Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution

John Le[1], Andy Edmonds[2], Vaughn Hester[1], Lukas Biewald[1]

[1]CrowdFlower
455 Valencia Street
San Francisco, CA, 94103
+1-415-621-2343
{john, vaughn, lukas}@crowdflower.com

[2]eBay Search Science
2065 Hamilton Avenue
San Jose, CA, 95125
+1-206-619-0100
aedmonds@ebay.com

ABSTRACT

The use of crowdsourcing platforms like Amazon Mechanical Turk for evaluating the relevance of search results has become an effective strategy that yields results quickly and inexpensively. One approach to ensure quality of worker judgments is to include an initial training period and subsequent sporadic insertion of predefined gold standard data (training data). Workers are notified or rejected when they err on the training data, and trust and quality ratings are adjusted accordingly. In this paper, we assess how this type of dynamic learning environment can affect the workers' results in a search relevance evaluation task completed on Amazon Mechanical Turk. Specifically, we show how the distribution of training set answers impacts training of workers and aggregate quality of worker results. We conclude that in a relevance categorization task, a uniform distribution of labels across training data labels produces optimal peaks in 1) individual worker precision and 2) majority voting aggregate result accuracy.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Systems and Software–performance evaluation

General Terms
Performance, Design, Experimentation, Human Factors.

Keywords
Crowdsourcing, search relevance evaluation, quality control.

1. INTRODUCTION

Crowdsourcing is the use of large, distributed groups of people to complete microtasks or to generate information. Because traditional search relevance evaluation requiring expert assessment is a lengthy process [2, 3, 5], crowdsourcing has gained traction as an alternative solution for these types of high volume tasks [2, 1]. In some cases, crowdsourcing may provide a better approach than a more traditional, highly-structured judgment task because it facilitates the collection of feedback from a wide variety of viewpoints for the same comparison. Feedback from varying viewpoints naturally captures the myriad interpretations a particular problem may have.

Quality assurance is a major challenge of crowdsourcing [8, 9]. Without a rigorous quality control strategy, workers often produce an abundance of poor judgments. Poor judgments by a worker can occur when a worker is ethical but misinterprets the designer's intent for the task. This is a case of a worker's bias introducing error. Unethical workers, who do not attempt to honestly complete tasks but simply answer as many questions as quickly as possible, are another source of erroneous judgments. In the first case we are left with some erroneous judgments which can adversely affect our results. The second case we are left with completely unusable judgments which invalidate our results.

To deal with each of these cases, we train our workers on previously defined gold standard data (training data) in an dynamic learning environment that gives instant feedback for why the answer they chose was incorrect. If a worker answers too many questions incorrectly, suggesting they are an unethical worker, we ban them from returning to the task.

By running tasks like this we saw that worker responses were influenced by the distribution of correct answers in the training data. Ethical workers naturally developed notions on how the data was distributed and actively applied what was learned to future questions. This situation is similar to that in machine learning, where classifiers develop bias towards the training data. When testing machine learning algorithms, training data must be fairly sampled from the underlying population distribution to ensure minimal bias. Unethical workers optimize their responses to maximize revenues while minimizing the time spent making judgments. For example, if a worker perceives 80% of the answers are of label A then they will answer A every time.

In this paper we attempt to quantify the influence of the dynamic learning environment by examining how the distribution of correct answers in the training data affects worker responses. We hypothesize that training data in which the distribution of correct answers is more uniform yields optimal results with respect to worker quality and aggregate majority vote result quality. We test this hypothesis on a task where we ask workers to categorize query results into four categories. We compiled a test set that had a skewed underlying distribution (a higher proportion of one label), and then trained different sets of workers on five different training sets. This will be explained in more detail in Sections 3 and 4.

2. BACKGROUND

Amazon Mechanical Turk (AMT) is a platform offered

SIGIR '10, July 19-23, 2010, Geneva, Switzerland

by Amazon Web Services that facilitates online work between job requesters and workers from around the world. CrowdFlower is a labor-on-demand product that facilitates the completion of online microtasks among a number of labor channels including AMT. CrowdFlower provides the infrastructure for training workers via training data. In combination, these two products allow high throughput while ensuring judge quality.

Current strategies for evaluating and ensuring quality in crowdsourced tests include measurement of agreement, qualification questions, and worker trust algorithms [10, 7, 6]. When measuring quality with agreement, either by majority vote or similar methods, it is important to consider that high agreement among multiple judges may reflect a variety of factors, particularly:

1. Correctness of the chosen label
2. Cultural bias of the workers
3. Interpretation/ambiguity of the question
4. Cheating and collusion [3, 9].

