Differences in Search Engine Evaluations Between Query Owners and Non-Owners

Alexandra Chouldechova
Stanford University
Stanford, CA 94305
achould@stanford.edu

David Mease
Google
Mountain View, CA 94043
dmease@google.com

ABSTRACT

The query-document relevance judgments used in web search engine evaluation are traditionally provided by human assessors who have no particular association with the specific queries selected for the evaluation. Most commonly, queries are randomly sampled from search logs and in turn randomly assigned to the human assessors. In this paper, we consider a very different approach in which we instead ask the human assessors to provide their own queries from their recent search experiences. Using these queries as our sample, we compare the relevance judgments from the "owners" of the queries to the relevance judgments of the non-owners.

We conduct experiments which reveal that query ownership has a substantial and beneficial impact on the accuracy of relevance judgments. In particular, we observe that owners are more consistently able to distinguish a higher quality set of search results from a lower quality set in a blind comparison. The implication for web search evaluation is that query owners provide more valuable relevance judgments than non-owners, presumably due to the background knowledge associated with their queries. We quantify the benefit of using owner assessments versus non-owner assessments in terms of sample size reduction. We also touch on some of the practical challenges associated with using query owners as assessors.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics—*experimental design*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Measurement

Keywords

Evaluation; experiment design; search engines; user queries

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WSDM'13, February 4–8, 2013, Rome, Italy.
Copyright 2013 ACM 978-1-4503-1869-3/13/02 ...$15.00.

1. INTRODUCTION

Traditionally web search engines are evaluated based on queries sampled from the logs of the search engine. The human assessors who are employed to provide relevance judgments of documents returned for these queries may have very little subject matter background knowledge for certain queries. The assessors will usually be required to be fluent in the query language, but other than that there are often no assurances made to be certain that the assessors are familiar with the query's meaning. Conversely, the person who originally issued the query would generally have more background knowledge and additionally would often be aware of the specific information or types of web content that would best satisfy the query. In short, there is reason to believe that the person who issues a query should be a better judge of relevance than a random assessor. In this paper we design experiments to test for and measure the magnitude of this effect.

For obvious practical reasons it would be impossible to learn the identities of search engine users who issued particular queries found in the logs and ask those users to assess the relevance of documents for those queries. Thus, in this paper we take a different approach. We begin with assessors and ask these assessors to recall queries they have issued in the past for their own personal reasons. Using this approach we are able to pair queries with their "owners". We then construct two sets of 10 search results for each query, one set which is known to have better quality (on average) than the other. We ask the assessors to choose the more relevant set of 10 from the pair, and we measure the extent to which their preferences for the queries they own are more accurate than their preferences for queries which they do not own. Accuracy in this context refers to the selection of the higher quality result set.

The paper is organized as follows. Section 2 provides a review of some relevant literature. Section 3 outlines the experimental design. In particular, in that section we provide details pertaining to how queries were elicited from assessors as well as how we constructed the two sets of 10 results for each query. Sections 4 and 5 summarize the results of our data analysis. In Section 6 we describe the inference carried out in order to establish the statistical significance of the results. Section 7 outlines the implications that these results have with regard to the sample size needed for experiments using human assessors. Conclusions and discussion of the results are presented in Section 8.

2. LITERATURE REVIEW

As mentioned earlier, it is most common in web search evaluation that assessors are randomly assigned to queries. This is generally the case, for instance, with the extremely popular TREC [5] assessments.

In addition to TREC, other authors have collected relevance judgments in a variety of ways. For example, Alonso and Mizzaro [1] examined the use of crowdsourcing via Mechanical Turk as an alternative to TREC assessors, and Al-Maskari et al. [10] collected relevance judgments from participants in an interactive IR experiment they conducted. Both of these studies are similar to each other (as well as to the vast majority of the literature dealing with human assessors) in that the assessors are randomly assigned to queries, and there is no notion of query ownership. This is also generally the case with evaluation experiments carried out by commercial search engines such as the studies by Chapelle et al. [3], by Singla et al. [14], and by Huffman and Hochster [7].

While the random assignment of assessors to queries is most common, there is a fair amount of literature that examines more intelligent and sophisticated assignments. We review this literature in the following two subsections. In particular, we first consider assessor assignments in which the assessors have some expertise on the topics to which they are assigned. Secondly we consider literature involving assessments by topic "owners", which is more similar to the experiments that will be presented in this paper.

It should be noted that the literature reviewed here considers *absolute* assessments of *single* documents, while our experiments deal with *relative* measures of *sets* of documents. For this reason, it is difficult to compare our quantitative results to those in the literature; however, the general findings are very much related and relevant.

2.1 Expert Assessments

In Kinney et al. [8], the authors compared generalist assessors to assessors with domain expertise ("experts") for two specific classes of queries: computer-programming queries and biomedical queries. The authors concluded that the assessors with domain expertise were in fact providing higher quality and more in-dept relevance judgments, while the generalist assessors seemed to focus too much on query-keyword prominence.

