CROWDSOURCING FOR DOCUMENT METADATA: RELIABLY AUGMENTING DOCUMENTS WITH CROWD CONTRIBUTIONS

BY

PETER ORGANISCIAK

DISSERTATION PROPOSAL

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy at the Graduate School of Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Michael Twidale, Professor
Miles Efron, Associate Professor
Stephen J. Downie, Associate Dean for Research
Jaime Teevan, Senior Researcher, Microsoft Research

CONTENTS

Chapter 1

INTRODUCTION

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

The internet is growing increasingly interactive as it matures. Rather than merely 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 can 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 adept 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, I am proposing a study in two parts, separated by their focus on *collecting* descriptive metadata reliably, and on *using* it in an appropriate information retrieval context.

1. In the first half, I will look at the effect of different designs for collection of descriptive metadata on the intercoder reliability of the collected data. This is a study motivated by prior work done by myself and others, with a problem often mentioned but, to my knowledge, not pursued formally.

2. In the second half of this dissertation, I will look at improving retrieval using already-collected crowdsourcing data. I focus on the system *Pinterest*, which is a valuable resource of human-encoded descriptive metadata, while sparse in its other textual content. It is also a example of the loosely constrained form of crowdsourcing contribution that is often required to encourage participation, a trade-off that is less structured than would be preferable for retrieval model.

I argue that the reliability of crowdsourced data can be improved by making an assumption that crowd contributors are honest-but-biased[1]. This is an assumption supported by prior work and not uncommon in research on classification, 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 a) more algorithmically valuable crowdsourced description and b) a greater proportion of useful contributions.

1.0.1 Approach

The study of collecting and modeling document metadata through crowdsourcing will be done in two different sites of crowdsourcing:

- in the design of effective contribution tasks, and

[1] In assuming that humans are biased, the biases referred to are the inclinations, leanings, and tendencies (*bias, adj., n., and adv.* 2014) of individuals, quirks that affect their worldview and how they understand and perform tasks. By this definition, such perceptual differences contribute to a greater statistical variance than if all contributions were expected to be identical, and should not be confused with the statistical definition of 'bias', referring to a model that is overfit, or overly 'biased,' to a particular dataset. In information science, this is closely related to *intercoder reliability*, the measure of how similar multiple coders will perform in a given parameterization of a task (Neuendorf 2002). When discussing the processes of humans –and only when doing so– this study may refer to biases, but discussion of effects on data will solely use statistical and information science language.

- in modeling normative meaning of contributions after they have already been collected.

Doing so will both adopt work that I have performed during my doctoral studies and contribute new research.

This study cannot account for all possible situations and methods for crowdsourced document enrichment. Rather, each chapter will focus on a novel sub-problem within the area, providing a grounding from which I can more thoroughly explore the larger problem space.

In the first sub-study, the performance of workers completing the same document description task will be compared across different design interface. In the second sub-study, I develop strategies for making use of crowdsourced information. Here, an information retrieval model is developed that incorporates a crowdsourcing-heavy system's user contributions in retrieval. Both studies stay true to the assumption of honest-but-biased workers, focusing on the responsibilities of a system designer in managing and interpreting the crowd rather than the faults of individuals in the crowd.

1.0.2 Take Away

A reader of the proposed dissertation will understand:

- the issues related to using crowdsourcing contributions for improving document metadata, particularly for information retrieval indexing;
- the effect of different designs of crowdsourcing collection tasks on the resulting reliability and consistency of the collected data, particularly designs that train workers, give them feedback, or hurry them;
- ways to use loosely-structured crowd contributions for retrieval, particularly user-curated lists; and
- the tractability of making an assumption of honest-but-biased contributors.

1.1 Crowdsourcing

Crowdsourcing is the distributed, large-scale collaboration of users contributing to a common product. Significantly, the term describes the *act* of a system opening up for contributions from distributed users. Users do not necessarily collaborate directly with each other – though they can – so the crowd in the term refers broadly to the collective users of the system. Sourcing describes the act of soliciting user contribution, regardless of whether it is successfully executed or not.

Crowdsourcing is an umbrella term preceded by a number of more narrowly scoped concepts, such as commons-based peer production (Benkler 2006), open source software development (Raymond 1999; Lakhani and E. v. Hippel 2003), and human computation (L. v. Ahn 2006; Law and L. v. Ahn 2011).

Surowiecki discussed aggregate crowd intelligence as the 'wisdom of the crowds' (Surowiecki 2004); one way to interpret crowdsourcing is the process of trying to utilize that wisdom.

