Worker Types and Personality Traits in Crowdsourcing Relevance Labels

Gabriella Kazai
Microsoft Research
Cambridge, UK
v-gabkaz@microsoft.com

Jaap Kamps
University of Amsterdam
The Netherlands
kamps@uva.nl

Natasa Milic-Frayling
Microsoft Research
Cambridge, UK
natasamf@microsoft.com

ABSTRACT

Crowdsourcing platforms offer unprecedented opportunities for creating evaluation benchmarks, but suffer from varied output quality from crowd workers who possess different levels of competence and aspiration. This raises new challenges for quality control and requires an in-depth understanding of how workers' characteristics relate to the quality of their work. In this paper, we use behavioral observations (HIT completion time, fraction of useful labels, label accuracy) to define five worker types: Spammer, Sloppy, Incompetent, Competent, Diligent. Using data collected from workers engaged in the crowdsourced evaluation of the INEX 2010 Book Track *Prove It* task, we relate the worker types to label accuracy and personality trait information along the 'Big Five' personality dimensions. We expect that these new insights about the types of crowd workers and the quality of their work will inform how to design HITs to attract the best workers to a task and explain why certain HIT designs are more effective than others.

Categories and Subject Descriptors: H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation (efficiency and effectiveness)*

General Terms: Design, Experimentation, Human Factors

Keywords: Crowdsourcing relevance labels, BFI test, worker typology

1. INTRODUCTION

Crowdsourcing [7] is emerging as an economically viable approach to scaling up efforts that require large-scale human input. In particular, it has been considered as a feasible alternative to employing editorial judges to collect relevance labels for evaluation of search engines [1, 2, 6, 11, 12]. With the convenience of crowdsourcing platforms, such as CrowdFlower or Amazon's Mechanical Turk (AMT), such labels can now be collected with ease from an almost unlimited global workforce. However, this comes at a price. Crowd workers come from different socio-economic backgrounds, with different skills and motivations, and complete tasks with varying levels of success [18]. As a result, crowdsourcing can suffer from low quality output due to dishonest, random, and sloppy workers' behavior, e.g., [4, 11, 16, 21]. Thus, various methods have been proposed to detect unethical and sloppy workers, e.g., by analyzing patterns in the worker's output and time spent on a task [4, 14, 19, 21], or by observing agreement levels among the workers and with a 'gold standard' set of trusted labels [1, 6, 11, 12].

More recently, increasing attention is being paid to studying characteristics of the workers themselves and how these relate to the quality of their work. For example, [3] experiment with a range of hypothetical assessor archetypes and examine the effects of associated assessor errors on the evaluation of IR systems. They differentiate between unenthusiastic, optimistic, pessimistic, topic-disgruntled, lazy, fatigued and 'Markovian' assessor models and use these models to simulate assessors going through the process of making judgments in the TREC Million Query track test collections. The work of [20] presents a classification of crowd workers, differentiating between two types of ethical workers (sloppy and proper), and two types of unethical workers (random spammers and uniform spammers). A total of 850 votes obtained from 33 crowd workers for 10 topics from the TREC 2010 Web Track ad-hoc task are used to confirm this classification. Related works in the broader area of online behavior have studied Internet users' personality traits, which are relatively stable, enduring properties of people and are also key in influencing behavior [10]. Personality traits have been shown to relate to user behavior in the context of e-commerce, e.g., [8], social media, e.g., [5] and web browsing and searching behavior, e.g., [15].

In this paper, we conduct a study of the crowd workers characteristics and propose a typology that reflects their observable behavior with the crowdsourcing platform, e.g., average time a worker spent on a task or the number of useful labels they contributed. Specifically, we analyze the data collected from the workers engaged in a relevance labelling task over 2100 topic-document pairs from the INEX 2010 Book Track's test collection. We classify the workers according to our typology and examine the personality traits that characterize the different groups of workers. Our analysis aims to answer the following research questions:

- Are behavior and personality traits related to label accuracy?
- What can analysis by worker types reveal about groups of workers?

