Cognitive Biases in Crowdsourcing

Carsten Eickhoff
Dept. of Computer Science
Zurich, Switzerland
ecarsten@inf.ethz.ch

ABSTRACT

Crowdsourcing has become a popular paradigm in data curation, annotation and evaluation for many artificial intelligence and information retrieval applications. Considerable efforts have gone into devising effective quality control mechanisms that identify or discourage cheat submissions in an attempt to improve the quality of noisy crowd judgments. Besides purposeful cheating, there is another source of noise that is often alluded to but insufficiently studied: Cognitive biases.

This paper investigates the prevalence and effect size of a range of common cognitive biases on a standard relevance judgment task. Our experiments are based on three sizable publicly available document collections and note significant detrimental effects on annotation quality, system ranking and the performance of derived rankers when task design does not account for such biases.

CCS CONCEPTS

• **Information systems → Crowdsourcing**; **Relevance assessment**; • **Social and professional topics → Quality assurance**;

KEYWORDS

Crowdsourcing, Human Computation, Relevance Assessment, Cognitive Biases

ACM Reference Format:
Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining (WSDM 2018).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3159652.3159654

1 INTRODUCTION

Over the past decade, crowdsourcing has been established as a popular method for data collection, annotation and consolidation as well as conducting user studies and evaluating system performance. Instead of relying on local annotators contributing labor in controlled lab environments, human intelligence tasks, so-called HITs, are outsourced to a large anonymous crowd of workers. Academics and industrial practitioners alike praise this alternative paradigm's cost-efficiency, quick response times and access to workers of more diverse backgrounds than were commonly included in most lab-based studies. As a consequence of this distributed weakly supervised approach, quality control remains a key challenge [17].

The literature commonly assumes low label quality to stem from three possible reasons: **(1)** Unethical spammers submit imprecise or even arbitrary labels in order to maximize their financial efficiency [18] or due to external distractions. **(2)** Unqualified workers are, despite their best efforts, unable to produce an acceptable annotation quality [15]. **(3)** Malicious workers purposefully aim to undermine or influence the labelling effort [53][1].

While there certainly is strong evidence for all of the above reasons, we argue that there is a fourth fundamental reason for noisy label submissions. Cognitive biases are systematic patterns of deviation from norm or rationality in judgment, whereby inferences about other people and situations may be drawn in an illogical fashion [27]. Individuals create their own "subjective social reality" from their perception of the input. An individual's construction of social reality, instead of the objective input, may dictate their behaviour and lead to perceptual distortion, inaccurate judgment, illogical interpretation, or irrationality [5]. Recognizing systematic biases in data collection efforts is an important step towards countering their effect on those systems that are trained on the basis of this data and is a key enabling factor for algorithmic fairness [25].

In a series of document relevance assessment experiments, this paper strives to demonstrate that cognitive bias can indeed affect crowdsourced labor and leads to significantly reduced result quality. This performance detriment is subsequently propagated into system ranking robustness and machine-learned ranker efficacy. The common strategies of controlling the crowd by means of qualification tests, demographic filters, incentives, gold standards and sophisticated worker models may not be enough to overcome this new source of noise which is inherently caused by the HIT setup. Instead, we would like to advocate careful task and study design that takes into account cognitive biases to reduce the interface's susceptibility to this kind of label noise.

The remainder of this document is structured as follows: Section 2 gives an overview of related work on crowdsourced data collection and annotation, quality control, worker models and biases. Section 3 follows up with an overview of frequently encountered cognitive biases that can take effect in crowdsourced relevance assessment tasks. Our experimental evluation in Section 4 strives to answer three main research questions:

RQ1 What is the effect of cognitive biases on the quality of crowdsourced relevance labels?

RQ2 To what degree do such effects influence retrieval system evaluation?

RQ3 To which extent do cognitive biases influence the quality of machine-learned rankers?

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. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WSDM 2018, February 5–9, 2018, Marina Del Rey, CA, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5581-0/18/02…$15.00
https://doi.org/10.1145/3159652.3159654

[1] While we do not believe that truly malicious workers are a frequently-encountered class of workers, we include them in this overview for completeness.

Finally, Section 5 closes with a summary of our findings as well as an outlook on future directions of inquiry.

2 RELATED WORK

Beginning with the Cranfield experiments [8], test collections have been one of the pillars of IR system evaluation. Traditionally, such collections are created by trained professionals in controlled lab environments. In the IR community, the Text Retrieval Conference (TREC) [50] supports one of the most widely known efforts to creating such collections.

The emerging crowdsourcing paradigm [30] has inspired an extensive line of research dedicated to using this new channel for the creation and annotation of IR test collections. An early set of experiments [1, 22, 38, 39] note that aggregated labels of multiple untrained crowd workers can reach a quality comparable to that of a single highly trained NIST assessor.

While this alternative labour market has been shown to be time and cost efficient [18], researchers have less control over the circumstances under which relevance judgments are created. Traditionally, inaccurate judgments as well as spam submissions have been a major challenge in the crowdsourcing process. As a consequence, quality control is one of the most deeply researched question in crowdsourcing [40] with solutions ranging from aggregating the results of independent workers to the use of honey pot questions [17, 28, 32]. Marshall *et al.* [41] highlight the importance of engaging HIT design on result quality while Yan *et al.* [57] demonstrate the ability of active learning techniques to greatly improve the worker-task allocation step. Bansal *et al.* [2] exploit this notion for task selection in highly efficiency-driven crowdsourcing scenarios.

Significant effort has been made into estimating worker reliability based on various behavioural and demographic traits. Tang and Lease [48] introduce a semi-supervised method for reliability estimation based both on labeled as well as unlabeled examples. Kazai *et al.* [36, 37] organize workers in 5 different classes and study their respective judgment reliability and behaviour. Based on social media profiles, Difallah *et al.* [14] model worker topic affinities, enabling them to assign tasks to workers with matching interest, resulting in significantly improved result quality. However, as the authors further note [13], it can be hard to have the necessary degree of control over the ephemeral crowd workforce in applied crowdsourcing scenarios that involve online platforms such as Amazon Mechanical Turk and CrowdFlower. Karger *et al.* [34] propose a joint model for iteratively learning worker reliability and aggregating votes by means of approximate belief propagation. Recently, Davtyan *et al.* [11] show the use of inter-document similarities for the goal of label aggregation. They intuit that, given the same query, similar documents should show similar relevance labels. Blanco *et al.* [4] study the robustness of crowdsourced relevance assessments over time, finding that repeated labeling efforts produce stable results even as longer periods of time elapse.