1.2 Problem

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 production. 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 increasing 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_investigating_2006]. Such information is invaluable in indexing items 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.

Similarly, other uses of crowdsourcing capitalize on humans' abilities to spot when algorithmic attempts at understanding an information object have failed. ReCaptcha uses human contributions to transcribe transcriptions of OCR problem text from Google Books and the New York Times (*What is reCAPTCHA?* 2008) The National Library of Australia's Trove also crowdsources corrections of scanned text, by allowing readers of their scanned newspapers to edit transcript text when they come across problems (Holley 2009).

Humans are also being used to encode parsable text descriptions for non-text materials or higher-level latent concepts. In libraries, this approach is being adopted with crowd transcription of materials which are too difficult for computer vision, such as digitized letters. For example, the Bentham Project at University College London has a pilot project for crowdsourcing the transcription of Jeremy Bentham's letters (Moyle, J. Tonra, and V. Wallace 2010; Causer, Justin Tonra, and Valerie Wallace 2012).

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

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, where hundred of people re-filmed small snippets of Star Wars in a charming hodgepodge of styles. Coding challenges, like those seen on TopCoder, also are interesting for the widely varying ways that a programming language allows you to express a problem. 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.

The variability seen in human interpretations of complex tasks is not a novel issue. It is a problem that we call low *intercoder reliability*, and can result from a variety of issues. Four 'threats to reliability' that Neuendorf

(2002) lists echo issues in crowdsourcing document description: an insufficient coding scheme, inadequate training, fatigue, and problem coders.

Whereas much research has looked at the fourth problem, when the contributors are the source of low reliability (e.g. Sheng, Provost, and Ipeirotis 2008; Whitehill et al. 2009; Welinder and Perona 2010; Raykar et al. 2009; Organisciak, Efron, et al. 2012), this study looks at the improvements in crowdsourcing for descriptive metadata that can be recovered from external factors: assuming an honest but biased rater.

1.2.1 Assumption of Honesty

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 Organisciak, Efron, et al. (2012), we found that the fastest workers generally did not contribute worse labor, except for one case: when workers spent less time on the instructions and first task. The fact that time was only significant in this one case suggests that the effect for this particular dataset was not a result of 'cheaters' as much as workers that did not interpret the instructions close enough. Similarly, during the research for Organisciak, Teevan, et al. (2013) we found that slowing workers down resulted in lower quality contributions, both in terms of internal consistency by workers and quality of the data for training a recommendation algorithm.

Following from signals that low-quality results stem from a number of causes, this research assumes that the quality of a contribution is not only affected by the objective quality of the worker, but also due to subjective differences in the worker's perception of the task.

$$Contribution = truth + \text{quality error} + \text{perception error} \tag{1.1}$$

This simplifying assumption underlies this proposal. While keeping in mind the possibility that variance can stem from good or bad quality contributors, this study is pursuing an understanding of that second bias: when contributors introduce variance that is stimulated by differing interpretations of task, ones that deviate from the instructive or normative ways to approach the task.

This assumption is not novel in areas of social research, but is neglected in crowdsourcing research. In views on intercoder reliability in tradition social science settings, reliability is treated as the responsibility of both the designer of the work and the workers themselves. In fact, bad workers are one of the last considerations when there are data problems.

The inclusion of the researcher/coordinator as a responsible party has not been common in crowdsourcing research. Perhaps it is because participants in crowdsourcing are more abstract than a local worker or survey taker, or maybe because the history of the Internet has justifiably encouraged a level of aloofness against dishonesty, but this dissertation hopes to see if this oversight is detrimental.

1.2.2 Intercoder reliability

In crowdsourcing, increasing intercoder reliability is sometimes at odds with the collection strategy. The most effective crowdsourcing deals with large numbers of people, and part of maximizing the involvement of contributors, especially those which are volunteers, is to minimize the restrictions on a contribution. To enforce a strong coding scheme or training contributors will reduce the number of individuals willing to perform the task. Whether the improvements in quality are worth the losses in contributions will be looked at in the first chapter of new research in this dissertation.

Other times, controlling the circumstances under which the contribution is created is not possible, such as in information retrieval over web documents. For tasks where the contribution is numeric and ordinally or continuously coded, methods exist for interpreting when coders are similar but operating with different frame. These include using covariation instead of agreement (Neuendorf 2002), and normalizing by a user mean (Hofmann 2004; Bell, Koren, and Volinsky 2008). A later chapter of the proposed dissertation will look at modeling sparse textual data on the crowdsourced website Pinterest, by smoothing – among other approaches –document models against other users' 'interpretations' of the same items.

