An Analysis of Systematic Judging Errors in Information Retrieval

Gabriella Kazai[2] Nick Craswell[1] Emine Yilmaz[2] S.M.M. Tahaghoghi[1]
[1]Microsoft, Bellevue, WA, USA
[2]Microsoft Research, Cambridge, UK
{v-gabkaz,nickcr,eminey,stahagh}@microsoft.com

ABSTRACT

Test collections are powerful mechanisms for evaluation and optimization of information retrieval systems. There is reported evidence that experiment outcomes can be affected by changes in the judge population or in judging guidelines. We examine such effects in a web search setting, comparing the judgments of four groups of judges: NIST Web Track judges, untrained crowd workers and two groups of trained judges of a commercial search engine. Our goal is to identify systematic judging errors by comparing the labels contributed by the different groups. In particular, we focus on detecting systematic differences in judging depending on specific characteristics of the queries and URLs. For example, we ask whether a given population of judges, working under a given set of judging guidelines, are more likely to overrate Wikipedia pages than another group judging under the same instructions. Our approach is to identify judging errors with respect to a consensus set, a judged gold set and a set of user clicks. We further demonstrate how such biases can affect the training of retrieval systems.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation (efficiency and effectiveness)*

General Terms

Experimentation, Measurement, Performance

Keywords

Relevance labeling, assessor errors, crowdsourcing

1. INTRODUCTION

In information retrieval (IR), test collections, comprising documents, topics and human-generated relevance judgments, are typically used to evaluate and optimize the performance of IR systems [21]. To build a test collection that reflects real users, it is possible to select a corpus and a sample of queries that is representative of the real-world usage of some IR system, such as a Web search engine. The more difficult aspect is to design a relevance judging procedure that can lead to relevance labels that reflect real-world user satisfaction. Arguably, this requires maximizing the judges' understanding of the real users' information needs. For example, faced with the query *shania twain*, an assessor may reason that the user is looking for the Wikipedia page on the celebrity, while the real user need may have been to go the star's official site or to read the latest news or celebrity gossip, or vice versa.

Traditionally, relevance judging methods rely on a fixed group of expert judges, who are trained to interpret user queries as accurately as possible and label documents accordingly. However, the growing volume and diversity of the topics and documents to be judged is driving an increasing adoption of crowdsourcing methods, where relevance judgments are distributed over a large population of humans, a *crowd* [11]. Crowdsourcing platforms, such as Amazon's Mechanical Turk, allow thousands of crowd workers to be hired temporarily for micro-payments to complete small human intelligence tasks (HITs), such as labeling images or documents. Clearly, such a model is in stark contrast to the highly controlled methods that characterize the work of trained judges. For example, in a micro-task based crowdsourcing setup, worker training is usually minimal or non-existent. Furthermore, it is widely reported that labels provided by crowd workers can vary in quality, leading to noisy labels [9, 12, 1]. On the other hand, it may be argued that training procedures and judging guidelines may actually lead to biased judgments, e.g., consistent overrating of some document types, like Wikipedia pages. Untrained judges, as a group, are less likely to demonstrate such biases. Another potential benefit of crowdsourcing is the diversity of its workers that may better reflect that of real users in Web search.

Judging errors and their impact on the outcome of an evaluation experiment has been the subject of extensive study in the past, while the growing reliance on crowdsourcing prompted a renewed interest in the topic. For example, low levels of inter-assessor agreements are often reported both for trained judges and crowd workers [19, 1]. Furthermore, evaluation outcomes have been shown to be affected when using different judge populations [2, 14] or different judging guidelines [6].

In this paper, we study differences between four groups of relevance judges: NIST Web Track judges from 2009 and 2010, crowd workers, and two different groups of trained

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

judges of a commercial Web search engine. We characterize and compare the judge groups according to their inter-assessor agreement, their agreement with click-based gold sets and with label-based gold sets. Rather than overall judge agreement, much of our analysis focuses on characterizing differences with respect to three specific types of judgments: Judgments of Wikipedia pages, judgments where the query text matches the URL text, and judgments where the URL is the root page of a site (and thus likely a homepage). This allows us to study, for example, whether crowd judges or trained judges are more likely to overrate or underrate Wikipedia pages with respect to a given gold set. Future work might extend to many more than these three types, automatically identifying the most important types of judgment where judge groups differ, and then making measurable improvements in judging procedures. Our current scope is to develop analysis methods that could be used, for a given type of judgment, to identify systematic differences between judge groups. In addition, we study the impact that systematic errors can have on the training of a supervised ranker, demonstrating that systematic differences in labeling can lead to different learned rankers.

2. RELATED WORK

Relevance judgments are the core of most IR experiments. This has led to numerous studies of the judgments themselves and their impact on the evaluation experiments. For example, it is well known that human judgments are subjective, influenced by various situational, cognitive, perceptual and motivational biases [3, 16]. Indeed, a wide range of factors that can impact on the variability of relevance judgments have been identified in the literature [17], including, for example, document variables, judgment conditions and scales, and personal factors [7].

Despite the subjective nature of relevance and the judgment disagreements at the query-document level [19, 1], the study by Voorhees showed that the outcome of experiments, in terms of system orderings, was robust against changes in the group of assessors [19]. However, in her study, all assessors were from the same population, i.e., all judges were trained on the same task, using the same guidelines. In contrast, evaluation outcomes have been shown to be affected when using different judge populations [2, 14], different judging guidelines [6, 10], simulated noisier judgments [5] or even taking the earlier vs. the later judgments in cases of intra-judge inconsistency [18]. For example, the study by Bailey et al. [2] compared three groups of judges: Gold judges who were topic originators and domain experts, silver judges who were domain experts only, and bronze judges who were neither experts nor topic originators. They found that switching between judge types resulted in changes to the system ordering. A similar comparison of expert and non-expert judges in [15] showed that non-expert judges tended to give shallow, inaccurate ratings compared to experts. Non-experts also disagree on the underlying meaning of queries significantly more often than experts, and often appear to "give up" and fall back on surface features such as keyword matching.

Within a single judge population, there is some evidence that changing the judging guidelines can change the evaluation outcome. Voorhees [20], for example, instructed judges to label on more than a binary scale, by also identifying highly relevant documents and the best document for each query. The relative effectiveness of the evaluated runs then differed depending on how the different levels were used or collapsed in the evaluation, highlighting the effects of the instructions on the number of levels and their definitions. The judging instructions at the TREC Web Track have also changed over the years reflecting the varying usefulness of link evidence depending on the task [10], e.g., informational adhoc judgments vs. navigational homepage finding judgments. In this case, the different tasks had different judging instructions but also used different query sampling methods, so some combination of these factors led to the relative differences in the performance of systems with and without link evidence.

