An Analysis of Assessor Behavior in Crowdsourced Preference Judgments

Dongqing Zhu and Ben Carterette
Department of Computer & Information Sciences
University of Delaware
Newark, DE, USA 19716
[zhu | carteret]@cis.udel.edu

ABSTRACT

We describe a pilot study using Amazon's Mechanical Turk to collect preference judgments between pairs of full-page layouts including both search results and image results. Specifically, we analyze the behavior of assessors that participated in our study to identify some patterns that may be broadly indicative of unreliable assessments. We believe this analysis can inform future experimental design and analysis when using crowdsourced human judgments.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance Evaluation*

General Terms

Experimentation, Measurement, Human Factors

Keywords

crowdsourcing, preference judgements, experimental design

1. INTRODUCTION

The search engine results page (SERP) layout has a great impact on users' searching experience. With the emergence of images, ads, news, blog posts, and even micro-blog posts in the search results, how to smartly integrate them into web pages to assist searching becomes an interesting though challenging problem. As the first step, we explore how to make the search results layout more user-friendly by varying the positions of images relative to ranked results. Images can assist users in finding the information they want and make searching more effective. On the other hand, images may take up too much valuable space on the SERP due to their sizes.

This work describes a pilot study we performed to determine the optimal placement of images among search results. The objects to be assessed are *full page* layouts consisting of both ranked results and results from image search verticals. It is the positions of images and ranked results relative to one another rather than the relevance of individual items on the page that we are interested in assessing. Because we believe the quality of a full layout is difficult to assess on an absolute scale, we used preference judgments: assessors see two different possible layouts and choose the one they prefer. Over a large number of assessors, we should be able to determine the most preferred general layout, as well as specific queries that deviate from the overall most preferred.

Preference judgments can be made quickly and are less prone to disagreements between assessors than absolute judgments [2]. Using preference judgments to evaluate is also more robust to missing judgments than using absolute judgments to evaluate [1]. Preferences between full page layouts seem to correlate well to traditional evaluation measures based on absolute relevance judgments on documents [9]. Finally, preference judgments can be mapped to much finer-grained grades of utility than is possible with absolute judgments [8]. Preference judgments do have some problems: the number of pairs to be judged grows quadratically rather than linearly, and assessors seem to find the increased number much more tedious [2]. Furthermore, when assessing full layouts, the number of objects that need to be assessed can grow factorially as individual items are rearranged relative to one another. And depending on how much layouts are allowed to vary, there can be a "credit assignment problem": it is difficult to tell *why* an assessor prefers one layout to another.

Nevertheless, preference judgments seem an ideal tool for this task, and because we both need many of them and they can be made quickly, they seem to be an ideal candidate for crowdsourcing via Amazon's Mechanical Turk[1] (MTurk) or some other system. The fact that they can be made quickly, however, may lead assessors being paid very low rates per judgment to "cheat" or produce otherwise unreliable data in various ways so that they can make more money without expending much effort. When data is unreliable, it leads to bias in judgments and possibly severe errors in evaluations [4]. Experimenters naturally would like to be able to prevent, detect, and compensate it [6][10][7]. But this cheating creates an adversarial relationship with the experimenter, in that as experimenters learn how to detect cheating, the assessors find new ways to cheat them. In our pilot study, we discovered that assessors seem to be cheating in ways that are not initially obvious, and further that they will sometimes cheat on one task while seemingly taking another

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

[1] http://www.mturk.com

seriously.

The rest of this paper is organized as follows: in Section 2, we describe our experimental design. In Section 3, we provide some analysis on the behavior of MTurk workers, and in Section 4 we summarize our results and describe directions for future work.

2. EXPERIMENTAL DESIGN

We set up an online survey that asks assessors to give their preferences on variations of the Yahoo! SERP layout for 47 queries formed from 30 topics taken from the TREC 2009 Web track [5] and the TREC 2009 Million Query track [3]. To limit the space of possibilities, we selected only queries that have results from an image vertical but not from any other vertical. We keep the search results fixed (we always use the same 10 URLs with the same summaries), and we insert image vertical results into one of three places: above all search results (top), below the top five results (middle), and below all ten results (bottom). In addition to results from the image vertical, some URLs have an inline image associated with them as well. These we displayed to either the left or the right of the summary. In all, each query had up to six different layout variants: queries with only inline images had two variants, queries with image vertical results had three variants, and queries with both had six. Two layout variations are shown side by side to the assessors as illustrated in Fig. 1.