Another prominent source of evidence can be found in the analysis of systematic judgment behaviour. Following Dawid and Skene [12], who investigate disagreements between diagnoses posed by multiple individual medical doctors, there have been several successful attempts at harnessing similar methods for crowdsourcing quality assurance [29, 51]. In this way, reliable workers that make occasional mistakes can be accurately separated from spammers that select answers at random or follow other, more sophisticated, cheating strategies.

While some examples were known previously, the systematic study of cognitive biases was first established by Tversky and Kahnemann [49] and has since seen many realizations in decision theory and advertisement. Recently, there have been a number of studies tying observed user behavior in Web search [55] and general information systems [21] to cognitive biases. The body of existing crowdsourcing work acknowledges the importance of task and interface design [33, 35, 46], as well as execution parameters such as batch sizes [52]. There are several articles that mention presentation and order effects [7, 16] or irrational aesthetically-motivated annotator behavior [42]. In addition, a number of scholars, without a dedicated study of the reason for biased crowd labor, present potential solutions in the form of tailored incentive schemes [20], Bayesian bias mitigation [54] or active learning [60].

While several pieces of related work have alluded to the impact of cognitive biases on crowdsourced data collection and annotation efforts, to the best of our knowledge, there exists no dedicated study that systematically describes and quantifies such effects. This paper aims to close this gap in the literature.

3 COGNITIVE BIASES

This section begins with a brief description of our experimental corpora as well as an unbiased baseline setup for crowdsourced relevance label collection. Subsequently, we propose four experimental crowdsourcing conditions in which a range of prominent cognitive biases are likely to occur.

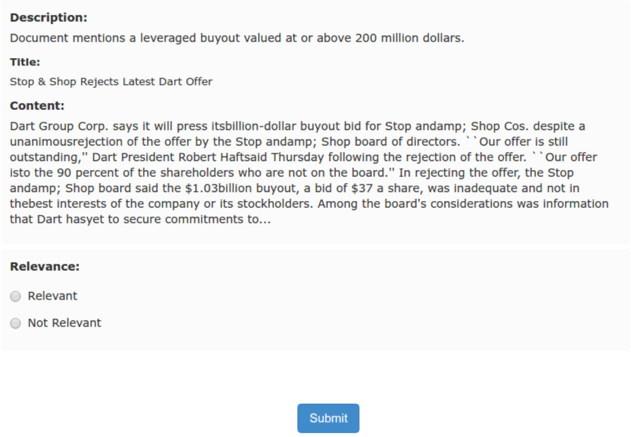
3.1 Datasets

Our experiments are based on three well-known document collections that were selected to span different domains, document types and collection sizes. As an example of a large-scale Web collection, we include the widely-used ClueWeb12 corpus[2], containing approximately 733M Websites. Corresponding test topics are taken from the 2013 and 2014 editions of the TREC Web Track [9, 10]. The TIPSTER collection [26] compiles newswire documents from a number of major outlets across the years 1987 – 1992 and represents an example of a traditional pre-Web retrieval task on a smaller set of individually well-curated documents. We include the corresponding 4 years of TREC adhoc topics (1992 – 1995) into our experiments. Finally, the TREC Clinical Decision Support track (CDS) investigates patient-centric literature retrieval. The documents are a subset of the PubMed repository of bio-medical journal articles. While the previous collections contained rather accessible information from the Web and news domains, this corpus heavily relies on technical jargon and significantly higher reading levels. Queries in this setting stem from the highly condensed patient summaries of all three editions (2014 – 2016) of TREC CDS [43, 44, 47]. Table 1 gives an overview of key statistics such as the number of documents, topics and official NIST-provided relevance judgments.

[2] http://lemurproject.org/clueweb12/

Table 1: Details of the three experimental corpora. For each corpus, we report the domain of origin, its size, the number of topics and the amount of official NIST-provided relevance judgments.

Corpus	Domain	# Documents	# Topics	# Qrels
ClueWeb12	Web	733M	100	91k
TREC CDS	Medical	1.3M	90	114k
TIPSTER 1–3	Newswire	1M	200	336k



Description:

Document mentions a leveraged buyout valued at or above 200 million dollars.

Title:

Stop & Shop Rejects Latest Dart Offer

Content:

Dart Group Corp. says it will press itsbillion-dollar buyout bid for Stop andamp; Shop Cos. despite a unanimousrejection of the offer by the Stop andamp; Shop board of directors. ``Our offer is still outstanding,'' Dart President Robert Haftsaid Thursday following the rejection of the offer. ``Our offer isto the 90 percent of the shareholders who are not on the board.'' In rejecting the offer, the Stop andamp; Shop board said the $1.03billion buyout, a bid of $37 a share, was inadequate and not in thebest interests of the company or its stockholders. Among the board's considerations was information that Dart hasyet to secure commitments to...

Relevance:

○ Relevant

○ Not Relevant

Figure 1: The baseline annotation interface presents workers with a brief topic description as well as document title and content and asks for binary relevance labels.

3.2 Baseline Setup

Let us consider a standard document relevance assessment task in which workers are first primed by a set of relevance guidelines before being presented with a brief topic description and a structured document representation. We include only the document title and main text content, with all remaining meta information hidden in this setup. Relevance assessment is made in binary fashion, distinguishing between relevant and non-relevant pairs of topics and documents. In order to prevent further sequence effects and presentation biases, we publish tasks in batches of one. Figure 1 gives an example of the judgment interface as seen by our workers. In an initial set of experiments, this baseline shows an encouraging individual accuracy of approximately 80% at matching NIST labels. In the further course of this document, we will discuss a number of interface variants used to demonstrate the effect of cognitive biases. To ensure legibility of the interface screenshots, here we truncate the, sometimes lengthy, document content field. This was exclusively done in the screenshots for this paper. Workers were always presented with the full document content.

3.3 Ambiguity Effect

The Ambiguity Effect occurs when missing information makes decisions appear more difficult and consequently less attractive [19]. While our baseline scenario includes only the title and main content fields, we further include a number of additional properties for each document: the document's age in days calculated from its publication date, its length in number of words and a random number sampled from the interval $[0, 5]$ that we dub the "k-Index" and the origin or meaning of which is never explained to the workers. It is intuitive that neither of these properties should have a strong implication on a document's relevance towards a given topic. We confirm this assumption by means of a Pearson product-moment correlation test. The only property that shows even mild correlation with the NIST relevance label is document length[3].

We randomly sample 50 relevant and 50 non-relevant pairs from each of our three collections. To test for traces of the Ambiguity Effect, we compare two conditions: The baseline setup in which no additional fields are shown, and an alternative including the additional fields. For 50% of the pairs in this second group, the values of all additional fields are present. For the remainder, we declare them as "unknown". The document title and content are always available.