With no change in the judge population or in the instructions, the work by Scholer et al. [18] demonstrated that the conclusions of an evaluation experiment can change by simply using judgments made at different times by the same judge. Webber et al. [22] studied the impact of judging errors on estimates of recall, where assessment pool sampling is stratified based on the results of the systems participating in the evaluation. They showed that assessor error is in general a greater source of inaccuracy than sampling error.

Our research contributes to the above studies by investigating the impact that changes in the population of judges or changes in the judging guidelines can have on the quality of the resulting relevance labels. We study both trained judges and untrained crowd workers, working under the same or different instructions and compare their judgments for documents with different characteristics. We measure label quality in terms of agreement with consensus, a gold set of labels and user clicks. This holistic approach allows us to analyze which types of errors are more likely by the different judge groups, to characterize their disagreements and, in some cases, to identify systematic biases.

3. TYPES OF JUDGING ERRORS

For the purposes of this study a judging error is a disagreement with some reference data, for example a gold set. This section describes some sources of judging error, what we mean by systematic bias in judgment and the three examples of bias that we use in our experiments.

Due to its subjectivity, the assessment of relevance can be particularly challenging by a third party. To mitigate such issues, assessors at TREC are often involved in the creation of the test topics. In addition, they are subjected to training, judging procedures and guidelines, aiming to aid them in providing consistent judgments. Measured in terms of consistency, label quality is thus predicated on the judges accurately understanding and conscientiously following a common set of guidelines. In practice, however, this is difficult to achieve and judges often make mistakes. Based on the analysis of a large sample of judgments collected from trained judges of a commercial search engine, we list some of the most commonly observed reasons for judging errors:

Intent mismatch: Judges may not understand the intent of a real user. For example, the query *circuit breaker* could relate to the electrical device, or to a program of this name run by the Illinois Department on Aging. On a 5-point scale, the one that is rated higher by a judge may not be the one that users would prefer. Beyond the ambiguity of a query, intent considerations may include aspects whether the user may prefer

a lay overview of a medical condition over specialist content; or the latest gossip about a celebrity over a peer-reviewed biography on Wikipedia.

Content quality: Aspects of quality such as the correctness or authoritativeness of a web page can be important for certain queries. For example, user-contributed song lyrics may contain errors; much online medical advice is questionable; and some sites simply copy or aggregate content from other sources. The challenge of assessing the quality of a page can be further exacerbated by a mismatch in the impression of quality made by the design of the page; "content farms" exploit this by framing often inane content on well-laid out pages with relevant keywords.

Judge biases: Judges may develop biases based on the training examples that they have seen, and then base their assessment on their own inferred rules. Judges may also have intrinsic biases and may rate too harshly or too generously, or keep their ratings within a narrow band of the rating scale.

The above is not intended as a comprehensive or dichotomous classification of judging errors. What is of interest to us is when judges make systematic errors in their judgments for any of the above reasons. For example, judges may consistently underrate navigational results in favor of results that better fit the informational interpretation of the query. We consider such consistent judging mistakes as examples of systematic error or bias, where an error means a disagreement with a given ground-truth, i.e., consensus or gold label. Moreover, we are not interested in the individual bias of a particular judge, but the systematic errors of a group of judges with common characteristics, e.g., trained judges working under a given set of instructions. In order to detect the presence of such systematic errors among a group of judges, we propose a method that compares across different groups of judges and judging conditions, i.e., trained vs. untrained judges, working under the same vs. different guidelines, measured against a diverse set of baselines, i.e., consensus, gold set labels and user clicks. To make our investigation feasible, we focus on three specific cases that may be prone to such judging errors. These are simple to calculate based on properties of the query-URL pairs, and have been observed as potentially interesting cases in a commercial search judging setting and/or at TREC [10, 13]:

Wikipedia bias: Can we see systematic errors in judges' treatment of Wikipedia pages? To answer this question we will compare labels by the different judge groups and guideline conditions across two subsets of the judged documents: Wikipedia (**WP**) vs. non-Wikipedia (**nWP**) pages. When the URL of the judged document contains the string "en.wikipedia.org", we classify it as a WP page and as an nWP page otherwise.

URL query bias: Can we find systematic errors in the judgment of documents whose URL contains the query words? Such cases may be indicative of a more navigational user intent, searching for a known site. Again, our approach is to split the judged corpus based on this document feature and compare the labels across the two subsets. When the query words appear in the URL (in the same sequence) then we tag the topic-document pair as **QiU** (short for "Query in URL") and with **nQiU** otherwise. For example, the query espn sports occurs in the URL `http://en.wikipedia.org/wiki/ESPN_Sports_Poll`, so this pair is tagged as QiU.

Homepage bias: Can we detect a systematically different judging of homepages in different judging conditions? Assuming that homepages are more likely to have root URLs that only contain domain information without a path, we split the corpus into root-URL (**RU**) and not-root-URL (**nRU**) subsets.

4. EXPERIMENT DESIGN

Our goal is to identify systematic judging errors by comparing the labels contributed by the different groups, working under the same or different judging guidelines. More specifically, we want to investigate if we can observe different types of systematic judging mistakes, e.g., a consistent overrating of homepages by one group of judges – as described in the previous section – in the different judging conditions.

In order to analyze the effects of a change in the guidelines, our case study is the ad hoc task of the TREC Web Track. In 2009, the track used three judging levels: Not Relevant (0), Relevant (1) and Highly Relevant (2). Non-English documents were to be judged non-relevant, and assessors were instructed that pages with misleading or malicious content should also be considered non-relevant. In 2010, in an effort to better support the assessment of queries with navigational intent, and to separate spam results from non-relevant, the Track adopted the following 5-level judging scale:

Nav (3) This page represents a homepage of an entity directly named by the query; the user may be searching for this specific page or site, e.g., `www.yahoo.com` for the query *yahoo*.

Key (2) This page or site is dedicated to the topic; authoritative and comprehensive, it is worthy of being a top result in a web search engine.

Rel (1) The content of this page provides useful information on the topic.

Non (0) The content of this page does not provide useful information on the topic.

Junk (-2) This page does not appear to be useful for any reasonable purpose; it may be spam or junk.