We take the advantage of Amazon Mechanical Turk as a platform to involve search engine users with different backgrounds around the world. In Mechanical Turk, "requesters" submit "HITs" (Human Intelligence Tasks) to be completed by "MTurkers" (Mechanical Turk Workers). We redirect MTurkers from our HIT (Human Intelligent Task) question to our own survey website, which allows us to show each MTurker a sequence of preferences and to log additional information such as time-on-task. MTurkers complete the survey, submit the confirmation code at the end of the survey via the HIT, and get paid US$0.13 for every 17 preferences they complete once their submissions are validated.

There are three different batches of survey questions, corresponding to queries with inline images only, image vertical results only, and both types. Each batch consists of 17 queries, and for each query there is a single preference judgment. We limited to one preference judgment per query per assessor to control possible learning on topic effects, even though it prevents us from acquiring all preference judgments for a query from a single assessor. The order of showing those 17 queries is randomized to control possible order effects and to allow analysis on whether assessors' behavior was changing as they learned about the task. For each query, two result pages are randomly selected and presented to the user, who may choose the "left" layout, the "right" variant, or express "no preference". Though each batch has different number of layout variants, we can easily control the data size required for each batch by letting user groups of different sizes to complete the survey.

We added an additional absolute-scale rating task for each query. Users not only give preferences for layouts but also rate the pictures by relevance on a ternary scale ("poor", "OK", "good"). In this way we can test whether the relevance of the pictures is also an important factor that influences the layout preference.

Out of the 17 queries, two are "trap" questions (following Sanderson et al. [9]) that have two identical result pages. We put them at the 6th and 15th respectively in the survey question sequence. Thus, assessors should have no preference for those pairs. The purpose of setting up the trap questions is to detect dishonest workers—if they do not se-

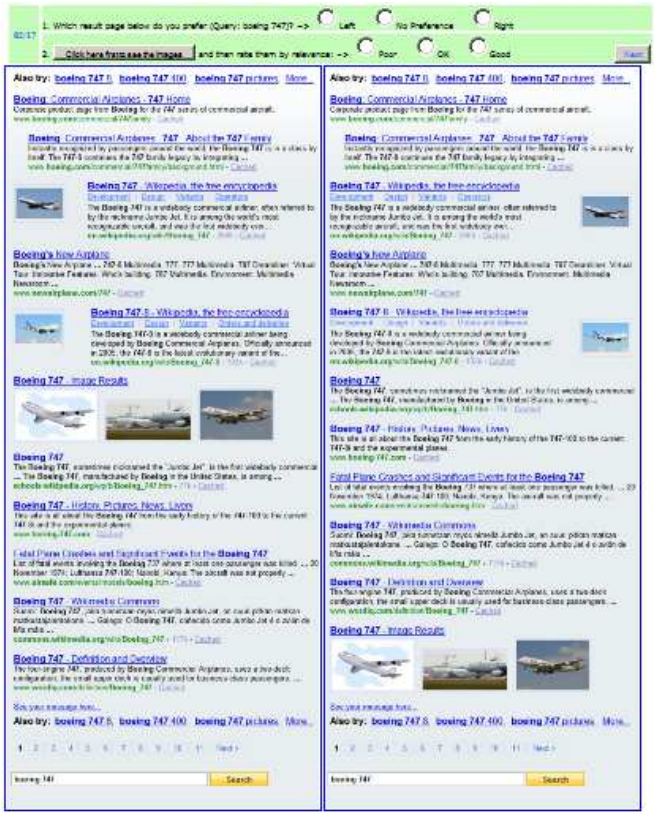


Figure 1: A sample survey question page. The first question asks assessors to make their perference. The second question asks assessors to rate the images in the SERP.

