Crowdsourcing Document Relevance Assessment with Mechanical Turk

Catherine Grady and Matthew Lease
School of Information
University of Texas at Austin
{cgrady,ml}@ischool.utexas.edu

Abstract

We investigate human factors involved in designing effective Human Intelligence Tasks (HITs) for Amazon's Mechanical Turk[1]. In particular, we assess document relevance to search queries via MTurk in order to evaluate search engine accuracy. Our study varies four human factors and measures resulting experimental outcomes of cost, time, and accuracy of the assessments. While results are largely inconclusive, we identify important obstacles encountered, lessons learned, related work, and interesting ideas for future investigation. Experimental data is also made publicly available for further study by the community[2].

1 Introduction

Evaluating accuracy of new search algorithms on ever-growing information repositories has become increasingly challenging in terms of the time and expense required by traditional evaluation techniques. In particular, while the Cranfield evaluation paradigm has proven remarkably effective for decades (Voorhees, 2002), enormous manual effort is involved in assessing topic relevance of many different documents to many different queries. Consequently, there has been significant recent interest in developing more scalable evaluation methodology. This has included developing robust accuracy metrics using few assessments (Buckley and Voorhees, 2004), inferring implicit relevance assessments from user behavior (Joachims, 2002), more carefully selecting documents for assessment (Aslam and Pavlu, 2008; Carterette et al., 2006), and leveraging crowdsourcing (Alonso et al., 2008).

We build on this line of work to investigating crowdsourcing-based relevance assessment via MTurk. While MTurk has quickly become popular as a means of obtaining data annotations quickly and inexpensively (Snow et al., 2008), relatively little attention has been given to addressing human-factors involved in crowdsourcing and their impact on resultant cost, time, and accuracy of the annotations obtained (Mason and Watts, 2009). The advent of crowdsourcing has led to many researchers, whose work might otherwise fall outside the realm of human-computer interaction (HCI), suddenly finding themselves creating HITs for MTurk and thereby directly confronting important issues of interface design and usability which could significantly impact the quality or quantity of annotations they obtain. A similar observation has been made recently regarding the importance of effective HCI for obtaining quality answers from users in a social search setting (Horowitz and Kamvar, 2010).

Our overarching hypothesis is that better addressing human factors in HIT design can yield significantly reduce cost, reduce time, and/or increase accuracy of the annotations obtained via crowdsourcing. Such improvement could come through a variety of complimentary effects, such as attracting more or better workers, incentivizing them to do better work, better explaining the task to be performed and reducing confusion, etc. While the results of this study are largely inconclusive with regard to our

[1] http://aws.amazon.com/mturk

[2] http://www.ischool.utexas.edu/∼ml/data

Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 172–179,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

experimental hypothesis, other contributions of the work are identified in the abstract above.

2 Background

To evaluate search accuracy in the Cranfield paradigm (Voorhees, 2002), a predefined set of documents (e.g., web pages) are typically manually assessed for relevance with respect to some fixed set of *topics*. Each topic corresponds to some static *information need* of a hypothetical user. Because language allows meaning to be conveyed in various ways and degrees of brevity, each topic can be expressed via a myriad of different *queries*. Table 1 shows the four topics used in our study which were generated by NIST for TREC[3]. We do use the paragraph-length "narrative" queries under an (untested) assumption that they are overly complex and technical for a layman assessor. Instead, we use (1) the short keyword "title" queries and (2) more verbose and informative "description" queries, which are typically expressed as a one-sentence question or statement.