The findings that will be presented in this paper can be thought of as a generalization of the expertise research in Kinney et al. [8] in the sense that assessors who have issued a query in the past for their own personal reasons (i.e., the owners) will generally have more domain expertise for that query than would a random assessor (i.e., a non-owner.) From this viewpoint, our research extends the work of Kinney et al. [8] beyond computer-programming queries and biomedical queries to a general sample of user-issued queries.

Law et al. [9] also compared random assessors to assessors with expertise. In that study, the expertise was discovered by allowing assessors to choose a set of queries from a given list. For each query selected, these assessors were asked to judge the relevance of five web pages. These assessors were then compared to assessors who were instead randomly assigned the same subsets of queries. The comparison revealed that the assessors who were permitted to choose queries did in fact have a higher (self-reported) degree of expertise than the assessors who were randomly assigned queries. No comparison was made in terms of the accuracy of the relevance judgments; however, the study did show that the assessors who were permitted to choose queries were less likely to make modifications to their original interpretations of the query intentions after providing their relevance judgments as compared to the randomly assigned assessors.

2.2 Owner Assessments

Next we discuss two studies which examined the impact of ownership (as opposed to merely expertise). In that regard, these two studies are more similar to ours.

The first of these is described by Voorhees [16]. In that paper, the author considers TREC data and examines the difference in relevance judgments between TREC topic authors and non-authors. She concludes that the authorship (ownership) does *not* have a substantial impact on the evaluation of the relative performance of the retrieval systems.

The second of these is described by Bailey et al. [2]. That research deals with three types of assessors labeled as "gold", "silver" and "bronze". Gold assessors are topic originators (owners). Silver assessors are non-owners but experts in the topics. Bronze assessors are neither owners nor experts. Again, the authors did not directly compare the accuracy of the relevance judgments, but did observe that the agreement between the gold and silver assessors was stronger than the agreement between the gold and bronze assessors.

2.3 Search Behavior

In addition to the research above dealing with how *evaluations* of owners are different from those of non-owners, we conclude this review of the literature by mentioning that there is also a substantial amount of work showing how *search behavior* can be quite different based on the user's familiarity with the query or topic. Most of these papers [4, 6, 17] focus generally on the differences in search behavior between experts and non-experts in the spirit of the Kinney et al. [8] paper mentioned above; however, the paper by Russell and Grimes [13] deals specifically with ownership in our sense. In particular, Russell and Grimes [13] found that study participants behaved differently when carrying out their own search tasks as opposed to experimenter assigned tasks. The behaviors found to differ included the amount of time spent completing the task as well as the number of distinct queries issued in the process. Since ownership leads to differences in search behavior, it is not suprising that it also leads to differences in search evaluation as will be demonstrated in this paper.

3. EXPERIMENTAL DESIGN

In this section we describe the design of our experiments in detail. We begin by describing the method by which we collected the queries from the assessors. Next we describe the nature of the evaluation tasks presented to the assessors in each experiment. Finally, we describe the construction of the search result sets being compared in each experiment.

3.1 Query Collection

At seven different time points over a 19 week period we contacted the members of a large pool of paid assessors in the United States to invite them to contribute queries. At each time point, any individual assessor was only permitted to contribute at most one single query. In this manner, we collected 23,530 queries, not all of which are unique.

The instructions to the assessors asked them to visit their Google query history at http://www.google.com/history and to select a single query from their history which they had issued for their own personal reasons (as opposed to a query issued as part of a previous assessment task). The assessors were told that their queries would be used in future evaluation tasks for both themselves as well as for other assessors in the pool. For this reason, they were encouraged not to include any highly personal or confidential information.

From here on, we will refer to the assessor who submitted a query as the "owner" of the query.

3.2 Task Description

Using subsets of these 23,530 queries, we conducted six evaluation experiments over a seven week period following the inital nineteen week query collection period. In each of these six experiments, two sets of 10 search results were shown side-by-side along with the query. The instructions to the assessors stated, "Please pick the side you would prefer if you were the user issuing the query." The assessors could indicate a preference for the right side, the left side or no preference.

In all six of our experiments, each assessor was always assigned any queries in the subset that he or she owned, and additionally five other queries that were randomly selected from the subset. For example, in Experiment 1 we used a subset of 5,000 queries from the set of 23,530. Thus, there were 30,000 total assessment tasks assigned, since each of the 5,000 queries was assigned to the owner plus five more random assessors from the same pool of query owners.

It is important to note that there was no indication provided to signal to the assessor whether he or she was evaluating a query he or she owned or a query that belonged to a different assessor. The only way an assessor could identify his or her own query would be from memory.

In an effort to make the comparison of the two sets of 10 search results easier, any search results which appeared on both sides of the comparison were marked accordingly. Figure 1 shows an example of the actual presentation to the assessors from the fourth of our six experiments. For all assessor tasks, the presentation of the two sides was randomized so that there was no left/right bias.