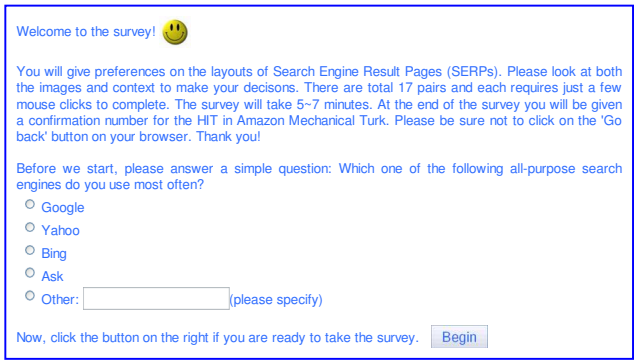


Figure 2: Welcome page of the survey with A simple question to identify user's favorite search engine.

lect "no preference", they are probably not paying attention.

Finally, we add a simple question at the beginning of the survey to identify a user's favorite search engine (Fig. 2). The purpose of this is to determine whether expectation plays any role in preference: Yahoo!'s default for inline images is to display them on the right, while Google's is to display them on the left (Bing's seems to vary by query). Our hypothesis is that Yahoo! users may prefer the right while Google users may prefer the left, at least initially.

3. DATA ANALYSIS

First we rejected the HITs that failed our trap question. After that, this pilot study produced a total of 25 approved HITs for which we had timing information (seconds per judgment), preferences, image ratings, and search engine preference. 24 of 25 preferred Google, so we were not able to test our hypothesis about expectations.

3.1 Time analysis

We plot the time in seconds that assessors spend on each preference judgment against the query sequence. Three general types of assessors are found and three representative curves are shown in Fig. 3. Fig. 3(a) shows the **Normal** pattern: the assessor starts out slow, quickly gets faster as he or she learns about the task, and then roughly maintains time. This is expected because assessors who never did this survey before require more time on the first few questions, but as they get more familiar with the questions, they make quicker responses. 17 out of the 25 assessors (68%) fall into this category.

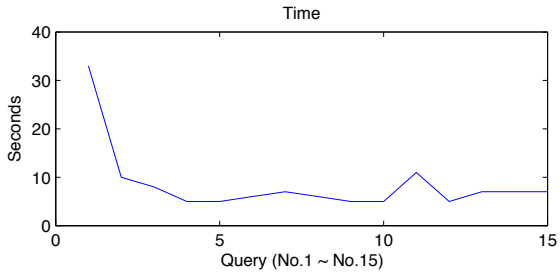
Fig. 3(b) shows the **Periodic** pattern which indicates very strange user behavior: some judgments are made fast and some are made slow, but the fast and slow tend to alternate. One possible explanation might be that these assessors were not fully dedicated to doing the survey. They are probably not purposely cheating, but they might be absent-minded periodically, calling their data into question. 6 out of the 25 assessors (24%) fall into this Periodic category.

Fig. 3(c) shows the **Interrupted** pattern in which occasional peaks appears under the background of Normal pattern. Users who have this kind of pattern might be interrupted in the middle of the survey, e.g., receiving a phone call while doing the survey; it seems unlikely they are cheating or allowing the interruption to affect their results significantly. Only 2 assessors fall into this category.

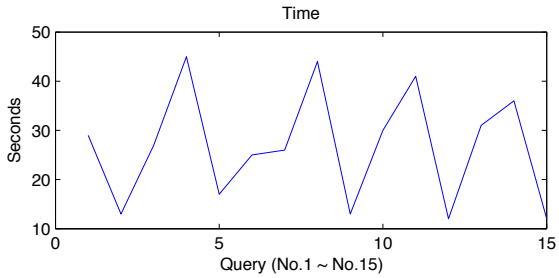
3.2 Image rating analysis

As to the image ratings in Fig. 4, they also exhibit 3 different patterns. Fig. 4(a) shows the **Normal** pattern. Users give rating 2 or 3 most often and give rating 1 occasionally (suggesting that Yahoo!'s image results are pretty good, as we expect). 21 of 25 assessors (84%) have this Normal time pattern.

