The Face of Quality in Crowdsourcing Relevance Labels: Demographics, Personality and Labeling Accuracy

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

Information retrieval systems require human contributed relevance labels for their training and evaluation. Increasingly such labels are collected under the anonymous, uncontrolled conditions of crowdsourcing, leading to varied output quality. While a range of quality assurance and control techniques have now been developed to reduce noise during or after task completion, little is known about the workers themselves and possible relationships between workers' characteristics and the quality of their work. In this paper, we ask how do the relatively well or poorly-performing crowds, working under specific task conditions, actually look like in terms of worker characteristics, such as demographics or personality traits. Our findings show that the face of a crowd is in fact indicative of the quality of their work.

Categories and Subject Descriptors: H.4.m [**Information Systems Applications**]: Miscellaneous

General Terms: Design, Experimentation, Human Factors

1. INTRODUCTION

The recent industrialization of online crowdsourcing, popularized by commercial crowdsourcing platforms, e.g., CrowdFlower or Amazon's Mechanical Turk (AMT), is turning millions of Internet users into crowd workers. This enables to scale up tasks that require large scale human involvement and rapid data collection. One such task is the gathering of relevance labels for the evaluation of search engines, where crowdsourcing offers a feasible alternative to employing editorial judges [1, 5]

However, the output of crowdsourcing, especially in the case of micro-tasks and when monetary incentives are involved, often suffers from low quality. This is typically attributed to worker error, dishonest behaviour, problems with the design of the human intelligence task (HIT), or mismatch in the incentives or in the task requirements and the workers' abilities [4, 12, 18, 21]. To address this issue, a range of quality assurance and control techniques have been proposed in the literature, aiming to reduce deliberate or unintended worker error during or after task completion. Common to these methods is that they treat workers equally, without differentiating between workers from different backgrounds. However, there is already evidence in the literature that workers from different countries or cultures work and behave differently [6, 15]. In this poster, *we explore the relationship between characteristics of the workers and the quality of their work*. In particular, we ask how do the relatively well or poorly-performing crowds, working under specific task conditions, actually look like in terms of demographics or personality traits. We take an empirical approach and run experiments, collecting relevance labels, using Amazon's crowdsourcing platform. More specifically, we devise HITs to capture demographics and personality information about the workers who perform the same relevance labeling task in one of two different conditions, differing in the richness of the employed quality control methods.

Our analysis of the characteristics of the workers reveal that diversity matters: Both the demographics and personality profiles of the workers are strongly linked to the resulting label quality. Among the observed factors, workers' location (region) has the greatest correlations with label quality.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

2. RELATED WORK

The evaluation of IR systems, based on the Cranfield method, requires the gathering of relevance labels, which is traditionally performed by trained editorial judges [19]. However, this is an expensive process that does not scale to meet today's demands due to the ever growing size and diversity of test corpora and the volume of queries for which to label documents. Crowdsourcing has proven particularly useful for labeling large data sets, such as acquiring relevance assessments for search results [1]. However, several papers raised concerns about the poor quality of crowdsourced data [17, 21]. Indeed, quality assurance has surfaced as a major challenge in crowdsourcing and instigated a number of research efforts to understand and remedy the problems, e.g., [7, 13, 17, 18]. Common approaches include the use of gold standard data [5, 18] and building in redundancy, i.e., ensure that the task is performed by multiple individuals [7]. While research and empirical explorations of these methods have led to useful guidelines for effective crowdsourcing, e.g., [4, 7, 11], it is not clear what sorts of crowds are better at providing the desirable outcomes. Is it possible that better workers share some common characteristics or do more diverse groups produce better labels?

Recent studies have shown that the same HIT design can have different effects across the worker population. For example, while the level of pay can affect both the quantity [13] and the quality of the work [10], it may influence different types of workers differently. Higher pay can encourage better work in some workers but it may also attract unethical workers motivated solely by the higher financial gain. Similarly, [3] found that US and non-US workers responded differently to information provided about the purpose of the task, as in the case of labeling cancer cells in images. Other work focused on observing the characteristics of the

engagement with the task, such as the rate of labeling documents, in order to identify behavioral patterns among crowd workers and classify them into several types [20, 21].We extend on these works by providing a side by side comparison of the characteristics of the crowds involved in a relevance labeling task, working in one of two HIT designs.

