Crowd Worker Strategies in Relevance Judgment Tasks

Lei Han+, Eddy Maddalena‡, Alessandro Checco‡, Cristina Sarasua*, Ujwal Gadiraju*, Kevin Roitero*, and Gianluca Demartini+

+University of Queensland, Australia. ‡University of Southampton, UK. ∗University of Sheffield, UK. ∗University of Zurich, Switzerland. +L3S Research Center, Germany. △University of Udine, Italy.

ABSTRACT
Crowdsourcing is a popular technique to collect large amounts of human-generated labels, such as relevance judgments used to create information retrieval (IR) evaluation collections. Previous research has shown how collecting high quality labels from a crowdsourcing platform can be challenging. Existing quality assurance techniques focus on answer aggregation or on the use of gold questions where ground-truth data allows to check for the quality of the responses.

In this paper, we present qualitative and quantitative results, revealing how different crowd workers adopt different work strategies to complete relevance judgment tasks efficiently and their consequent impact on quality. We delve into the techniques and tools that highly experienced crowd workers use to be more efficient in completing crowdsourcing micro-tasks. To this end, we use both qualitative results from worker interviews and surveys, as well as the results of a data-driven study of behavioral log data (i.e., clicks, keystrokes and keyboard shortcuts) collected from crowd workers performing relevance judgment tasks. Our results highlight the presence of frequently used shortcut patterns that can speed-up task completion, thus increasing the hourly wage of efficient workers. We observe how crowd work experiences result in different types of working strategies, productivity levels, quality and diversity of the crowdsourced judgments.

KEYWORDS
Crowdsourcing, IR Evaluation, Relevance Judgment, User Behavior

1 INTRODUCTION
Microtask crowdsourcing has provided unprecedented opportunities to readily acquire human input at scale [5]. It is now possible to build large-scale ground-truth data to train machine learning models, evaluate system effectiveness, and harness the “wisdom of crowds” for a variety of applications [22]. One such popular application can be found in the field of Information Retrieval (IR), where the subjective task of acquiring document relevance judgments is assigned to crowd workers in the form of human intelligence tasks (HITs) [1, 3]. However, it is unclear how the crowd sample involved in these tasks has an effect on the quality of the data being collected. More specifically, how crowd worker experience can have an impact on the collected data has been unexplored. In this paper, we address this knowledge gap by looking at the strategies adopted across different worker populations (e.g., experienced and less-experienced workers) to understand how crowdsourced collections can be affected by worker experience and their strategies.

Although several works have explored various aspects of crowdsourcing task and workflow design for the problem of relevance judgments, few works have studied human factors and individual differences between workers in terms of the strategies they adopt to complete such HITs. There has been sufficient evidence to indicate that different workers complete tasks differently. In a relevance judgment task, for example, Kazai et al. used behavioral observations such as HIT completion time, and the accuracy of workers to define different worker types [19]. Identifying and characterizing these differences can enable us to route tasks to suitable workers, thereby improving quality outcomes and managing costs – as shown in the task of information finding on the Web [9]. Rzeszotarski and Kittur proposed to rely on the differences in how workers completed tasks rather than explicitly assessing the quality of their work on gold-standard questions to separate good workers from the crowd [30, 31]. Finally, worker demographic attributes [20, 29], varying monetary rewards [18, 28], as well as the constant influx of new inexperienced workers in the crowd [6, 7] can lead to differences in marketplace dynamics and thereby influence the methods or strategies adopted by workers to complete tasks.

Prior works have established worker differences both at a low-level by using worker behavioral traces and at the high-level by assessing quality of work produced. However, there is a lack of understanding on how these differences manifest [10]. Marketplace dynamics are often influenced greatly by large batches of similar tasks [7] which recur over time, providing an opportunity to workers to learn [11] and adapt to effective working strategies. Even within one task, there exist positive learning effects that enable workers to develop strategies to do the task faster and better as they progress over steps [13]. Does worker experience influence how the tasks are being performed, and in turn the quality of crowdsourced outcomes? To understand their strategies and how IR evaluation
collected by means of crowdsourcing can be affected, we make original contributions to the following research questions.

