The Effects of Choice in Routing Relevance Judgments

Edith Law
Carnegie Mellon University
Pittsburgh, USA
elaw@cs.cmu.edu

Paul N. Bennett
Microsoft Research
Redmond, USA
pauben@microsoft.com

Eric Horvitz
Microsoft Research
Redmond, USA
horvitz@microsoft.com

ABSTRACT
The emergence of human computation systems, including Mechanical Turk and games with a purpose, has made it feasible to distribute relevance judgment tasks to workers over the Web. Most human computation systems assign tasks to individuals randomly, and such assignments may match workers with tasks that they may be unqualified or unmotivated to perform. We compare two groups of workers, those given a choice of queries to judge versus those who are not, in terms of their self-rated competence and their actual performance. Results show that when given a choice of task, workers choose ones for which they have greater expertise, interests, confidence, and understanding.

Categories and Subject Descriptors
H.1.2 [**User/Machine Systems**]: Human Information Processing; H.3.3 **[Information Search and Retrieval]**: Relevance Feedback

General Terms
Design, Experimentation, Human Factors

Keywords
Task Routing, Human Computation, Relevance Judgments

1. INTRODUCTION
Human computation systems have become an increasingly popular platform for distributing tasks, including search relevance judgments. To date, tasks are routed to individual workers in a random manner. Yet, judging the relevance of webpages is a knowledge-intensive task, requiring workers to not only understand the meaning of search queries, but also the different possible intentions that led searchers to issue the queries. We explore the use of a relevance judgment system to improve the match of judgment tasks to workers' interests and capabilities.

Obtaining relevance judgments to guide the evaluation, design, and optimization of retrieval methods has been a long-standing challenge in information retrieval [4]. These challenges include among others, obtaining quality judgments [3], controlling the cost of judgments [2], and collecting a sufficient number of judgments [8]. Studies have shown poor agreement among judges about the intentions behind queries [6,7]. Several efforts have focused specifically on the expertise of judges, showing that expertise can vary widely, and that there is poor agreement among judges with different levels of expertise [1]. In other related work, Russell & Grimes [5] concluded that users conducting searches for information were more engaged when choosing their own task, issuing fewer queries per task but engaging in longer search sessions. They hypothesized that self-selected tasks are clearer in the minds of searchers and noted that participants who provided input have a better preconceived notion of the task as they demonstrate a smaller range of possible alternate queries.

SIGIR, July 24–28, 2011, Beijing, China.

2. EXPERIMENT
We conducted an experiment with 26 subjects to investigate the effectiveness of routing relevance judgment tasks *by choice* to workers. In the *choice* condition, subjects are given a choice of five search queries, of which they choose one query to judge. Each set of queries are selected from a log of queries, input by real-world users to a major search engine. Note that unlike TREC, queries drawn from a log do not have a description of query intent. Mitigating such ambiguity has been discussed in other work [3]. The queries are balanced for length, difficulty (in terms of ambiguity as measured by click entropy, a metric capturing the variation of pages clicked on following a query), and topics. Thus, task distribution in the choice routing condition is aimed at providing a balanced set of queries leaving expertise and interests as the deciding factors for choosing a query. In the *yoked* condition, each participant is assigned exactly the same queries that a particular choice subject has selected. Pairing the yoked and choice subjects allows us to compare the two user groups using the same set of queries to control for query variance.

The participants performed 15 relevance judgment tasks in total (one task=one query). In each task, participants judge the relevance of five webpages (image captures) returned for a particular search query, rated using a 5-point scale: "perfect," "excellent," "fair," "good," and "bad."

For each query, before seeing and judging the webpages, subjects are given a pre-judgment questionnaire about the query, which asks the following: (1) Specify up to three intentions behind the search query (i.e., what the user was looking for?) and mark the most likely; (2) How confident are you that the most likely intention is the user's goal? (3) How knowledgeable are you in the topic of the search query? and (4) How interested are you in the topic of the search query? The latter three questions are assessed on a four-point scale.

After judging the relevance of results for a query, participants are given an identical post-judgment questionnaire, on which they can report any modifications to their previous answers, based on the knowledge they may have gained from reviewing the webpages.

We sought answers to two questions. First, does the routing of relevance judgment tasks to people by allowing them to choose (choice condition) result in a set of judges who see themselves as

having higher expertise, interests, and confidence in the tasks than those who cannot choose tasks (yoked condition)? Second, beyond self-rated expertise and confidence, is there evidence that the choice subjects have a better understanding of the search query than the yoked subjects? We now discuss the results.

2.1 Expertise, Interests and Confidence

Figure 1 summarizes the differences in self-ratings of expertise, interests, and confidence of choice subjects versus yoked subjects.