3.3 Experiment Descriptions

In this section, we describe how we constructed the two sets of 10 search results for each query. With the exception of Experiment 6, which will be discussed later, the two sets of results for each query were constructed such that one set would be known to be of higher quality (on average). In order to do this, we drew on the basic principle that Google's search results are on average ranked better than any re-ranking. For example, switching the first and second Google results will on average lead to a worse overall ranking than Google's original ranking. This type of experimentation is similar to studies such as those in [12]. In fact, our Experiment 2 is very similar to one used there.

The subsections below describe specifically how the two sets of 10 search results were constructed in Experiments 1 to 5. These are also summarized in Table 1. Some limited details regarding Experiment 6 are also provided.

Experiment 1

For Experiment 1, the higher quality search result set was constructed by taking Google's first search result for each query followed by its 12th through 20th in order. The lower quality search result set also used Google's first search result on top but then followed it with Google's 42nd to 50th results in order. Using the notation Gn to denote the nth result retrieved by Google as in Table 1, the higher quality result set for Experiment 1 is [G1, G12, G13, G14, G15, G16, G17, G18, G19, G20] and the lower quality result set is [G1, G42, G43, G44, G45, G46, G47, G48, G49, G50].

One advantage of constructing Experiment 1 in this manner is that neither of the two result sets is equivalent to Google's original search result ranking. Additionally, both result sets differ from Google's original ranking for exactly results 2 through 10. This helps eliminate any memory effect originating from previous experiences with the same queries on Google. To say this a different way, if one of the two result sets were more similar to Google's original search result ranking than the other result set, the ownership effect we are trying to measure could be confounded with a memory effect, since all owners have necessarily previously issued their queries on Google.

The queries selected for Experiment 1 were a random sample of 5,000 queries from the total set of 23,530. Experiments 2, 3 and 4 described below also use this same subset of 5000 queries.

Experiment 2

For Experiment 2, the higher quality result set is the top 10 Google search results in order, and the lower quality search result set is the same except two results between ranks 2 and 5 are randomly swapped with two results between ranks 6 and 10. This is a minor modification of the SWAP2 procedure described in [12].

Experiment 3

For Experiment 3, the higher quality result set is again the top 10 Google search results in order, and the lower quality result set is the same except Google's 11th and 12th results are inserted into the second and third positions (pushing Google's 9th and 10th results out of the set). Using Table 1 notation, the lower quality search result set is [G1, G11, G12, G2, G3, G4, G5, G6, G7, G8]. We refer to this type of change as an *insertion*.

Experiment 4

Experiment 4 is also an insertion experiment. The higher quality search result set is the same as that for Experiment 1: [G1, G12, G13, G14, G15, G16, G17, G18, G19, G20]. The lower quality search result set is obtained by inserting G21 into the fifth position of that set to produce [G1, G12, G13, G14, G21, G15, G16, G17, G18, G19]. Figure 4 shows exactly what an assessor would see for an example query ('Tyler Perry") in Experiment 4.

Experiment 5

Unlike Experiments 1-4 which used a random subset of 5,000 queries, for Experiment 5 we used all queries with five or more words. This yielded 4,870 queries from the complete set of 23,530 queries.

The higher quality search result set is once again the top

Figure 1: The two sets of 10 results as presented to the assessors in Experiment 4 for the query "Tyler Perry"

10 Google search results in order. The lower quality search result set consists of the top 10 Google search results for the query with the third word removed. For instance, the lower quality search result set for the query "Chase Freedom Mastercard double warranty" would be the top 10 Google search results for the query "Chase Freedom double warranty".

Experiment 6

Experiment 6 is unique in that instead of artificially creating higher quality and lower quality result sets, we used an actual experiment concerning a particular aspect of the Google search algorithm that was being investigated at the time of writing this paper. As a result, it is not known which set is actually of higher quality on average, and the positive/negative sign for the Experiment 6 display later in Figure 2 is arbitrary.

Unfortunately, we are not able to provide any details for either of the two result sets in Experiment 6. We are, however, able to say that the experiment was run using 4,420 of the 23,530 queries as the other queries were not impacted. Further, we will note that this experiment was of particular interest due to contradictory data from a number of various evaluation procedures that otherwise had proven to be quite consistent with one another historically across a variety of experiments.

4. RESULTS

In this section we summarize our findings concerning the effect of query ownership in our data. We begin by describing the data and scoring system used in our analysis, and then provide a discussion of the results.

4.1 Method

Due to nonresponse, many of the tasks assigned to the assessors were not completed during the study period. Our analysis is restricted to the subset of queries for which we obtained an assessment from the query owner and at least one of the five random assessors assigned. Across the six experiments the percentage of the queries we were able to keep ranged from 39% to 55%, with the variation in this number due mainly to the fact that some experiments were kept open to the participants for a longer period of time than others.