(RQ#1) **What is the role of crowd worker experience in shaping worker strategies to complete tasks?** Through a qualitative study comprising interviews of ten expert crowd workers, we found that over time workers acquire skills that make them more efficient in selecting and completing tasks effectively. In addition, through a survey addressing 100 experienced workers we found that the use of keyboard shortcuts is very common.

(RQ#2) **What are the strategies that experienced workers have developed to complete tasks efficiently?** By a data-driven analysis of behavioral logs collected from workers completing relevance judgment tasks, we found that copy/paste action is very popular among crowd workers. Compared to less-experienced workers, experienced workers write less typed text and copy more previously generated content to speed-up task completion. This indicates the existence of task completion strategies involving the reuse and adaptation of textual content created previously.

(RQ#3) **How can worker experience affect the quality of the crowdsourced results?** Based on the outcomes of a large scale relevance judgment experiment, we observed that the judgments given by less-experienced workers showed a comparable quality to those provided by experienced workers.

2 RELATED WORK

2.1 Human Factors in Relevance Judgments
The task of performing relevance judgment is critical to generate high quality IR evaluation collection. With the rising popularity of crowdsourced relevance judgments, researchers started to look at how to deal with the quality challenges intrinsic to such an approach. Alonso and Mizzaro [2] looked at how crowdsourced relevance judgments compare to editorial judgments produced by TREC assessors showing good correlation in IR evaluation outcomes. Maddalena et al. [23] looked at the impact of limiting the time available for the judgment task showing how introducing a reasonable time limit may even increase the quality of the collected judgments. McDonnell et al. [26] showed how asking for a justification more than just a judgment of relevance intrinsically makes crowd workers think more about the judgment and improves its quality. Recently, Goyal et al. [12] showed how behavioral traces in relevance judgment tasks can be used to predict judgment quality.

As compared to this body of previous work, in this paper we aim at understanding different crowdsourcing task completion strategies rather than predicting quality or filtering workers. Our goal is to understand how experienced crowd workers complete tasks more efficiently as compared to other workers.

2.2 Participation Bias in Crowd Work
While crowd work is a very individual activity, workers have self-organized into online communities where they discuss their experiences and share HITs to work on [39]. Workers share information and discuss about a variety of topics including HITs, requesters, earning, and scripts/tools to support their crowd work. Sharing HITs on forums has shown to be affecting crowdsourcing market dynamics with the HIT batch throughput significantly increasing after a task is shared on forums [38]. This creates a participation bias in the type of workers who complete the batch. Difallah et al. [6] showed how the propensity of participation in certain HITs may create biases in the type of workers who participate in a study, for e.g., biases on workers from certain countries. Hube et al. showed how workers’ opinion biases can influence their judgments [14].

Related to this, we look at which strategies help workers complete tasks more efficiently and thus possibly creating a participation bias by completing more tasks than other workers.

2.3 Tools and Scripts to Support Crowd Work
Efficient crowd work is often supported by tools and scripts (e.g., browser extensions and plug-ins). One of the most popular tools for crowd workers to select tasks to work on is TurkOpticon [15], which provides peer-reviewed requester reputation scores to inform workers about prior experiences with a requester. An overview of popular scripts and extensions used by MTurk workers is presented by [8]. Such scripts can provide significant reduction in time needed to search for tasks, for example, by notifying workers of new tasks being published by their favorite requesters. They can also provide functionalities to monitor earnings and requester reputation metadata that, again, supports their task choice. Most of such scripts are built for the MTurk platform and its functioning and thus do not focus supporting work on specific HITs. El Maarry et al. [8] also looked at the concept of Super Turkers showing high correlation between engagement with external crowd worker forums and longevity on the crowdsourcing platform. It has also been shown that workers who earn a lot of money on MTurk use significantly more extensions and scripts [17].

In our work we look at which task-level work strategies crowd workers use to be more efficient in completing tasks rather than looking at platform-level dimensions like requester reputation and task selection biases. Our work still addresses the general issue of population bias in crowdsourcing platforms as highly efficient workers have more work opportunity than less efficient workers who can perform fewer HITs before a HIT batch gets completed.