1.2.3 Contribution variance

Variance that exists between different contributors adds noise both to tasks that make a subjective assumption and tasks that make an objective assumption.

In subjective tasks, it is assumed that there is no universally correct form of contribution. For example, when crowd contributions are used to inform recommendations, such as for music or film, it often assumed that different types of people enjoy different products. We thus see approaches to recommendation such as collaborative filtering, where users are matched to similar users based on the overlap between their tastes rather than a global definition of 'good' or 'bad' products. In such a case, inter-rater consistency is still important, to make it possible to identify similar users. Modern approaches to collaborative filtering commonly normalize ratings against a user-specific bias (i.e. "how does this rating compare this user's average rating") and sometimes against an item-specific bias (i.e. "how does this rating compare to what the rest of the community thinks about the item").

For objective tasks, Neuendorf (2002) differentiates between two types: manifest and latent.

In a simplified comparison, tasks with manifest content are ones where there is a clear correct contribution. Correcting or transcribing text from a scanned image would be grouped in the category.

In contrast, latent tasks are assumed to have a theoretical truth, but one that is not outwardly stated. When a person tags a photograph with a free-text label or a worker classifies the sentiment of an opinionated tweet, they are interpreting the content. As Neuendorf (2002) notes, "objectivity is a much tougher criterion to achieve with latent than with manifest variables".

1.2.4 Benefits of Recovering Error from Crowd Contributions

Why try to account for human error in crowdsourcing collection? With large enough numbers, it doesn't matter. Problems of user quality get smoothed over when enough honest people collaborate, while problems stemming from perception biases in many cases will converge on the normative understanding of the task. However, by recovering a cleaner signal from human contributions, a system is reliant on less workers. Doing so thus helps keep system less affected by the ebbs and flows of motivating volunteers, or the costs of paying workers. Since the attention that contributors is not uniform across all items in a system, usually resembling an inverse power-law distribution, understanding crowdsourced information with less aggregation means more of the middle of the distribution can be represented.

In other words, accounting for individual biases seeks to make each individual contribution more valuable.

1.3 Relevance

The contribution of this work is in the application of corrective techniques to the crowd-based encoding of metadata about existing information objects, and the broader understanding of the nature of such contributions.

There are many ways to apply a lens to such research. This study reflects my own field of information retrieval, and more broadly in information science.

Information science deals with the representation of information objects, giving crowdsourcing considerable potential as a tool for item description.

By way of example, consider crowd curation of materials. In the presence of large collections of information objects, information-seeking and discovery can be aided by user-curated lists of thematically-similar objects. Sites like Amazon, LibraryThing and the new Delicious let people create lists of products, books, and websites, respectively. The themes binding the lists are also user-defined, so a list can be about quality (e.g. “favorites”, “worst of”), thematic (e.g. “teen vampire romance novels”), or administrative (e.g. “to buy”, “read this year”). This crowdsourced information is useful to users directly, but it also provides high-quality information for understanding the content in a collection and its relationship to other materials.

Inversely, this can return value to users curating content themselves: consider a system that can discover further items for a user that are thematically in line with a group that they have compiled.

New OPACs are increasing giving users the ability to classify and curate content, connecting to user habits that are commonly associated with public libraries. For example, BiblioCommons – employed at the Edmonton and New York Public Libraries – positions list-making as a “curated topic guide,” a way to “share your expertise with others” (*Lists* 2011). According to one study of social OPACs, the list feature in BiblioCommons is heavily used, many times greater than commenting and more than ratings (Spiteri 2011).

Similarly, cultural heritage collections have reported past success in using crowd contributions for increasing discoverability to content, improving metadata quality, or even contributing to item description. For example, after a pilot partnership with Flickr, the Library of Congress implemented a workflow for review public comments on images for research or information to integrate back into item records (Springer et al. 2008).

Crowd curation is just one example of a use of crowdsourcing to create information. Table 1.3 shows a number of different actions that have been observed for collecting descriptive metadata.

While crowdsourcing has shown itself as a useful method for enriching information objects, there remains the question of how the method of collection affects the way the data can be used. Contributors are self-selected and often without verified reliability, training or expertise. Agreement is sometimes a useful metric for objective information, but sometimes there is meaning in disagreement, such as in collaborative filtering.