In order to carry out a quantitative analysis, we converted the assessor feedback to a numerical score of $+1$, 0 or -1 depending, respectively, on whether they preferred the higher

Experiment 1 Higher Quality Search Result Set	G1, G12, G13, G14, G15, G16, G17, G18, G19, G20
Experiment 1 Lower Quality Search Result Set	G1, G42, G43, G44, G45, G46, G47, G48, G49, G50
Experiment 2 Higher Quality Search Result Set	G1 , G2 , G3 , G4 , G5 , G6 , G7 , G8 , G9 , G10 (Google's top 10)
Experiment 2 Lower Quality Search Result Set	Google's top 10 with two results between ranks 2 and 5 randomly swapped with two results between ranks 6 and 10 - for example, G1 , G9 , G3 , G6 , G5 , G4 , G7 , G8 , G2 , G10
Experiment 3 Higher Quality Search Result Set	G1 , G2 , G3 , G4 , G5 , G6 , G7 , G8 , G9 , G10 (Google's top 10)
Experiment 3 Lower Quality Search Result Set	G1 , G11, G12, G2 , G3 , G4 , G5 , G6 , G7 , G8
Experiment 4 Higher Quality Search Result Set	G1, G12, G13, G14, G15, G16, G17, G18, G19, G20
Experiment 4 Lower Quality Search Result Set	G1, G12, G13, G14, G21, G15, G16, G17, G18, G19
Experiment 5 Higher Quality Search Result Set	G1, G2 , G3, G4, G5, G6, G7, G8 , G9 , G10 (Google's top 10)
Experiment 5 Lower Quality Search Result Set	Google's top 10 for query with 3rd word removed
Experiment 6 Higher Quality Search Result Set	(no details given)
Experiment 6 Lower Quality Search Result Set	(no details given)

Table 1: Descriptions of the search results for the six experiments (Gn denotes the nth search result returned by Google)

quality result set, had no preference, or preferred the lower quality result set. Again, for Experiment 6 this assignment was arbitrary so we just chose one of the two ranking algorithms being considered to be +1 and set the other as −1.

4.2 Discussion of Results: Experiments 1 - 5

Figure 2 shows a plot of the mean scores for owners and non-owners for each experiment. This analysis reveals that for Experiments 1-5 query owners did in fact demonstrate a stronger preference for the higher quality results than did the randomly assigned non-owners.

These results argue in favor of the idea that owners are better able to assess relevance than non-owners since they more often chose the higher quality set of results. We will examine the statistical significance of the difference between owners and non-owners in Section 6.

4.3 Discussion of Results: Experiment 6

As discussed earlier, Experiment 6 is an actual experiment concerning a particular aspect of the Google search algorithm that was being investigated at the time of writing this paper. The experiment previously had tested fairly negatively on human assessors (consistent with what is seen with the non-owner data in Figure 2) but relatively much more neutral on live traffic metrics. This discrepancy made the experiment a bit of a curiosity and raised questions about the extent to which human assessors could be trusted to represent real users for this particular type of experiment.

Since the owner assessments for Experiment 6 in Figure 2 also appeared to be quite neutral, this gave the experimenters increased confidence in the value of owner evaluations over those of non-owners. In other words, the owner data seemed to agree more with live traffic experiment data than the non-owner data.

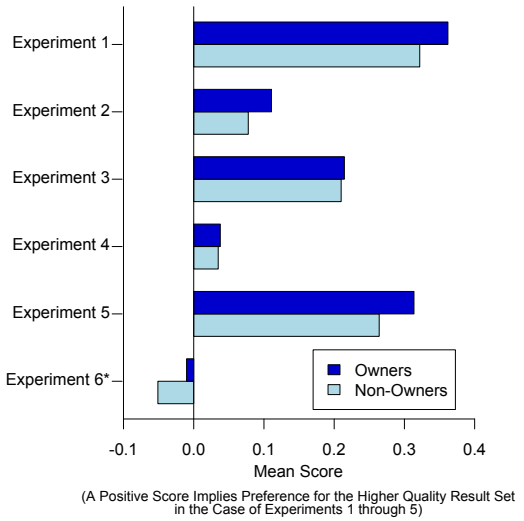


Figure 2: Comparision of owner and non-owner experiment mean scores for Experiments 1-6

Admittedly, we are not able to make this argument at all convincing to the reader since we are not revealing the details of the two search algorithms in Experiment 6 nor the accompanying live traffic experiment data. However, for the purposes of this paper, at the very least Experiment

6 represents an interesting example of how owner and non-owner evaluations can differ substantially for an experiment.

5. FURTHER ANALYSIS OF RESULTS

While it is clear in Figure 2 that the owner's scores are on average more positive than the non-owner's scores for Experiments 1 through 5, the mean alone does not tell us why this change is occurring. For example, two possible explanations are that (a) owners give roughly the same amount of -1's but fewer 0's or (b) owners give fewer -1's but roughly the same amount of 0's. These two explanations would suggest different underlying effects for query ownership.

