based on this, we evaluate two simple methods for collaborative judging: a practical method prioritizing highly-ranked documents for trusted judging because of their significant impact on ranking metrics (Section 3.1), and an oracle method which prioritizes documents assuming omniscient knowledge of disagreement for each topic (Section 3.2).

3.1 Descending Rank Based Ordering (DRBO)

Documents at higher ranks more significantly impact rank-based evaluation metrics of IR system performance (e.g., MAP). Therefore, an intuitive method for collaborative judging would be to assign these more important documents to trusted judges.

Specifically, we calculate the weight of each document-topic pair using statAP method [28]'s weighting function, which assigns higher scores to documents at higher ranks. Subsequently, we rank the documents based on their weights in descending order, for each topic. The first K documents of each topic are assigned to the trusted judges, while the rest are judged by crowd workers.

DRBO is an easy-to-implement method in real-life scenarios. We can rank the documents by just using ranked lists of multiple IR systems, and determine the number of documents to be judged by trusted judges based on the available test collection budget.

3.2 Oracle Topic-Based Scheduling (Oracle TBS)

As discussed in Section 2.4, judging disagreement exhibits high variance across topics. Therefore, quality of judgments might be improved by assigning to trusted judges those topics for which the most disagreement is expected.

In practice, predicting which topics are easier to judge for crowd workers is challenging [10, 30, 35]. Prior work may offer some hints to identifying such topics. For example, Kutlu et al. [23] analyze topic specific disagreement reasons and find that ambiguous topic definitions and topics requiring a certain level of topical expertise cause high disagreements. Therefore, test collection builders might

use their own judgments to hypothesize which topics merit expert judges. We leave such prediction for future work.

In this paper, we instead adopt a simpler oracle model as a proof-of-concept which perfectly predicts judging agreement for each topic. Using this oracle, we then prioritize documents for expert judging starting with the lowest agreement topics first. Documents for the same topic are ordered randomly.

4 EVALUATION

Our experiments compare the proposed collaborative judging methods on our test collections. We also report the *random* method, which assigns randomly-selected K documents to the trusted personnel for each topic, as a baseline.

Because the number of documents judged per topic greatly varies, we vary the ratio of NIST judgments used per topic, instead of using a fixed constant. Given our incomplete judgments, we evaluate IR systems using *bpref* [6], which ignores the documents for which no judgment is available. We adopt Soboroff's corrected bpref formulation [32], as implemented in `trec_eval`[4]. We assume that ground-truth ranking of systems comes from ranking systems based on their bpref scores using NIST judgments. We calculate Kendall's τ to measure correlation between the ground-truth ranking and the ranking induced by collaborative judging. We run the random method 50 times and the Oracle TBS method 10 times and report average Kendall's τ. By convention, $\tau = 0.9$ is assumed to constitute an acceptable correlation level for reliable IR evaluation [34].

Results are shown in **Figure 3**, with area-under-curve (AUC) reported as a simple summary statistic. We offer several observations. Firstly, the Oracle TBS method achieves the best overall results, reaching $\tau = 0.9$ score by assigning 55% and 15-20% of the judgments to the trusted personnel in MQ'09 and WT'14, respectively. This suggests that if we could predict which topics will more likely exhibit judging disagreement, then we might maintain NIST quality judging at lower cost through collaborative judging. Secondly, DRBO consistently outperforms the random baseline in WT'14, but not in MQ'09. This may be due to lower quality crowd judgments in MQ'09 (See Table 1). With higher quality crowd judgments, however, DRBO seems to be a simple and effective method. Overall, our results suggest that collaborative judging is a promising method to efficiently build high-quality test collections.

5 CONCLUSION AND FUTURE WORK

In this work, we investigated when crowd workers disagree with NIST assessors and proposed one oracle and one practical collaborative judging approach. Based on our experiments conducted on two different test collections, we offer several observations.

Firstly, higher agreement with NIST on documents NIST judges to be relevant appears to be consistent with prior findings [33] that when in doubt, inexpert crowd judges may be more liberal in judging uncertain documents as relevant. Secondly, we also reaffirm prior work's finding of greater judging agreement at very high and low ranks, suggesting documents whose relevance is not borderline [24, 34]. Thirdly, we do see high variance in agreement across topics, suggesting further confirmation of judging differences between primary and secondary assessors [1, 11, 34].

[4] http://trec.nist.gov/trec_eval/

In regard to collaborative judging, the oracle predicting assessor disagreement also achieved the highest rank correlation, suggesting that a model which could effectively predict judging disagreement [10, 30, 35] could be usefully applied toward collaborative judging. Practical prediction without such an oracle thus remains for future work. Alternatively, our DRBO approach often outperformed the random baseline across collections and aggregation algorithms, and especially with higher quality crowd judgments.