3 EXPERIENCED CROWD WORKERS

First, we report on two studies aimed at understanding the behaviors of experienced workers on crowdsourcing platforms. We run (i) focused interviews with crowd workers recruited through worker forums and (ii) a survey run on a crowdsourcing platform targeted to experienced workers.

3.1 A Qualitative Study to Understand Experienced Crowd Workers
With an aim to advance our understanding of how experienced crowd workers complete work on microtask crowdsourcing platforms, we recruited and interviewed ten workers chosen among the top influencers in two of the main MTurk worker forums: reddit.com/r/mturk and Turkernation. In order to find the top influencers, we considered the cumulative amount of received feedback in the forum as the main metric. In this way we not only consider the most prolific worker in terms of amount of work completed, but also the ones that have had a significant positive response from the community. The interviews were semi-structured, with questions corresponding to workers’ experience, challenges faced, and
strategies they typically adopt to complete work. In this section, we report the main observations from a thematic analysis of responses collected in this study. The interviewees described the main skills gained as their experience grew in the crowdsourcing platforms:

**Ability to recognize attention checks.** Often attention checks are repeated among different tasks [25], some questions of personality or logic have become infamous in the crowdsourcing community for how often they appear. A few excerpts from workers referring to this are quoted here: *W9—"questions like I’m having a heart attack right now, it’s your fault"*. *W10—"I’ve done enough of those that I don’t even read the whole paragraph anymore [...] they need to just to stop recycling those! W1—I have no issue with test questions [...] I notice them. Everyone writes the same test questions. W7—Yeah they’re pretty obvious to me, and most of the experienced workers."* Workers also learn to look for test questions in common hidden places: “we’ll have requesters hide something in the consent form.”

**Ability to recognize gold questions.** With experience, workers learn to recognize subtle differences between gold set and real data: *W8—"[requesters] have like an idealized version of the world that’s not quite as sloppy as real world data. And so, when you see their examples, a lot of times they’ll use, if it’s an image HIT, for example, they might use an image that’s basically one from the examples.”*

**Ability to estimate hourly wage.** Workers learn to estimate the time required to complete a task, thus increasing the ability to estimate the hourly wage: *W2—"I’ll accept it, look at it, and then spend 20 seconds and if it’s like, yeah that HIT that said it’s going to take ten minutes and it pays ten cents, sometimes HITs that say they’re going to take ten minutes were clearly copy paste, and it’s actually going to take ten seconds.”*

**Scripts to speed up some common tasks.** Workers learn to use keyboard scripts to: (i) reduce strain on or speed up repetitive tasks: *W1—" [...] rebinds some mouse clicks to one and two so that I can avoid carpal tunnel ". W8—" [...] tasks that I like to do is just simple one response to three response batch work. It’s just, if something looks like it can be just answered quickly with a script.", W7—“I’ve learned how to use my computer to the most efficient way. I’ve learnt how to do a little programming, I can write, easy scripts”; (ii) automate web retrieving tasks: *W8—" [...] a lot of requesters will request basic web scraping tasks [...] that type of stuff is totally scriptable, all you do is just verify, look at it, you know you just push one button as long as everything’s okay.”*

**Scripts to be faster in the recruiting phase.** This is considered one of the main problem of crowdsourcing platforms. Over time, users learn to use scripts to assist them in the task selection phase.

**Use of forums to improve recruiting quality.** Workers monitor forums to find reviews of tasks and requesters, to inform them in their task selection and avoid bad quality tasks.

**Ability to estimate the risk of rejection.** Workers learn to estimate the risk of rejection in a task based on “gut feelings”, requester reputation, and batch size. *W10—"I’m not going to take a chance on a rejection.”*

---

1. A primary crowdsourcing platform, https://www.figure-eight.com/
can perform) than workers who start by reading the instructions and examples provided by requesters (25 out of 78) (see Fig. 2g).

When accomplishing work (see Fig. 2h), 32 out of 84 workers mentioned that they are careful and dedicate time to understand instructions properly and curate their answers, while 16 out of 84 responded that they try to accomplish tasks as quickly as possible. Moreover, 27 out of 84 workers said they use shortcuts while working on tasks.