To address this question, we analyzed histograms of the assessors -1, 0 and $+1$ scores. Interestingly, no single pattern emerged and in fact we observed slightly different trends from one experiment to the next. For example, Figure 3 shows the distribution of scores for owners and non-owners in the cases of Experiments 1 and 5. Experiment 1 seems to be an instance of (a), and Experiment 5 better matches (b).

6. STATISTICAL INFERENCE

In this section we investigate the question of the statistical significance of the owner versus non-owner differences seen in Figure 2. In other words, we examine whether the differences between owner and non-owner experiment mean scores are too large to be attributed simply to variability in human assessment. We will conclude that indeed the differences are statistically significant for Experiments 1, 2, 5 and 6 but not for Experiments 3 and 4. This would be a logical guess after looking at Figure 2, but the purpose of this section is to formally validate this.

To verify statistical significance, we will construct 95% confidence intervals for the differences in the mean scores of owners minus non-owners. By definition, if a 95% confidence interval does not include the value zero, then we have established statistical significance (at the $\alpha = .05$ level). We compute these 95% confidence intervals for each of the five experiments as

$$\text{mean of score differences} \pm 1.96 \cdot \frac{\text{s.d. of score differences}}{\sqrt{\text{number of queries}}}.$$

In this expression, the "number of queries" refers to the total number of queries in each experiment for which we have obtained an assessment from the query owner and at least one random assessor. We discard queries in the confidence interval calculation for which this is not true. The value of "score differences" is obtained by subtracting the mean of the random assessor scores for each query from the (single) owner score. Using that quantity, the values of "mean of score differences" and "s.d. of score differences" are obtained by taking the mean and standard deviation respectively.

These confidence intervals are displayed in Figure 4. We note that for Experiments 1, 2, 5 and 6 these confidence intervals do not include zero, thus establishing the statistical significance.

7. SAMPLE SIZE IMPLICATIONS

In this section we discuss some of the implications that our findings have for the number of assessors (sample size) required in human evaluation experiments of web search engine quality. We show that by collecting assessments from

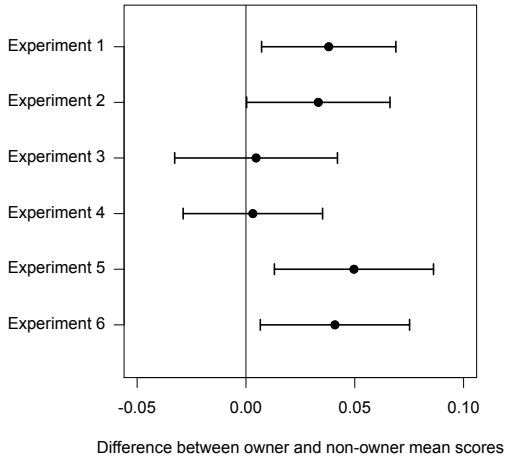


Figure 4: 95% confidence intervals for the difference between owner and non-owner mean scores

query owners instead of non-owners we are able to determine whether an experiment reflects a (statistically significant) positive or negative change to the search engine with fewer assessors than we would otherwise require. Depending on how large an effect the given experiment has on the quality of the search engine, the difference in the number of assessors required to detect it can be very large.

For the purpose of this discussion, we assume that the human evaluation experiment is conducted in an attempt to determine whether a particular change has a positive or negative mean effect on the search engine. To simplify the analysis, we will further assume from here on that we collect exactly one owner and one non-owner assessment per query. Note that the non-owner datasets for our experiments do not follow this assumption; however, it is valuable to consider this design since it turns out to be the most efficient (for a fixed budget of assessments) in addition to being the most simple to analyze.

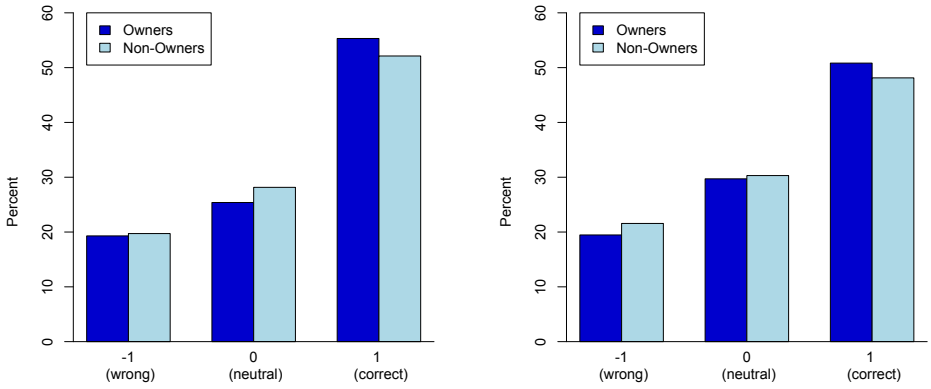
The main question which we seek to address can be stated as follows: How many assessors (or equivalently, how many queries) are required to determine (with a fixed statistical significance level) whether a particular change to the search engine has a positive or negative effect, and how does this quantity depend on whether the assessors are query owners or non-owners?