In order to study the effect of a changes in the judge population, we collect relevance labels from both trained and untrained judges. For our untrained judge group we recruit crowd workers – from the US only – through Clickworker, a labor market similar to Amazon's Mechanical Turk. To participate, workers are first qualified via a basic Web judging task, but receive no training. We will refer to these judges as "Crowd". For our trained judges, we recruited two groups of professional assessors, 10 in each group, employed and trained by a commercial search engine. One group is trained in Web search judging (we will call them "ProWeb") and thus would be aware of aspects of search engine usage, for example the importance of navigational results. The other group is trained for a different judging task, not for Web search ("ProNonWeb"). The NIST judges of the Web Track 2009/2010 campaigns comprise the fourth judge group.

In our analysis we identify judging errors with respect to a consensus based ground-truth set, a judged gold set and

a set of user clicks, extracted from the query log of a commercial web search engine. In the following, we describe the different datasets we use for analysis in more detail.

4.1 Experiment Data

4.1.1 WT (Web Track) set

We designate the Web Track ad hoc task relevance judgments provided by the judges hired and trained by NIST as our **WT** (Web Track) set, with WT 2009 and WT 2010 identifying the labels collected for a given year of the track. This set consists of nearly 50k relevance labels for 100 topics[1], roughly 25k labels per year. In this set we only have a single label per topic-document pair. As discussed in the previous section, the labels in 2009 span a 3-level relevance scale, and a 5-level scale in 2010. We note that while the document corpus used in these two years of the track was the same, the sample of documents assessed by NIST judges may have been different, which could have given rise to collection bias. There may also have been differences in the query sampling procedure.

4.1.2 ST (Sampled TREC) set

In addition to the WT set, we collected relevance labels from both trained and untrained judges. We first took the 100 Web track 2009/2010 topics and using these as queries we scraped the top 10 search results from Google and Bing for each query. This gave us a total of 1,603 unique query-URL pairs over the 100 topics. We chose this method of re-sampling documents for the TREC topics, instead of relabeling a sample of the WT set, in order to ensure up to date coverage of the topics and high overlap with the query-document pairs that appear in the query logs of the commercial search engine, which we aim to use in our analysis later on. To collect relevance labels for our sample of documents for the 100 Web track queries, we designed a simple judging interface that showed judges a query and a web search result (in an iframe) and asked judges to rate the search result's usefulness to the query using a five-point scale[2]:

Ideal (4) User seeing this result would think it ideal.
Very happy (3) User seeing this result would be very happy.
Happy (2) User seeing this result would be happy.
Somewhat happy (1) User would be somewhat happy.
Unhappy (0) User seeing this result would be unhappy as this is a useless result for the current search.

We collected relevance labels from our Crowd, ProWeb and ProNonWeb judge groups in separate experiments using the same judging interface. While we recruited 20 professional judges (10 ProWeb and 10 ProNonWeb), our Crowd judge group is made up of the 45 US workers who chose to work on our HITs on Clickworker (workers were pre-filtered by the platform). We gathered up to 3 labels per query-document pair, per judge group. We stopped at 2 labels in case of agreement, or collected a third label when the first two disagreed. We obtained a total of 12,879 relevance labels across the three judge groups. We designate the obtained labels set as our **ST** (Sampled TREC) label set.

[1]The actual total number of topics with NIST judgments is 98; topics 95 and 100 have no labels.

[2]Judges could skip a query-document pair (by selecting the "I can't tell" option). They could also research a query by viewing the top 10 results from Google and Bing.

4.1.3 HQ (Hard Queries) set

Finally, we collected relevance labels from trained Web judges and untrained crowd workers for a hand-picked set of 689 query-URL pairs, sampled from the commercial search engine's query log, that is representative of the types of judging errors we described in Section 3. This set has of a total of 74,530 labels, 72,463 of which was contributed by 193 professional Web judges (average 375 labels per judge or 105 labels per query-URL pair) and 2,067 labels by 18 crowd workers (115 labels per judge, 3 labels per query-URL pair). We will call this data our **HQ** (hard queries) set.

The relevance data and judge groups used in the experiments are summarized in Table 1.

Table 1: Datasets and judge groups

Set	Queries	Query-URLs	Labels	Judge groups
WT	98	48,501	48,917	NIST judges
ST	100	1,603	12,879	10 ProWeb; 10 ProNonWeb; 45 Crowd
HQ	665	689	74,530	193 ProWeb; 18 Crowd

4.1.4 Click data set

The click log data set contains raw click count information for sets of query-URL pairs from 18 months of user traffic for a commercial Web search engine, with a minimum threshold of at least 10 clicks. 1,583 of the 1,603 unique query-URLs in our ST dataset and 1,367 of the 48,501 query-URLs in the WT set have click information in the logs (1,231 in common across ST and WT). As the HQ set contains mostly tail queries, we have no corresponding click data.

4.2 Analysis Methodology

When collecting multiple ratings from judges, a consensus label may be derived, e.g., using simple majority vote or methods that take into account judge reliability [8]. However, it is not always feasible to rely on consensus; with a semi-skilled pool of judges, incorrect rating may be the rule rather than the exception. Hence, it is important that ratings provided by the judges be not only internally *consistent*, but also *correct*. One way to assess correctness is to sample the labels and evaluate each one against a known "gold" rating. The generation of gold sets is however expensive and it may not be practical to test each judge on all the gold data. Alternative gold sets may be derived from click evidence. For example, one can create pairs of URLs for a query with known click preference relations and then check if the judgments agree with the click preference.

In our analysis we use all of these methods, i.e., consensus and gold sets based on labels and clicks, in order to identify systematic errors in the judging behavior of the four different groups of assessors: NIST, ProWeb, ProNonWeb, and Crowd. We start with the consensus based analysis in the next section.

5. CONSENSUS ANALYSIS

5.1 Inter-assessor Agreement Analysis

In Table 2, we report values of Fleiss' kappa, a well-known technique that estimates the degree of consensus between n

judges by correcting for agreement by chance alone, calculated over the labels contributed by the different groups of judges in our ST set.