In the next section we detail our experimental methodology and define the worker typology comprising five categories: spammers, sloppy, incompetent, competent, and diligent workers. Section 3 studies the behavior and personality of workers in relation to their accuracy, and Section 4 analyzes the worker types. We conclude in Section 5.

2. STUDY SETUP AND TYPOLOGY

In this section we discuss the setup of our crowdsourcing experiment on AMT and define a typology of crowd workers.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

Table 1: BFI test (1=Disagree strongly, . . . , 5=Agree strongly)

I see myself as someone who ...	
1. ... is reserved	(1)(2)(3)(4)(5)
2. ... is generally trusting	(1)(2)(3)(4)(5)
3. ... tends to be lazy	(1)(2)(3)(4)(5)
4. ... is relaxed, handles stress well	(1)(2)(3)(4)(5)
5. ... has few artistic interests	(1)(2)(3)(4)(5)
6. ... is outgoing, sociable	(1)(2)(3)(4)(5)
7. ... tends to find fault with others	(1)(2)(3)(4)(5)
8. ... does a thorough job	(1)(2)(3)(4)(5)
9. ... gets nervous easily	(1)(2)(3)(4)(5)
10. ... has an active imagination	(1)(2)(3)(4)(5)

Scoring the BFI-10 scales (averages of two rows): Extraversion: 1R, 6; Agreeableness: 2, 7R; Conscientiousness: 3R, 8; Neuroticism: 4R, 9; Openness: 5R; 10 (R = item is reversed-scored). Reproduced from [17].

2.1 Labeling Task and Data

For our crowdsourcing experiments, we use a relevance labeling task and ask workers on AMT to label pages of digitized books from the INEX Book Track's test collection [13]. We use 21 topics and 3,557 page level relevance judgments collected from the INEX participants as our gold set (GS).

For each topic, we create 10 HITs, each containing 10 book pages of which on average 4.5 are GS pages with known labels of which roughly half, 2.2, are known relevant pages. This results in a total of 210 HITs where each HIT is assigned to 3 workers, resulting in 630 HIT assignments. We payed $0.25 per HIT assignment, costing a total of $173.25, including commission fees. This HIT design is that of the Simple Design (SD) in [12].

2.2 Personality Traits Survey

As part of the HIT design, we included a questionnaire to collect personality traits information on the workers. We used a standard 10 question test [17] used to measure the five personality dimensions by Goldberg [10], aka the 'Big Five' (BFI), see Table 1: *Openness* is the tendency to be imaginative, independent, and interested in variety (high score) vs. practical, conforming, and interested in routine (low score). *Conscientiousness* is the tendency to show self-discipline, act dutifully, be organized, careful, and disciplined (high score) vs. disorganized, careless, and impulsive (low score). *Extraversion* is the tendency to be sociable, fun-loving, and affectionate (high score) vs. retiring, somber, and reserved (low score). *Agreeableness* is the tendency to be compassionate and cooperative, trusting, and helpful (high score) vs. self-interested, suspicious, antagonistic and uncooperative (low score). *Neuroticism* is the tendency to be calm, secure, and self-satisfied (low score) vs. anxious, insecure, emotional instable and self-pitying (high score).

2.3 Behavioral Observations

We calculate the following measures that characterizes workers' behavior based on observations on the collected data: the number of HITs completed by a worker (#HITs), the average amount of time a worker spent on a HIT (AvgTime), the percentage of useful labels (i.e., relevant/non-relevant vs. missing labels or labels like 'cannot judge') contributed (%Rel), and the ratio of correct labels based on agreement with the gold set (Accuracy).

2.4 Data Analysis Methods

In all our analysis, we calculate statistics per worker and aggregate these to groups of workers sharing a characteristic. For survey style questions, we take the most frequent response over the different HITs per worker. For quantitative measures, e.g., time spent per HIT, we take the arithmetic mean.