NIST has typically invested significant time training annotators, something far less feasible in a crowdsourced setting. NIST has also typically employed a single human assessor per topic to ensure consistent topic interpretation and relevance assessment. One downside of this practice is limited scalability of annotation, particularly in a crowdsourced setting. When multiple annotators have been used, previous studies have also found relatively low inner-annotator agreement for relevance assessment due to the highly subjective nature of relevance (Voorhees, 2002). Thus in addition to reducing time and cost of assessment, crowdsourcing may also enable us to improve assessment accuracy by integrating assessment decisions by a committee of annotators. This is particularly important for generating reusable test collections for benchmarking. Practical costs involved in relevance assessment based on standard pooling methods is significant and becoming increasingly prohibitive as collection sizes grow (Carterette et al., 2009).

MTurk allows "requesters" to crowdsource large numbers of HITs online which workers can search, browse, preview, accept, and complete or abandon.

3. Joint Ventures. Document will announce a new joint venture involving a Japanese company.

13. Mitsubishi Heavy Industries Ltd. Document refers to Mitusbishi Heavy Industries Ltd.

68. Health Hazards from Fine-Diameter Fibers. Document will report actual studies, or even unsubstantiated concerns about the safety to manufacturing employees and installation workers of fine-diameter fibers used in insulation and other products.

78. Greenpeace. Document will report activity by Greenpeace to carry out their environmental protection goals.

Table 1: The four TREC topics used in our study. Topic number and <title> field are shown in bold. Remaining text constitutes the description (<desc>) field.

With regard to measuring the impact of different design alternatives on resulting HIT effectiveness, MTurk provides requesters with many useful statistics regarding completion of their HITs. Some effects cannot be measured, however, such as when HITs are skipped, when HITs are viewed in search results but not selected, and other outcomes which could usefully inform effective HIT design.

3 Methodology

Our study investigated how varying certain aspects of HIT design affected annotation accuracy and time, as well as the relationship between expense and these outcomes. In particular, workers were asked to make binary assessments regarding the relevance of various documents to different queries.

3.1 Experimental Variables

We varied four simple aspects of HIT design:

- **Query**: <title> vs. <desc>
- **Terminology**: HIT title of "binary relevance judgment" (technical) vs. "yes/no decision" (layman)
- **Pay**: \$0.01 vs. \$0.02
- **Bonus**: no bonus offered vs. \$0.02

The **Query** is clearly central to relevance assessment since it provides the annotator's primary basis for judging relevance. Since altering a query can have enormous impact on the assessment, and because we were testing the ability of Mechanical Turk workers to replicate assessments made previously by TREC assessors, we preserved wording of

[3] http://trec.nist.gov

the queries as they appeared in the original TREC topics (see §2). We hypothesized that the greater detail found in the topic description vs. the title would improve accuracy with some corresponding increase in HIT completion time (longer query to read, at times with more stilted language, and more specific relevance criteria requiring more careful reading of documents). An alternative hypothesis would be that a very conscientious worker might take longer wrestling with a vague title query.

Terminology: the HIT title is arguably one of a HIT's more prominent features since it is one of the first (and often the only) description of a HIT a potential worker sees. An attractive title could conceivably draw workers to a task while an unattractive one could repel them. Besides the simple variation studied here, future experiments could test other aspects of title formulation. For example, greater specificity as to the content of documents or topics within the HIT could attract workers that are knowledgeable or interested in a particular subject. Additionally, a title that indicates a task is for research purposes might attract workers motivated to contribute to society.

Pay: the base pay rate has obvious implications for attracting workers and incentivizing them to do quality work. While anecdotal knowledge suggested the "going rate" for simple HITs was about $0.02, we started at the lowest possible rate and increased from there. Although higher pay rates are certainly more attractive to legitimate workers, they also tend to attract more spammers, so determining appropriate pay is something of a careful balancing act.

Bonus: Two important questions are 1) How does knowing that one could receive a bonus affect performance on the current HIT?, and 2) How does actually receiving a bonus affect performance on future HITs? We focused on the first question. When bonuses were offered, we both advertised this fact in the HIT title (see Title 4 above) and appended the following statement to the instructions: "[b]onuses will be given for good work with good explanations of the reasoning behind your relevance assessment." If a worker's explanation made clear why she made the relevance judgment she did, bonuses were awarded regardless of the assessment's correctness with regard to ground truth. Decisions to award bonus pay were made manually (see §5).

