Reliable Task Design for Descriptive Crowdsourcing

Peter Organisciak
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Champaign, IL

Abstract

Crowdsourcing offers a valuable method to improve information retrieval indexing by using humans to improve the indexable data about documents or entities. Human contributions open the door to latent information, subjective judgments, and other encoding of difficult to extract data. However, such contributions are also subject to variance from the inconsistencies of human interpretation. The proposed dissertation studies the problem of such variance in crowdsourcing for information retrieval, and investigates how it can be controlled both in already collected data and in collecting new data.

This paper outlines a corresponding study where the effect of different contribution system designs on the resulting data is compared in paid crowdsourcing environments. At the heart of this study is the assumption of honest-but-biased contributors. Rather than focusing on finding dishonest or unreliable contributors, a well-studied problem in crowdsourcing, this study focuses on strategies that understand the quirks and inconsistencies of humans in trying to account data reliability problems.

Introduction

> In these democratic days, any investigation in the trustworthiness and peculiarities of popular judgments is of interest – Francis Galton (1907)

The internet is growing increasingly interactive as it matures. Rather than simply transmitting information to readers, web pages allow their audience to react and interact with their information. The products of these interactions are a trove of qualitative judgements, valuable to modeling information objects. In recent years, this form of creation through collaboration has been studied as *crowdsourcing*.

Effective information retrieval depends on reliable, detailed information to index. Crowdsourcing has the potential to improve retrieval over web documents by having humans produce descriptive metadata about documents. Humans are able to provide latent information about documents that would not be possible to ascertain computationally, such as quality judgments or higher-level thematic description. They are also good at critical actions such as correcting, describing in different language, or inferring relationships with other documents. More importantly, crowdsourcing looks at human contribution at scales that are potentially useful for retrieval.

However, humans have predictable and unpredictable biases that make it difficult to systematically adopt their contributions in an information system. How do we control and interpret qualitative user contributions in an inherently quantitative system? This study looks at crowdsourcing for document metadata, which I refer to by the shorthand of *descriptive crowdsourcing*, and how to interpret this form of human contributed metadata in information retrieval.

Concretely, the proposed work is in two parts, separated by focus on *collecting* descriptive metadata reliably, and on *using* it in an appropriate information retrieval context. In line with the expertise at HCOMP, this paper focuses on the first part: looking at the effect of different collection interface designs on the intercoder reliability of the collected data. This is a study motivated by prior work, with a problem often mentioned but, to my knowledge, not pursued directly.

I argue that the reliability of crowdsourced data can be improved by making an assumption that crowd contributors are honest-but-biased. This assumption is not uncommon in tradition research on classification and information access, such as the literature on intercoder reliability, but is understudied in crowd research. The proposed study follows the hypothesis that such an assumption leads to more algorithmically valuable crowdsourced description and a greater proportion of useful contributions.

A reader of the proposed dissertation will understand:

- the issues related to using crowdsourcing contributions for information retrieval indexing;
- the effect of designing crowdsourcing collection tasks that encourage different contribution behaviours on a paid crowdsourcing platform, with a sense of how this information generalizes to different tasks or collection spaces; and
- the tractability of making an assumption of honest-but-biased contributors.

Motivation

The growth of digital collections has outpaced the ability to comprehensively clean, transcribe, and annotate the data. Similar roadblocks are affecting born-digital information,

where the rapid creation of documents often follows from passive or unrestricted forms of creation. The lack of strong descriptive metadata poses an obstacle for information retrieval, which must infer the aboutness of a document in order to surface it for an interested user. Crowdsourcing is increasingly being used to address this problem.

Many of the benefits of crowdsourcing follow from the fact that humans approach tasks in qualitative and abstract ways that are difficult to emulate algorithmically. A human can respond to complex questions on a Q&A website, judge the quality of a restaurant/product/film, or decipher a sloppy piece of handwriting. Since many information systems are intended to serve an information-seeking user, the information that crowdsourcing collects can better reflect the needs of users. For example, a user-tagged image in a museum collection can fill in terms that are more colloquial than the formal vocabulary employed by a cataloguer (Springer et al. 2008; Trant and Wyman 2006). Such information is invaluable in indexing documents for information retrieval, where the goal is commonly to infer what a user is searching from their textual attempt to describe it in a query.

More than typical description, additional useful information can be reactionary or critical. Indexing human judgments of a document's quality, for example, can enable a information retrieval system to rank the best version of multiple similarly relevant document.