Table 2: Fleiss' kappa agreement levels per judge group and per feature for the ST set (p<0.001, Q=unique queries, QU=unique query-URL pairs)

SubSet	#Q	#QUs	Labels	Kappa
Crowd	100	1603	4442	**0.33**
WP	78	138	401	0.28
nWP	100	1465	4041	0.34
QiU	71	690	1917	0.32
nQiU	96	913	2525	0.34
RU	89	472	1281	0.35
nRU	99	1131	3161	0.33
ProWeb	100	1603	4047	**0.59**
WP	78	138	365	0.51
nWP	100	1465	3682	0.60
QiU	71	690	1721	0.63
nQiU	96	913	2326	0.56
RU	89	472	1195	0.54
nRU	99	1131	2852	0.61
ProNonWeb	100	1603	4390	**0.33**
WP	78	138	390	0.26
nWP	100	1465	4000	0.33
QiU	71	690	1891	0.33
nQiU	96	913	2499	0.33
RU	89	472	1304	0.27
nRU	99	1131	3086	0.35

It is easy to see that professional Web judges have a much higher level of within-group agreement than crowd workers (κ of 0.59 for ProWeb vs. 0.33 for Crowd). This is even more evident in the HQ set (not shown), where a ProWeb agreement level of $\kappa = 0.57$ is matched only by $\kappa = 0.16$ for the Crowd. Interestingly crowd workers do as well as professional judges who are not trained in Web search judging ($\kappa = 0.33$ for both groups). This can be taken as evidence in support of the need for judge training and calibration.

Although disagreement analysis is informative, it may well be that not all disagreements are bad. In some cases disagreements among judges may reflect the diverse opinions of real-world users. For example, for the query *gps*, we obtain an average σ of 0.68 among the judgments of the Crowd per judged URLs, compared with $\sigma = 0.14$ for ProWeb and $\sigma = 0.47$ for ProNonWeb. The URLs for this query include the Wikipedia page on GPS, the `www.gps.gov` and various shopping result pages. Similarly, for the query *kcs*, judges may rate the homepage of the Kansas City Southern Railway and the related Wikipedia page differently. Thus, the diversity of the judgments can reflect the diversity of the associated user needs.

With the aim to gain deeper insights into judging disagreements, we calculate agreement levels over subsets of the data set, divided based on one our analysis features, e.g., Wikipedia, query URL, homepage. We see that inter-assessor agreement is lower for Wikipedia pages across all judge groups (0.28, 0.51, 0.26 for Crowd, ProWeb, ProNonWeb, respectively), compared with nWP judgments (0.34, 0.6, 0.33, resp.). This suggests that Wikipedia pages are inherently more controversial and are harder to judge consistently.

Dividing the sets of relevance labels based on whether the query words occur in the URL, we observe different behavior by the different judge groups. ProNonWeb judges show the least influence of this property on their judgments ($\kappa = 0.33$ for both QiU and nQiU). Crowd judges agree slightly more in the nQiU subset (with 0.02 difference in κ). However, most strikingly, the ProWeb judges show much better agreement when the query does appear in the URL ($\kappa_{QiU} = 0.63$ while $\kappa_{nQiU} = 0.56$). The latter may be an indication of possible judging guideline bias for ProWeb judges, but it could equally signal a noise-reduction and calibration effect of the guidelines on the judges.

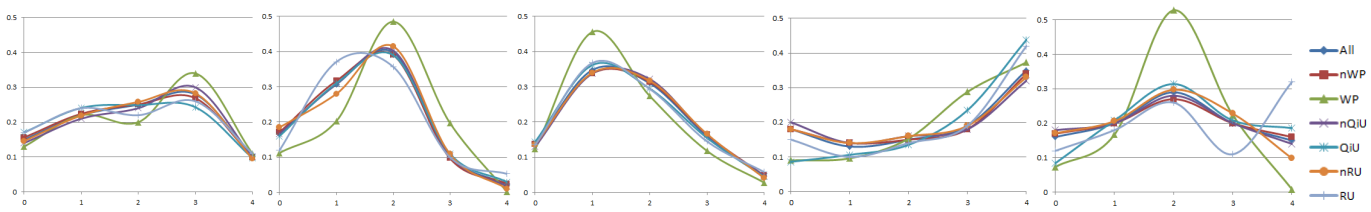
We also see inconsistent behavior between the judge groups on agreeing about the relevance of root vs. non-root URLs. Crowd workers agree slightly more in the RU subset (0.02 increase in κ over nRU), while both ProWeb and ProNonWeb judges have higher agreement levels on nRU (0.07-0.08 differences in κ). We found the same trends in the HQ set (not shown).

Based on the observations so far, we can conclude that judge training on a specific judging task does indeed increase assessors' judging consistency. Furthermore, we may argue that observed changes in judgment behavior across different datasets with different properties can signal healthy judging procedures, especially when the same changes are mirrored across different groups of judges. On the other hand, we also see that different groups of judges exhibit different behaviors on the same dataset. Can, in such cases, disagreement signal judging bias? Indeed, what is not clear from this agreement analysis is whether higher agreement actually reflects higher label quality. Can higher agreement simply be a result of skewed label distributions that are not necessarily correct? For example, a group of judges who are evaluated and paid based on their inter-assessor agreement may learn to optimize their judging by reducing the use of risky labels (e.g., adopting a middle of the road approach), thus skewing the distribution of labels. In the following, we aim to answer this question by comparing the judgments to gold labels.

5.2 Probability of a Label

Figure 1 shows the label probability distributions in the ST and HQ data sets for the different judge groups. We see very different label probability patterns across the groups. We observe evidence of judge calibration among the trained judges, whereby the top relevant grades are assigned very sparingly, while untrained judges tend to be more liberal. We see that professional judges are characterized by a skewed distribution, peeking at the middle grades (or below for ProNonWeb). Such a skew can mean that the 5 point scale is in essence reduced and can also boost agreement statistics. Crowd workers produce a more uniform distribution with a flatter bell shape on the ST set. Interestingly, crowd judgments show a top grade heavy distribution for the HQ set. This may be a result of the crowd's leniency and possible unawareness of a query's potentially diverse intents. For example, while trained judges may grade results that satisfy a less likely informational intent for a query lower than a navigational result, crowd workers may rate both types of results equally highly. This could be addressed by gathering labels using a HIT with multiple results per query.