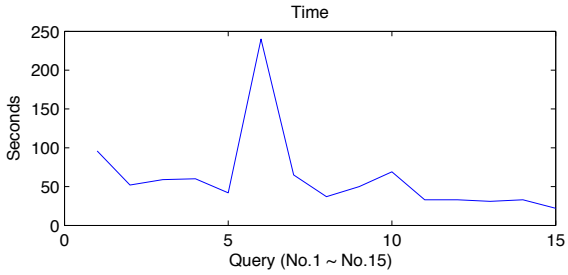
Fig. 4(b) shows a **Periodic** pattern similar to that observed for timing: a user shows a tendency to alternate between a subset of ratings. These assessors may be cheating, but exhibiting an advanced cheating behavior by purposefully trying to "randomize" their responses so that it would be



(a) **Normal** assessors start out slow and quickly get up to a consistent judgment speed.



(b) **Periodic** assessors vacillate between relatively fast judgments and relatively slow judgments.



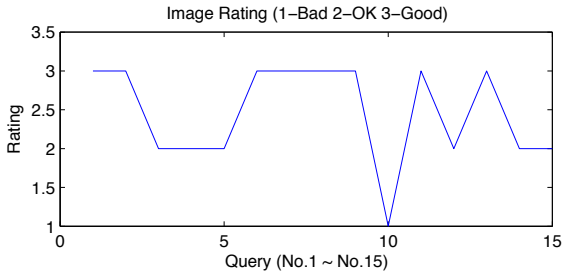
(c) **Interrupted** assessors look like **normal** assessors except for a large spike in time.

Figure 3: Each plot shows the time an assessor took to make a preference judgment for each of 15 queries in a randomized order. "Trap" queries have been excluded. The plots are not averages; each is an exemplar of one of the cases.

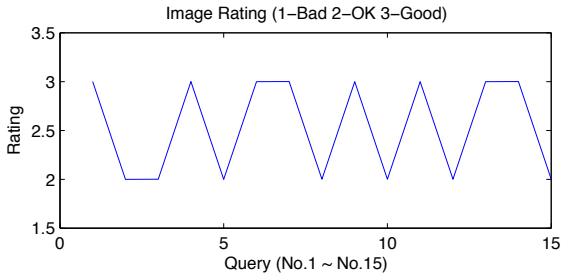
difficult for requesters to discover their dishonesty. 2 out of 25 assessors (8%) fall into this category and we find that one of these two assessors also shows a periodic time pattern.

Note that Figures 4(a) and 4(b) actually have roughly equal numbers of "good" judgments. If indeed images for 8 of the 15 queries can be given the highest rating, then because the order is randomized, there is some chance that an honest assessor will actually produce such a periodic pattern. However, the probability of any periodic or even quasi-periodic (i.e., with some short repetitions) pattern being observed due to chance alone is very low—we estimate it to be less than one in a thousand. It therefore seems safe to conclude that both assessors produced invalid data.

Fig. 4(c) shows the **Fixed** pattern in which all or most of the images rating are the same. 2 out of 25 assessors (8%) have the Fixed pattern. This type of assessor seem clearly not interested in rating the images and thus almost always



(a) **Normal** assessors demonstrate no clear pattern in their image relevance judgments.



(b) **Periodic** assessors vacillate between two or more ratings in a consistent way.

(c) **Fixed** assessors give every image set the same judgment.

Figure 4: Each plot shows the rating an assessor gave to the set of images retrieved by the image vertical.

give a fixed rating.

3.3 Preference judgment analysis

In this section we analyze preference judgments for the mixed set (both inline and vertical image results) to determine whether we can identify one or the other as the primary factor in the assessor's preference. If so, we may be able to (partially) address the "credit assignment problem" described in Section 1.

We separately analyzed the inline placement and vertical placement preferences for the time being. We assign TMB (top/middle/bottom vertical variants) and LR (left/right inline variants) scores to indicate the layout preferences according to the following method:

- Given a T-B pair, if the user prefers T, we assign 1.5 as TMB score of that pair. Otherwise, we assign -1.5.
- Given an M-B pair, if the user prefers M, we assign 1 as TMB score of that pair. Otherwise, we assign -1.
- Given a T-M pair, if the user prefers T, we assign 1 as TMB score of that pair. Otherwise, we assign -1.
- Given an L-R pair, if the user prefers L, we assign 1 as LR score of that pair. Otherwise, we assign -1.