3.2 Experimental Constants

Various factors kept constant in our study could also be interesting to investigate in future work:

- **Description**: the worker may optionally view a brief description of the task before accepting the HIT. For all HITs, our description was simply: "(1) Decide whether a document is relevant to a topic, 2) Click 'relevant' or 'not relevant', and 3) Submit".
- **Keywords**: HITs were advertised for search via keywords "judgment, document, relevance, search"
- **Duration**: once accepted, all HITs had to be completed within one hour
- **Approval Rate**: workers had to have a 95% approval rate to accept our HITs
- **HIT approval**: all HITs were accepted, but approval was not immediate to suggest that HITs were being carefully reviewed before pay was awarded
- **Feedback to workers**: none given

More careful selection of high-interest Keywords (e.g., "easy" or "fun") may be a surprisingly effective way to attract more workers. It would be very interesting to analyze the query logs for keywords used by Workers in searching for HITs of interest.

Omar Alonso suggests workers should always be paid (personal communication). Given the low cost involved, keeping Workers individually happy avoids the effort of having to justify rejections to angry Workers, maintains one's reputation for attracting Workers, and still allows problematic workers to be filtered out in future batches.

3.3 Experimental Outcomes

With regard to outcomes, we were principally interested in measuring accuracy, time, and expense. Base statistics, such as the completion time of a particular HIT, allowed us to compute derived statistics like averages per topic, per Worker, per Batch, per experimental variable, etc. We could then also look for correlations between outcomes as well as between experimental variables and outcomes.

Accuracy was measured by simply computing the annotator mean accuracy with regard to "ground truth" binary relevance labels from NIST. A variety of other possibilities exist, such as deciding binary annotations by majority vote and comparing these to ground truth. Recent work has explored ensemble methods for weighting and combining anno-

Topic	Relevant	Non-Relevant
3	48, 55, 84, 120	85
13	28, 30	*193*, 84, 117
68		157, 163, 170, 182, 186
78	*9978*	134, 166, 167,*0062*

Table 2: Documents assessed per topic, along with "true" binary relevance judgments according to official TREC NIST annotation. Document prefixes used in table: (3 and 13) `WSJ920324-`, except `*WSJ920323-0193*`, (68 and 78) `AP901231-` except `*FBIS4-9978*` and `*WSJ920324-0062*`. Only one document, `84`, was shared across queries (3 and 13).

#	Name	Query	Term.	Pay	Bonus
1	Baseline	title	BRJ	$0.01	-
2	P=0.02	title	BRJ	$0.02	-
3	T=yes/no	title	yes/no	$0.01	-
4	Q=desc.	desc.	yes/no	$0.01	-
5	B=0.02	title	yes/no	$0.01	$0.02

Table 3: Experimental matrix. Batches 2 and 3 changed one variable with respect to Batch 1. Batches 4 and 5 changed one variable with respect to Batch 3. *Terminology* varied as specified in §3. For batch 5, 23 bonuses were awarded at total cost of $0.46.

tations (Snow et al., 2008; Whitehill et al., 2009) which also could have been used like majority vote.

As for time, we measured HIT completion time (from acceptance to completion) and Batch completion time (from publishing the Batch to all its HITs being completed). We only anecdotally measured our own time required to generate HIT designs, shepherd the Batches, assess outcomes, etc.

Cost was measured solely with respect to what was paid to Workers and does not include overhead costs charged by Amazon (§2). We also did not account for the cost of our own salaries, equipment, or other indirect expenses associated with the work.