Looking at Wikipedia vs. non-Wikipedia judgments, we can observe clear shifts in the label probability distributions for all three judge groups. Crowd judges show an in-



a) ST, Crowd b) ST, ProWeb c) ST, ProNonWeb d) HQ, Crowd e) HQ, ProWeb

Figure 1: Label probability distributions

creased probability level for grade 3 (shift from grade 2) for Wikipedia pages ($p(3) = 0.34$ for WP vs. 0.27 for nWP), while ProWeb judges have increased probability levels for both $p(2) = 0.48$ and $p(3) = 0.20$ for WP vs. $p(2) = 0.39$ and $p(3) = 0.10$ for nWP. Interestingly, ProNonWeb judges are more likely to assign lower grades to Wikipedia than to non-Wikipedia pages. These more skewed judging patterns for WP, combined with the observed lower inter-assessor agreement levels, suggest that the shift in judging patterns is not consistent across the individual judges and the documents. Unlike the judging patterns for the WP/nWP feature, no such clear patterns are observed for the QiU/nQiU feature across the judge groups. Ratings for homepages (RU) show some irregularities compared to the other patterns, especially on the HQ set. Paired t-tests over all feature pairs confirmed significant differences for WP/nWP judgments in both the ST and HQ sets and for all judges ($p_t < 0.001$, apart from ProNonWeb on ST with $p_t < 0.005$). On the ST set significant differences were found for QiU/nQiU judgments by the Crowd ($p_t < 0.001$) and ProNonWeb judges ($p_t < 0.05$), and for RU/nRU labels by ProWeb judges ($p_t < 0.001$). On the HQ set all feature pairs showed statistically significant judging patterns for both Crowd and ProWeb ($p_t < 0.001$).

So, thus far, we saw that trained judges are more reliable in terms of inter-assessor agreement, but that they show label probability distributions skewed toward the middle grades. The latter may be a result of the combined effects of a stricter interpretation of the value associated with high relevance grades (inherent or trained) as well as their intentional or unintentional optimization for agreement, a metric often used to evaluate judges. We also observed that Wikipedia pages receive spacial treatment by all groups, but lead to the lowest levels of inter-assessor agreement overall.

We conduct a similar analysis of the WT set, comparing NIST judgments from 2009 and 2010, but note that the observed differences in this case cannot be clearly attributed to changes in judging behavior due to the different query sets and pools in the two years. We find that NIST judges in 2009 were twice as likely to label a Wikipedia page as highly relevant than a non-Wikipedia page: $p(2) = 0.16$ for WP vs. 0.08 for nWP. In 2010, we observe mixed effects: the probability of a WP being labeled as Key (2) is 0.09 vs. 0.03 for nWP. However, the probability of a WP labeled as Rel (1) is only 0.13 compared with 0.16 for nWP. The probability of a WP result being irrelevant has also shifted in the two years. In 2009, it was more likely that a nWP result would be labeled irrelevant, while in 2010 this is reversed. Regarding the assessment of URLs where the query occurs in the URL, we find that NIST judges were more likely to label QiU documents as relevant than nQiU documents: $p(> 0) = 0.36$ and 0.24 for QiU, and 0.26 and 0.19 for nQiU, in 2009 and 2010, respectively. Paired t-tests confirmed significant differences between the means of two distributions for all feature pairs in both years and over the whole WT dataset, apart from RU/nRU in 2009 with $p_t = 0.4013$.

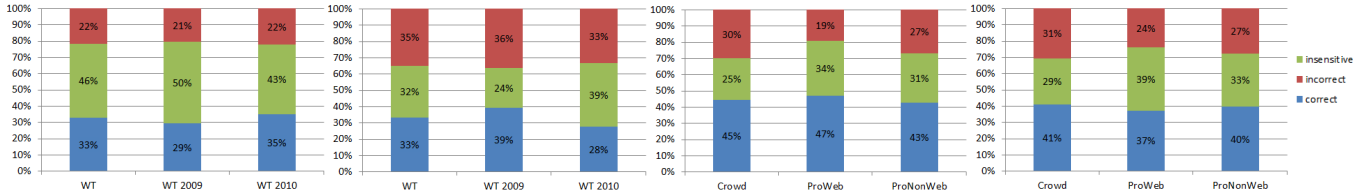
6. GOLD SET ANALYSIS

An alternative method of measuring label quality to inter-assessor agreement is via the use of gold judgments. We consider two types of gold data: 1) a set of labels provided by a group of highly trusted judges (super judges) and 2) a click-based gold set reflecting real Web users.

6.1 Click-based Pairwise Preferences

From our click data set, we first construct pairs of URLs per query and remove pairs where the difference in click counts is less than the median over the whole set (25 clicks for the WT set and 52 clicks for the ST set). For each URL pair, the relative click volume gives us a click preference relation. We then compare the click preference relations with the preferences derived from the collected relevance labels – by a given group of judges – for the respective query-URLs in each pairing. We consider label preferences that point in the same direction as the user click preferences to be correct, and label preferences that point in the opposite direction to be incorrect. We use an insensitive category when the click preference relation is not detected in the labels, i.e., the same label is assigned to both URLs. We repeat the same analysis using only the top two most-clicked URLs per query. Figure 2 shows the percentage of correct, incorrect and insensitive label vs. click preferences for the WT set (also broken down by year) and for the ST data set (broken down by judge group).

As we can see in Figure 2a), a high ratio of NIST label pairs do not distinguish between URLs with different volumes of clicks: 46% of the label preferences are insensitive. Comparing the two years, we can see a lot of improvement in 2010 over 2009 in the labels better reflecting real Web users: in 2010, 35% of the label preferences agree with click preferences, compared with only 29% in 2009. This demonstrates the impact of the change in guidelines from one year to the next, i.e., moving from a three point scale to a more sensitive five point scale. However, over the subset with the top two clicked URLs per query, see Figure 2b), we see that 2010 judgments have actually lower ratio of correct (down to 28% from 39% in 2009) and higher levels of insensitivity (39%, up from 24% in 2009). A detailed look at the differences reveals that most insensitivity errors occur for Key (2) and Non (0) labels, providing support for a finer grained relevance scale. Alternatively, it could also mean that judges are less able or willing to be more discriminative in their ratings for these types of query-document pairs.



a) WT (all click data) b) WT (top 2 clicked) c) ST (all click data) d) ST (top 2 clicked)

Figure 2: Ratios of correct/incorrect/insensitive label preferences when compared to click preference relations for the different data sets and judge groups

Ratios of correct preferences per our feature set are summarized in Table 3 (over all click preferences). It is interesting to note that NIST assessors perform differently in relation to the different document features. For example, they show high agreement with click preferences in 2010 when one or both URLs are Wikipedia pages (WP1=40%, WP2=51%, compared with nWP=31%). Lowest agreements with clicks is found in 2009 when judging similar results, e.g., when both results were WP (17% correct), QiU (22%) or RU (15%). This suggests that the 3 point scale was not sufficient to differentiate between such results or that judges developed internal judging rules based on document types, leading to possible bias.