From the question related to highlighting issues and reporting their experience in crowd work (see Fig. 2i), we found that 24 workers (out of 48) do not tend to provide feedback on the work they complete. Off those who do, the majority of workers use direct communication that may be specific to crowd work (e.g., ticketing system / email to the platform’s support team, or the field in task asking for general feedback about the task) or general-purpose channels that people already use to communicate with their peers (e.g., Whatsapp, social media). Only 5 mentioned that they discuss their work in forums and 1 said (s)he relies on other ways to report tasks. We observed that all the answers providing details on reporting actions referred to documenting errors and negative properties of tasks, instead of outstanding task designs or positive experience.

### 3.3 Discussion

From the previous two studies (i.e., focused interviews with forum leaders and survey with crowdsourcing workers) we have observed different crowd work strategies to efficiently identify and complete tasks including the use of shortcuts and skills in identifying test questions and attention checks in tasks.

Given the frequent use of shortcuts revealed in these studies and the differences that we have observed between users who have a clear strategy (e.g., first assess task completion time/effort and only then proceed with task completion) and users who just proceed with reading the instructions and complete the task, we decided to closely analyze the behavioral patterns of workers in a data-driven study presented in Section 4, with the comparison of experienced workers with their less experienced counterparts.
4 A DATA-DRIVEN ANALYSIS ON TASK COMPLETION STRATEGIES

In this section we present the results of a study aiming at comparing behavioral differences between experienced and less-experienced crowd workers. We use a document relevance judgment task where we log mouse clicks and keystrokes data to observe commonly exhibited behavioral patterns and to infer task completion strategies.

4.1 Task Design

To spot out behavioral differences between experienced and less-experienced crowd workers, we ran a relevance assessment task, in which workers were asked to judge the relevance of a series of four documents with respect to a given topic. Documents were shown sequentially, and for each of them, workers were required to provide: (i) a relevance score on a four-level scale (Not Relevant, Partially Relevant, Relevant, and Highly Relevant), (ii) a free text justification of the given assessment, and (iii) a summary of the document content (max. 300 characters). To study workers’ interaction in depth, we embedded a logger in our task that keeps track of the meaningful high-level events taking place during task completion. These include navigation between documents, or the answer and assessment provided. In addition, we tracked important low-level events identified during the studies presented in Section 3, including: (i) hot-keys combinations involving CTRL and CMD keys, (ii) editing of text (in justification and summary fields), (iii) events of find, cut and copy/paste, and (iv) text selections. We deployed the task on FigureEight addressing 300 HITs, divided in six equal-cardinality groups, each one assigned to a different topic of the TREC8 collection [36]. Each worker was rewarded $0.30 and allowed to complete each topic only once. To create a balanced population of workers among different levels of experience, we selected workers belonging to each one of the four crowdsourcing platform levels proportionately (from unleveled to Level 3), assuming a correlation between worker levels and experience. Before starting the relevance assessment activity, workers were required to copy their platform work statistics from a frame showing a portion of their personal dashboard and paste it into a text area where a regular expression checked the conformity of the pasted string. For each worker, such text provided us the overall number of test questions answered so far in the platform, which allows us to profile the worker’s experience to generate the break-down analysis reported in the following. We use the number of test questions answered as far as a proxy for worker experience since this number is directly related to one’s accuracy rate in the platform, and thus the workers are likely to be careful doing the tasks involving test questions. We embedded two quality checks in our task, consisting of: (i) an initial test question about the topic to ensure the worker read and understood it, and (ii) a time-based check to ensure the worker spent at least 20 seconds in at least two of the four documents (as a minimum expected threshold for a trustworthy worker).

4.2 Results

Overall, we collected 1200 relevance judgments submitted by 154 distinct workers. In order to divide these workers in experienced and less-experienced groups and to keep a balanced population in each group, we set a threshold of 5000 completed test questions to split the whole population into two groups. Table 1 presents the threshold by means of the number of answered test questions to classify worker experience and the number of workers in each group (i.e., 79 less-experienced and 75 experienced workers). In order to investigate the impact of worker experience on the quality of their submitted judgments (see Section 4.2.4), we further divide each group into two sub-groups: lower and upper sub-groups, by the threshold of 300 and 20,000 answered test questions, respectively.