3.4 Additional Details

Assessment was performed on XML documents taken from the TREC TIPSTER collection of news articles. Documents were simply presented as text after simple pre-processing; a better alternative for the future would be to associate an attractive style sheet with the XML to enhance readability and attractiveness of HITs. Relatively little pre-processing

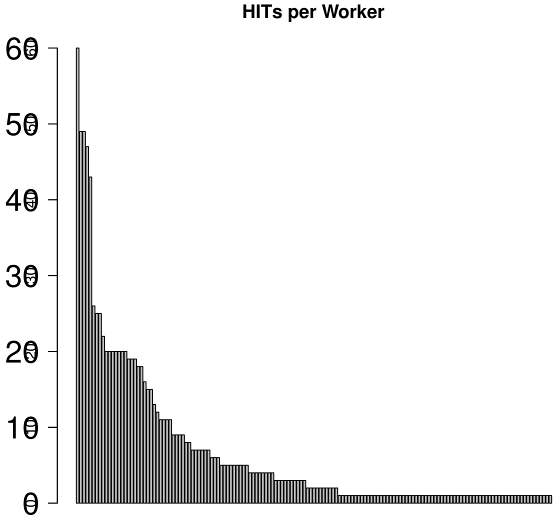

Figure 1: Number of HITs completed by each worker

was performed: (1) XML tags were replaced with HTML, (2) document ID, number, and TREC-related info was commented out, and (3) paragraph tags were added to break up text.

Our basic HIT layout was based on a pre-existing template for assessing binary relevance provided by Omar Alonso (personal communication). This template reflected several useful design decisions like having HITs be self-contained rather than referring to content at an external URL, a design previously found to be effective (Alonso et al., 2008).

4 Evaluation

We performed five batch evaluations, shown in Table 3. For each of the four topics shown in Table 1, five documents were assessed (Table 2), and ten assessments (one per HIT) were collected for each document. Each batch therefore consisted of $4 * 5 * 10 = 200$ HITs, for an overall total of 1000 HITs. Document length varied from 162 words to 2129 words per document (including HTML tags and single-character tokens). Each HIT required the worker to make a single binary relevance judgment (i.e. relevant or non-relevant) for a given query-document pair. In all cases, "ground truth" was available to us in the form of prior relevance assessments created by NIST. 149 unique Workers com-

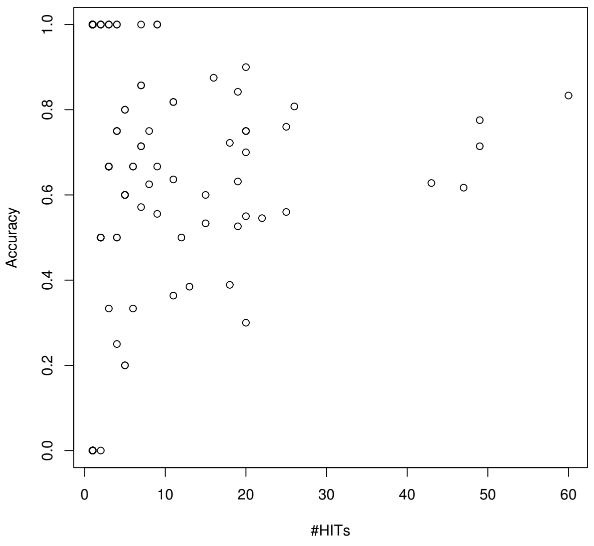

Figure 2: HITs completed vs. accuracy achieved shows negligible direct correlation: Pearson $|\rho| < 0.01$.