7.1 The 5 Experiments

Before addressing this question in generality, we demonstrate how it can be answered in each of the first five experiments we conducted. For the purpose of simplifying this analysis, we downsample our dataset such that we retain exactly one owner assessment and one non-owner assessment per query.

Figure 5 shows a plot of confidence intervals for the owner and (downsampled) non-owner data for Experiments 1 to 5. These confidence intervals are computed as

$$\text{mean score} \pm 1.96 \cdot \frac{\text{s.d. of scores}}{\sqrt{\text{number of assessors}}}.$$



(a) Experiment 1

(b) Experiment 5

Figure 3: Histograms of scores for Experiments 1 and 5

Note that we use "number of assessors" here as opposed to "number of queries". These two counts are of course equivalent as a result of our downsampling, but we prefer to use the former in order to emphasize that the main cost is the assessment as opposed to the query sampling.

Since all 10 of the confidence intervals lie entirely above zero, each of the experiments is sufficient for us to conclude that the given change had a positive effect on the search engine regardless of whether we use owner or non-owner assessments and even in spite of the downsampling. This is a consequence of our large sample sizes for these experiments. However, for smaller sample sizes the width of the confidence intervals will widen and the sample size benefits of using owner assessments will become clear.

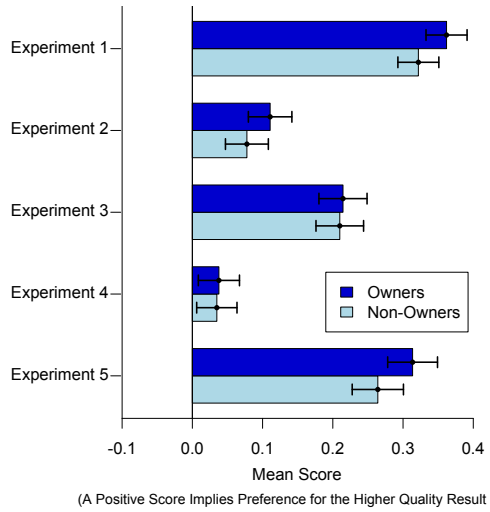


Figure 5: 95% Confidence intervals for owner and non-owners from the downsampled data

In particular, based on the data we collected we are able to deduce the minimal number of assessors that would be required in order to obtain statistical significance. We do so by computing the mean scores and the standard deviations of the scores for the owners and non-owners in each experiment, and then determining the minimal number of assessors for which the corresponding 95% confidence interval lies above 0. The minimal number of assessors required for statistical significance for each experiment is tabulated in Table 2. We can see in Table 2 that for both Experiments 2 and 5 the sample size needed for statistical significance is reduced by more than 40% as a result of using owner assessments.

	Owners	Non-Owners	% Reduction
Experiment 1	19	24	21%
Experiment 2	191	322	41%
Experiment 3	60	60	0%
Experiment 4	1381	1545	11%
Experiment 5	24	43	44%

Table 2: The minimal number of assessors needed to establish statistical significance for Experiments 1 through 5

7.2 General Case

In this section we generalize our observations from the five experiments by identifying the features of the experiment which are most relevant in determining how many assessors are required to detect an effect. A common feature of confidence intervals, regardless of their construction, is that their width decreases at a rate proportional to the square root of the number of assessors used in the study. That is to say they take the form

$$\text{mean score} \pm \frac{C}{\sqrt{\text{number of assessors}}}$$

where C is some constant which typically depends on the desired confidence level and the variability of the data. In our analysis above, we took $C = 1.96 \cdot \text{s.d.}$ of scores.

From this expression for the confidence interval, we can deduce that in order to detect that the given experiment

had an effect, we require that

$$\text{number of assessors} > C^2/\text{mean score}^2.$$

In particular, if the experiment results in only a minor change, corresponding to a small mean score, we require a very large number of assessors.

Assuming that an experiment truly is positive (as is the case with our Experiments 1 through 5), we saw in our data that the owners consistently had higher mean scores than non-owners. Pairing that fact with the above expression for the minimal number of assessors reveals why using owners instead of non-owners in human evaluation studies requires fewer assessors in order to determine whether a given change to the search engine did in fact have a positive effect. How much fewer depends on the difference between owner and non-owner assessments for the given experiment.

In general we can state that the minimum number of assessors required to detect a change depends primarily on two quantities: the severity of the change induced by the experiment (experiment mean score) and the difference between owner and non-owner mean scores. In our Experiments 1 to 5, the experiment mean score ranged from 0.03 to 0.37, and the difference between owner and non-owner mean scores ranged from 0.003 to 0.049.