From these workers, we obtained approximately 80,000 behavioral log entries. In the analysis we removed the “copy/paste” actions that we required them to perform to report their work experience in the platform and only considered those actions performed on the relevance judgment work. In the following, we first report on the actions performed by all workers, and then present the results with a break-down over worker experience.

4.2.1 Action and Shortcut Popularity. First, we report on the overall distribution of shortcut use in our experiment. Table 2 shows the overall number and proportion of workers who have performed a certain action as well as the overall observed action frequency in the experiment. In the table, copy (and paste, respectively) refers to all possible ways to copy some selected text in the task (e.g., shortcuts, menu bars, contextual menus, etc.) while “Ctrl (Cmd) + C” refers to the copy action using a keyboard shortcut and it is then a subset of the copy observations.

We can see that most of the participating workers use copy/paste actions to speed-up their work. We can also note that while copy and paste are common actions, only about half of those who perform such actions do it using keyboard shortcuts (i.e., “Ctrl (Cmd) + C” and “Ctrl (Cmd) + V”) while the other half perform them using other ways such as contextual menus triggered by mouse clicks. Another observation that can be made is the significantly lower popularity of the “Ctrl (Cmd) + F” shortcut that allows to search and highlight certain keywords in the task (e.g., the topic title in the document to

---

2The Figure Eight platform levels are defined as having completed at least 100 jobs and having obtained an accuracy over the test questions of at least 70% (Level 1), 80% (Level 2), or 85% (Level 3).

3The accuracy rate in the platform is computed as the number of correctly answered test questions divided by the total number of test questions attempted so far.

---

```markdown
Table 1: Classification of workers by experience.

<table>
<thead>
<tr>
<th>#Test Questions Answered</th>
<th>0-300</th>
<th>301-5000</th>
<th>5001-20000</th>
<th>&gt; 20000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Groups</td>
<td>lower</td>
<td>upper</td>
<td>lower</td>
<td>upper</td>
</tr>
<tr>
<td>#Workers</td>
<td>39</td>
<td>40</td>
<td>37</td>
<td>58</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Action</th>
<th>Number of Workers (and ratio)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>copy</td>
<td>131 (85%)</td>
<td>129</td>
</tr>
<tr>
<td>Ctrl (Cmd) + C</td>
<td>57 (37%)</td>
<td>596</td>
</tr>
<tr>
<td>paste</td>
<td>136 (88%)</td>
<td>1379</td>
</tr>
<tr>
<td>Ctrl (Cmd) + V</td>
<td>77 (50%)</td>
<td>776</td>
</tr>
<tr>
<td>Ctrl (Cmd) + F</td>
<td>39 (25%)</td>
<td>160</td>
</tr>
<tr>
<td>text selection</td>
<td>142 (92%)</td>
<td>2486</td>
</tr>
<tr>
<td>text change</td>
<td>152 (99%)</td>
<td>62056</td>
</tr>
</tbody>
</table>
```
be assessed), in comparison to copy/paste shortcuts (Mann-Whitney \( U \) test \[24\] revealed statistical significance; \( p < 0.05 \)). Selecting text has been a dominant action in this experiment and it is even more prevalent than copy/paste actions. Some workers may select text as a means of support while reading digital content even if they do not intend to copy/paste it \[32\].

Next, we look at where crowd workers copy text from. Figure 3 shows the percentage of text copying actions over the 6 topics split by the target area where the text has been copied from. We see that the two largest sources of text where crowd workers perform copy actions are documents and topics, since this content is displayed to them as per our task design, while justifications and summaries are constructed by workers themselves. More specifically, most of the copy actions are done over the content of the document to be judged (e.g., to provide a justification for the judgment done on the document or to construct the required document summary). Copying text from the topic description is also a common strategy. This can be useful to workers who want to highlight topic/query keywords in the document to assess its relevance faster. Less common but still present is the behavior of copying text from the justification and summary areas which may indicate willingness to reuse/adapt it for future assessments. We can observe how textual content is being reused and possibly adapted for other parts of the task and other documents being judged in the task. Such strategy is clearly a step towards more efficient crowd work.