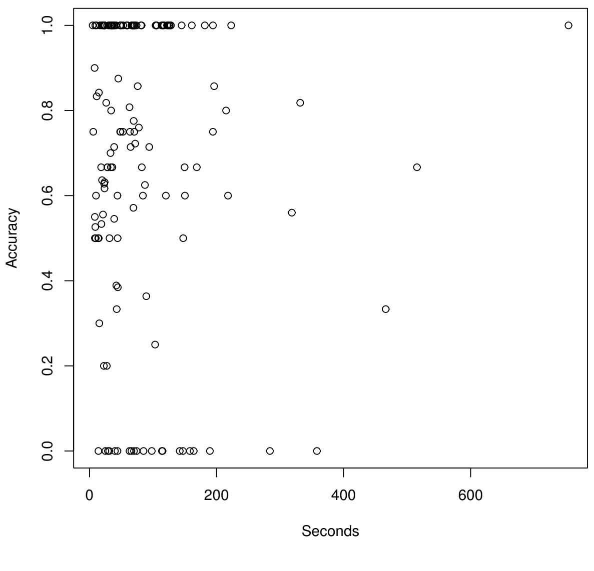

Figure 3: HIT completion time vs. accuracy achieved shows negligible direct correlation: Pearson $|\rho| \approx 0.06$.

pleted the 1000 HITs, with some Workers completing far more HITs than others (Figure 1).

We did not restrict Workers from accepting HITs from different batches, and some Workers even participated in all 5 Batches. Since in some cases a single Worker assessed the same query-document pair multiple times, our results likely reflect unanticipated effects of training or fatigue (see §5).

Statistical significance was measured via a two-tailed unpaired t-test. The only significant outcomes observed were increase in comment length and number of comments for higher-paying or bonus batches. We note p-values < 0.05 where they occur.

Maximum accuracy of 70.5% was achieved with Batch 3, which featured use of Title query and yes/no response. Similar accuracy of 69.5% was also achieved in both Batch 1 and 2. Accuracy fell in Batch 4 (using the Description query) to 66.5%, and fell further to 64% in Batch 5, which featured bonuses. With regard to varying use of Title vs. Description query (Batches 1-3,5 vs. 4), accuracy for the Title query HITs was 68.4% vs. the 66.5% reported above for Batch 4. Thus use of Description queries was not observed to lead to more accurate assessments. HIT completion time was also highest for Batch 4, with workers taking an average of 72s to complete a HIT, vs. mean HIT completion time of 63s over the four Title query batches.

The number of unique workers (UW) per Batch gives some sense of how attractive a Batch was, where a high number could alternatively suggest many workers were attracted (positive) or incentives were too weak to encourage a few Workers to do many HITs (negative). UW in batches 1-4 ranged from 64-72. This fell to 38 UW in Batch 5 (bonus batch), perhaps indicating that workers were incentivized to do more HITs to earn bonuses. At the same time that the number of workers went down, the accuracy per worker went up, with the average worker judging 3.37 documents correctly, compared to a range of 2.10 - 2.20 correct answers per average worker for Batches 1-3 and 1.85 correct answers per average worker for Batch 4 (which, interestingly, had slightly more UWs than the other batches).

Subset			**Cost**		**Batch Completion Time**				**HIT Completion Time**			
	#B	HITs	noB	withB	Total	MeanH	MeanB	sdB	Total	MeanH	MeanB	sdH
Query 3	5	250	$3.00	$3.14	N/A	N/A	N/A	N/A	16127	64.50	3225.4	92.48
Query 13	5	250	$3.00	$3.14	N/A	N/A	N/A	N/A	17148	68.59	3429.6	139.08
Query 68	5	250	$3.00	$3.06	N/A	N/A	N/A	N/A	14880	59.52	2976	111.23
Query 78	5	250	$3.00	$3.12	N/A	N/A	N/A	N/A	17117	68.46	3423.4	122.89
Pay=$0.01	4	800	$8.00	$8.46	1078821	1348.52	269705.25	47486.7	54379	67.97	13594.75	123.57
Pay=$0.02	1	200	$4.00	$4.00	386324	1931.62	386324	N/A	10893	54.465	10893	88.87
Title	4	800	$10.00	$10.46	1227585	1534.48	306896.25	67820.58	50968	63.71	12742	117.14
Desc.	1	200	$4.00	$4.00	237560	1187.8	237560	N/A	14304	71.52	14304	119.20
No Bonus	4	800	$10.00	$10.00	1124799	1405.99	281199.75	70347.43	51966	64.95	12991.5	111.32
Bonus	1	200	$2.00	$2.46	340346	1701.73	340346	N/A	13306	66.53	13306	139.97
Batch 1	1	200	$2.00	$2.00	249921	1249.60	249921	N/A	13935	69.67	13935	130.66
Batch 2	1	200	$4.00	$4.00	386324	1931.62	386324	N/A	10893	54.46	10893	88.87
Batch 3	1	200	$2.00	$2.00	250994	1254.97	250994	N/A	12834	64.17	12834	102.01
Batch 4	1	200	$2.00	$2.00	237560	1187.8	237560	N/A	14304	71.52	14304	119.20
Batch 5	1	200	$2.00	$2.46	340346	1701.73	340346	N/A	13306	66.53	13306	139.97
All	5	1000	$12.00	$12.46	1465145	1465.14	293029	66417.07	65272	65.272	13054.4	117.54