Figure 6 helps to illustrate how the minimum number of assessors required to detect a change depends on the experiment mean score for three values of owner minus non-owner mean differences within our observed range. These curves are plotted by taking C equal to 1 for simplicity so that the required sample size for the non-owners becomes $1/\text{mean score}^2$ compared to that for the owners which is $1/(\text{mean score} + \Delta)^2$ where Δ denotes the difference between the owner and non-owner mean scores. The three plots correspond to owner minus non-owner mean score differences of $\Delta = 0.005$, 0.010 and 0.040 which are within the range of our observed values (0.003 to 0.049 as mentioned above.) The x-axis in all three plots ranges from 0.03 to 0.37 which matches the range of our observed experiment mean scores. These plots show that the sample size savings from using owners' assessments can be quite substantial for small experiment mean scores. This is especially true for large values of Δ such as $\Delta = .040$ similar to what was observed in Experiments 1 and 5.

8. CONCLUDING REMARKS

The results of this paper show that in general owners provide more accurate assessments than non-owners with regard to choosing higher quality result sets. The magnitude of the difference between owners and non-owners varies depending on the experiment, and can be quite substantial in some cases. One benefit from the more accurate assessments is a reduction in the sample size needed to obtain statistical significance for experiment mean scores.

Despite the desirability of using owner assessments over non-owner assessments, there are a number of practical reasons which make this difficult. We mention three of these below.

First, as mentioned earlier, it is not possible to sample queries from the search engine logs and subsequently locate the corresponding users who issued the queries. Rather, some method of sampling users who are willing to do assessment and subsequently eliciting queries from these users must be employed. We do not pretend to have developed a good method for doing this.

Second, once a set of queries with owners is established by some means, there is the additional problem that some of the assessors will eventually drop out of any long-term study. Thus, maintaining a query set with current relevance assessments over any extended period of time becomes difficult. Either the query set will constantly decrease in size or "replacement" owners must be found for queries which have been orphaned.

Finally, when eliciting queries from assessors, it is difficult to ensure and control representativeness of the query sample. Some analysis comparing the 23,530 queries used in this paper to a random sample from Google's US query logs suggests that there are some pronounced differences between the nature of the queries in the two samples. For example, the rate of spelling errors in our 23,530 queries is roughly half of that in a random sample from the Google US query logs. (The spelling error rate here was computed using the triggering of Google's "Showing results for" spelling correction.) One possible explanation for this large difference is that when visiting http://www.google.com/history to select queries, many misspelled queries are often immediately followed in the user session by a correctly spelled query (either corrected by the user or by Google). Since the assessor selects only one query each time, it is plausible the assessors gravitated toward the correctly spelled query. Note that when sampling from query logs it is possible to programmaticly choose the first query, last query or a random query from a session; whereas, when asking assessors to volunteer queries it is somewhat difficult to control this type of selection.

While it will never be as simple to use owner assessors as it is to use non-owner assessors, we believe this is a technique that will become more popular in the future. We believe the main driving force behind this will not necessarily be the increased accuracy illustrated in this paper, but rather this increased accuracy will be a beneficial byproduct and that the use of query owners as assessors will become necessary as search engines continue to personalize results based on a user's location, query history, social graph and other data. The result sets considered in the experiments in this paper were by construction unpersonalized, and it is interesting to imagine the equivalent experiments with results personalized for the owners to better match the true user experience. For such personalized results, the difference between owner assessments and non-owner assessments would be even more substantial, perhaps to the extent that non-owner assessments become completely useless for many experiments. We note that the research on personalization techniques already follows the practice of using owner assessments. Examples include [11] and [15].

9. ACKNOWLEDGMENTS

The authors would like to thank Daniel Russell and Ya Xu for their help with the preparation of this paper.