2.5 Worker Typology

Since a key challenge in crowdsourcing is to attract workers with desirable characteristics and behavior and deter those with undesired behavior, we define a simple typology of workers based on combination of the observable variables:

- *Diligent workers* take care in completing their HITs and may thus be characterized by a high ratio of useful labels, longer average time spent per HIT, and a high label accuracy.
- *Competent workers* may be skilled workers who provide many useful labels and obtain a high accuracy, but work fast, making them very efficient and effective workers.
- *Sloppy workers* care little about the quality of their work. They may still provide a high fraction of useful labels, but they work as fast as possible, spending relative little time per HIT. As a result, their Accuracy is expected to be low.
- *Incompetent workers* may also provide many useful labels, but despite spending considerable time per HIT only obtain a low accuracy, plausibly due to lacking skills or competencies such as a poor understanding of the task.
- *Spammers* may come in different shapes and forms, but those workers that deliver very few useful labels are an obvious case of malicious workers.

We attach no particular importance to the exact label of an individual worker, rather, we view these types as approximations of groups of workers, useful for analysis and for making sense of the inherently noisy crowdsourcing data. For example, another typology may class subsets of sloppy or incompetent workers as spammers.

3. ANALYSIS OF CROWD WORKERS

In this section we analyze the behavior, the personality traits and accuracy of the crowd workers who took part in our experiment.

3.1 Behavioral Observations

A total of 155 workers completed 630 HITs, yielding an average of 4.1 HIT per worker, but with median of 1 and maximum of 86 HITs per worker, see Figure 1a. Similarly to other studies, e.g., [1, 9], HIT uptake follows a power law like distribution—over 54% of the workers is only observed during a single HIT. The average HIT completion time has a mean of 560 seconds but with a large standard deviation of 390 seconds, shown in Figure 1a. Figure 1b shows relatively high ratio of useful relevance labels (%Rel), with a mean of 0.76 and standard deviation (std.) of 0.27, and much lower and varied levels of label accuracy measured on the gold set (Accuracy), with a mean of 0.53 and std. 0.32.

Looking at the average label accuracy achieved by a worker, we find no significant relation between the number of HITs completed by a worker and their accuracy. This is not surprising given that most workers complete only a single HIT. The average time per HIT is, however, significantly related to accuracy (Pearson correlation r of 0.22, $p < 0.01$), see Figure 1d. Moreover, we find that the fraction of useful labels obtained is strongly correlated to accuracy, the relation is highly significant (Pearson $r = 0.66$, $p < 0.001$), shown in Figure 1e.

3.2 Personality Traits

Figure 1c shows the distribution of responses to the BFI questionnaire. For *Openness* we observe a mean score of 3.44 (std. 0.71, median 3.0) suggesting a tendency toward artistic interests and active imagination. The mean score of *Conscientiousness* is higher with 3.80 (std. 0.82, median 4) suggesting that workers do a thorough job. *Extraversion* has a mean 3.14 (std. 0.98, median 3)

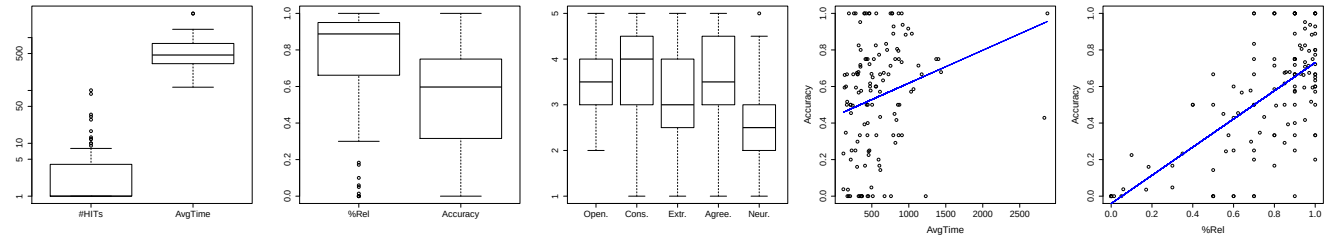

Figure 1: Behavioral observations and personality traits per worker: (a) the number of HITs (#HITs) and average time per HIT (AvgTime), (b) ratio of useful labels (%Rel) and label accuracy (Accuracy), (c) average scores on BFI personality traits, (d) correlation between AvgTime and Accuracy, and (e) correlation between %Rel and Accuracy.