In the future, we plan to use and extend disagreement prediction modeling [10, 30, 35] to further improve collaborative judging. Since our current DRBO model prioritizes highly-ranked documents without considering expected disagreement, it would be of further interest to integrate both strategies into a single, joint model for collaborative judging.

ACKNOWLEDGMENTS

We thank the many talented crowd contributors and NIST relevance assessors who provided the data for our study, as well as Ellen Voorhees and Ian Soboroff for their many years of leading TREC to support the IR research community. 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. 2015. Practical Lessons for Gathering Quality Labels at Scale. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1089–1092.
[3] Omar Alonso, Daniel E Rose, and Benjamin Stewart. 2008. Crowdsourcing for relevance evaluation. In *ACM SigIR Forum*, Vol. 42. ACM, 9–15.
[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 *SIGIR*. ACM, 667–674.
[5] Chris Buckley, Matthew Lease, and Mark D. Smucker. 2010. Overview of the TREC 2010 Relevance Feedback Track (Notebook). In *TREC Conference Notebook*.
[6] Chris Buckley and Ellen M Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 25–32.
[7] Robert Burgin. 1992. Variations in relevance judgments and the evaluation of retrieval performance. *Info. Processing & Management* 28, 5 (1992), 619–627.
[8] Ben Carterette, Virgiliu Pavlu, Hui Fang, and Evangelos Kanoulas. 2009. Million Query Track 2009 Overview.. In *TREC*.
[9] 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.
[10] Praveen Chandar, William Webber, and Ben Carterette. 2013. Document Features Predicting Assessor Disagreement. In *36th ACM SIGIR Conference on Research and Development in Information Retrieval*. 745–748.
[11] Alexandra Chouldechova and David Mease. 2013. Differences in search engine evaluations between query owners and non-owners. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 103–112.
[12] 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.
[13] Kevyn Collins-Thompson, Craig Macdonald, Paul Bennett, Fernando Diaz, and Ellen M Voorhees. 2015. TREC 2014 web track overview. In *TREC*.
[14] Alexander Philip Dawid and Allan Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Stat.* (1979), 20–28.
[15] Ryan Drapeau, Lydia B Chilton, Jonathan Bragg, and Daniel S Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.
[16] Efthimis N. Efthimiadis and Mary A. Hotchkiss. 2008. Legal discovery: Does domain expertise matter? *Proceedings of the American Society for Information Science and Technology* 45, 1 (2008), 1–2.