Table 3: Percentage of label preferences that agree with user click preferences for pairs of URLs per query-URL feature (WP1/WP2/nWP: one/both/none of the URL(s) in the pair is a Wikipedia page, etc.)

WT dataset	All	2009	2010
WP1/WP2/nWP	38/39/30	34/**17**/28	40/51/31
QiU1/QiU2/nQiU	35/30/36	35/**22**/34	36/34/38
RU1/RU2/nRU	35/20/34	32/**15**/32	37/29/35
ST dataset	Crowd	ProWeb	ProNonWeb
WP1/WP2/nWP	44/42/44	45/37/45	43/46/42
QiU1/QiU2/nQiU	42/44/45	45/45/46	42/40/44
RU1/RU2/nRU	44/45/43	47/50/43	42/41/42

Turning our attention to the ST dataset and the differences between the three groups of judges, we observe that the ProWeb judges have the highest agreement with click preferences (47%), see Figure 2c), i.e., they are the most successful at interpreting the real user needs. Interestingly, crowd workers perform better (45%) than the ProNonWeb judges (43%). On the other hand, we can also see that ProWeb judges show the highest insensitivity levels (34%), while crowd judges appear to be the most opinionated, resulting in 25% insensitive but also 30% incorrect preferences. Focusing only on the top two clicked URLs per query, we find that crowd workers are actually the best at agreeing with real users (41%), see Figure 2d). Similarly to NIST judges, we observe a high ratio of insensitivity for ProWeb judges (39%), which can signal the need for a finer grade scale or the judges' unwillingness to be more discriminative in their ratings. The latter is supported by the findings of skewed label probability distributions, favoring middle relevance grades (see Section 5.2).

Overall, for our cross-judge group experiments, we find so far that trained judges perform best both according to inter-assessor agreement levels and agreement with user clicks. At the same time, we also see emerging judging patterns that can indicate bias, as in the case of rating Wikipedia pages.

6.2 Ratings for Most Popular URLs

Next we focus our investigation on the most popular, in terms of clicks, URLs per query and calculate how likely it is that a given judge group rates the most clicked URL highly. We calculate two probabilities. The first one is the probability p_a of a judge assigning one of the two highest labels, e.g., Key (2) and Nav (3) in the WT set, to the URL that has the highest volume of clicks for the given query, where we limit to URLs with a minimum of 100 clicks for the WT set and over 1000 clicks for the ST set. Since for some queries, the most clicked URL may not actually be included in the judged set, it may not be reasonable to expect a high relevance grade. Thus, we calculate a second metric, p_r, which is the probability that a top clicked judged URL is labeled highest among the judged URLs for the query. Table 4 reports the obtained probability values.

On the WT data, we see that although judges in 2009 were less likely to rate popular URLs highly ($p_a = 0.24$) in absolute terms than in 2010 (0.48), they actually did better in labeling them relative to the less clicked URLs for the given query ($p_r = 0.68$ vs. 0.62). Broken down by WP/nWP, we find that the rating of highly clicked Wikipedia pages improved greatly in 2010 ($p_a = 0.53$ in 2010, up from 0.23; and $p_r = 0.75$, up from 0.71). Interestingly, the likelihood of rating a highly clicked nWP result highly in relative terms drops in 2010 to 0.53 from 0.67 in 2009. Thus, it seems that while NIST judges may have become better at rating popular Wikipedia pages in 2010, their accuracy in rating popular non-Wikipedia pages reduced.

On the ST data, we observe that both Crowd and ProWeb judges consistently label popular URLs highly. However, while crowd judges tend to give high grades in absolute terms to the most clicked URLs of a query ($p_a = 0.69$), they are less accurate in rating results relative to each other ($p_r = 0.59$). This is partly a result of crowd workers' leniency to assign high grades. ProWeb judges show a different tendency: they are more conservative in their ratings ($p_a = 0.61$), but their relative ratings correctly reward the popular URLs ($p_r = 0.77$). Judges trained in other tasks perform the worst against these measures. When broken down by WP/nWP, we see no differences in rating accuracy for crowd judges: $p_a \approx 0.70$ and $p_r \approx 0.60$ for both WP and nWP. ProWeb judges show a different behavior: they are more likely to assign top grades to highly clicked non-Wikipedia pages than for Wikipedia ($p_a = 0.66$ for nWP and 0.54 for WP). Despite this, they are more likely to assign the relative best label for a query to the highly clicked

WP results than to nWP results ($p_r = 0.82$ for WP and 0.75 for nWP). This suggests a slight tendency of ProWeb judges to underrate WP results compared to click evidence.

Table 4: Probability of a high grade relevance label (p_a) and (/) probability of relative highest grade label per query (p_r) being assigned to the highest clicked judged URL for the query

WT dataset	2009	2010	
All	0.24/0.68	0.48/0.62	
WP	0.23/0.71	0.53/0.75	
nWP	0.25/0.67	0.44/0.53	
ST dataset	Crowd	ProWeb	ProNonWeb
All	0.69/0.59	0.61/0.77	0.45/0.50
WP	0.70/0.60	0.54/0.82	0.38/0.51
nWP	0.68/0.59	0.66/0.75	0.50/0.50

6.3 Gold Label Set based Analysis

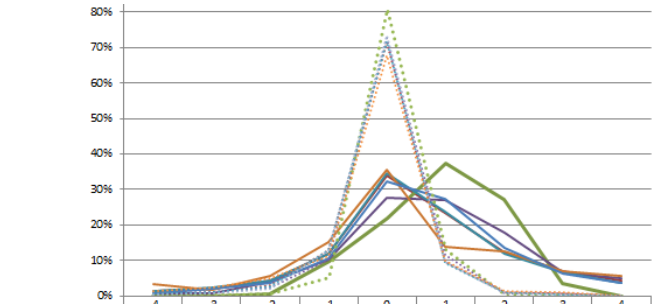
In this section we analyze the labels collected in our HQ set with respect to a gold label set contributed by highly trained Web (super) judges. Given a gold label, we calculate judging error as the distance between the relevance grades assigned by a judge and the gold label. Thus, for our five point scale, error can take a value in the range of [-4,4], where an error of -3 means that a judge underrated the given query-URL pair by 3 grades, e.g., labeling it as grade 1 when the gold label is grade 4. Figure 3 summarizes the error distributions per judge groups and features.