4.2.2 Time to Complete Tasks. We first present the analysis of the time that the workers have spent in completing tasks with a breakdown over worker experience (i.e., 79 less-experienced and 75 experienced workers). Since there are four documents to judge in each HIT, we define judgment on each document as a step. Therefore, each step starts from presenting the document in the current task page and ends when the worker clicks “Next” button. Moreover, we further divide each step into two phases: (i) taking a judgment decision, and (ii) writing the judgment justification and document summary. This allows us to distinguish the behavioral differences exhibited by less-experienced workers in reading documents and writing text, compared to experienced ones.

Figure 4 shows the time spent in taking a decision \( U \) test to examine the differences in shortcut popularity, time spent to complete tasks (see Section 4.2.2) and the length of text that workers provide (see Section 4.2.3), because the distribution of the data is not interval scaled.

4.2.3 Text over Steps and Content Reuse. In order to understand how experienced workers differ from less-experienced ones in writing behaviors, we analyze the judgment justification text and document summaries provided by all workers from Step 1 to 4. Given the result that most of the copied text is from documents and topics (see Fig. 3), we measure the similarity of the text provided in each HTML text area (i.e., justification and summary) to the content in the document and topic. We use the Ratcliff-Obershew similarity metric \[27\] to match the longest common sub-strings, and define the similarity as the proportion of matched characters to the total length of the text provided by the worker. Thus, according to this definition, similarity scores go from 0 (nothing copied) to 1 (completely copied). Based on these scores, we assume workers have copied text from the source area (i.e., document or topic) bearing the highest similarity value with the submitted justifications. As the document summary field appears after the judgment justification field, while writing summaries workers may additionally copy content from the justification they have entered. Thus, for document summaries we consider highest similarity with document, topic and justification to the summary to detect the copied content source.
Figure 5a shows the length (measured in number of characters) of newly entered text by all workers in addition to the content resulting from “copy/paste” actions at each step in task. We can observe that, as compared to experienced workers, less-experienced workers submit content by typing significantly (Mann-Whitney U p < 0.05) more new text in addition to what they have copied and pasted from the task. When comparing the length of copied text between the two groups of workers, we do not observe a statistically significant difference. Therefore, we can conclude that due to typing longer text for justifications and summaries, less-experienced workers are slower than experienced ones in providing text content, which has an impact on their overall task completion time.

Other than writing shorter new text, we look at other factors contributing to the experienced workers’ faster task completion time. Since there are two pieces of mandatory input text (i.e., justification and summary) to complete for each document judgment, we further analyze the similarity between justification and summary at a given step, to understand how much content is reused in constructing these two text inputs. Figure 5b shows the similarity between justification and summary at each step. The results show that experienced workers reuse significantly (t-test p < 0.05 at Step 2 and 3) more content across the two text inputs as part of their strategies to reduce task completion time.

4.2.4 Quality and Diversity of Crowdsourced Relevance Judgments Collections. Finally, we present an analysis of judgments quality and of contribution diversity of the judgment collections, aiming to understand how crowd worker experience can affect IR system evaluation. To measure the quality of the judgments given by workers, we use Krippendorff’s Alpha [21] coefficient to measure the agreement with expert labels previously created using the same 4-level relevance scale [33]. The score goes from −1 (complete disagreement) through 0 (agreement equivalent to random guessing) to 1 (complete agreement). Since only a subset of the documents were pooled for expert judgment, we assume the unjudged documents to be not relevant. This is in line with the sampling strategy adopted when creating the expert collection which avoided to judge those documents already judged as not relevant by TREC assessors [33].

We measure the quality of workers from two perspectives: (i) the overall quality of the submitted judgments, and (ii) the quality at each step in the task. For a given step, we examine all the judgments from the beginning of the task up to this step, which allows us to understand the evolution of worker quality as they progress in task. Figure 6a presents the evolution of quality over steps and the quality of the final submitted judgments. We can observe that the quality of all workers (both experienced and less-experienced workers) increases as they progress in the HIT, showing that their judgments become better when they judge more documents. This result is consistent with previous work where all workers who submitted tasks showed a positive learning effect as they progress in a relevance judgment task [13]. On average, the submitted judgments display highest quality. This is because when the workers fail the quality checks they are allowed to revise their judgments, so their judgments become more comparable with those by experts. Yet, there is no significant difference between the quality of less-experienced and experienced workers at each step. With a further experience breakdown (i.e., the 4 levels of experience defined in Tab. 1), Figure 6b shows the quality of submitted judgments provided by workers in the four sub-groups. Although the judgment quality is positively correlated with work experience level, only the difference between the lower less-experienced group (i.e., novice workers) and the upper experienced group (i.e., highly experienced workers) is statistical significant (t-test p < 0.05).