Table 4: Preliminary analysis 1. Column labels: #B: Number of Batches, # HITs, noB: Cost without bonuses, withB: Cost with bonuses, Total, MeanH/B: Mean per-HIT/Batch, sdB/H: std-deviation across Batches/HITs.

Recall that bonuses were awarded whenever Workers provided clear justification of their judgments (whether or not those judgments matched ground truth). In 74% of these cases (17 of the 23 HITs awarded bonuses), relevance assessments were correct. Thus there may be a useful correlation to exploit provided practical heuristics exist for automatically distinguishing quality feedback from spam.

Feedback length might serve as a more practical alternative to measuring quality while still correlating with accuracy. Mean comment length for Batches 2 and 5 was 38.6 and 28.1 characters per comment, whereas Batches 1, 3, and 4 had mean comment lengths of 13.9, 12.7, and 19.3 characters per comment. The mean difference in comment length between Batch 2 and Batch 1 was 24.7 characters ($p<0.01$), 25.9 characters between Batches 2 and 3 ($p<0.01$), and 19.3 characters between Batches 2 and 4 ($p<0.01$). Batch 5 and Batch 1 had a mean comment-length difference of 14.2 characters ($p<0.01$), and Batches 5 and 3 differed by 15.4 characters ($p<0.01$). Thus higher-paying HITs or HITs with bonus opportunities may correlate with greater Worker effort. Batches 2 (pay=$0.02) and 5 (bonus batch) garnered the highest number of comments, with each averaging 0.37 comments per HIT. In contrast, Batches 1, 3, and 4 averaged only 0.21, 0.18, and 0.23 comments per HIT, or a difference of 0.16 ($p<0.01$), 0.19 ($p<0.01$), and 0.14 ($p<0.01$) comments, respectively.

5 Discussion

How to control for the same worker participating in multiple experiments. We found many of the same workers completed HITs in multiple batches, compromising our experimental control and likely introducing effects of training or fatigue. It does not appear that MTurk provides an easy way to preventing this; one can block a worker from doing jobs, but blocking is more of a tool to prevent poor performance. It is also construed as a punishment: workers' ratings can be negatively affected by blocking. Because of this, blocking is not a substitute for a mechanism that simply allows requesters to hide HITs or otherwise disallow repeat workers from completing HITs. It would be nice to develop a simple mechanism for automatically ensuring each experiment involves a different set of workers.

Automatic HIT validation. MTurk does not appear to automatically ensure a submitted HIT was actually completed, i.e. a worker can submit a HIT without having actually done anything. While the submitted HIT can be rejected and re-requested, building some trivial validation of HITs to catch such cases automatically appears worthwhile.