Overall, ProWeb judges are much more accurate while crowd workers are much more noisy against our gold set: crowd judges are less likely to have 0 error (only 31% on their labels are correct, compared with 73% for ProWeb). Crowd workers are also skewed towards overrating results: 25% of their labels fall in the +1 error category, compared with 10% for ProWeb judges.

Interestingly both judge groups show a tendency to overrate Wikipedia pages. For Crowd the label distribution peaks at +1 error for WP (37%). The probabilities to underrate, overrate or be correct are shown in Table 5. This confirms crowd workers' tendency to overrate, with especially high probabilities for WP, QiU and nRU results (p=0.68, 0.56 and 0.51, resp.). In contrast, ProWeb judges are more likely to underrate results overall. The only case where ProWeb judges are more likely to overrate is when the URL is a Wikipedia page (p(under)=0.05 and p(over)=0.14 for WP, compared with p(under)= 0.14 and p(over)=0.13 for nWP). This is interesting since as we saw earlier ProWeb judges tend to pick middle grade labels for WP results (Section 5.2), which then led to insensitivity compared to click evidence (Section 6.1), but show up as slight tendency to overrate according to the gold set. On the other hand, crowd workers are much more likely to grade leniently and overrate WP results according to the gold set, but achieve better agreement with click preferences. The different behaviors suggests possible guideline bias whereby ProWeb judges may learn to treat Wikipedia pages in a specific way, which, in this study, disagrees with the click evidence.

6.4 Regression analysis

Using the HQ gold set as target we fit a linear regression model over our query-URL feature set to determine which features have significant impact in the two judge groups.



	-4	-3	-2	-1	0	1	2	3	4
Crowd nWP	1%	2%	4%	12%	34%	23%	12%	7%	4%
Crowd WP	0%	0%	1%	10%	22%	37%	27%	3%	0%
Crowd QiU	1%	1%	4%	10%	28%	27%	18%	6%	5%
Crowd nQiU	1%	2%	4%	12%	34%	24%	12%	7%	4%
Crowd RU	3%	2%	6%	15%	36%	14%	12%	7%	6%
Crowd nRU	1%	2%	4%	11%	32%	27%	14%	6%	4%
ProWeb nWP	1%	1%	3%	13%	71%	10%	1%	1%	0%
ProWeb WP	0%	0%	0%	5%	81%	13%	1%	0%	0%
ProWeb QiU	1%	1%	2%	10%	73%	12%	1%	1%	0%
ProWeb nQiU	1%	1%	3%	13%	72%	9%	1%	0%	0%
ProWeb RU	1%	2%	5%	11%	68%	10%	1%	1%	0%
ProWeb nRU	0%	1%	2%	13%	73%	10%	1%	0%	0%

Figure 3: Levels of judging error per query-document feature and judge group on the HQ set

Table 5: Probability to underrate, overrate or be correct when compared to gold label set per judge group and query-URL feature

Feature	p(under)	p(over)	p(correct)
Crowd			
WP/nWP	0.10/0.20	**0.68**/0.47	0.22/0.34
QiU/nQiU	0.16/0.20	**0.56**/0.46	0.28/0.34
RU/nRU	0.26/0.17	0.39/**0.51**	0.36/0.32
ProWeb			
WP/nWP	**0.05**/0.18	**0.14**/0.11	0.81/0.71
QiU/nQiU	0.14/0.17	0.13/0.11	0.72/0.68
RU/nRU	0.20/0.16	0.13/0.11	0.68/0.73

Table 6 shows that for ProWeb judges both WP/nWP and RU/nRU are significantly correlated to the gold label, although WP/nWP positively correlates while RU/nRU negatively. For Crowd, both QiU/nQiU and RU/nRU features have significant impact, QiU positively while RU negatively. Due to the bias in crowd workers to overrate Wikipedia pages, this feature is less predictive of the gold labels.

Looking at the predictive power of the various features for estimating the actual label that different judges would assign, we find that for both Crowd and ProWeb judges, the RU/nRU feature has the greatest impact(coefficient of 81 and 84, respectively), followed by WP/nWP (71 and 72) and finally QiU/nQiU (70 and 68), all significant.

Table 6: Regression coefficients per query-URL feature on the HQ set, * indicates $p_t < 0.001$**

Feature	Estimate	Std. Err.	p_t
Crowd			
WP/nWP	-0.079	0.167	0.637
QiU/nQiU	0.253	0.108	0.019 *
RU/nRU	-0.306	0.114	***
ProWeb			
WP/nWP	0.0124	0.017	***
QiU/nQiU	0.021	0.010	0.0467
RU/nRU	-0.184	0.010	***

7. IMPACT ON RANKER TRAINING

Even though judgments may be biased or noisy, they may not impact the training or evaluation of retrieval systems. To test whether judgments obtained from the different judges affect the training of ranking systems, we use the LambdaRank [4] learning to rank algorithm to train different rankers using judgments from all different judge groups and report the test set performance of each ranker when the judgments of the different judge groups are used to create the test set. We focus only on the WP/nWP feature here, which was found to be the most problematic to judge in the previous sections.

For our training experiments, we use five fold cross-validation, where the data is divided into five different parts. For each fold, three parts are used for training, and one part each for validation and testing. In Table 7, we report the average performance obtained over the five folds using NDCG at rank 10 as the test metric. Each row corresponds to training LambdaRank using judgments obtained from a given judge group. For each judge group, we train three different rankers using three different feature sets: (1) BM25, (2) WP, and (3) BM25 and WP. In the case of using just one feature such as BM25 or WP, our goal is to identify whether these features get a positive or negative weight during training. We use the ranker obtained using BM25 as the ranking feature as our baseline ranker and everything that is significantly different than this baseline is marked with a * (t-test, $p = 0.05$).

We see that rankers trained on the WP feature alone achieve a reasonable NDCG@10 performance for all training sets. If we focus on training and testing on judgments obtained from the NIST judges, it can be seen that the test set performance obtained using the WP feature alone is significantly better than test set performance obtained using BM25 (0.5518 vs. 0.5257). Using BM25 together with the WP feature in training does not result in much further improvement over when using the WP feature alone (0.5518 vs. 0.5564). This suggests that there is an extremely strong Wikipedia bias for NIST judges.