While the complex qualitative actions of human contributions are the cornerstone of such contributions' usefulness, they present a challenge for algorithmic use because they can be highly variable. A task becomes more open to interpretation the more complex it becomes. Some projects revel in the broad interpretive nature of complex tasks. We see large art projects like Star Wars Uncut embrace the quirkiness of humans. However, in cases where there is a goal to find either an objective truth, manifest or latent, or to gauge the subjective approaches and opinions of people in a comparable way, the breadth of interpretations possible for a task presents a problem for reliably understanding it in aggregate.

Background

The variability seen in human interpretations of complex tasks is not a novel issue. It is a problem known as low intercoder reliability, and can result from a variety of issues. Threats to reliability echo common issues seen in crowdsourcing document description: an insufficient coding scheme, inadequate training, fatigue, and problem coders (Neuendorf 2002).

Whereas much research has looked at the Neuendorf's fourth threat to reliability, when the contributors are the source of low reliability (Sheng, Provost, and Ipeirotis 2008; Whitehill et al. 2009; Welinder and Perona 2010; Raykar et al. 2009), the inclusion of the researcher/coordinator as a responsible party has not been common in crowdsourcing research. This study looks at the improvements in crowdsourcing for descriptive metadata that can be recovered from external factors, assuming an honest but biased rater and focusing instead on design issues like codebooks, training, and fatigue.

Much crowdsourcing research makes an adversarial assumption, focusing on removing variability by detecting or smoothing over cheaters. For example, Eickhoff and Vries (2012) note that a significant proportion of Mechanical Turk workers sacrifice correctness for speed, in order to maximise their profits. However, 'sacrificing quality for speed' is not always the case. For example, in past work we found that the fastest workers generally did not contribute worse labor (Organisciak et al. 2012), and in some cases we found that slowing workers down resulted in *lower* quality contributions (Organisciak et al. 2013).

There is some precedent, however, for looking at issues related to designing effective crowdsourcing tasks. Particularly, Grady and Lease explored the effect of changing human factors on information retrieval relevance judging through Mechanical Turk (2010). They considered four factors: terminology, base pay, offered bonus, and query wording. Though their findings were inconclusive, their study provides guidance on the issues related to this form of study. The proposed dissertation builds upon Grady and Lease's work, as well as other parameterization studies like Mason and Watts (2010), by evaluating more drastic deviations from the core structure of a paid crowdsourcing task.

The effect of wording and terminology, one of Grady and Lease's focal points, has often been alluded to as a factor in crowdsourcing. In writing about The Commons, a successful museum crowdsourcing project with Flickr, the Library of Congress reported that the "text announcing the Commons ('This is for the good of humanity, dude!!') struck just the right chord" (Springer et al. 2008). Jeff Howe relays a similar empirical story about citizen journalism, about a contribution button that went through two iterations of text that went unnoticed, before finding that "'Get Published' were the magic words" (Howe 2008). These empirical observations allude to more qualitative factors is collecting crowdsourcing contributions.

Alonso and Baeza-Yates have also written about the effect of different parameterizations of paid crowdsourcing tasks, considering the quality of relevance judgments with varying numbers of contributors evaluation each task, topics per task, and documents per query. In doing so, they cite interface design as the most important part of experimental design on Mechanical Turk and recommend following survey design guidelines and provided clear, colloquial instructions (Alonso and Baeza-Yates 2011). This study agrees with their sentiment, and strives to formally understand and articulate the differences that interface design influences in crowdsourcing.

The research in this study also follows as a logical progress from findings in my earlier research, and has been on multiple occasions been a 'future direction' worthy of direct focus.

In Organisciak et al. (2012) we found evidence that at least some error in crowdsourced relevance judgments stems from differing but not necessarily malicious interpretations of the task, suggesting that improved quality can follow from tweaks in design. Ways to encourage this behaviour were not pursued, but the results suggest that doing so might require workers to be more aware of their performance relative to

the codebook or norms, and to reassess their understanding of the task when it is necessary.

While performing other research (Organisciak et al. 2013), we found that asking people to reflect on their response changed the nature of their response, with less internal consistency than when they did not have to explain their choices. In another small study comparing the space of incidental crowdsourcing across two systems (Organisciak 2013), it was found that an 'easy' rating interface – one that puts up less hurdles to contribution – results in a shifted distribution of ratings than a 'hard' interface. Finally, in unpublished recent research on low grader consistency in the ground truth of a Music Information Retrieval Exchange (MIREX) task, one of the results found that redesigning the task to attach finer instruction to the rating improved the quality of judgments by crowdsourced judges.