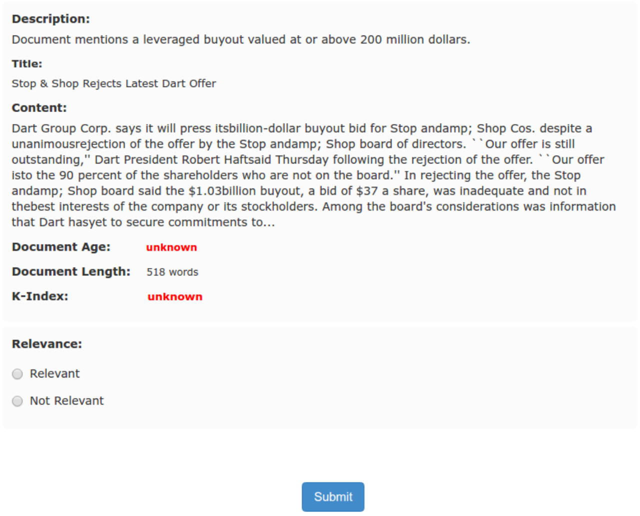
In the baseline setting, the observed likelihood of a worker assigning a label of 'relevant' is 58%. When showing additional fields with complete information, we observe a comparable likelihood of relevancy of 59%. When, however, these additional values are missing, the likelihood of a positive vote plummets to 44%, irrespective of true document relevance. Despite the fact that the potentially missing information has, at best, a weak connection to document relevance, and even though it can in some cases be inferred (*i.e.*, document length is observable from the displayed main content.), the missing values create an illusion of uncertainty and lets workers doubt the quality of a document.

Let us, in a second experiment assume a gradual occurrence of missing data that is more representative of real-life missing-value scenarios. Instead of hiding either all or none of the additional values per document, we are introducing a masking probability $p_m \in [0, 1]$ that is applied to each of the additional fields independently, allowing some values to be missing while others are present. Figure 2 shows the judgment interface for this setup.

As we plot the overall likelihood of document relevance in response to p_m (Figure 3), we note a considerable decline in the likelihood of relevance up to $p_m \approx 0.6$. At this point the likelihood of relevance stabilizes and eventually partially recovers. We believe that this final trend is due to training effects under which workers learn that hardly any of the documents exhibit complete values, making the perceived uncertainty appear less severe. It is furthermore interesting to note that at $p_m = 1.0$ when none of the additional fields are ever shown, there is still a significantly lower likelihood of arbitrary document relevance than at $p_m = 0.0$, indicating that even uniformly missing information may make workers more hesitant to assign positive labels.

In summary, we see a clear negative impact of showing missing values to workers, even if these missing properties were uninformative to start with. This observation should serve as a word of caution when designing crowdsourcing experiments. In settings that can experience missing values it may be beneficial to omit a property altogether.

[3] Age: $r = -0.16$, Length: $r = 0.23$, K-Index: $r = -0.07$



Description:

Document mentions a leveraged buyout valued at or above 200 million dollars.

Title:

Stop & Shop Rejects Latest Dart Offer

Content:

Dart Group Corp. says it will press itsbillion-dollar buyout bid for Stop andamp; Shop Cos. despite a unanimousrejection of the offer by the Stop andamp; Shop board of directors. ``Our offer is still outstanding,'' Dart President Robert Haftsaid Thursday following the rejection of the offer. ``Our offer isto the 90 percent of the shareholders who are not on the board.'' In rejecting the offer, the Stop andamp; Shop board said the $1.03billion buyout, a bid of $37 a share, was inadequate and not in thebest interests of the company or its stockholders. Among the board's considerations was information that Dart hasyet to secure commitments to...

Document Age: unknown

Document Length: 518 words

K-Index: unknown

Relevance:

○ Relevant

○ Not Relevant

Figure 2: The ambiguity interface additionally presents workers with three uninformative meta attributes. Some of these attributes are declared "unknown" to create the impression of incomplete information.

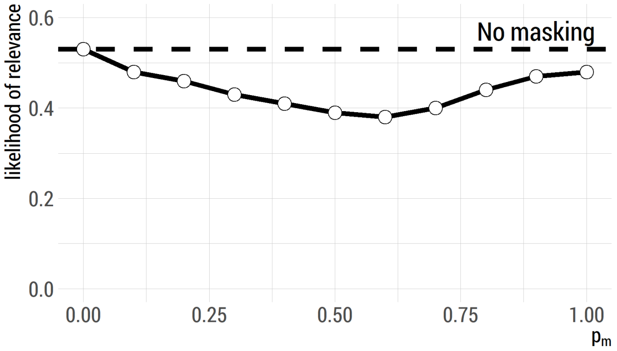
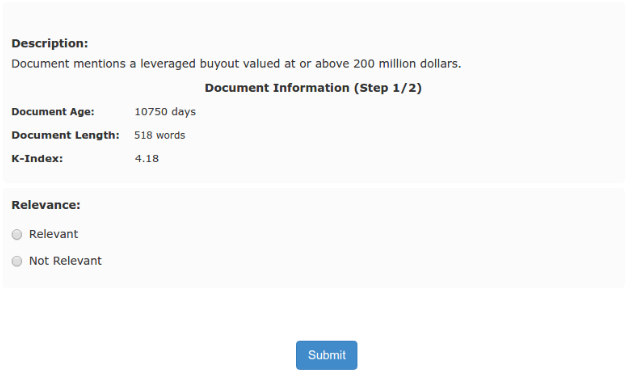


Figure 3: The masking probability p_m controls the amount of missing field values. The overall likelihood of relevance drops as p_m approaches 0.6. After this point it stabilizes and eventually recovers.

In the further course of this document, when referring to the Ambiguity Effect, we will use the above fields and individually declare their values unknown with $p_m = 0.5$.

3.4 Anchoring

Anchoring or Focalism describes situations in which workers disproportionately focus on one piece of information (often the first one presented to them) even as additional contradicting evidence becomes apparent [49]. In organic crowdsourcing experiments such effects are likely to take place as information is revealed to the user in subsequent stages. We simulate this situation by structuring the judgment process in two phases. At first, the worker is presented exclusively with the topic description and the three additional information fields described in the previous section (document age, document length and a random number introduced as the "k-Index"). Note that this time all information is present at all times and there are no missing values. The worker is asked to assign a relevance label solely based on these three properties which, as we showed earlier have no systematic association with the true relevance label. In the second stage of the experiment we additionally reveal the document title and content to the worker and ask them to update their relevance vote. Figure 4 shows an example of the corresponding judgment interface.



Description:

Document mentions a leveraged buyout valued at or above 200 million dollars.

Document Information (Step 1/2)

Document Age: 10750 days

Document Length: 518 words

K-Index: 4.18

Relevance:

○ Relevant

○ Not Relevant

Figure 4: The anchoring interface presents document-related information in two steps, beginning with uninformative meta data.