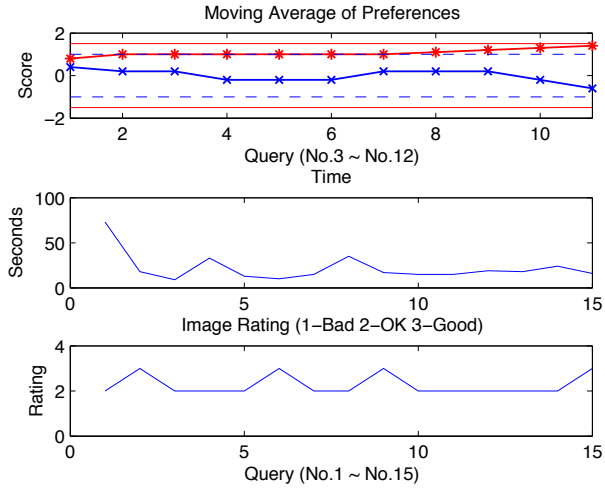
In the example in Figure 1, if an assessor preferred the left layout to the right, the judgment would have a TMB score of 1 because the assessor prefers vertical results in the middle rather than the bottom, and an LR score of 1 because the assessor prefers inline images on the left to the right. Note that the first three items above are mutually exclusive. Thus, higher TMB and LR scores indicate the preference for higher vertical images position and left inline image position.

Figures 5 and 6 show moving averages of the TMB and LR scores (red and blue lines, respectively) against query number; we call these curves layout preference curves. They are moving averages to control for variance in which layout variants the assessor saw. There are roughly two patterns of layout preference curves: either only one of the scores changes over time, or both do. Fig. 5 shows two representative curves of the first pattern along with judgment time and image rating curves (note that both assessors are more-or-less **normal** types as defined above). In Fig. 5(a), the inline image preference gradually goes from a left-preference to a right-preference while the vertical image preference remains high. In Fig. 5(b), the inline image preference stays fixed on the right while the vertical image preference seems to be changing periodically (with a relatively long wavelength). We infer from this that it is the layout preference associated with the varying curve that is the leading factor in making preferences: assessor 5(a) definitely prefers vertical results towards the top, so bases his or her judgments on the position of the inline images, while assessor 5(b) definitely prefers inline images on the right, so bases his or her judgments on the position of the vertical results. The alternative possible explanation, that the assessor simply doesn't care about one or the other type, is probably not the case: since each variant for each type was equally likely to occur in either the left or right position, the assessor had to consciously choose their consistent preference every time.

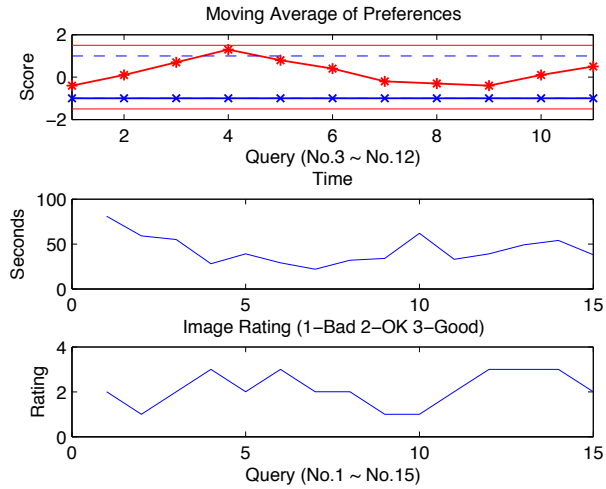
Fig. 6 shows two representative curves of the second pattern along with judgment time and image rating curves (again, both assessors seem **normal** w.r.t. time, though there is a sense that assessor 6(a) stopped doing the image rating task while assessor 6(b) has somewhat periodic ratings). In this pattern, both TMB and LR curves vary over time, so we cannot conclude that one or the other is responsible for the preference. However, we may take this as an indicator of expectations shifting as the assessor becomes more familiar with the different layouts. Assessor 6(a) starts out with a preference for inline images on the right but gradually comes to have no preference; he or she also starts with no preference for vertical image placement but gradually comes to prefer them on the top. In these cases we may need to look at each SERP pair individually to determine if TMB and LR positions have a combinational effect on layout preference.