Automatic bonus pay. For Batch 5 (which included bonus pay), one of the authors spent an hour manually processing/evaluating worker annotations and feedback, distributing bonus pay for 23 of the 200 HITs. While some time is certainly well spent in manually analyzing annotations and feedback, the

Subset	Accuracy				Unique Workers				HPW	Feedback Given		Feedback Length	
	#Correct	MeanH	MeanB	sdH	Total	MeanH	MeanB	Acc	Mean	Total	MeanH	MeanH	sd
Query 3	144	0.58	28.8	0.50	84	0.34	16.8	1.71	2.98	60	0.24	17.91	42.44
Query 13	191	0.76	38.2	0.43	88	0.35	17.6	2.17	2.84	71	0.28	25.04	55.82
Query 68	183	0.73	36.6	0.44	83	0.33	16.6	2.20	3.01	69	0.28	21.08	44.69
Query 78	162	0.65	32.4	0.48	83	0.33	16.6	1.95	3.01	76	0.30	26.02	56.52
Pay=$0.01	541	0.68	135.25	0.47	137	0.17	34.25	3.95	5.84	201	0.25	18.49	42.52
Pay=$0.02	139	0.70	139	0.46	64	0.32	64	2.17	3.13	75	0.38	38.60	71.53
Title	547	0.68	136.75	0.47	132	0.17	33	4.14	6.06	229	0.29	23.32	51.23
Desc.	133	0.67	133	0.47	72	0.36	72	1.85	2.78	47	0.24	19.29	46.40
No Bonus	552	0.69	138	0.46	121	0.15	30.25	4.56	6.61	201	0.25	21.12	50.26
Bonus	128	0.64	128	0.48	38	0.19	38	3.37	5.26	75	0.38	28.09	50.21
Batch 1	139	0.70	139	0.46	66	0.33	66	2.11	3.03	42	0.21	13.90	37.62
Batch 2	139	0.70	139	0.46	64	0.32	64	2.17	3.13	75	0.38	38.60	71.53
Batch 3	141	0.71	141	0.46	64	0.32	64	2.20	3.13	37	0.19	12.67	31.96
Batch 4	133	0.67	133	0.47	72	0.36	72	1.85	2.78	47	0.24	19.29	46.40
Batch 5	128	0.64	128	0.48	38	0.19	38	3.37	5.26	75	0.38	28.09	50.21
All	680	0.68	136	0.47	149	0.15	29.8	4.56	6.71	276	0.28	22.51	50.30

Table 5: Preliminary analysis 2. Column labels: HPW: HITs per worker, MeanH/B: Mean per-HIT/Batch, sd(H): std-deviation (across HITs), Acc: mean worker accuracy. Feedback length is in characters.

disparity in cost of our own salaries vs. bonus expenses suggests decisions on bonus pay should be automated if possible (and it likely pays to err on the side of being generous). Of course, automated bonus distribution may negatively affect quality of work if, for example, any string of characters in the feedback box yields bonus pay and workers catch on to this. Similarly, automation may fail to reward truly valuable qualitative feedback from workers which is harder to automatically assess than simply evaluating worker accuracy on known examples.

6 Future Work

Assessing relevance of Web pages. In the near-term, we will be using MTurk to evaluate search accuracy of systems participating in the TREC 2010 Relevance Feedback Track. This will involve addressing several significant challenges: (1) achieving scalable evaluation, (2) protecting workers from malicious attack pages while maintaining assessment accuracy, (3) addressing issues of Web spam, and (4) handling issues of unknown mature content workers may encounter during assessment.