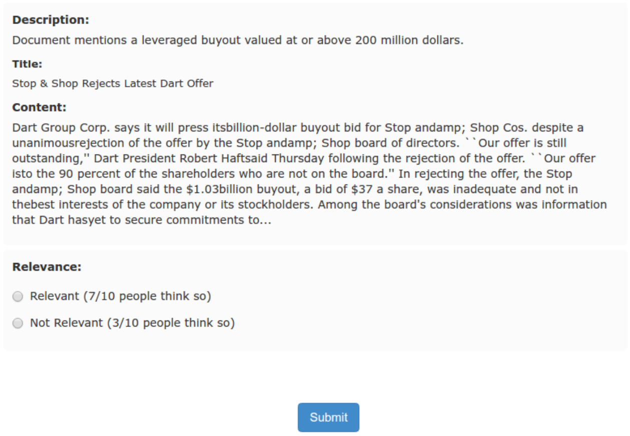
Table 2: A direct comparison of two-stage and single-stage crowdsourcing processes demonstrates significantly lower overall label accuracy due to Anchoring when initially presenting uninformative data.

	Stage-1	Stage-2	Single-Stage
Relevant	0.42	0.43	0.54
Non-Relevant	0.58	0.57	0.46
Accuracy	0.49	0.67	0.81

Let us now consider an experiment under which we randomly draw 50 relevant and 50 non-relevant pairs from each of the three corpora and request crowd labels via the process described above. In a separate control group we request labels for the same data in a single-stage fashion, presenting the same information at once, rather than sequentially.

In expectation, a rational worker would have to blindly guess at stage one of the judgment process and would be correct 50% of the time[4]. In the next stage, when document title and content are revealed, this same rational worker would be expected to disagree with their earlier vote in 50% of the cases. Instead of such unbiased behavior, Table 2 notes significant evidence of anchoring.

[4] A highly informed worker that is privy to the mild systematic connection between document length/age and relevancy could do somewhat better. Such effects would be very subtle and are disregarded here.



Description:

Document mentions a leveraged buyout valued at or above 200 million dollars.

Title:

Stop & Shop Rejects Latest Dart Offer

Content:

Dart Group Corp. says it will press itsbillion-dollar buyout bid for Stop andamp; Shop Cos. despite a unanimousrejection of the offer by the Stop andamp; Shop board of directors. ``Our offer is still outstanding,'' Dart President Robert Haftsaid Thursday following the rejection of the offer. ``Our offer isto the 90 percent of the shareholders who are not on the board.'' In rejecting the offer, the Stop andamp; Shop board said the $1.03billion buyout, a bid of $37 a share, was inadequate and not in thebest interests of the company or its stockholders. Among the board's considerations was information that Dart hasyet to secure commitments to...

Relevance:

○ Relevant (7/10 people think so)

○ Not Relevant (3/10 people think so)

Figure 5: The bandwagon interface additionally presents workers with the historic distribution of relevance labels assigned to this topic-document pair by their peers.

While the single-stage annotators perform at the same level of accuracy of approximately 80% that was earlier observed in the baseline setting, workers under the two-stage condition updated their prior arbitrary beliefs too infrequently, resulting in a considerably reduced stage-2 accuracy. To exclude outside experimental effects, we conduct the same experiment in inverse temporal order, *i.e.*, first revealing title and content and subsequently showing the additional fields. In this experimental condition, workers show the same accuracy rates as in the single-stage case, suggesting that the detrimental effect was indeed induced by anchoring on conclusions drawn from uninformative data.

3.5 Bandwagon Effect

The Bandwagon Effect (sometimes referred to as Groupthink or Herd Behavior) occurs when workers forego their own reasoning in favor of following an existing group's behavior [3]. In crowdsourced relevance assessment tasks, we can easily simulate this effect by disclosing to workers how their peers judged a given topic-document pair. The assessment interface is identical to the baseline with the only exception of binary relevance choices additionally displaying the number of previous users who made this choice for this pair. A screenshot of the resulting interface can be found in Figure 5.

In our first experiment, instead of using actual historic votes, we show the user artificial statistics according to which *n* out of 10 previous users judged this pair relevant. From each of our three collections, we randomly draw 50 relevant and 50 non-relevant pairs and have crowd workers judge them once more. Figure 6 plots the likelihood of workers assigning a label of 'relevant' to an arbitrary pair as *n* ranges from 0 to 10.

We can observe a balanced tendency of workers following perceived herd behavior. Irrespective of the true relevance label, topic-document pairs were more likely to be judged relevant when we claim a high number of prior votes indicating relevance. Similarly, although somewhat less strongly expressed, we note the probability of relevance decreasing as the perceived prior consensus becomes more negative. In the center of the plot, for choices of *n* between 4-6, we do not find considerable bias in either direction.

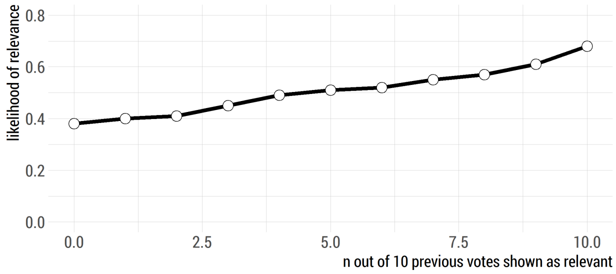


Figure 6: The likelihood of relevance monotonically follows the number of artificially reported previous positive judgments irrespective of true document relevance.

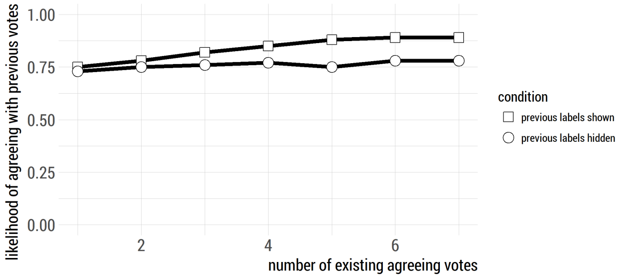


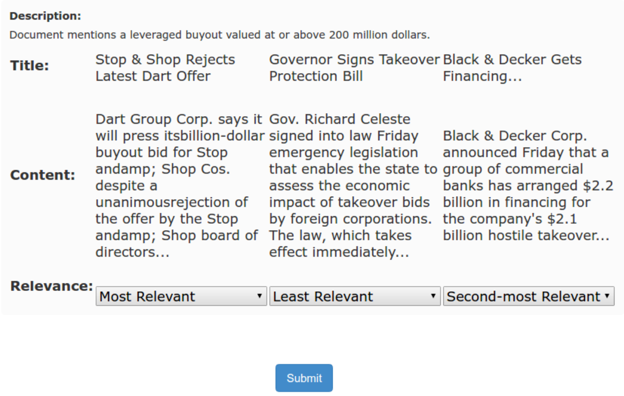
Figure 7: The likelihood of agreeing with an existing majority gently increases when votes are hidden from workers but sharply rises with the amount of existing votes when revealed prior to assessment.