3. EXPERIMENT DESIGN

Our goal is to study the characteristics of groups of workers and the quality of their work. We experiment with two HIT designs of the same relevance labeling task that differ in the richness of the employed quality control methods:

- ▷ *Full design (FD)*: This design includes various quality controls, such as trap and qualification questions, and challenge-response tasks. We also restrict participation to workers who completed over 100 HITs at a 95+ % approval rate. We set pay at the rate of $0.50 per HIT.
- ▷ *Simple design (SD)*: This design is a stripped down version of FD with a reduced set of quality control mechanisms, e.g., no challenge-response tests or qualifications requirements, and no pre-filtering on workers. Pay is $0.25 per HIT.

Due to the weaker quality control in SD, we expect that the group of workers in this condition will demonstrate higher levels of undesired behavior, while FD is designed to deter such workers.

To characterize workers, we collect the following information as part of the HITs:

- ▷ Demographics: location (region), gender, age, education and book reading frequency. we also include an indicator of reading habits.
- ▷ Personality traits: measures of the five personality dimensions [8], aka. the 'Big Five', see Table 1.

The data used in our experiments consists of the online books, search topics, participants' official retrieval runs, and relevance judgments of the INEX 2010 Book Track [9]. We use the same 21 topics that were used in the official evaluation of the track and the relevance labels that were collected from the INEX participants (our gold set). This set includes 3,357 page level judgments, 169 judged pages per topic on average. For each topic, we selected a random set of 100 book pages from the pool of pages in the participants' runs, ensuring that each such set contains at least 10 relevant pages. We grouped 10 pages per HIT, using the same data set in both designs. We collected labels from 3 workers each in both FD and SD. As part of the HIT design in both FD and SD, we administered a questionnaire to collect worker information, see Figure 1. For the personality traits, we used a standard 10 question test [14].

To measure the quality of work, we calculate the ratio of correct labels *per worker* based on agreement with the gold set (Accuracy) and *aggregate these to groups of workers*, e.g., per design. We exclude the 6 workers from our analysis who contributed HITs to both the FD and SD batches, thus removing data for 15 assignments. For survey style questions, we take the most frequent response or mode over the different HITs per worker. For quantitative measures, e.g., Accuracy, we take the arithmetic mean.

4. CROWD DIVERSITY AND ACCURACY

In this section, we present the results of our analysis of the relationships between the workers' characteristics and the resulting task performance measured by the workers' labeling accuracy. We calculate statistics over 3 groups of workers, the 117 workers who completed FD HITs, the 146 workers in SD, and the total 263 workers. For each of the groups, we plot the distributions of worker accuracy scores per demographic or personality category and test the correlation between these.

Figure 1: Survey administered to gather information on worker characteristics (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)

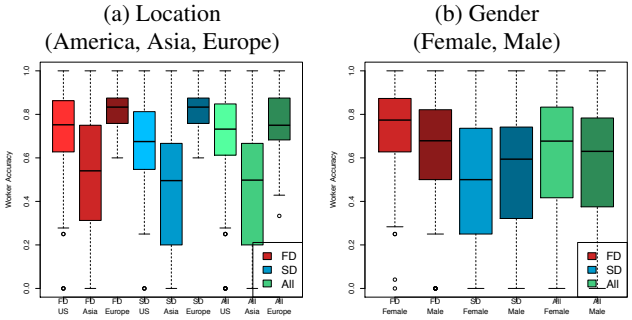

Figure 2: Worker accuracy over location and gender demographics in FD, SD and All HITs

4.1 Demographics

Among the demographic data, we see that location has a strong relation to accuracy over FD, SD and All data (one way ANOVA $p < 0.001$), see Figure 2a. A reason for the significant drop in Asian workers' performance may be that they lacked the necessary language skills for the task. Such a lack of understanding of the task or of the documents that needed to be labeled led to increased effort (well-performing Asian workers spent significantly longer time on the HITs than others) on the one hand and increased cheating or sloppy behavior on the other. Looking at the distribution of workers by location, see Figure 4a, we see that the majority of workers who contributed to FD HITs were from America (64%, 61% of which in the US), SD HITs were completed mostly by workers in Asia (61%, 53% of which in India). This difference could possibly be a result of the applied pre-filtering in FD, which would then suggest that most Asian workers do not have sufficiently high AMT reputation scores. However, regardless of the reason, based only on the differences in the accuracy (Figure 2a), similarly to [3, 16], we can conclude that Asian workers produced in our task lower quality work than American or European workers.