Action	Examples
Rating	Rating helpfulness of online comments or reviews (e.g. Amazon), rating the quality of online content (e.g. items on Youtube, Netflix, LibraryThing, etc)
Classification / Curation	tagging (e.g. Delicious), labeling, adding to lists
Saving / Recommending	Starring, liking/recommending (i.e. Facebook), adding to favourites (e.g. Flickr)
Editing	Translations (e.g. Facebook), Corrections (e.g. National Library of Australia)
Feedback	Marking online comments as inappropriate (e.g. ABC News), “Did you find this helpful?” (e.g. Edmunds)
Other	Commenting, sharing, encoding

Table: Types of actions seen in descriptive crowdsourcing

Action	User Use	System Use
Tagging a photo / bookmark	Easy personal retrieval, appeal of collecting, item grouping for easy sharing	Improved search, improved browsing
Rating an product	Sharing opinion	improved recommendations, prioritize good values
Rating a digitally digested item i.e. video, Comment	sharing opinion, communicating approval	Identifying and promoting quality
Flagging content	cleaning windows for the community, catharsis	Higher signal-to-noise in editorial maintenance
Starring	communicating approval, saving for future reference	Identifying quality content
Sharing	showing items to friends, referring or curating content	Identifying popular/interesting content
Feedback	sharing personal knowledge and opinions, altruism	Correct problem data, discover system issues

Table 2: Chart comparing user and system uses for a selection of incidental crowdsourcing actions

1.4 Definitions

Before proceeding, the terminology of this study should be established. As this work spans multiple domains, and makes reference to recently introduced concepts, it is important to establish a shared understanding of language within these pages.

Note that the treatment here is cursory; a more in-depth look can be found in the literature review.

1.4.1 Descriptive crowdsourcing

This paper focuses on crowdsourcing for descriptive metadata.

The distinction here is that the human contributions are reactive. There is an information object that already exists, and crowdsourcing workers add information about it. The response can be subjective, such as ratings or interpretations, or objective, such as descriptions or corrections.

Crowdsourcing descriptive metadata stands in contrast to crowdsourcing that *creates*, introducing new information objects into the world. One example of this is T-shirt design contests on Threadless[2].

This approach to crowdsourcing was looked at in Organisciak (2013) when defining the concept of *incidental crowdsourcing*. Incidental crowdsourcing is an approach to crowdsourcing that is unobtrusive and non-critical. This form of peripheral collection of data was noted to favour descriptive activities.

1.4.2 Human computation

Human computation is a separate but closely related concept to crowdsourcing. It refers to activities where humans perform work in a paradigm reminiscent to computing, and which could conceivably one day be done by computers (Law and L. v. Ahn 2011; Quinn and Bederson 2011). Human computation does not need to be crowdsourced, but many such tasks benefit from crowdsourcing. Likewise, while a notable portion of crowdsourcing tasks are creative, such as writing or commenting, human computation represents a large portion of the types of crowdsourcing seen in the wild.

[2] http://www.threadless.com

1.4.3 Worker, volunteer, contributor

The space of crowdsourcing is large and the incentives for contributors are varied. The most significant distinction within crowdsourcing is in comparing uses that pay their contributors and those that do not. It's valuable to make this distinction because paying a person changes they way that they perform, while also simplifying some of the concerns that are necessary in retaining volunteers.

In general, I refer to crowd individuals as *contributors*. When the distinction is necessary, paid contributors are referred to as *workers*, while elective contributions are made by *volunteers*.

1.5 Chapter Outline

The proposed dissertation follows the below structure, delineated by chapters.

Introduction

The first chapter will introduce the use of crowdsourcing for information retrieval and outline the problems of variance and low intercoder reliability in crowdsourced data. The scope will be drawn out as this document has done, making clear that the focus within crowdsourcing is on uses that augment our knowledge of existing information objects, and the focus in information retrieval is in improving retrieval, rather than evaluation, in a reliable manner. Subsequently, the assumption of honest but biased contributors will be outlined, and the hypothesis on this assumption will be outlined along with the two studies that will be pursued to test it.

Literature review

The literature review will serve as a comprehensive review of the field around the research. It consists of the following sections, some of which have already been performed for this proposal:

- define all the necessary concepts in crowdsourcing and provide their history,
- provide an extensive taxonomy of crowdsourcing,
- discuss the existing research into using crowdsourcing information for improving retrieval,
- discuss the most notable research in library and information science,
- list the most notable projects to utilize crowdsourcing, and
- provide a "greatest hits" list crowdsourcing research that any interested reader should be familiar with.