While one might argue that such strong Bandwagon Effects are artificial and may not often occur unintentionally in organic crowdsourcing efforts, let us consider a second experiment. This time, we do not introduce false information but, instead, truthfully disclose the prior vote distribution. While the baseline experiment shows a good accuracy of 80%, the same setup significantly drops in accuracy (to a level of 76%) when disclosing prior vote statistics. Further analysis suggests that this error is introduced by early incorrect votes that are subsequently followed by other workers.

This assumption is confirmed when we inspect the distribution of likelihood of workers disagreeing with existing majority votes. Figure 7 plots the observed likelihood of disagreement with current majority vote[5]. While consensus is generally high, we note that in the baseline case, where vote distributions are not shown, agreement ratios only gently rise, reflecting the general degree of controversy in each pair while, for the exact same pairs, they sharply increase in case of disclosed vote statistics.

To summarize, the Bandwagon Effect shows clear expression in crowdsourced relevance assessment efforts both when artificially

[5] Ties are not broken in this analysis and are simply excluded from likelihood computation.



Description:
Document mentions a leveraged buyout valued at or above 200 million dollars.

Title:	Stop & Shop Rejects Latest Dart Offer	Governor Signs Takeover Protection Bill	Black & Decker Gets Financing...
Content:	Dart Group Corp. says it will press itsbillion-dollar buyout bid for Stop andamp; Shop Cos. despite a unanimousrejection of the offer by the Stop andamp; Shop board of directors...	Gov. Richard Celeste signed into law Friday emergency legislation that enables the state to assess the economic impact of takeover bids by foreign corporations. The law, which takes effect immediately...	Black & Decker Corp. announced Friday that a group of commercial banks has arranged $2.2 billion in financing for the company's $2.1 billion hostile takeover...
Relevance:	Most Relevant ▾	Least Relevant ▾	Second-most Relevant ▾

Figure 8: The decoy interface asks workers to assign relative preference labels for three documents.

primed but also when merely revealing true previous votes. In the further course of this document we will only consider this second, more realistic realization of the Bandwagon Effect.

3.6 Decoy Effect

The Decoy Effect (sometimes also referred to as the Asymmetric Dominance Effect) occurs when workers' preference between options A and B changes in favor of option B when an option C is presented, which is similar but clearly inferior to option B [31]. It is a well-known and frequently exploited bias studied in marketing and advertising where decoy products are introduced to seemingly increase the actual product's viability in comparison to actual market alternatives. We simulate this situation by asking workers rank ("most relevant", "second-most relevant", "least relevant") three documents with respect to the same topic. A screenshot of the resulting interface can be found in Figure 8.

From each of the three corpora, we randomly draw 50 relevant topic-document pairs that will serve as our choice A. For each such pair, we draw a non-relevant document B and additionally select an additional document C that is similar to B ($cos(B,C) \geq 0.7$) but less relevant ($BM25(B) > BM25(C)$). Finally, for the same topic, we draw another non-relevant document D that shares no considerable similarity with either A or B ($max(cos(A,D), cos(B,D)) \leq 0.3$). If, for a given topic these requirements cannot be fulfilled, we reject the current pair and draw another one until our pool of 50 unique tuples $\langle A,B,C,D \rangle$ per corpus is filled.

We now compare two experimental conditions: First, we present workers with options A, B, C, a setup in which we expect to see some expression of the Decoy Effect. In our control group, workers see options A,B,D, where no such effect should occur. The relative ordering of options A, B, C, and D is chosen randomly for every task and does not reflect relevance in any way. Figure 9 plots the various likelihoods of being assigned a given relevance rank across options and experimental conditions. As expected, the single relevant option, A, shows the highest probability of receiving the "most relevant" rank in both settings. If we, however, compare the observed likelihood of option B being "most relevant", we note a considerable increase in likelihood when displaying A and B alongside decoy option C ($p(B) = 0.35$) rather than the unrelated D ($p(B) = 0.12$). This trend shows that there is a considerable risk of suffering from Decoy Effects in real-world crowdsourcing scenarios when multiple options are shown for relative ranking. This becomes especially relevant since several studies (*e.g.*, [6, 58, 59]) find human assessors to be more reliable when assigning relative preferences rather than absolute relevance levels to documents.

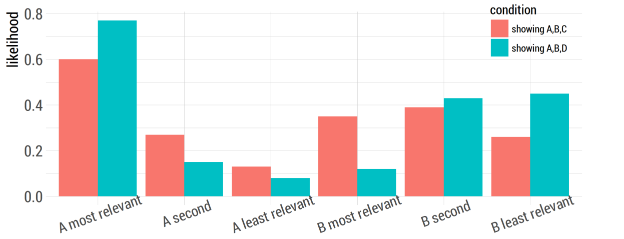


Figure 9: The observed likelihood of assigning higher relevance ranks to non-relevant option B increases when a similar but less relevant option C is presented.

For the remainder of this paper, to ensure consistency with the labels collected under the other biased processes, we translate preference ranks into binary relevance labels. To do so, we use the fact that in each tuple $\langle A,B,C,D \rangle$, only a single document A is in fact relevant. In consequence, the rank "most relevant" will be assigned a relevant vote and all remaining ones receive a non-relevance vote.

4 EXPERIMENTS

Building on the initial evidence presented in the previous section, we conduct a series of experiments in which the effect of cognitive biases is quantified in terms of label accuracy, retrieval system evaluation and machine-learned ranker effectiveness.

4.1 Label Accuracy

Let us begin by establishing a side-by-side quality comparison of relevance labels contributed by biased and unbiased crowd labeling processes. In this section, we measure the accuracy of the various crowdsourcing methods at reproducing NIST expert judgments. To this end, we consider 6 crowdsourced relevance label collection processes:

AE
Crowd labels obtained under influence of the Ambiguity Effect with likelihood p_m of hiding additional uninformative fields of 0.5.

A
Crowd labels obtained under influence of Anchoring when using a two-stage judging process.

BE
Crowd labels obtained under potential influence of the Bandwagon Effect by revealing the true distribution of previous workers' votes.

DE
Crowd labels obtained under influence of the Decoy Effect

Table 3: Accuracy of crowdsourced relevance votes at matching the existing NIST expert labels.