Agreement assessment is often used in conjunction with worker error estimation on a previously defined gold standard [10, 7, 9]. In Snow [10] and Ipeirotis [7], gold standard answers are hidden from the workers and used in post-processing to estimate the true worker error-rate. For a movie rating categorization task, Ipeirotis showed that weighting worker responses by their error-rate on a previously defined gold standard improved accuracy from 95% to 99.8% [7].

As stated in Section 1, we use previously-defined gold standard data in an dynamic learning environment to provide instant feedback to workers when they answer these questions incorrectly. The gold standard data used in training will be referred to hereinafter as training data (interchangeably with training set). Analogously, testing data (test set) is the gold standard data against which results are reported.

3. DATA

The dataset came from a major online retailer's internal product search projects. It consisted of 256 queries with 5 product pairs associated with each query. In other words, the dataset contained 1,280 search results. We will refer to each batch of five search results to a query as a result set. There were 164 distinct queries which included product queries such as: "LCD monitor," "m6600," "epiphone guitar," "sofa," and "yamaha a100."

The training data was sampled from a dataset previously judged by a set of experts from the online retailer. The test set was taken from the same dataset without repetition. We ran five tasks where the test set had a highly skewed distribution towards "Not Matching" results; 82.67% of results were "Not Matching", 14.5% "Matching", 2.5% "Off Topic" and 0.33% "Spam".

We varied the distribution of answers in the training set from one skew to the other, particularly as seen in Table 1. We attempted to vary "Matching" and "Not Matching" results as symmetrically as possible, but as we decreased the number of "Not Matching" results, the number of "Off Topic" results increased significantly as well.

Table 1: Training Data Skew

Experiment	1	2	3	4	5
Matching	72.7%	58%	45.3%	34.7%	12.7%
Not Matching	8%	23.3%	47.3%	56%	84%
Off Topic	19.3%	18%	7.3%	9.3%	3.3%
Spam	0%	0.7%	0%	0.7%	0%

4. EXPERIMENTAL DESIGN

4.1 Amazon Mechanical Turk

We set up five tasks via CrowdFlower to be run in parallel on Amazon Mechanical Turk. The task instructions, layout, title, pricing, and design were all the same, and hence all appeared as the same task on Mechanical Turk. We paid workers 20 cents to judge the relevance of five result sets, or 25 search results.

4.2 Task Design

When a worker comes to our task or HIT, they see a set of instructions followed by five queries with five corresponding search results. Each query is accompanied by the category in which it was searched, if available.

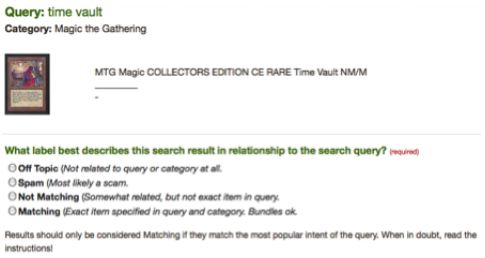

Figure 1: Example of one query-product pair in a HIT

The instructions detail how to label search results as one of four categories: "Off Topic", "Spam", "Not Matching", and "Matching". The instructions include examples and reasons to guide workers as they make judgments. We tell workers that search results in the above categories should follow the following guidelines (shortened here for brevity):

- Matching is a result that matches the "most likely intent" of the query. We define this as the core product in the search. For instance a search for "iphone" may yield results for iPhone cases; only results for an actual iPhone are matching.
- Not Matching is a result that does not match the most likely intent of the query, but is still relevant. In the above example, we would consider iPhone cases non-matching to the query "iPhone."
- Spam is a result that appears to be a solicitation or pornographic in nature. An image which does not picture the product but instead scantily clad women usually indicates spam.
- Off Topic is a result that is completely unrelated to the given query, though the query may appear in the result. A query for "iPod" may have a result for a car

where "iPod" is in the result string. The main product in the result is a car which has nothing to do with the iPod.

The full instruction set and task interface are available here: <http://crowdflower.com/judgments/mob/13838>.

4.3 Dynamic Learning for Quality Control

This experiment used the training data first in an entry training module, in which each worker has to complete 20 query-result pairs successfully before proceeding to test-set questions. The workers are notified that only upon passing this section will they receive payment. We inform workers of their mistakes. After this training period, training data is used as periodic screening questions [4] to provide live feedback when workers err. The feedback explains what the correct answer should be and why. For every 20 query-result pairs a worker saw, they also were exposed to five training data questions in periodic screening.