3.4 Analysis of rejected data

Finally, we looked at the assessors that failed the "trap" question by expressing a preference when the layouts were

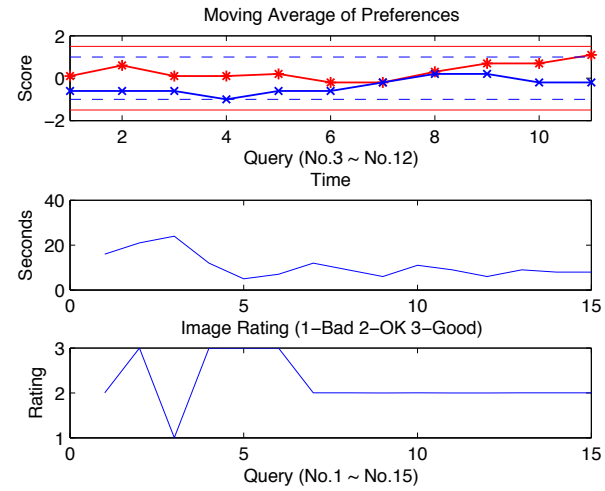


(a) Varying inline image preference curve (LR curve).

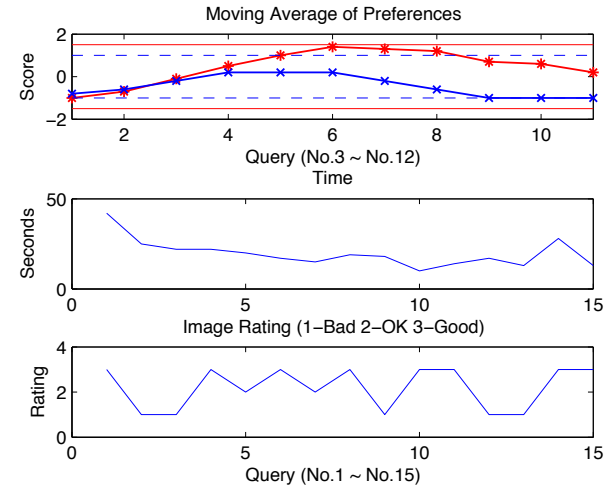


(b) Varying vertical image preference curve (TMB curve).

Figure 5: One layout preference curve (top plots) changes dramatically over time while the other changes slightly. The solid red line with $*$ indicates TMB score (for vertical image preference) and the solid blue line with $\times$ indicates LR score (for inline image preference). The solid red lines are the upper and lower bounds for TMB score while the dashed blue lines are bounds for LR score. The moving window size is 5. Thus, the curve starts at Query No.3 and ends at No.12. The middle and bottom plots show the judgment time and image rating curves respectively.



(a)



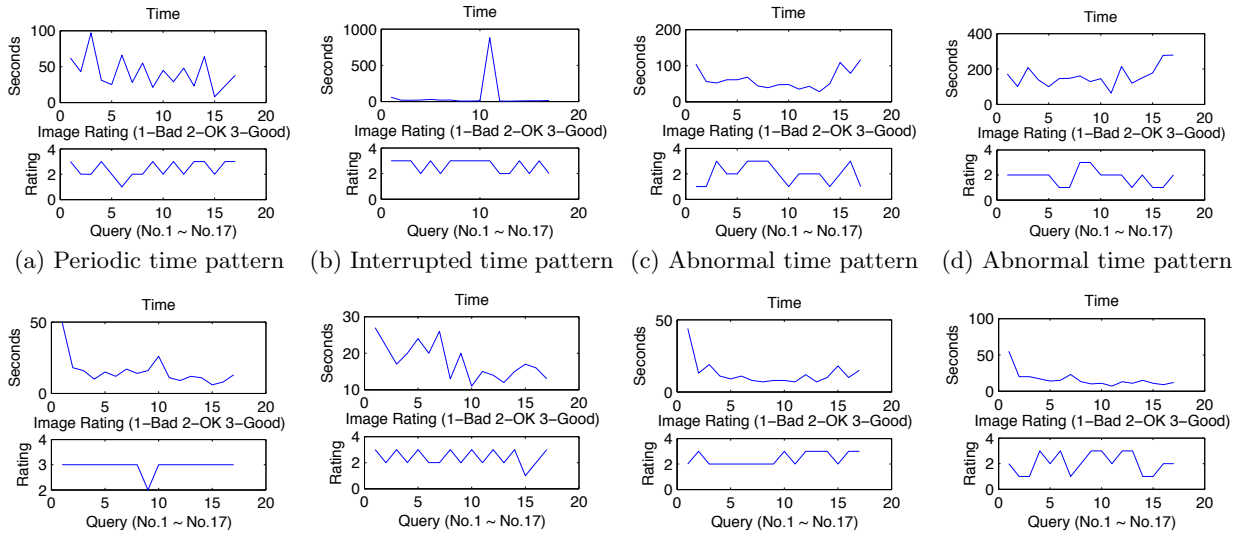
(b)