With regard to (1), we will be scaling up Cranfield-based relevance assessment to support search evaluation on the massive ClueWeb09 Web crawl[4]. As for (2), many Web pages containing attack code designed to compromise the viewer's computer, and in a crowdsourced environment we cannot ensure all workers have installed the latest security patches for their Web browsers. Various tradeoffs may be involved between security and usability in pre-rendering Web pages to assess as static images, creating a "safe-viewer" applet, etc. Web spam (3) can be annoying to workers and thereby impact the quality of their work, wastes time and money since spam is never relevant to any query by definition, and spam detection is conceptually a distinct task and ought to be handled as such. In the short term, we may simply ask workers to not only decide relevance vs. non-relevance, but to simultaneously differentiate non-relevant content from non-relevant spam, but a better solution would be preferable. Mature content (4) is similar to spam but can be far worse than annoying to workers, touches on legal issues, and inability to filter it could significantly reduce the number of workers willing to accept HITs which may contain it. Our short-term solution will likely be to perform some simple pre-filtering and simply warn workers they may encounter such content, but this solution is not ideal.

Varying number of annotations in proportion to annotator agreement. While we collected a fixed number of relevance assessments for each query-document pair, it may be both more efficient and more effective to collect few assessments when inner-annotator agreement is high and proportionally more assessments when greater disagreement exists between annotators (Von Ahn et al., 2008).

[4] http://boston.lti.cs.cmu.edu/Data/clueweb09

Graded vs. binary relevance. We want assessors to be both maximally informative and maximally consistent, and there is an inherent tradeoff here. Allowing assessors to make graded relevance judgments corresponds to the intuitive notion that relevance is typically not a binary proposition. Evaluation of commercial search engines today often reports use of a five-point graded scale, and such graded feedback allows us to better distinguish relative effectiveness of different search algorithms at a finer scale. However, the right number of relevance levels to assess is unclear, and too many would likely involve making overly nuanced judgments that could overwhelm assessors and lead to low inner-annotator agreement. We may similarly ask assessors to further differentiate relevance judgment from cases of "I don't know" and "this HIT seems broken". There is also the possibility of inducing graded relevance levels from binary judgments, such as by averaging and rescaling. The utility could be measured by comparing benchmark algorithms using the explicit or induced assessments.

Evaluating annotation accuracy with regard to ground-truth labels vs. task accuracy. While much research with MTurk has measured accuracy in terms of reproducing a ground-truth label, ultimately we are not interested in the labels themselves but rather in what we can do with them. Relevance assessment in particular suffers from notoriously low inner-annotator agreement. Consequently, one alternative to comparing against "ground-truth" labels would be to evaluate the ability of crowdsourced labels for effectively distinguish between different benchmark algorithms.

Crowd demographics. While it is typically suggested that experts produce superior annotations, there are important questions of effects from who is judging the annotations. For example, if you want to know if the general public will think a particular web page is relevant to a particular query, more useful assessments might be obtained from a layman than from someone who builds search engines for a living. This also suggests another reason why it may even be preferable in some circumstances for crowdsource annotations to disagree with "ground-truth" expert labels. It also raises questions about generality of system comparisons based on expert labels when systems are to be used by the general public.

References

Omar Alonso, Daniel E. Rose, and Benjamin Stewart. 2008. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15.

J. Aslam and V. Pavlu. 2008. A practical sampling strategy for efficient retrieval evaluation. Technical report, Northeastern University.

C. Buckley and E.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*, pages 25–32. ACM New York, NY, USA.

B. Carterette, J. Allan, and R. Sitaraman. 2006. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 268–275.

B. Carterette, V. Pavlu, E. Kanoulas, J.A. Aslam, and J. Allan. 2009. If I Had a Million Queries. In *Proceedings of the 31st European Conference on Information Retrieval*, pages 288–300.

D. Horowitz and S.D. Kamvar. 2010. The Anatomy of a Large-Scale Social Search Engine. In *Proc. of the 19th international conference on World wide web (WWW).*

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the 8th SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142.

W. Mason and D.J. Watts. 2009. Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 77–85. ACM.

R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. 2008. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465.

E.M. Voorhees. 2002. The philosophy of information retrieval evaluation. *Lecture Notes in Computer Science*, pages 355–370.

J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. *Proceedings of the 2009 Neural Information Processing Systems (NIPS) Conference.*