Collection	Year	BC	AE	A	BE	DE	UC
CW'12	2013	0.67*	0.68*	0.65*	0.70*	0.69*	0.83
	2014	0.77*	0.66*	0.63*	0.75*	0.67*	0.85
CDS	2014	0.58*	0.65*	0.52*	0.70	0.66*	0.71
	2015	0.75	0.71*	0.71*	0.60*	0.68*	0.79
	2016	0.65	0.58*	0.59*	0.67	0.65	0.68
TIPSTER	1992	0.78	0.78	0.66*	0.65*	0.67*	0.82
	1993	0.69*	0.68*	0.61*	0.71*	0.79	0.84
	1994	0.66*	0.78	0.59*	0.78	0.81	0.81
	1995	0.68*	0.80	0.68*	0.60*	0.77	0.82
Overall		0.69*	0.70*	0.63*	0.69*	0.71*	0.79

when consdering only the single most highly ranked document relevant.

BC
General biased crowd labels obtained via aggregation of all previously listed processes.

UC
A baseline crowdsourcing process designed to trigger none of the previously discussed biases.

We randomly sample 500 topic-document pairs from the NIST qrels of each TREC task edition (4500 pairs in total). For each of these pairs we collect a crowd label using the five organic crowdsourcing processes (AE, A, BE, DE, UC), giving our experimental dataset for this stage an overall volume of 9 tasks × 5 crowd processes × 500 labels × 3 workers = 135k labels. BC, finally, is merely the aggregation of all biased crowd collections (AE, A, BE, DE) in which final labels are determined via uniform majority voting with coin tosses as tie breaker.

All labels were collected on the Amazon Mechanical Turk crowdsourcing platform[6] at a rate of 0.03 US$ per label. Table 3 lists per-task and overall accuracies for the various crowdsourcing processes. Ground truth labels are given by NIST expert annotations. Statistically significant performance losses between biased crowd processes and the bias-free reference process (UC) are indicated by an asterisk. Statistical significance is determined using a Wilcoxon signed-rank test at $\alpha < 0.05$-level.

We note that, while news and Web topics can be reliably answered by UC crowd judges, the biomedical topics seem to be considerably more difficult to solve due to the high degree of medical jargon as well as the required domain-specific training. An exception to this general trend is the 2015 edition of the CDS track, the topics of which seem to be easier to solve. This observation has been previously made by CDS participants [23]. Despite some local variance, we note significant and substantial overall performance losses when considering biased crowdsourcing processes. Most notably, Anchoring shows dramatic effects with relative accuracy losses of up to 28%. In case of the already difficult CDS topics, for example, this lowers worker accuracy to a level that barely surpasses random guessing.

Table 4: Historic runs submitted to TREC tracks are ranked by descending nDCG according to the various crowdsourced label collection processes. Spearman's ρ describes their correlation to the original NIST ranking.

Collection	Year	# runs	# qrels	ρ_{BC}	ρ_{AE}	ρ_{A}	ρ_{BE}	ρ_{DE}	ρ_{UC}
CW'12	2013	61	47k	0.71	0.76	0.69	0.76	0.73	0.85
	2014	30	44k	0.84	0.71	0.64	0.83	0.74	0.93
CDS	2014	102	38k	0.61	0.70	0.53	0.79	0.70	0.79
	2015	178	38k	0.84	0.75	0.72	0.67	0.69	0.88
	2016	115	38k	0.71	0.65	0.61	0.70	0.66	0.74
TIPSTER	1993	38	63k	0.78	0.70	0.63	0.73	0.82	0.90
	1994	40	97k	0.73	0.81	0.64	0.82	0.86	0.89
	1995	39	87k	0.70	0.87	0.70	0.65	0.84	0.86

With respect to **RQ1**, we conclude that all investigated forms of cognitive bias show significant negative impact on the quality of crowdsourced relevance assessments.

4.2 Consequences for System Evaluation

Aside from the previously studied effect of cognitive biases on the immediate quality of crowdsourced relevance assessments, it is conceivable that such quality differences should influence retrieval system evaluation. To test for this hypothesis, we use the relevance judgments produced by the various biased crowdsourcing processes discussed earlier to rank a wide range of retrieval systems according to their retrieval performance. In a second step, we compare these system rankings to the one induced by the labels that were obtained from NIST experts (NE) and measure degree to which the relative ordering of retrieval systems is perturbed by cognitive label collection biases.

For this comparison, we consider all historic submissions to the relevant TREC adhoc, Web and CDS tracks and rank them in terms of nDCG. We can now compute the Spearman rank correlation coefficient ρ between system orderings induced by the various crowdsourced label collections and the ground truth NIST ordering. Values of ρ close to 1[7] show robust rankings in which no major differences to the original NIST ranking can be found. The further ρ's value differs from 1, the stronger the ranking perturbation, leading to faulty relative performance comparisons between retrieval systems. As ρ approaches 0, there is no systematic association between original and crowdsourced rankings any more.

Table 4 lists, for all three corpora and their respective editions, the number of historic runs, the amount of available NIST relevance assessments and the resulting values of ρ. Note that we were unable to obtain TREC-1 results from the NIST archive and therefore only include the 1993–1995 editions.

We notice that unbiased crowd labels (UC) show only minor perturbations from the original NIST rankings and generally yield robust system orderings. Again, for the 2014 and 2016 editions

[6] https://www.mturk.com/

[7] Values close to -1 are similarly desirable as they would reflect a near-perfect inverse system ranking that would, again, allow correct interpretation. In practice, however, we do not expect to observe such dramatic perturbations.

of CDS, the previously noted low crowd accuracy translates into somewhat reduced ranking coherency. As we turn towards the various biased crowd processes, we can observe varying degrees of perturbation, reaching, in the most severe case, a ρ of 0.53, signifying considerably different system rankings based on NIST and biased crowd judges. The various bias types appear to equally affect ranking robustness with Anchoring, again being a negative outlier.

With respect to **RQ2**, we note considerable differences in relative system orderings when using biased crowdsourcing labels as ground truth relevance indicators. In contrast, unbiased crowdsourced labels result in rankings that are largely identical to original NIST rankings.

4.3 Consequences for Derived Systems

In the previous sections, we discussed the impact of cognitive biases on relevance label accuracy as well as its propagated influence on retrieval system evaluation. In this final experimentation section, we are interested in the impact that such biases have on the quality of derived systems such as machine-learned rankers. To this end, we use the collected relevance assessments to train a range of data-driven rankers and compare three experimental conditions: Rankers based on biased crowds (BC), rankers based on unbiased crowds (UC) and rankers based on NIST expert labels (NE). As before, the biased condition is subdivided into its four constituent biases.