Figure 6: Both layout preference curves (top plots) change over time. The solid red line with $*$ indicates TMB score (for vertical image preference) and the solid blue line with $\times$ indicates LR score (for inline image preference). The solid red lines are the upper and lower bounds for TMB score while the dashed blue lines are bounds for LR score. The moving window size is 5. Thus, the curve starts at Query No.3 and ends at No.12. The middle and bottom plots show the judgment time and image rating curves respectively.

identical. Eight assessors were rejected for this reason; of these, one showed a periodic time pattern (Fig. 7(a)) and one showed an interruption (Fig. 7(b)). Two showed an *abnormal* time pattern of taking longer on the last few queries (Fig. 7(c), Fig. 7(d)); this was not observed among assessors that passed the trap question. One showed a fixed rating pattern (Fig. 7(e)) one showed a partially periodic rating pattern (Fig. 7(f)), and two seemed normal in both time and rating (Fig. 7(g) and 7(h)).

4. CONCLUSIONS

We performed a pilot study to determine whether Amazon Turk could produce useful preference judgments for distinguishing between different layouts including both search engine results and image results. Though we did not discuss it in the main body of this work, the overall results were overwhelmingly in favor (by a factor of roughly 2:1) of vertical image results near the top of the page and inline image re-



(a) Periodic time pattern (b) Interrupted time pattern (c) Abnormal time pattern (d) Abnormal time pattern

(e) Fixed image rating pattern (f) Partially periodic image rating pattern (g) Normal (h) Normal

Figure 7: Analysis of assessors that failed the "trap" question.

sults on the left. However, we believe our analysis provides some interesting data for designing future studies:

1. Unreliable assessors may reveal their unreliability in different ways: some through periodic timings, some through abnormal timings, some through periodic ratings, some through fixed ratings.
2. When asked to perform more than one task, assessors may be reliable on one without necessarily producing reliable data for the other.
3. Trap questions are useful for identifying unreliable assessors, as 6 of the 8 responses rejected for failing the trap question also exhibited unusual behavior patterns.
4. Trap questions alone may not identify all the unreliable assessors, as 6 of 25 passed the trap question but showed periodic timings and 4 of 25 passed the trap question but showed strange image rating behavior.
5. The use of periodic image ratings may suggest MTurkers learning how to avoid being detected when cheating.

Certainly there is more analysis that can be done, particularly in terms of the total number of assessments needed, whether it is "safe" to have a single assessor make multiple preference judgments for the same query, and how to aggregate preferences over assessors to learn about particular queries. These are all directions we are pursuing currently.

Acknowledgements

This work was supported in part by a gift from Yahoo! Labs and in part by the University of Delaware Research Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

5. REFERENCES

[1] B. Carterette and P. N. Bennett. Evaluation measures for preference judgments. In *Proceedings of SIGIR*, 2008.

[2] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: Preference judgments for relevance. In *Proceedings of ECIR*, pages 16–27, 2008.

[3] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas. Overview of the trec 2009 million query track. In *Proceedings of TREC*, 2009.

[4] B. Carterette and I. Soboroff. The effect of assessor errors on ir system evaluation.

[5] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the trec 2009 web track. In *Proceedings of TREC*, 2009.

[6] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 453–456, New York, NY, USA, 2008. ACM.

[7] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In L. Bottou and M. Littman, editors, *Proceedings of the 26th ICML*, pages 889–896, Montreal, June 2009. Omnipress.

[8] M. E. Rorvig. The simple scalability of documents. *JASIS*, 41(8):590–598, 1990.

[9] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *Proceedings of SIGIR*, 2010.

[10] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*, pages 254–263, Morristown, NJ, USA, 2008. Association for Computational Linguistics.