To understand how worker experience can affect crowdsourced IR system evaluations, we look at the correlation between IR system rankings generated by crowd judgments and those contributed by experts. We aggregate the submitted judgments per workers in different groups, by taking median value for each document. Thus, we obtain a set of relevance judgments for each level of worker experience. In addition, we use two editorial judgment collections by

![Figure 5: (a) Length of newly entered text; (b) Similarity of document summary to justification over steps for both experienced and less-experienced workers.](image)

![Figure 6: (a) Quality over steps for both experienced and less-experienced workers; (b) Overall quality over different levels of work experience.](image)

<table>
<thead>
<tr>
<th>Crowd Judgments</th>
<th>Expert Judgments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worker Group</td>
<td>TREC8</td>
</tr>
<tr>
<td>Less-Experienced</td>
<td>lower</td>
</tr>
<tr>
<td></td>
<td>upper</td>
</tr>
<tr>
<td>Experienced</td>
<td>lower</td>
</tr>
<tr>
<td></td>
<td>upper</td>
</tr>
</tbody>
</table>

5The average α agreement of submitted judgments with experts is 0.75 for experienced (0.67 for less-experienced) workers, while it is 0.73 for experienced (0.66 for less-experienced) workers at Step 4.
TREC8 [36] and Sormunen [33] to create system rankings using expert judgments. By measuring IR system effectiveness using NDCG [16], we are able to compute the Kendall’s τ correlation coefficient to quantify the quality of crowd judgments from each group compared to those given by experts. Table 3 presents the results. We can observe that: (i) compared to less-experienced workers, judgments given by experienced workers generate more similar IR system rankings to that from expert judgments; (ii) highly experienced workers (upper experienced group) do not necessarily produce better judgments than the workers in the lower experienced group; (iii) there is little difference between the results obtained with judgments by workers in the lower and upper less-experienced groups. This is confirmed by the fact that the Wilcoxon signed-rank test shows no statistical difference over the six topics for each group.

These results imply that previous crowd work experience does not affect significantly the quality of crowdsourced data. Because of the presence of quality checks, the judgments that can be successfully submitted at the end of the HIT achieve a comparable quality across experience levels. This confirms the ability of quality checks to implicitly select a sample of workers who can provide high quality judgments.

Figure 7 shows the impact of worker experience on the diversity of crowdsourced data collections. We observe that in our experiment 95 (61.7% of 154) workers have completed only one task (see Fig. 7a), while the other 59 (38.3% of 154) workers have submitted at least two tasks each. There are 9 workers (5.8% of 154) completing tasks for every topic (i.e., 6) (see Fig. 7a). This result has implications on the diversity of crowdsourced collections: Few workers complete a large number of tasks and thus dominate the judgment contribution to the generated collection. In our experiment, 24 (15.6% of 154) workers have completed 122 (40.7% of 300) tasks (see Fig. 7a, workers completing 4, 5 and 6 tasks). Figure 7b shows that experienced workers have done significantly more tasks than less-experienced workers ($t$-test $p < 0.05$). Most of the workers who have done multiple tasks are experienced workers. Therefore, we can conclude that the diversity of the collected data is limited as experienced workers contribute most judgments.

5 DISCUSSION

From the previous sections we have observed how different workers in the crowd demonstrate different work behaviors in terms of the use of different task completion strategies and shortcuts. This type of results are very valuable for a number of different reasons. Firstly, the findings in this paper advance our understanding of low-level behavior that is exhibited by workers in microtask crowdsourcing marketplaces. Moreover, by revealing the disparity between how experienced and less-experienced workers complete work, our findings can inform future training and scaffolding strategies for novice crowd workers or workers with less experience. As we have observed that some strategies (e.g., efficient reading) can be developed within tasks, less-experienced workers are able to benefit from knowing (e.g., when reading the instructions) such strategies to speed-up the task completion process, which may increase their hourly earnings.