Proposed Research

Humans don't operate with the formality of computers. Many of the benefits of crowdsourcing follow from that fact: human contributions are valuable specifically because they are not easily automated. However, when using crowd contributions to inform an algorithmic system, as in information retrieval, the inconsistencies of human work present a challenge.

In a controlled set up, crowdsourcing in information retrieval usually follows a typical design: a task, description, and a set of one or more documents that are reacted to. This type of design is common for creating custom evaluation datasets through relevance judgments (Alonso, Rose, and Stewart 2008), but has been used for encoding and verifying indexing information (Chen and Jain 2013).

Evidence suggests that the design of a data collection interface affects the quality and distribution of user contributions (Alonso, Rose, and Stewart 2008; Howe 2008; Organisciak 2013). The manner to improve on a basic task/description/items interface design is not immediately clear, though: some success has been attained by slowing workers down, while other times it has been beneficial to encourage cheaper, more impulsive contributions in larger numbers.

This study compares the effect of task design on collected information retrieval data. Scoped to a reasonable parameterization of crowdsourcing as it is commonly practiced in information retrieval – a typical encoding task performed by paid crowds, the following questions will be pursued:

- **RQ1**: Which approaches to collection interface designs are worth pursuing as alternatives to the basic design commonly employed in paid crowdsourcing?
- **RQ2**: Is there a significant difference in the quality, reliability, and consistency of crowd contributions for the same task collected through different collection interfaces?
- **RQ3**: Is there a qualitative difference in contributor satisfaction across different interfaces for the same task?
- **RQ4**: Do the questions above generalize to different tasks, task types, and contexts (i.e. outside of paid platforms)?

RQ1 is the question of design, on synthesizing prior work and brainstorming directions to explore. It is a partially subjective question, but one still worth pursuing with diligence. As recent research found, the effects seen in traditional user studies are still present in online crowd markets (Komarov, Reinecke, and Gajos 2013). Their finding suggests that non-crowdsourcing research in human-computer interaction is informative for our purposes.

RQ2 and RQ3 are the primary questions being explored in this chapter of the proposed dissertation, on quality for computational use and on satisfaction. While this dissertation is explicitly pursuing the former question, collecting computationally useful contributions needs to be understood in the context of contributor satisfaction. The trade-off between contributions that crowds want to make and the reliability of the data is a central consideration for fostering sustainable, or alternately affordable, crowdsourcing.

RQ4 is the question of generalizability. It is a broadly scoped question, but one that should be addressed as thoroughly as possible.

Design overview

This study will evaluate multiple interfaces for encouraging less deviation between human contributors. Motivated by earlier work, the particular focus will be on designs that slow down workers and make them aware of how their perception of the task deviates from the standard, and alternately designs that encourages quicker, "gut" responses.

I adopt an established information retrieval problem to control for the task: enriching terse microblogging messages through paid crowdsourcing. What is being completed is not as central to this study as how it is done, but this is a task that is structured similar to many on-demand crowdsourced information retrieval tasks.

Workers will identify the topic of a microblogging message from Twitter – a tweet. This is a task where the information object is sparse and the topics are often short-lived and previously unseen, making crowdsourcing a promising approach to improve information retrieval across the data. It is also a realistic task that has been attempted with crowdsourcing in the past.

Task

Microblogging messages, in this case from Twitter, are notably brief, often missing context and heavily abbreviated. This creates problems for parsing the topic of an individual message. The use of microblogging is so ephemeral and diverse that many information retrieval needs are completely new when introduced and only exist for a short period of time (Chen and Jain 2013). Due to the sparse information and novel needs of microblog retrieval, crowdsourcing has been used in this area, both for augmenting tweets and for creating datasets to train classifiers specific to the corpus.

The task in this study is a topic identification task: "Is this tweet about topic X?" Workers are shown a tweet that contains the terms of a query, Q, where Q represents an extracted entity. Their task will be to describe whether the entity is the topic of the tweet, or simply mentioned. Such a

task is useful, but potentially easy to misinterpret by contributors conflating a term being the topic of a tweet with merely being mentioned in the tweet. To keep focused, all designs will pursue this topic identification task.

A second task will be a more difficult summarization task: "Find the most self-explanatory tweet from Set A." Workers will be shown ten microblogging messages from a trending topic and asked to identify the one tweet that best communicates the topic.