Collecting crowdsourcing data: the effect of task design on resulting quality

Not all crowdsourcing designs are equal, and the resulting differences in contributions from different designs is an often neglected issue in crowdsourcing. The third chapter will look at the ways that task design changes the nature of contributions. Focusing on an existing use of crowdsourcing for improving information retrieval – annotating microblogging messages – this chapter will run multiple controlled studies with different interface design options.

Using crowdsourcing data: modeling documents from crowdsourced information

To complement a treatment on collecting contributions, the fourth chapter considers how to use them. This chapter will be structured around an information retrieval study around the website Pinterest. Viewed one way, Pinterest is a website of commentary about web images or pages: potentially a trove of additional 'aboutness' information, but sparse and unstructured when viewed as individual contributions. In line with where the value of the proposed dissertation lies, the Pinterest study is intended as a foil to the larger discussion of interpreting crowd contributions.

Discussion and conclusions

The last chapter of the dissertation will summarize the findings of the previous two chapters, and take a high-level look at descriptive crowdsourcing in information science. This chapter will serve as a shorthand reference of the full document, which future readers can consult for the main points. Doubtless, there will likely be a number of new research questions which emerge from the proposed dissertation; they will also be collected in the final chapter.

Details of the proposed study designs for chapters 3 and 4 are provided later in this proposal, and an initial literature review is provided next.

Chapter 2

LITERATURE REVIEW

Crowdsourcing is a simple concept that has received considerable research attention in the past few years, alongside a realization of the power of the internet for effectively connecting people in large numbers.

Perhaps unsurprisingly, the concept of crowdsourcing precedes the language that has developed around it in recent years, and the research in crowdsourcing has been uneven.

This section provides an initial overview of crowdsourcing and the notable research within it.

2.1 Definitions and History

Crowdsourcing broadly describes the use of distributed crowds to complete a task that would otherwise be by one or a few people. Beneath this large umbrella, there are many concepts that are either in its purview or overlap with it.

2.1.1 Crowdsourcing

Crowdsourcing refers to "groups of disparate people, connected through technology, contributing to a common product" (Organisciak 2010). It broadly captures the abilities of the internet, as a communications medium, to efficient connect people.

Nothing about crowdsourcing is fundamentally tied to the internet. It is entirely possible to bring together large groups of people in different ways, but the access and efficiency of the internet is both what makes the concept seem so novel and what gives it value in the various realms where it is applied. Whereas crowds have long been noted for their collective simplicity (Le Bon

1896) or irrationality (Mackay 1852), through the internet, one can perform human-specific tasks at a scale usually only seen for computational tasks.

The term is recent and has an unambiguous source, but immediately upon its introduction, it was adopted and expanded on through public discourse.

The term *crowdsourcing* comes from a 2006 Wired article by Jeff Howe (J. Howe 2006b). Howe was writing from a labor perspective, looking at online marketplaces for people to solve problems and create content. His focus was on systems like InnoCentive, a site for companies to outsource research and development problems for a bounty, and iStockPhoto, a website that allowed amateur photographers to sell their images as stock photos. The article briefly looked at user-generated online content, though in the context of television programs that use online video as content.

Despite the narrow initial definition, the term *crowdsourcing* struck a chord more broadly and was culturally co-opted. the definitional appropriation happened very quickly: within nine days Howe noted a jump from three Google results to 189,000 (Jeff Howe 2006). Within a month, Howe addressed the co-opting of the term, "noticing that the word is being used somewhat interchangeably with Yochai Benkler's concept of commons-based peer production" (J. Howe 2006a). He gives his definition, but also notes that language is slippery, and he is "content to allow the crowd define the term for itself (in no small part because [he is] powerless to stop it.)."

Thus, crowdsourcing was adopted to refer broadly to a series of related concepts, all related to people being connected online. These concepts included the 'wisdom of the crowds'(Surowiecki 2004), human computation(L. v. Ahn and Dabbish 2004), commons-based peer production(Benkler 2006), and free and open-source development (Lakhani and E. v. Hippel 2003; Raymond 1999).

While there have been occasional semantic attempts to redefine crowdsourcing again as a more granular term, its colloquial adoptions seem to have cemented its use as a broad concept.

2.1.2 Wisdom of the Crowds

The Wisdom of the Crowds(Surowiecki 2004) is a book written by journalist James Surowiecki in 2004. The book observes the strength of human decisionmaking when done in aggregate, and the term ‘wisdom of the crowds’ has survived the book to refer to cases that make use of this.