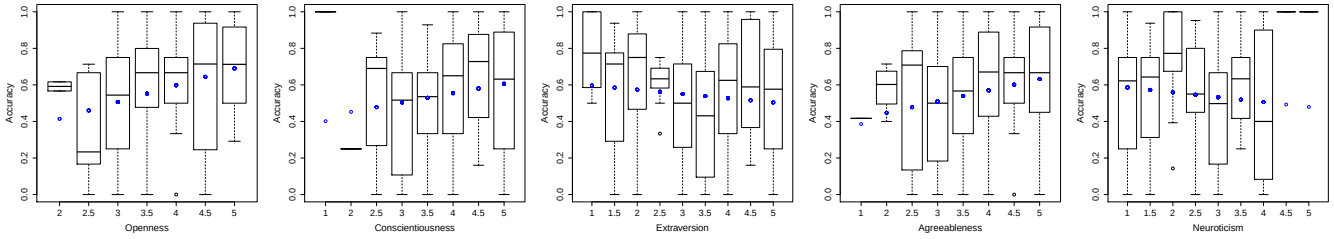

Figure 2: Accuracy over personality traits

exhibiting no particular disposition on this trait. We observe again a high score for *Agreeableness* with a mean of 3.58 (std. 0.83, median 3.5), suggesting that worker are generally trusting and helpful. Finally, we find a low mean score for *Neuroticism*, 2.56 (std. 0.90, median 2.5) which suggests a relaxed nature of the workers.

Among the personality traits, we find that *Openness* significantly relates to accuracy (Spearman r = 0.19, $p < 0.05$), see Figure 2. *Conscientiousness* and *Agreeableness* also have a positive relation to accuracy (r = 0.10, not significant and r = 0.16, not significant, resp.). *Extraversion* and *Neuroticism* exhibit a negative relation (r = -0.11, not significant and r = -0.08, not significant, resp.). In general, the results suggest that for high label accuracy, one should attract workers that score high on openness, conscientiousness, and agreeableness, and score low on extraversion and neuroticism.

So far, we related behavioral observations and personality traits to accuracy and found strong relations with %Rel and AvgTime, but mostly weak relations with personality traits. This suggests that the behavioral observations are more effective at distinguishing low quality work for the job at hand, but the personality characteristics, which reflect the person rather than the precise task performance, can be useful to distinguish between good and better workers.

4. ANALYSIS BY WORKER TYPES

In this section we use the typology of workers introduced in Section 2.5 with a 60% threshold for %Rel and Accuracy, and the median time of 468 seconds as threshold for AvgTime, see Table 2.

4.1 Behavior by Worker Type

Our segmentation resulted in a strong correlation between workers types and accuracy (ANOVA, $p < 0.001$). Figure 3a shows the mean accuracy of workers fitting a particular type. Spammers are clearly differentiated from the rest of the workers with very low accuracy while diligent workers obtain the highest label accuracy against the gold set. Although we used label accuracy to define competent and diligent workers relative to incompetent and sloppy workers, it is interesting to note the differences within these two groups. Competent workers underperform diligent ones on average, and show higher performance variation. Among the competent workers, those who work more diligently and spend longer on the task achieve higher label accuracy (Pearson correlation r of 0.40, $p < 0.05$). This suggests that a more granular definition of worker types may be useful in the analysis, e.g., differentiating 'diligent and competent' workers. On the other hand, among diligent workers, AvgTime is only weakly correlated with Accuracy (r=0.15, not significant), suggesting that this group of workers is more consistent in their behavior. Interestingly, the relation between AvgTime and Accuracy is much weaker among incompetent workers (r=0.03, not significant) and spammers (r=0.21, not significant), and negative for sloppy workers (r=-0.30, not significant). This suggests that time is less useful for detecting these types of workers.