Exploring the design space

Before parameterizing the designs of the microblogging task to be studied, a brief exploration of the design space will help discussion.

Commonly, a paid crowdsourcing worker goes through the following steps:

1. Worker w arrives at task page
2. w is shown a preview of task t
3. Worker w accepts the task t
4. Work performs task t and submits
5. A new task t' is chosen and, worker is taken back to *step 2* or *step 3*

The above steps are the model used by Amazon Mechanical Turk when a task is followed through to completion. Workers are also given escape options, to skip, reject or return tasks.

Metadata encoding tasks generally consist of the following parts:

- **Goal** statement/question. *e.g. "Is this page relevant to query q?", "Find the topic of a tweet."*
- **Instructions** for performing the task.
- one or more **Items** that worker responds to. *e.g. webpage snippets, microblogging messages*
- **Action**, one per item: the data collection mechanism.

Within this framework, we can see a number of factors that may potentially affect how our microblog encoding task is completed. First are the parameterizations of the task within its existing structure. A task may change with different payment, bonuses, and quantities of tasks available. Instructions can differ on clarity, length, and restrictiveness versus interpretability. Items can differ in different task design in how many items are offered per task. Finally, the actions can differ by their complexity, such as the granularity of rating actions.

Of course, we're not constrained to the task structure provided above. We can add elements to the task design before the task is accepted, at the start of the task, during or in response to individual interactions, or after the task is completed. Taking away elements might also be possible. The possibilities are endless; to inspire useful ones, it is helpful to consider one naturalistic set of factors that may affect the outcome of a paid crowdsourcing task: worker behaviours.

A worker's contribution may be affected by a myriad of factors. Some possibilities include *experience*, *skill*, *self-confidence and decisiveness*, *attentiveness and fatigue*, *perceived importance of task*, and *time spent on each task*.

In a moment I'll rein in discussion to a smaller set of design interfaces to test. However, an exercise to think through the possibilities afforded to us by the features in the previous section will be helpful, in the style of gedanken experiments.

Consider this study's Twitter encoding task. How would the contribution change if tasks were 100 items long? 200? 1000? Only 1? What if instructions were written very tersely? Verbosely, with many examples? What if contributors were tested on the instructions at the beginning of the task? If there were gold label items throughout the task? If everything had a known answer and workers were inconvenienced (e.g. with a time delay) when they got an answer wrong?

How about if contributors were asked to volunteer their time? Were paid 1c per task, or 10c, or paid by the hour? What if contributors were paid bonuses for performance against a ground truth or internal consistency? What would be the effect of tasks/time quotas to meet for bonuses, and if they were forced into these quotas (with tasks automatically moving forward)? What if a timer showed how long until their task disappeared? Would bonuses be helpful for encouraging continued task completion, or based on some qualified difficulty of the task? Maybe contributors could be shown their performance? What if they were ranked against other workers? What if they gained levels or earned badges for performance? Contributors were told when they got something wrong? What if you lie to them?

Some of these ideas of exciting, others are unfeasible. Designs to encourage longer engagement from individuals do not appear to be a promising direction. Worker experience was previously measured (Organisciak et al. 2012) and found to not be significant for simple tasks. Other areas are already well-tread. The effect of incentive structures, payment and bonuses, has been studied frequently, notable by Mason and Watts (2010).

Proposed designs

So what interface designs will this study measure?

It is still unclear as to whether simple encoding tasks benefit more from workers using their brain or gut. Designs that can change a worker's attentiveness address an interesting problem and may bring potential improvements. Also, it should be seen whether a task can push a worker into internalizing the codebook rather than interpreting it. Finally, in assuming that many reliability errors are introduced by honest workers that intend to do well, it may also be important to keep workers informed of their performance, at least when they are not performing well.

With those considerations in mind, this study will compare the following interfaces to study for crowdsourced data collection: a basic interface, a training interface, a feedback interface, and a time-limited interface.

The *basic interface* will resemble an archetypal task, following conventions of Mechanical Turk tasks described earlier. This will be the baseline task.

In the *training interface*, a worker is walked through their first task slowly. As they complete the tasks, their answers are evaluated against a gold standard and they are informed

immediately if they completed it correctly or incorrectly. Incorrect answers will also be given an explanation of why the actual answer is correct.