As a worker answers these training data questions, we calculate their accuracy and use it as an estimate for the worker's "true" accuracy. We rely on a simplified metric, the percent of correct responses, as described by Snow et al [10]. Workers are blocked from continuing on a task if their accuracy is poor. Before being blocked, a worker will receive a warning that their accuracy is too low and that they should reread the instructions to correct mistakes. Unlike in [10, 7], we did no post-processing to refine worker error estimation.

5. RESULTS

5.1 Workers

There were 255 unique workers who participated in these five experiments. There were no AMT qualifications to exclude certain workers from this task. The workers were split randomly into one of the tasks that were live simultaneously such that test group sizes were uniform. We stored the task assignment of each worker on our servers; if a worker had previously been working on a task and then resumed the session, he/she would be returned to the same task.

Routing to a task stopped if the task fulfilled its judgment needs of five trusted judgments per result. The distribution of unique workers across each task is affected by a variety of factors: individual worker output, untrusted workers, the number of judgments needed to complete each task, changes in the routing of workers away from tasks that had fulfilled their judgment needs, etc. Table 2 shows the distribution of worker involvement.

Table 2: Worker Distribution

Experiment	1	2	3	4	5
Came to the task	43	42	42	87	41
Did Training	26	25	27	50	21
Passed Training	19	18	25	37	17
Failed Training	7	7	2	13	4
Percent Passed	73%	72%	92.6%	74%	80.9%

5.2 Individual Worker Quality

In the experiments where the underlying distribution skewed toward "Not Matching," individual worker test accuracy increased as the training set more closely reflected the underlying distribution. Optimal worker accuracy is achieved when training distributions match the population distribution. But when the test set is highly skewed, other measures may be more effective since a worker can achieve 82% accuracy by answering all "Not Matching." Worker precision on "Not Matching" labels peaked when the training answers were uniform over the labels (Table 3).

Table 3: Average Worker Performance Measures

Worker \ Experiment	1	2	3	4	5
Accuracy (Overall)	0.690	0.708	0.749	0.763	0.790
Precision (Not Matching)	0.909	0.895	0.930	0.917	0.915
Recall (Not Matching)	0.704	0.714	0.774	0.800	0.828

5.3 Aggregate Result Quality

The aggregate majority vote results have the greatest accuracy (87.67%) when the distribution of training data answers is the most uniform. This accuracy is 5% greater than baseline accuracy (82.67%) and 12.77% greater than individual worker accuracy as shown in Table 3. Baseline accuracy for the aggregate results would be defined as a majority of workers answering all questions as "Not Matching." When the training data distribution is the same as the underlying distribution accuracy was 85% (3% above baseline accuracy).

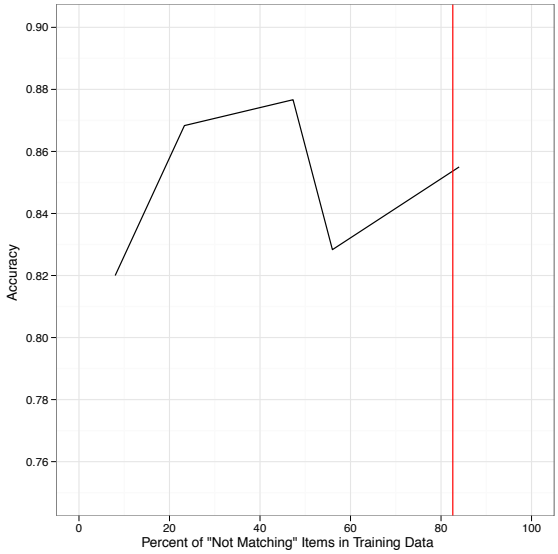

Figure 2: Aggregate Accuracy on Test Data

In these experiments the underlying distribution is so heavily skewed towards one label that we may want to optimize other performance metrics. As seen in Table 4, each measure is maximized in Experiment 3 (which contained a uniform training set distribution).

Table 4: Aggregate Performance Measures

Experiment	1	2	3	4	5
Precision	0.921	0.932	0.936	0.932	0.912
Recall	0.865	0.917	0.919	0.863	0.921

Note that as the training distribution more closely reflects the skewed population distribution in experiment 5, recall exceeded precision. A majority of workers were optimizing for the "Not Matching" label, labeling more items as "Not Matching" at the cost of being less precise.