In each condition, we train a LambdaMART classifier [56] using a range of standard learning-to-rank features[8] and a neural network-based ranker [24]. As an additional point of reference, we include a static BM25 retrieval model [45] that does not require any training data[9]. The available topics are split into 10 non-overlapping folds and each model is trained on 9 varying folds and evaluated on the remaining one in cross-validation fashion. Performance evaluation is always based on the NIST-provided expert judgments. Table 5 shows the results of this experiment in terms of nDCG. Statistically significant differences between a biased run and its corresponding UC counterpart are indicated by an asterisk. Statistical significance is determined using a Wilcoxon signed-rank test at $\alpha < 0.05$-level.

The general trend is for both machine-learned systems to outperform simple BM25 rankers trained on NIST or UC labels. In this comparison LambdaMART is in most cases superior to DRMM. Systems trained on biased crowd labels, however, deteriorate rapidly in performance and often reach a level of ranking quality that is below that of non-parametric exact-matching models such as BM25. Again, this tendency is observed to be most severe for labels produced under the influence of Anchoring or the Bandwagon Effect.

With respect to **RQ3**, we find that cognitive biases as modelled in this paper have a significant negative effect on the performance of derived systems such as machine-learned rankers.

5 CONCLUSION

This paper investigates the effect of cognitive biases on the quality of submissions to standard document relevance assessment tasks. We concentrate on four common types of bias and, in a series of experiments on well-known TREC research corpora, demonstrate the significant detrimental influence that biased label collection can have on label quality, retrieval system evaluation and ranker training. Especially the more subtle forms of bias, *e.g.*, the Bandwagon or Decoy effects can occur unintentionally in crowdsourcing experiment protocols and should be carefully checked for in order to avoid label degradation.

Table 5: Retrieval performance of rankers trained on biased (BC) or unbiased crowd data (UC) as well as NIST expert labels (NE) in terms of nDCG. BC aggregates the various individual bias types.

Collection	Model	BC	AE	A	BE	DE	UC	NE
CW'12	LambdaMART	0.27	0.25*	0.22*	0.25*	0.23*	0.28	0.34
	DRMM	0.22*	0.25*	0.24*	0.20*	0.24*	0.30	0.32
	BM25	0.24	0.24	0.24	0.24	0.24	0.24	0.24
CDS	LambdaMART	0.18*	0.21*	0.16*	0.23	0.22	0.24	0.32
	DRMM	0.15*	0.19	0.17*	0.15*	0.18	0.21	0.24
	BM25	0.21	0.21	0.21	0.21	0.21	0.21	0.21
TIPSTER	LambdaMART	0.24*	0.26*	0.21*	0.28*	0.26*	0.32	0.38
	DRMM	0.20*	0.27	0.22*	0.23*	0.24*	0.28	0.32
	BM25	0.25	0.25	0.25	0.25	0.25	0.25	0.25

There are several exciting direction for future inquiry. In this paper, we focus on a range of newly conducted crowdsourcing initiatives that are instrumented to induce cognitive bias and demonstrate its effect on result quality. In the future, instead, it would be interesting to devise formal tests to analyze existing historic labeling efforts such as done in the course of the TREC Crowdsourcing track in terms of their susceptibility to bias. While this paper concentrates on crowd judges, a class of human laborers that has previously been shown to be prone to distraction and noisy submissions, there is no reason why test subjects in controlled lab environments should not suffer from the same effects. In the future, we suggest further confirming this assumption with a comparable lab study.

REFERENCES

[1] Omar Alonso, Daniel E Rose, and Benjamin Stewart. 2008. Crowdsourcing for relevance evaluation. In *ACM SigIR Forum*, Vol. 42. ACM, 9–15.

[2] Piyush Bansal, Carsten Eickhoff, and Thomas Hofmann. 2016. Active Content-Based Crowdsourcing Task Selection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management.* ACM, 529–538.

[3] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy* 100, 5 (1992), 992–1026.

[4] Roi Blanco, Harry Halpin, Daniel M Herzig, Peter Mika, Jeffrey Pound, Henry S Thompson, and Thanh Tran Duc. 2011. Repeatable and reliable search system evaluation using crowdsourcing. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.* ACM, 923–932.

[5] Herbert Bless, Klaus Fiedler, and Fritz Strack. 2004. *Social cognition: How individuals construct social reality.* Psychology Press.

[6] Ben Carterette, Paul Bennett, David Chickering, and Susan Dumais. 2008. Here or there. *Advances in Information Retrieval* (2008), 16–27.

[7] Dana Chandler and John Joseph Horton. 2011. Labor allocation in paid crowdsourcing: Experimental evidence on positioning, nudges and prices. *Human Computation* 11 (2011), 11.

[8] Cyril Cleverdon. 1997. Readings in Information Retrieval. In *Readings in Information Retrieval*, Karen Sparck Jones and Peter Willett (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, Chapter The Cranfield Tests on Index Language Devices, 47–59. http://dl.acm.org/citation.cfm?id=275537.275544

[8] We use all non-proprietary features (1–130) described at https://www.microsoft.com/en-us/research/project/mslr/. For newswire and medical domains, we omit Web-based features such as PageRank.

[9] While one can use training data to tune the model's parameters, we refrain from doing so and use robust common choices of $k_1 = 1.5$ and $b = 0.75$ throughout all experiments.

[9] Kevyn Collins-Thompson, Paul Bennett, Charles LA Clarke, and Ellen M Voorhees. 2014. *TREC 2013 web track overview.* Technical Report. MICHIGAN UNIV ANN ARBOR.

[10] Kevyn Collins-Thompson, Craig Macdonald, Paul Bennett, Fernando Diaz, and Ellen M Voorhees. 2015. *TREC 2014 web track overview.* Technical Report. MICHIGAN UNIV ANN ARBOR.

[11] Martin Davtyan, Carsten Eickhoff, and Thomas Hofmann. 2015. Exploiting Document Content for Efficient Aggregation of Crowdsourcing Votes. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management.* ACM, 783–790.

[12] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics* (1979), 20–28.

[13] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, and Philippe Cudré-Mauroux. 2014. Scaling-up the Crowd: Micro-Task Pricing Schemes for Worker Retention and Latency Improvement. In *Second AAAI Conference on Human Computation and Crowdsourcing.*

[14] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2013. Pick-a-crowd: tell me what you like, and i'll tell you what to do. In *Proceedings of the 22nd international conference on World Wide Web.* International World Wide Web Conferences Steering Committee, 367–374.

[15] Carsten Eickhoff. 2014. Crowd-powered experts: Helping surgeons interpret breast cancer images. In *Proceedings of the First International Workshop on Gamification for Information Retrieval.* ACM, 53–56.

[16] Carsten Eickhoff and Arjen de Vries. 2011. How crowdsourcable is your task. In *Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the fourth ACM international conference on web search and data mining (WSDM).* 11–14.

[17] Carsten Eickhoff and Arjen P de Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval* 16, 2 (2013), 121–137.

[18] Carsten Eickhoff, Christopher G Harris, Arjen P de Vries, and Padmini Srinivasan. 2012. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval.* ACM, 871–880.