In the case of ProWeb judges, training on the WP feature alone obtains significantly worse performance than training on the BM25 feature (0.6623 vs. 0.7335). When the WP feature is used together with BM25, we see only small improvements over the baseline.

The effect of Wikipedia bias seems the weakest for crowd judges: using the WP feature alone results in worse performance than the baseline (though difference is not statistically significant). However, using the BM25 and WP features together does not seem to make much difference in the test set performance.

These results show that ranker training and evaluation can indeed be affected by the different judgment biases. For example, a Web Track run that ranks Wikipedia documents containing query terms at the top is likely to achieve good performance when evaluated on the NIST judgments.

When NDCG values across different datasets are compared, one can see that the range of NDCG scores obtained can vary highly depending on which judge group is used in the test set (higher NDCG values are obtained using the ProWeb judgments in the test set). On the other hand, when training on judgments obtained from different judge groups is compared, the test set values do not seem to differ much. So, the judges seem to be replaceable for the purposes of creating training data.

Table 7: NDCG10 results using different ground-truth sets by the different judge groups, * indicates significant with p<0.05

Training labels	BM25	WP/nWP	BM25+WP/nWP
NIST labels as test set			
NIST	0.5257	0.5518*	0.5564*
ProWeb	0.5257	0.5518*	0.5583*
Crowd	0.5257	0.5518*	0.6011*
ProWeb labels as test set			
NIST	0.7335	0.6623*	0.7409*
ProWeb	0.7335	0.6623*	0.7335
Crowd	0.7335	0.6623*	0.7409*
Crowd labels as test set			
NIST	0.6566	0.6369	0.6582
ProWeb	0.6566	0.6369	0.6566
Crowd	0.6566	0.6369	0.6583

8. CONCLUSIONS

Past studies have shown that changes in judging procedures can change the outcome of experiments. With an increasing use of crowdsourcing for IR relevance judging, it becomes even more important to consider judge quality. When judging procedures do affect outcomes, tools are needed not only to detect such differences, but to understand which procedure was correct, what systematic errors occurred, and how to make progress on improving procedures.

In this paper, we investigated several methods for comparing different populations of judges and their judgments. Agreement with consensus can be used, and broken down into sub-cases based on features of the query-document, like whether the document is from Wikipedia. However a greater agreement on Wikipedia documents could either indicate reduced noise or could mean for example that judges adopted a rule of thumb or used a smaller range of labels for Wikipedia pages. For this reason we introduced gold set analysis, using clicks and labels.

Using a click-based gold set, we can identify cases where real users of a system have a preference, and check the level of agreement with judges. If a test collection is seen as a user model, and if click preferences on average reflect the preferences of real users, then judgments that agree more often with click preferences give us a better user model. In our comparison of judgments with the top two clicked URLs, we found agreement most often from crowd workers. This result bears more investigation in future, whether the population of crowd judges are representative of the population of clicking users.

Using a labeled gold set, we can identify difficult cases where judges are likely to make mistakes, then take care to identify the label we would like to see from judges. We can measure the calibration and noisiness of judges with respect such a set. We can also repeat the breakdown by features. In our HQ gold set experiments we identified crowd judges as being more likely to overrate Wikipedia pages. If we trust our gold judgments, that the Crowd really have this systematic bias, then we can change the interface or instructions to try and correct for this.

Our ranking experiments demonstrated that systematic differences in labeling can lead to differences in ranker training and evaluation. For a ranking system optimized against training data, this allows us to draw a connection between

differences that exist in the relevance judging procedures and differences in the characteristics of the ranking seen by end users.

Overall this paper has demonstrated a number of approaches for analyzing relevance judgments, beyond just inter-judge agreement. We have analyzed multiple populations of judges and found different cases of systematic judging error. In the context of machine learned ranking, which is powerful enough to reproduce biased results if given biased training judgments, reducing systematic errors is particularly important. We have demonstrated empirically that different characteristics of training sets lead to different learned rankers.

Through methods such as those introduced in this paper, it may be possible to track the ability of judges to agree with real users, and follow to judging guidelines. This could lead to better training and measurement of real-world IR systems, but also IR research conducted with the confidence that the test collection models a real-world search task.

9. REFERENCES

[1] O. Alonso and R. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *Proc. of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 153–164. Springer-Verlag, 2011.

[2] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *Proc. of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 667–674. ACM, 2008.

[3] P. Borlund. The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10):913–925, May 2003.

[4] C. J. C. Burges, R. Ragno, and Q. V. Le. Learning to rank with nonsmooth cost functions. In *NIPS*, pages 193–200. MIT Press, 2006.

[5] B. Carterette and I. Soboroff. The effect of assessor error on IR system evaluation. In *Proc. of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 539–546. ACM, 2010.

[6] C. L. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 75–84. ACM, 2011.

[7] C. Cuadra and R. Katter. The relevance of relevance assessment. *Proc. of the American Documentation Institute*, page 95âĂŞ99.

[8] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, pages 20–28.

[9] C. Grady and M. Lease. Crowdsourcing document relevance assessment with mechanical turk. In *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 172–179, 2010.

[10] D. Hawking and N. Craswell. The very large collection and web tracks. In E. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[11] J. Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Publishing Group, New York, NY, USA, 1 edition, 2008.

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

[13] K. H. Jiyin He, Krisztian Balog and E. Meij. Heuristic ranking and diversification of web documents. In *TREC*, 2009.

[14] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling. Crowdsourcing for book search evaluation: impact of HIT design on comparative system ranking. In *Proc. of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 205–214. ACM, 2011.

[15] K. A. Kinney, S. B. Huffman, and J. Zhai. How evaluator domain expertise affects search result relevance judgments. In *Proc. of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 591–598. ACM, 2008.

[16] E. Pronin. Perception and misperception of bias in human judgment. *Trends in cognitive sciences*, 11(1):37–43, 2007.

[17] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance. *J. Am. Soc. Inf. Sci. Technol.*, 58(13):2126–2144, Nov. 2007.

[18] F. Scholer, A. Turpin, and M. Sanderson. Quantifying test collection quality based on the consistency of relevance judgements. In *Proc. of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1063–1072. ACM, 2011.

[19] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proc. of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 315–323. ACM, 1998.

[20] E. M. Voorhees. Evaluation by highly relevant documents. In *Proc. of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 74–82. ACM, 2001.

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

[22] W. Webber, D. W. Oard, F. Scholer, and B. Hedin. Assessor error in stratified evaluation. In *Proc. of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 539–548. ACM, 2010.