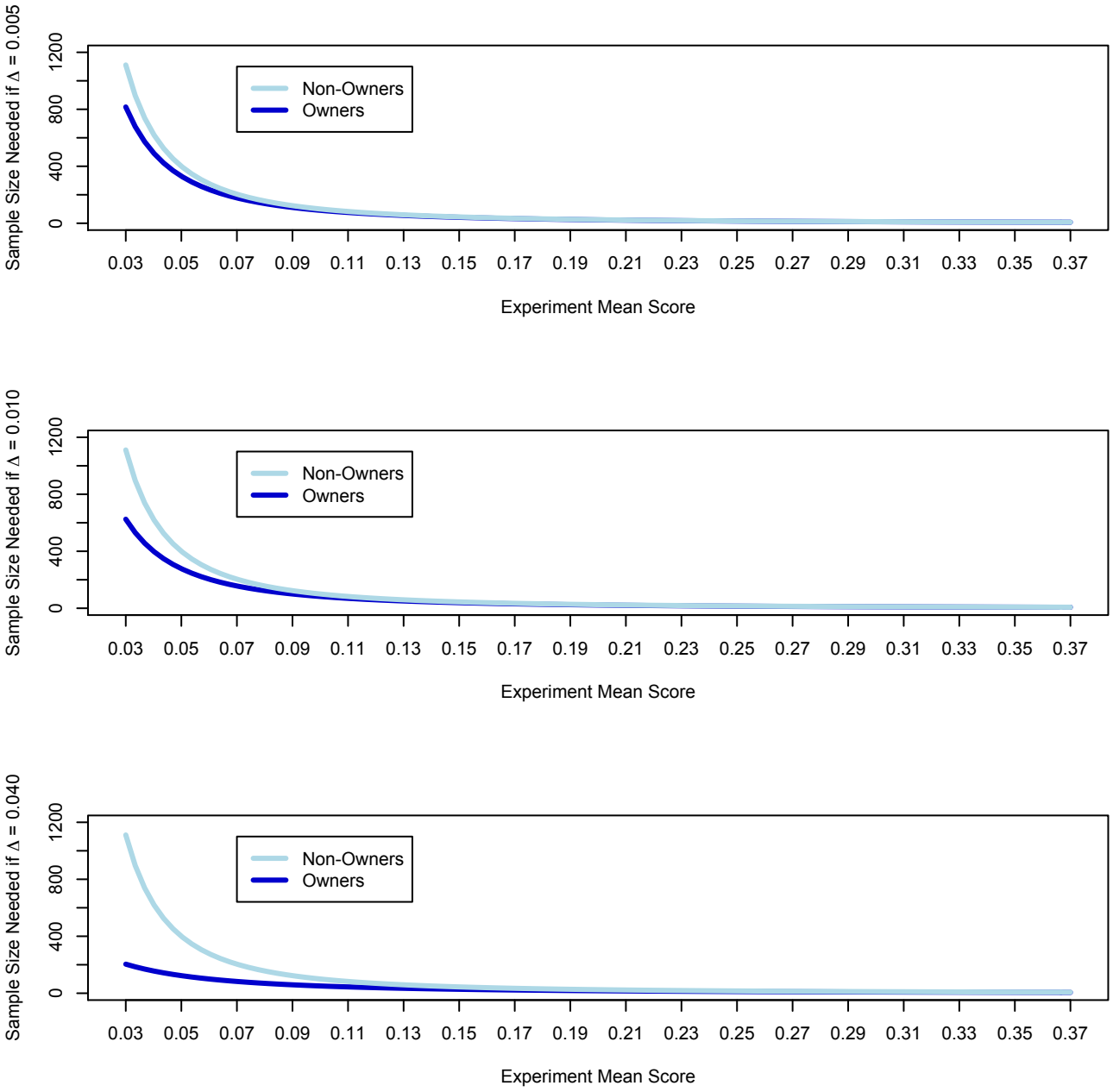


Figure 6: Comparison of the functions $1/(\text{mean score} + \Delta)^2$ and $1/(\text{mean score})^2$ for $\Delta = 0.005$, 0.010 and 0.040 respectively to illustrate the difference between owners and non-owners with respect to the minimal sample sizes required for statistical significance

10. REFERENCES

[1] O. Alonso and S. Mizzaro. Can we get rid of trec assessors? Using mechanical turk for relevance assessment. In *SIGIR '09: Workshop on The Future of IR Evaluation*, 2009.

[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 *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 667–674, New York, NY, USA, 2008. ACM.

[3] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM.

[4] G. B. Duggan and S. J. Payne. Knowledge in the head and on the web: using topic expertise to aid search. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 39–48, New York, NY, USA, 2008. ACM.

[5] D. Harman. Overview of the first trec conference. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 36–47, New York, NY, USA, 1993. ACM.

[6] C. Hölscher and G. Strube. Web search behavior of internet experts and newbies. *Comput. Netw.*, 33:337–346, June 2000.

[7] S. B. Huffman and M. Hochster. How well does result relevance predict session satisfaction? In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 567–574, New York, NY, USA, 2007. ACM.

[8] K. A. Kinney, S. B. Huffman, and J. Zhai. How evaluator domain expertise affects search result relevance judgments. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 591–598, New York, NY, USA, 2008. ACM.

[9] E. Law, P. N. Bennett, and E. Horvitz. The effects of choice in routing relevance judgments. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 1127–1128, New York, NY, USA, 2011. ACM.

[10] A. A. Maskari, M. Sanderson, and P. Clough. Relevance judgments between TREC and Non-TREC assessors. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 683–684, New York, NY, USA, 2008. ACM.

[11] N. Matthijs and F. Radlinski. Personalizing web search using long term browsing history. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 25–34, New York, NY, USA, 2011. ACM.

[12] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *International Conference on Information and Knowledge Management*, pages 43–52, 2008.

[13] D. M. Russell and C. Grimes. Assigned tasks are not the same as self-chosen web search tasks. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, HICSS '07, page 83, Washington, DC, USA, 2007. IEEE Computer Society.

[14] A. Singla, R. White, and J. Huang. Studying trailfinding algorithms for enhanced web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 443–450, New York, NY, USA, 2010. ACM.

[15] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 449–456, New York, NY, USA, 2005. ACM.

[16] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf. Process. Manage.*, 36:697–716, September 2000.

[17] B. M. Wildemuth. The effects of domain knowledge on search tactic formulation. *J. Am. Soc. Inf. Sci. Technol.*, 55:246–258, February 2004.