In the *feedback interface*, a worker is shown feedback about their estimated performance on past tasks. The first task that they complete is identical to the basic interface. Starting with the second task, however, the top of the interface informs users of their estimated performance, in terms of agreement with other workers, and provides a visualization of where they fall in the distribution of all workers.

Not all crowdsourcing contribution cases require more focus: sometimes a worker in a quicker mode of thinking contributes more consistent and reliable work. In contrast to the training and feedback interfaces, which will serve to slow down workers and make them more focused on their contributions, the final data collection interface will pursue the opposite approach. The *time-limited interface* encourages quicker interactions by giving users a limited amount of time to complete all tasks.

The experiments in this study will be run in a naturalistic setting: running directly on a paid crowdsourcing platform, Amazon Mechanical Turk, with real workers. There are trade-offs to this setting. It is easy to instrumentalize and properly captures the actual skills and attentiveness of paid crowd workers. However, working within the conventions of the system means that some parts cannot be controlled. Workers cannot be forced to perform multiple tasks, only encouraged. Also, the actual user pools testing the different interfaces are not necessarily the same individuals, making it important to pursue demographic similarities through geographic and temporal restrictions.

Issues and Challenges

In addition to the concerns discussed throughout this paper, it will be important to stay sensitive to the unforeseen. There are the unforeseen effects: both paid and volunteer crowdsourcing are motivated by a complex array of choices that may affect the user's interpretation and performance of a task. There are also unforeseen precedents: while this study is grounded in information science, human-computer interaction, and the growing crowdsourcing literature, it is important to stay aware of other fields that may offer relevant research, such as social psychology or marketing. Finally, there are the unforeseen unknowns, potential study pitfalls that are only learned through practice. For all of these, the best protection is to seek the advice of colleagues and mentors throughout the study.

References

Alonso, O., and Baeza-Yates, R. 2011. Design and implementation of relevance assessments using crowdsourcing. In Clough, P. e. a., ed., *Advances in Information Retrieval*, number 6611 in Lecture Notes in Computer Science. Springer Berlin Heidelberg. 153–164.

Alonso, O.; Rose, D. E.; and Stewart, B. 2008. Crowdsourcing for relevance evaluation. *SIGIR Forum* 42(2):915.

Chen, E., and Jain, A. 2013. Improving twitter search with real-time human computation.

Eickhoff, C., and Vries, A. P. 2012. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*.

Galton, F. 1907. Vox populi. *Nature* 75:450451.

Grady, C., and Lease, M. 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*, CSLDAMT '10, 172179. Stroudsburg, PA, USA: Association for Computational Linguistics.

Howe, J. 2008. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Business, 1 edition.

Komarov, S.; Reinecke, K.; and Gajos, K. Z. 2013. Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 207–216. New York, NY, USA: ACM.

Mason, W., and Watts, D. J. 2010. Financial incentives and the "performance of crowds". *SIGKDD Explor. Newsl.* 11(2):100108.

Neuendorf, K. A. 2002. *The Content Analysis Guidebook*. Thousand Oaks, CA, USA: Sage Publications.

Organisciak, P.; Efron, M.; Fenlon, K.; and Senseney, M. 2012. Evaluating rater quality and rating difficulty in online annotation activities. *Proceedings of the American Society for Information Science and Technology* 49(1):110.

Organisciak, P.; Teevan, J.; Dumais, S.; Miller, R. C.; and Kalai, A. T. 2013. Personalized human computation.

Organisciak, P. 2013. Incidental crowdsourcing: Crowdsourcing in the periphery. DH '13.

Raykar, V. C.; Yu, S.; Zhao, L. H.; Jerebko, A.; Florin, C.; Valadez, G. H.; Bogoni, L.; and Moy, L. 2009. Supervised learning from multiple experts: whom to trust when everyone lies a bit. ICML '09, 889896. New York, NY, USA: ACM.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. KDD '08, 614622. New York, NY, USA: ACM.

Springer, M.; Dulabahn, B.; Michel, P.; Natanson, B.; Reser, D. W.; Ellison, N. B.; Zinkham, H.; and Woodward, D. 2008. For the common good: The library of congress flickr pilot project.

Trant, J., and Wyman, B. 2006. Investigating social tagging and folksonomy in art museums with steve. museum. In *Proceedings of the WWW06 Collaborative Web Tagging Workshop*.

Welinder, P., and Perona, P. 2010. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 25–32. IEEE.

Whitehill, J.; Ruvolo, P.; Wu, T.; Bergsma, J.; and Movellan, J. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems* 22:20352043.