Gender shows no significant relation to accuracy over All data, but it is significant for SD and FD (ANOVA $p < 0.05$ for both), see Figure 2b. Interestingly, the relation is reversed for the designs: for FD, female workers outperform male workers, whereas for SD, male workers outperform female workers. Looking at the differ-

Table 1: Big Five personality traits, *OCEAN*

Trait	Description - High Score	Description - Low Score
Openness	Imaginative, independent, interested in variety	Practical, conforming, interested in routine
Conscientiousness	Self-disciplined, dutiful, organized, careful	Disorganized, careless, impulsive
Extraversion	Sociable, fun-loving, affectionate	Retiring, somber, reserved
Agreeableness	Compassionate, cooperative, trusting, helpful	Self-interested, suspicious, antagonistic, uncooperative
Neuroticism	Anxious, insecure, emotional instable, self-pitying	Calm, secure, self-satisfied

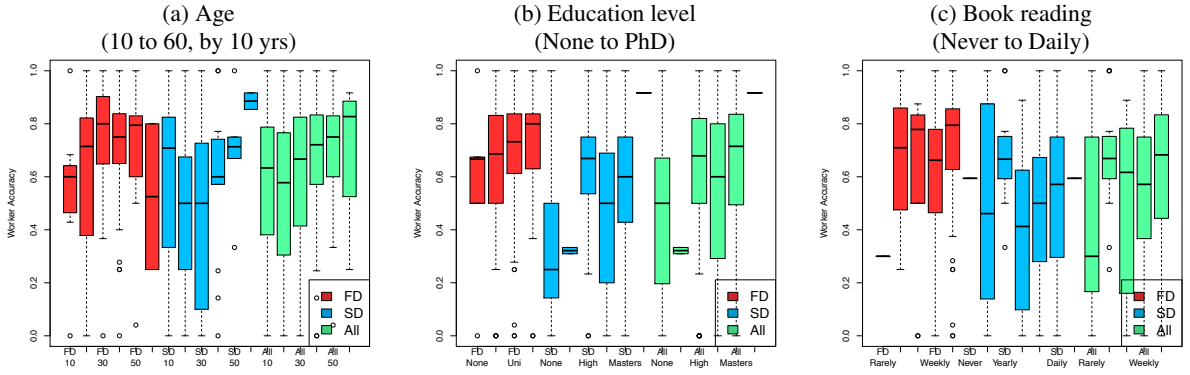

Figure 3: Worker accuracy over demographics in FD (red), SD (blue) and All HITs (green)

ences in gender distribution between FD and SD, Figure 4b, we see that more male workers contributed to SD HITs and more females to the FD HITs. These differences correlate with the regional distribution of workers: workers from India tend to be males (58%), while workers in the US are mostly female (59%, although 22% of US workers withheld this information) and are similar to those reported in [6, 15].

Age also has a significant relation with accuracy over All data (Spearman r = 0.17, $p < 0.05$), see Figure 3a. The relation is similar for FD (r = 0.16, not significant) but slightly negative for SD (r = -0.08, not significant). As we can see in Figure 4(c), SD workers are somewhat younger (average age of 26) than FD workers (average age is 31), and this difference resulted in a considerable impact on accuracy. Unlike the workers' age, and contrary to expectation, we find that education level is not significantly related to accuracy (Figure 3b), despite the correlation between age and education, and the expectation that more educated (skilled) workers would be better placed to tackle this high cognitive task [2]. Answers to the education level question, see Figure 4d, reveal that a high portion of SD workers are university students (61%, compared with 44% in FD), which corresponds with the reported age distributions. On the other hand, reading habits are significantly related to accuracy over All data (Spearman r = -0.19, $p < 0.01$), indicating that more frequent readers are more accurate, see Figure 3c. This suggests that reading habits are better predictors of accuracy than education. The relation is also significant for FD (r = -0.20, $p < 0.05$). There is a weak but insignificant relation for SD, which, given that SD workers claimed to be relatively more frequent readers, see Figure 4(c), may suggest a tendency of SD workers to give socially desirable responses.