Table 2: Distribution of workers over types

	Spammer	Sloppy	Incompetent	Competent	Diligent
%Rel	Low	High	High	High	High
AvgTime	–	Low	High	Low	High
Accuracy	–	Low	Low	High	High
Workers	37	29	22	25	42

Earlier we saw that the number of HITs completed by a worker was unhelpful in characterizing workers' performance in the entire crowd sample. However, when we break the sample down by worker types, see Figure 3b, we find that surprisingly, accuracy decreases with a higher number of completed HITs, not only for spammers (r=-0.14, not significant), but also for competent (r=-0.26, not significant) and diligent workers (r=-0.37, $p<0.05$). This may signal fatigue or insufficient ratio of pay verse effort, thus the more serious, but possibly digruntled, workers leaving the task. At the same time, we also see evidence of some competent and diligent workers being 'hooked' on the task, completing as many as 35 HITs (competent, max 24 HITs per diligent worker). The distribution of number of HITs completed is particularly marked for incompetent workers, most of whom do not return after (a not so successful) first try. Sloppy workers have the highest 'return-rate', where the performance of workers decreases with the more HITs they complete (r=0.37, p=0.05). Again, this suggests that in a refined typology such workers may be labelled as 'sloppy spammers'.

4.2 Personality Traits

Figure 4 shows the personality traits over worker types. For *Openness*, we see marked differences across worker types (ANOVA,

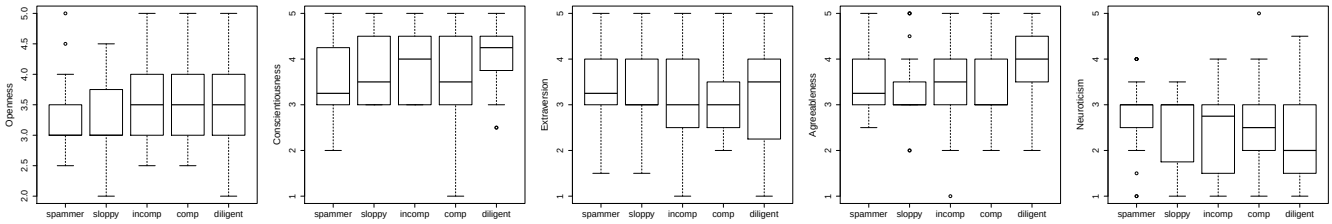

Figure 4: Personality traits over worker types: Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism

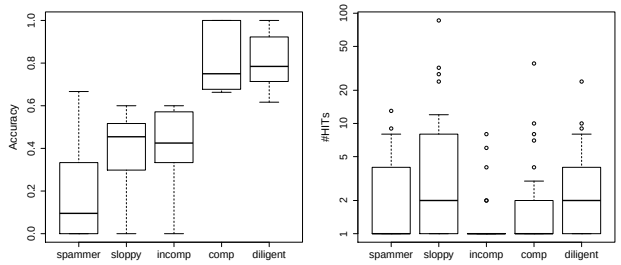

Figure 3: Accuracy and number of HITs by worker types

not significant), with diligent, competent, but also incompetent workers (who try but fail) having higher levels of openness. In terms of *Conscientiousness*, we see higher levels for the competent and diligent types, but the difference is not significant across types. Interestingly, among competent workers there is a relatively large difference between highly conscientiousness workers (with a score of >3.5), who achieve an average accuracy of 92% vs less conscientiousness workers who obtain an average of 74% accuracy. We find no correlation between *Extraversion* and worker types; this trait is not useful for distinguishing between desirable and undesirable workers. In terms of *Agreeableness* we see that faster workers (i.e., sloppy and competent) have noticeably lower scores than the others on this personality trait. Finally, for *Neuroticism*, we see lower levels for the competent and especially for the diligent workers, but the difference is not significant across types. We note that the distribution of BFI scores for spammers is markedly different from other types, with high peaks around the score of 3 for all five traits.