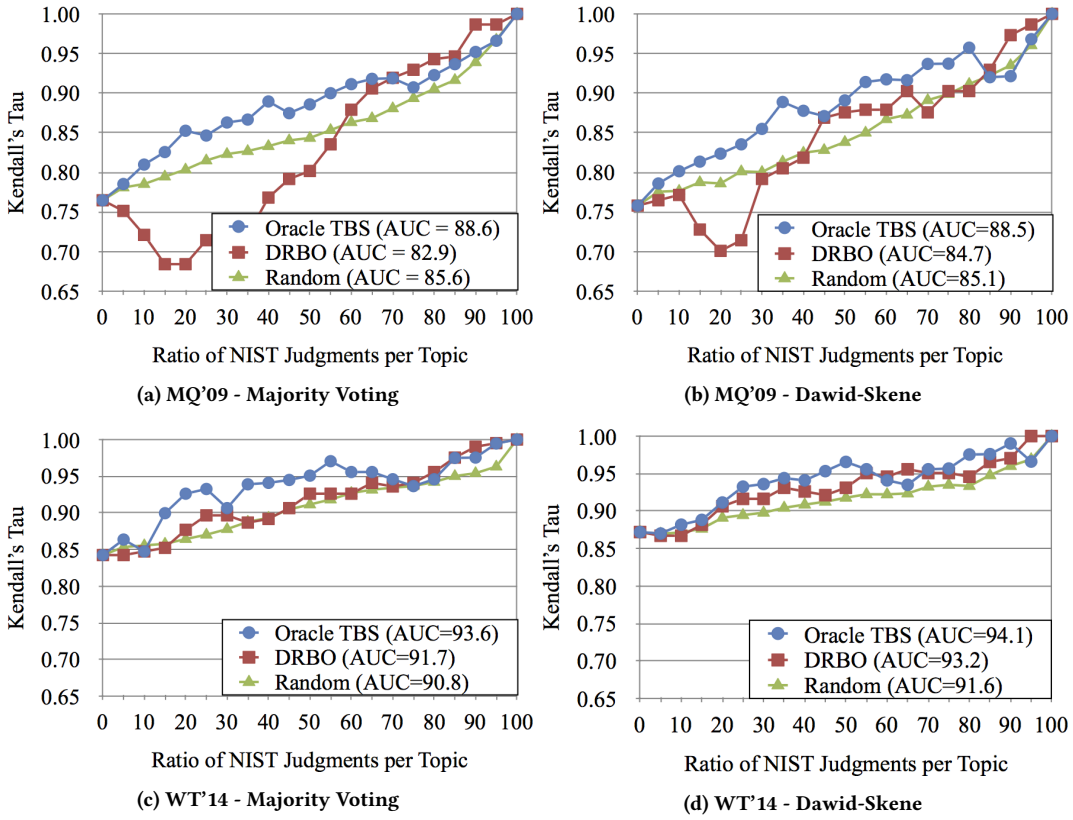


Figure 3: Comparing cost vs. correlation of Oracle TBS, DRBO, and Random methods for collaborative NIST and crowd judging.

[17] Mark E Funk and Carolyn A Reid. 1983. Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association* 71, 2 (1983), 176.

[18] Tanya Goyal, Tyler McDonnell, Mucahid Kutlu, Matthew Lease, and Tamer Elsayad. 2018. Your Behavior Signals Your Reliability: Modeling Crowd Behavioral Traces to Ensure Quality Relevance Annotations. In *Proceedings of the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

[19] Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. 2013. An evaluation of aggregation techniques in crowdsourcing. In *International Conference on Web Information Systems Engineering*. Springer, 1–15.

[20] Hyun Joon Jung, Yubin Park, and Matthew Lease. 2014. Predicting Next Label Quality: A Time-Series Model of Crowdwork. In *The 2nd AAAI Conference on Human Computation & Crowdsourcing (HCOMP)*. AAAI.

[21] Gabriella Kazai, Nick Craswell, Emine Yilmaz, and Seyed MM Tahaghoghi. 2012. An analysis of systematic judging errors in information retrieval. In *21st ACM intl. conference on Information and knowledge management (CIKM)*. 105–114.

[22] Evgeny Krivosheev, Fabio Casati, Valentina Caforio, and Boualem Benatallah. 2017. Crowdsourcing Paper Screening in Systematic Literature Reviews. In *5th AAAI Conference on Human Computation and Crowdsourcing (HCOMP): Works-in-Progress Track*. 108–117.

[23] Mucahid Kutlu, Tyler McDonnell, Yassmine Barkallah, Tamer Elsayed, and Matthew Lease. 2018. Crowd vs. Expert: What Can Relevance Judgment Rationales Teach Us About Assessor Disagreement? *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval* (2018).

[24] Michael E Lesk and Gerard Salton. 1969. Interactive search and retrieval methods using automatic information displays. In *Proceedings of the May 14-16, 1969, spring joint computer conference*. ACM, 435–446.

[25] 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.

[26] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. In *4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

[27] An Thanh Nguyen, Byron C. Wallace, and Matthew Lease. 2015. Combining Crowd and Expert Labels using Decision Theoretic Active Learning. In *Proceedings of the 3rd AAAI Conference on Human Computation (HCOMP)*. 120–129.

[28] V Pavlu and J Aslam. 2007. *A practical sampling strategy for efficient retrieval evaluation*. Technical Report. Technical report, Northeastern University.

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

[30] Aashish Sheshadri. 2014. *A Collaborative Approach to IR Evaluation*. Master's thesis. Department of Computer Science, University of Texas at Austin. https://repositories.lib.utexas.edu/handle/2152/25910.

[31] Aashish Sheshadri and Matthew Lease. 2013. SQUARE: A Benchmark for Research on Computing Crowd Consensus. In *Proceedings of the 1st AAAI Conference on Human Computation (HCOMP)*. 156–164.

[32] Ian Soboroff. 2006. Dynamic test collections: measuring search effectiveness on the live web. In *Proc. SIGIR*. 276–283.

[33] 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.

[34] Ellen M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Info. Processing & Management* 36, 5 (2000), 697 – 716.

[35] William Webber, Praveen Chandar, and Ben Carterette. 2012. Alternative Assessor Disagreement and Retrieval Depth. In *Proc. 21st ACM CIKM*. 125–134.

[36] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: is the problem solved? *Proceedings of the VLDB Endowment* 10, 5 (2017), 541–552.