4.2 Personality Types

Among the personality traits, both openness and conscientiousness significantly relate to accuracy over All data (Spearman r = 0.24, $p < 0.001$ and r = 0.21, $p < 0.05$, resp.) and over FD (r = 0.22, $p < 0.05$ and r = 0.24, $p < 0.05$) but not over SD (r = 0.23 and r = 0.34), see Figure 5. Agreeableness has a weak (not significant relation with accuracy over All data (r=0.13, p=0.06). Interestingly, for SD none of the BFI traits are related to accuracy. This suggests that

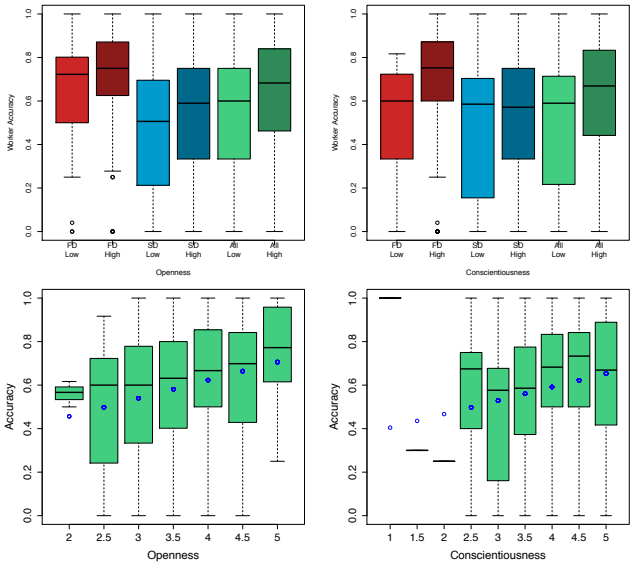

Figure 5: Accuracy over personality: openness (left), conscientiousness (right), top row shows accuracy for High (>3) and Low (<3.5) BFI values per FD, SD and All, bottom row shows calculated BFI values for All

personality characteristics are not very effective at distinguishing low quality workers but are useful to distinguish between good and better workers.

Figure 6 shows the distribution of responses to the BFI questionnaire in FD, SD and All. Again, we can observe clear differences between the workers in the FD and SD sets. FD workers are more open to new experiences while SD workers are more confirming to routine (low openness scores). FD workers are also more conscientious than workers in the SD group. SD workers are more extravert, self-interested (agreeableness), and more insecure (neuroticism) than FD workers (not significant).

Given the overall high quality of work by workers in FD, we may classify desirable workers, at least for a high cognitive relevance

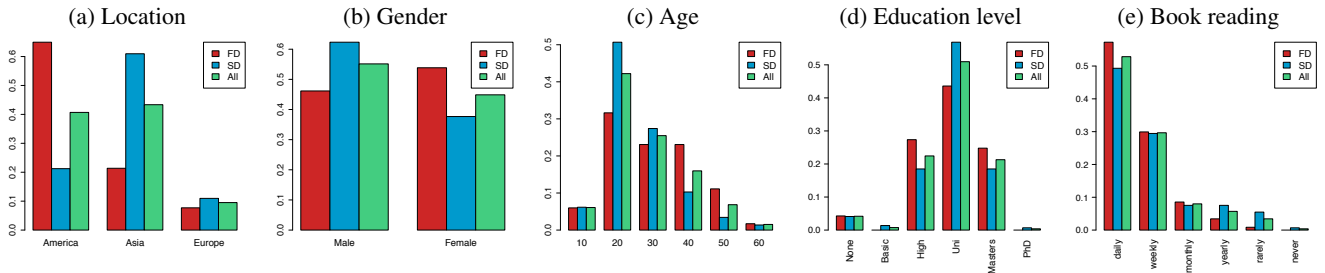

Figure 4: Distribution of responses to demographic questions in FD (red), SD (blue), and All HITs (green)