5. CONCLUSIONS

The main focus of this paper is the study of the crowd workers' characteristics, performed on the workers engaged in the crowdsourced evaluation of the INEX 2010 Book TrackâĂŹs *Prove It* task. We observed workers' behavioral patterns (number of HITs completed, average HIT completion time, fraction of useful labels contributed by a worker) and their personality profiles, based on the Big Five personality traits. We related these characteristics to the accuracy of the labels provided by the workers and found a strong correlation with HIT completion time, fraction of useful labels, and the Openness trait. Based on the behavioral patterns we defined five worker types (Spammer, Sloppy, Incompetent, Competent, Diligent) and demonstrated the benefits of such a segmentation for further analysis of the workers' data. By combining such behavioral typologies with the workers' personality traits, we can potentially identify the best workers for a given job and develop methods to attract them. Indeed, in our specific case, we showed that several workers' characteristics strongly correlate with the label quality. Such findings are useful not only for segmenting and analyzing the collected data but explaining how different HIT designs lead to different output quality by attracting different crowds.

REFERENCES

[1] O. Alonso and R. A. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *Proc. ECIR'11*, pages 153–164, 2011.

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

[3] B. Carterette and I. Soboroff. The effect of assessor error on ir system evaluation. In *Proc. SIGIR'10*, pages 539–546. ACM, 2010.

[4] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor. Are your participants gaming the system?: screening Mechanical Turk workers. In *Proc. CHI'10*, pages 2399–2402, 2010.

[5] S. D. Gosling, S. Gaddis, and S. Vazire. Personality impressions based on Facebook profiles. *Psychology*, pages 1–4, 2007.

[6] C. Grady and M. Lease. Crowdsourcing document relevance assessment with Mechanical Turk. In *Proc. CSLDAMT'10*, pages 172–179, 2010.

[7] J. Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Publishing Group, 2008.

[8] J.-H. Huang and Y.-C. Yang. The relationship between personality traits and online shopping motivations. *Social Behavior and Personality*, 38:673–680, 2010.

[9] P. G. Ipeirotis. Analyzing the Amazon Mechanical Turk marketplace. *XRDS*, 17:16–21, 2010.

[10] O. P. John, L. P. Naumann, and C. J. Soto. Paradigm shift to the integrative big-five trait taxonomy. In *Handbook of personality*, chapter 4, pages 114–212. Guilford Press, New York NY, 2008.

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

[12] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling. Crowdsourcing for book search evaluation: impact of HIT design on comparative system ranking. In *Proc. SIGIR'11*, pages 205–214, 2011.

[13] G. Kazai, M. Koolen, J. Kamps, A. Doucet, and M. Landoni. Overview of the INEX 2010 book track: Scaling up the evaluation using crowdsourcing. In *Proc. INEX'10*, pages 101–120, 2011.

[14] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *Proc. CHI'08*, CHI '08, pages 453–456, 2008.

[15] M. Kosinski, F. Radlinski, and P. Kohli. Personality and online behavior. In *Proc. CIKM'11*, 2011. ACM.

[16] J. Le, A. Edmonds, V. Hester, and L. Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *Proc. CSE'10*, pages 21–26, 2010.

[17] B. Rammstedt and O. P. John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41:203–212, 2007.

[18] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: shifting demographics in Mechanical Turk. In *Proc. CHI 2010*, pages 2863–2872. ACM, 2010.

[19] 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 *Proc. EMNLP'08*, pages 254–263, 2008.

[20] J. Vuurens, A. P. de Vries, and C. Eickhoff. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, pages 21–26, 2011. ACM.

[21] D. Zhu and B. Carterette. An analysis of assessor behavior in crowdsourced preference judgments. In *Proc. CSE'10*, pages 17–20, 2010.