While preserving worker anonymity in the crowd, we have shown it is possible to collect data allowing platforms and requesters to profile workers and to better understand the reasons behind their more or less efficient and effective performance. Following this type of research we can envision advanced models based on behavioral data; for example, process mining [35] or Markov models [37] that can be used for either predicting what workers will do next within different tasks, or to build supervised classification models of worker types based on their behaviors. This would allow to deal with cold-start issues (where no worker profiling data is available for new workers approaching a task) and to classify a worker’s type after observing just a few behavioral actions at the beginning of a task. This has broad implications on early filtering or pre-selection of crowd workers.

Our findings in this work also have strong implications on the use of crowdsourcing to collect human-labeled annotations. Judgments, labels, and content generated by crowd workers may be biased by the behavioral patterns we have observed in our study. More efficient crowd workers may introduce a bias in the generated collections rather than these being the result of a democratic process where members of a population are equally represented and equally contributing to it. As we have observed that experienced workers are more efficient and indeed have completed more tasks than less-experienced workers, they also have the opportunity to contribute more than others to a crowdsourced data collection project. On the other hand, the fact that these workers perform a lot of content reuse and adaptation actions has implications on the type and the diversity of the collected data. The intensive use of copy/paste actions we have observed may result in annotation collections with many duplicated content which may or may not be desirable depending on the specific application. For the specific crowdsourcing task of relevance judgments to evaluate IR systems, the benefits of less-experienced workers giving richer justification and summary text shed some lights on the use of such text to reduce or remove idiosyncratic bias, which is introduced when workers assign scores based on subjective perceptions and their own preferred implicit judgment scales [34]. Because of the ability of quality checks presented in tasks to implicitly select a population of high quality workers for specific tasks, a more diverse data collection contributed by less-experience workers would allow us to explore more insights of the annotations provided, for example to learn from their self-constructed justifications, which may act as an indicator to understand workers trustworthiness in answer aggregation.
6 CONCLUSIONS
In this paper we have studied crowd worker behavioral patterns in document relevance judgment tasks. We conducted different types of studies including surveys (n = 100) and qualitative interviews with crowd worker community leaders (n = 10) to understand how experienced crowd workers complete tasks. We delve into the task completion strategies of experienced workers that make their work more efficient in juxtaposition to their less-experienced counterparts. We also carried out a quantitative data-driven experiment (n = 154), and analyzed the behavioral traces left by workers completing relevance judgment tasks on the FigureEight crowdsourcing platform. Our results show that (i) when efficient reading skills could be improved with short-term practice, efficient text generation strategies are established in the long term, (ii) the use of shortcuts and reuse of existing text is a very common strategy to reduce task completion time, and thereby increase workers’ hourly wages, (iii) the “Ctrl (Cmd) + F” shortcut that can support efficient identification of textual content has a very limited popularity, (iv) experienced workers write shorter typed text and reuse more content than less-experienced workers, and (v) experienced workers do not necessarily provide higher quality judgments, but they are faster and complete more tasks as compared to less-experienced ones, which can influence the diversity of crowdsourced collections.

These results have strong implications on the use of IR evaluation as crowd worker selection bias may affect the way in which labels are generated. We concluded from our mixed-methods research that not all workers complete relevance judgment tasks in the same way. Thus, depending on who participates in a batch of HITs, the collected data may be more or less reliable. In our future work, we will conduct experiments to observe crowd worker strategies in other types of tasks (e.g., image labelling). We can also envision the use of the identified worker strategies to both improve learning through achievement priming in crowdsourced information finding microtasks. In Proceedings of the Seventh International Learning Analytics and Knowledge Conference. ACM, 105–114.


Gabriella Kazai, Jaap Kamps, and Natalia Milic-Frayling. 2011. Worker types and personality traits in crowdsourcing relevance labels. In CIKM ’11. ACM.


Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*(1947), 50–60.


REFERENCES