6. DISCUSSION

The task environment is one in which workers can learn what is expected of them as they progress through the task. This enables the task designer to incorporate more detailed instructions with the expectation that workers can and will skip the instructions to immediately begin answering questions. We anticipate that many workers will only reference the instructions upon notification of their mistakes.

The training method is analogous to training a classifier with a machine learning algorithm. Most machine learning algorithms are applied using a randomly selected training set, which would be expected to approximate the underlying distribution.

We found that workers yielded greater precision on identifying "Not Matching" items when they were trained on a training set with a more uniform distribution of correct answers. Results aggregated by majority vote had greater accuracy even though the test set had a skewed distribution towards "Not Matching" items.

Workers are very adept at realizing that items are heavily skewed to a certain label (an anchoring effect) and may be predisposed to select the label with the highest prior. Workers would then be more likely to miss items that deviate from their expectations. Thus in this learning environment, training questions should predispose no bias.

This phenomenon may be due in part to workers' ability to learn testing data as they are exposed to it. Machine-learned classifiers generally cannot learn from test data as it is processed, which is why it is so important to have robust training sets. Humans are not machines, so when doing machine-learning-like tasks where we use humans as classifiers, we must apply different techniques to train them. Tong et al [11] noted that incorporating active learning methods in training machine-learned classifiers may offer improvements to traditional methods. This result may also imply that strategies for training humans could inform future research on machine learning algorithms.

This experiment suggests broader implications for practitioners; namely that a dynamic learning environment can be used strategically to: 1) identify unethical workers and 2) train ethical workers more effectively. However, the attributes of the learning environment are critical. In particular the choice of training examples will affect worker output. Further development and application of these principles will enable us to approach search relevance tasks involving ambiguous queries or even more complex tasks that require domain-specific knowledge.

7. FUTURE WORK

We shall run more experiments to further validate these results. Future research should also extend the learning environment, possibly by incorporating active learning methods to train workers on similar examples of items they got incorrect and by developing a more refined model for estimating the "true" error rate of workers using a full multinomial model [10]). Having such a model for worker responses may better show why workers are getting questions wrong and may also point to difficulty and ambiguity in our task. If we differentiate workers by demographics we may also be able to identify cultural differences, which could in turn improve task design.

8. NOTES

We have run this type of task numerous times over the past year on AMT, and as such workers may have entered the job with expectations as to what answers allow them to pass training questions. On previous runs of this task, items were overwhelmingly "Matching" (about 80%). Because repeat workers can learn the training data through repeated exposure, our experiments skewed the distribution of items towards "Not Matching." We point out that the training data used for these experiments had not been used previously on any crowdsourcing platform.

We priced this task at an extremely high rate for an AMT task. An unusually high price tends to attract many opportunistic or untrustworthy workers. Part of the goal of this experiment was to engage a diverse set of both ethical and unethical workers.

9. ACKNOWLEDGMENTS

We would like to thank Brian Johnson (eBay), James Rubinstein (eBay), Aaron Shaw (Berkeley), Alex Sorokin (CrowdFlower), Chris Van Pelt (CrowdFlower) and Meili Zhong (PayPal).

10. REFERENCES

[1] O. Alonso. Guidelines for designing crowdsourcing-based relevance evaluation. In *ACM SIGIR*, July 2009.

[2] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[3] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Here or there: Preference judgments for relevance. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, 2008.

[4] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor. Are your participants gaming the system?: screening mechanical turk workers. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pages 2399–2402. ACM, 2010.

[5] G. Dupret, V. Murdock, and B. Piwowarski. Web search engine evaluation using click-through data and a user model. In *Proceedings of the Workshop on Query Log Analysis WWW2007*, May 2007.

[6] C. Grady and M. Lease. Crowdsourcing document relevance assessment with amazon's mechanical turk. In *NAACL/HLT 2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk (at the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics)*, 2010.

[7] P. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *KDD-HCOMP '10*, 2010.

[8] J. Kamps, S. Geva, C. Peters, T. Sakai, A. Trotman, and E. Voorhees. In *Report on the SIGIR 2009 workshop on the future of IR evaluation*, volume 43, pages 13–23. ACM, 2009.

[9] G. Kazai and N. Milic-Frayling. On the evaluation of the quality of relevance assessments collected through crowdsourcing. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 21–22, 2009.

[10] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics, 2008.

[11] S. Tong, D. Koller, and P. Kaelbling. Support vector machine active learning with applications to text classification. In *Journal of Machine Learning Research*, pages 999–1006, 2001.