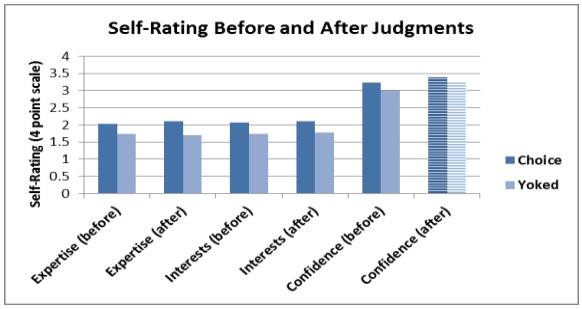

Figure 1. Results before and after judging webpages. Groups with solid bars indicate a statistically significant difference.

Results show that the choice subjects have higher self-rated expertise (difference before judging = 0.30, $p < 0.006$; difference after judging = 0.41, $p < 0.0003$), interests (difference before judging = 0.33, $p < 0.003$; difference after judging = 0.33, $p < 0.004$) and confidence in the intention of the search query (difference before judging 0.23, $p < 0.013$; difference after judging = 0.12, $p < 0.20$) than the yoked subjects. All differences are statistically significant by the paired *t*-test, except for the confidence in the intention behind the search query assessed *after* judging webpages, perhaps because both choice and yoked subjects gained an understanding about the meaning of the search query following review of the candidate webpages.

2.2 Understanding

Although self-rating may be a useful in routing, it only reveals the workers' *perception* of their own competence with solving a task and not necessarily their actual performance. For example, it is possible that by being granted a choice, workers feel more confident, regardless of *which* choice they made. One way to confirm that choice subjects have a deeper understanding of the task at hand than the yoked subjects, is to compare the number of modifications each group makes to their assessed intentions for the search query *after* judging the webpages, i.e., the differences in answers to Question 1 of the pre-judgment and post-judgment questionnaires. We assume that subjects who do not understand a query are more likely to make changes to the intentions that they had originally specified when they receive additional information.

We measured the percentage of queries for which modifications (addition, deletion, or revision) were made to (i) any of the intentions, and (ii) the most likely intention in the post-judgment questionnaire. Intentions were inspected by hand to remove superficial changes (e.g., re-wording, or spelling corrections). In both cases, the yoked subjects modified the intentions more often (29.23% of any intentions, and 19.85% of the most likely intention) than the choice subjects (15.38% of any intentions, and 14.50% of the most likely intention). This suggests that the yoked subjects are not as certain as the choice subjects about the meaning of the search query prior to reviewing the webpages.

Commonly Selected	Not Selected
weather vienna austria 10 day (3/3)	usb-uart controller nokia 6610i (0/8)
soccer (4/7)	th50px500u (0/7)
36 weeks pregnant baby size (2/3)	Webtender (0/7)

Table 1. Examples of commonly selected and non-selected queries and ratio of selected vs. presented.

Table 1 displays several examples of selected and unselected queries by choice subjects, providing additional support that people tend to select queries they understand. We note that this does not indicate that people select queries that are *universally* understood (i.e. non-ambiguous), as in such a case, yoked subjects judging the chosen queries would be expected to have similar performance rather than the differences observed in Figure 1. Table 1 also suggests, that the number and proportion of times a query is selected are measures of query-understanding and interest (e.g. "soccer" is easily understood with a high number of selections but not all judges have an interest in the query); a potential future task routing system using the number of times a query has not been selected as part of the balance criterion may lead to more uniform completion rates for all queries as well as higher quality judgments. Verifying this is an area of future work.

3. CONCLUSION

Using relevance judgment of webpages as a case study, we investigate whether task routing by choice can lead to work being done by workers with greater self-rated expertise, interests, confidence, and understanding of the search queries. When given the option of selecting queries participants provided higher ratings for their expertise, interests, and confidence about the intentions behind queries. The choice cohort also modified assessments of intentions less than yoked subjects following review of additional information, suggesting they have greater expertise and confidence in their knowledge. We introduce statistics on the modification of intent assessment as a proxy for expertise on a relevance judgment task. We believe such statistics warrant additional study as potential signals of expertise. Overall, we view task routing systems and policies that take into account expertise as promising directions for crowdsourcing relevance.

4. REFERENCES

[1] Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A. and Yilmaz, E. Relevance assessment: are judges exchangeable and does it matter? In *SIGIR 2008*.

[2] Carterette, B., Allan, J., & Sitaraman, R. Minimal Test Collections for Retrieval Evaluation. In *SIGIR 2006*.

[3] Carterette, B., Bennett, P. N., Chickering, D. M., & Dumais, S. T. Here or there: preference judgments for relevance. In *ECIR 2008*.

[4] Cleverdon, C. The cranfield tests on index language devices. *Readings in IR*, Morgan Kaufmann, 1997.

[5] Russell, D. and C. Grimes. Assigned tasks are not the same as self-chosen web search tasks. In *HICSS 2007*.

[6] Teevan, J., Dumais, S. and Horvitz, E. Characterizing the Value of Personalizing Search. In SIGIR 2007

[7] Teevan, J., Morris, M. and Bush, S. Discovering and Using Groups to Improve Personalized Search. In *WSDM 2009*.

[8] Yilmaz, E. and Aslam, J. Estimating average precision with incomplete and imperfect judgments. In *CIKM 2006*.