[19] Daniel Ellsberg. 1961. Risk, ambiguity, and the Savage axioms. *The quarterly journal of economics* (1961), 643–669.

[20] Boi Faltings, Radu Jurca, Pearl Pu, and Bao Duy Tran. 2014. Incentives to counter bias in human computation. In *Second AAAI conference on human computation and crowdsourcing.*

[21] Marvin Fleischmann, Miglena Amirpur, Alexander Benlian, and Thomas Hess. 2014. Cognitive biases in information systems research: a scientometric analysis. (2014).

[22] Catherine Grady and Matthew Lease. 2010. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk.* Association for Computational Linguistics, 172–179.

[23] Simon Greuter, Philip Junker, Lorenz Kuhn, Felix Mance, Virgile Mermet, Angela Rellstab, and Carsten Eickhoff. 2016. ETH Zurich at TREC Clinical Decision Support 2016.. In *TREC.*

[24] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management.* ACM, 55–64.

[25] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: from discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 2125–2126.

[26] Donna Harman. 1992. The DARPA tipster project. In *ACM SIGIR Forum*, Vol. 26. ACM, 26–28.

[27] Martie G Haselton, Daniel Nettle, and Damian R Murray. 2005. The evolution of cognitive bias. *The handbook of evolutionary psychology* (2005).

[28] Matthias Hirth, Tobias Hoßfeld, and Phuoc Tran-Gia. 2010. Cheat-detection mechanisms for crowdsourcing. *University of Würzburg, Tech. Rep* 4 (2010).

[29] Mehdi Hosseini, Ingemar J Cox, Nataša Milić-Frayling, Gabriella Kazai, and Vishwa Vinay. 2012. On aggregating labels from multiple crowd workers to infer relevance of documents. In *Advances in information retrieval.* Springer, 182–194.

[30] Jeff Howe. 2008. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business* (1 ed.). Crown Publishing Group, New York, NY, USA.

[31] Joel Huber, John W Payne, and Christopher Puto. 1982. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of consumer research* 9, 1 (1982), 90–98.

[32] Panos Ipeirotis. 2011. Crowdsourcing using mechanical turk: quality management and scalability. In *Proceedings of the 8th International Workshop on Information Integration on the Web: in conjunction with WWW 2011.* ACM, 1.

[33] Panagiotis G Ipeirotis and Praveen K Paritosh. 2011. Managing crowdsourced human computation: a tutorial. In *Proceedings of the 20th international conference companion on World wide web.* ACM, 287–288.

[34] David R Karger, Sewoong Oh, and Devavrat Shah. 2014. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research* 62, 1 (2014), 1–24.

[35] Gabriella Kazai, Jaap Kamps, Marijn Koolen, and Natasa Milic-Frayling. 2011. Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.* ACM, 205–214.

[36] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2011. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management.* ACM, 1941–1944.

[37] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information retrieval* 16, 2 (2013), 138–178.

[38] Gabriella Kazai and Natasa Milic-Frayling. 2009. On the evaluation of the quality of relevance assessments collected through crowdsourcing. In *SIGIR 2009 Workshop on the Future of IR Evaluation.* 21.

[39] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems.* ACM, 453–456.

[40] Matthew Lease. 2011. On Quality Control and Machine Learning in Crowdsourcing. *Human Computation* 11 (2011), 11.

[41] Catherine C Marshall and Frank M Shipman. 2011. The ownership and reuse of visual media. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries.* ACM, 157–166.

[42] Rohan Ramanath, Monojit Choudhury, Kalika Bali, and Rishiraj Saha Roy. 2013. Crowd Prefers the Middle Path: A New IAA Metric for Crowdsourcing Reveals Turker Biases in Query Segmentation.. In *ACL (1).* 1713–1722.

[43] Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, and William R Hersh. 2016. Overview of the TREC 2016 Clinical Decision Support Track.. In *TREC.*

[44] Kirk Roberts, Matthew S Simpson, Ellen M Voorhees, and William R Hersh. 2015. Overview of the TREC 2015 Clinical Decision Support Track.. In *TREC.*

[45] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, and others. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995), 109.

[46] Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines.. In *LREC.* 859–866.

[47] Matthew S Simpson, Ellen M Voorhees, and William Hersh. 2014. *Overview of the trec 2014 clinical decision support track.* Technical Report. LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS BETHESDA MD.

[48] Wei Tang and Matthew Lease. 2011. Semi-supervised consensus labeling for crowdsourcing. In *SIGIR 2011 workshop on crowdsourcing for information retrieval (CIR).*

[49] Amos Tversky and Daniel Kahneman. 1975. Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making.* Springer, 141–162.

[50] Ellen M Voorhees, Donna K Harman, and others. 2005. *TREC: Experiment and evaluation in information retrieval.* Vol. 1. MIT press Cambridge.

[51] Jeroen Vuurens, Arjen P de Vries, and Carsten Eickhoff. 2011. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11).* 21–26.

[52] Jing Wang, Siamak Faridani, and Panagiotis G Ipeirotis. 2011. Estimating the Completion Time of Crowdsourced Tasks Using Survival Analysis Models. *Crowdsourcing for Search and Data Mining (CSDM 2011)* (2011), 31.

[53] Tianyi Wang, Gang Wang, Xing Li, Haitao Zheng, and Ben Y Zhao. 2013. Characterizing and detecting malicious crowdsourcing. *ACM SIGCOMM Computer Communication Review* 43, 4 (2013), 537–538.

[54] Fabian L Wauthier and Michael I Jordan. 2011. Bayesian bias mitigation for crowdsourcing. In *Advances in neural information processing systems.* 1800–1808.

[55] Ryen White. 2013. Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval.* ACM, 3–12.

[56] Qiang Wu, Christopher JC Burges, Krysta M Svore, and Jianfeng Gao. 2010. Adapting boosting for information retrieval measures. *Information Retrieval* 13, 3 (2010), 254–270.

[57] Yan Yan, Glenn M Fung, Rómer Rosales, and Jennifer G Dy. 2011. Active learning from crowds. In *Proceedings of the 28th international conference on machine learning (ICML-11).* 1161–1168.

[58] Peng Ye and David Doermann. 2013. Combining preference and absolute judgements in a crowd-sourced setting. In *Proc. of Intl. Conf. on Machine Learning.* 1–7.

[59] Dongqing Zhu and Ben Carterette. 2010. An analysis of assessor behavior in crowdsourced preference judgments. In *SIGIR 2010 workshop on crowdsourcing for search evaluation.* 17–20.

[60] Honglei Zhuang and Joel Young. 2015. Leveraging in-batch annotation bias for crowdsourced active learning. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining.* ACM, 243–252.