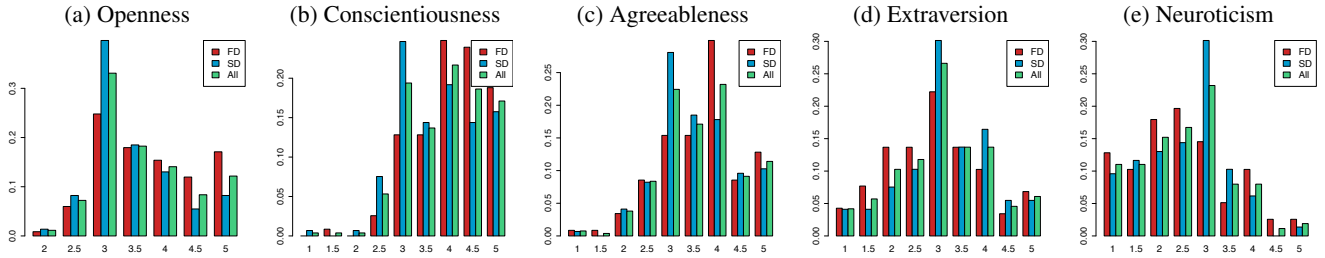

Figure 6: Average scores on the 'Big Five' personality traits for FD (red), SD (blue), and All HITs (green)

assessing task, as open and conscientious middle-aged females in the US, who read books on a regular basis.

5. CONCLUSIONS

Crowdsourcing is characterized by its large and anonymous workforce, yet the individual competencies of the workers are crucially determining the quality of work. This paper explored the relationship between certain characteristics of the workers and the quality of their work. Our main findings are as follows. First, in terms of demographics, we found that location has a very strong relation to accuracy, with Asian workers demonstrating significantly poorer performance than American or European workers. Interestingly, we also found a clear separation of the crowds by location across the two designs, with mostly Asian workers in SD and American workers in FD. Although the HITs were not restricted to geographical regions, the exact HIT conditions impacted on the geographical distribution of the workers who completed the HITs, which in part then explain the differences in the quality of work. Second, in terms of personality types, we found that openness and conscientiousness relate significantly to accuracy. Again, we saw that workers with notably different personality traits worked under the different HIT conditions. Our overall conclusion is that there is a complex interaction between the particular HIT conditions and the types of workers who engage in a task, and that these worker characteristics are related to the quality of their work. The face of a crowd is in fact indicative of the quality of their work and should be considered when designing crowdsourcing systems.

REFERENCES

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

[2] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *Proc. of SIGIR 2008*, pages 667–674, 2008.

[3] D. Chandler and A. Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. Technical report, Working paper, University of Chicago, 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. of CHI 2010*, pages 2399–2402. ACM, 2010.

[5] C. Grady and M. Lease. Crowdsourcing document relevance assessment with mechanical turk. In *Proc. of CSLDAMT'10*, pages 172–179, 2010.

[6] P. G. Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS*, 17:16–21, 2010.

[7] P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proc. of HCOMP'10*, pages 64–67, 2010. ACM.

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

[9] G. Kazai, M. Koolen, A. Doucet, and M. Landoni. Overview of the INEX 2010 book track: At the mercy of crowdsourcing. In *INEX 2010 Workshop Pre-proceedings*, pages 89–99, 2010.

[10] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling. Crowdsourcing for book search evaluation: Impact of quality on comparative system ranking. In *Proc. of SIGIR 2011*. ACM, 2011.

[11] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proc. of CHI 2008*, pages 453–456, 2008. ACM.

[12] J. Le, A. Edmonds, V. Hester, and L. Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, pages 21–26, 2010.

[13] W. Mason and D. J. Watts. Financial incentives and the "performance of crowds". In *Proc. of HCOMP'09*, pages 77–85, 2009. ACM.

[14] 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.

[15] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proc. of CHI 2010, Extended Abstracts Volume*, pages 2863–2872. ACM, 2010.

[16] A. Shaw, J. Horton, and D. Chen. Designing incentives for inexpert human raters. In *Proc. of CSCW'11*, 2011.

[17] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proc. of KDD'08*, pages 614–622, 2008.

[18] 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. of the EMNLP'08*, pages 254–263, 2008.

[19] E. M. Voorhees and D. K. Harman, editors. *TREC: Experimentation and Evaluation in Information Retrieval*. MIT Press, 2005.

[20] J. Vuurens, A. de Vries, and C. Eickhoff. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proc. of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, 2011. ACM.

[21] D. Zhu and B. Carterette. An analysis of assessor behavior in crowdsourced preference judgments. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, pages 17–20, 2010.