2.1.3 Human-Computation

Human Computation emerged from from the doctoral dissertation work by Luis von Ahn in 2005, popularized alongside the ESP Game (L. v. Ahn and Dabbish 2004; L. v. Ahn 2006). It refers to the process of computation – the “mapping of some input representation to some output representation using an explicit, finite set of instructions” (Law and L. v. Ahn 2011) – performed by humans.

In synthesizing the definition of human computation in relation to crowdsourcing, collection intelligence, and social computing, Quinn and Bederson note two characteristics of consensus: that “the problems fit the general paradigm of computation, and as such might someday be solvable by computers”, and that “the human participation is direction by the computational system or process” (Quinn and Bederson 2011).

As noted by Law and L. v. Ahn (2011), Turing defined the purpose of computers as carrying out operations that humans would normally do. Human Computation, then, refers to utilizing humans for operations that computers are not capable of performing yet.

By this definition, much human computation aligns with crowdsourcing, but large swaths of crowdsourcing are not relevant to human computation. For example, creative crowdsourcing projects like T-shirt design website Threadless and online encyclopedia Wikipedia are not human computation. Inversely, human computation does not have to be sustained by self-selected workers; a more traditionally hired closed system can suffice (Law and L. v. Ahn 2011).

2.1.4 Open-Source

An early model reflecting the properties of crowdsourcing is open source software development.

With open source, software’s underlying source code is freely accessible. As a consequence of this form of transparency, open-source development began to adopt some unique properties: users and distributed developers could jump into the code to fix a bug, or add a feature that they wanted to see. The significance of this became apparent when Linus Torvalds released Linux in 1992 with a development model that accepted external code contributions heartily, released early and often, and followed the pulse of user's needs. Eric Raymond compared this form of software development to a bazaar, "open to the point of promiscuity", and contrasted it to the traditionally managed 'cathedral' style seen in the commercial world and earlier open source projects (Raymond 1999).

The many hands approach to open-source demonstrated that technologically-connected crowds can coherently delegate and create works. Like with crowdsourcing, open source software development often does not discriminate on credentials or background; if a contributor can make an adequate contribution, it can be used.

The roots of crowdsourcing in open source are noted Jeff Howe's short definition formulated after his Wired article: "the application of Open Source principles to fields outside of software."

2.1.5 User Innovation and Commons-Based Peer Production

It should not be surprising that recent cultural observers have noted the behaviours seen in crowdsourcing through various lens. Crowdsourcing emerges from various affordances of modern information networks. Such as seen with open-source software development, networked society encourages new forms of cultural creation, not by intention but by consequence of the type of connectedness it allows.

As networked society has developed and the internet has grown ubiquitous, numerous scholars have noted the cascading consequences in how individuals interact with culture and participate in the creation of cultural objects. Two such streams of study are von Hippel's work on *user innovation* and Benkler's study of the networked information economy, including his concept of *commons-based peer production.* Both of these borrow from economic and market-driven theory rather than sociological theory, but they offer valuable language for understanding crowdsourcing as a cultural phenomenon.[1]

[1] One might argue for the term *consequence* rather than *phenomenon*, because it positions crowdsourcing as neither an accident nor a product of intention, but acknowledges

If crowdsourcing is a generalized version of open source principles, von Hippel's work on user innovation(E. v. Hippel 1988; E. v. Hippel 2006) was an early observation of the trend toward a greater user focus in computer tools and services.

With user innovation, new information products or physical products are generated by users – those that benefit from using rather than selling the product. Notably, von Hippel focuses on 'lead users,' users with specific needs that precede broader trends. These users either develop new products to fill their needs or modify existing products.

Not all crowdsourcing creation is user innovation, though there are echoes of von Hippel's work in companies that turn to the Internet for help in conducting their business, whether it is soliciting feedback and suggestions (e.g. MyStarbucksIdea [2]), bug reports, or even work at a bounty (e.g. advertising contests). User sharing of work performed for themselves is another similar area: for example, when a music service allows users to share their playlists publicly, their realization of a personal need has potential value to other users.

Benkler's work takes a political economy view on what he calls the 'networked information economy', but arrives at a very similar place to von Hippel. He argues that the unique landscape of the 'networked information economy' empowers individuals to do more for themselves and in collaborative groups outside of established economic spheres (Benkler 2006). This agency allows commons-based peer-production: for innovation and creation to rise out of the commons rather than from firms.

Benkler(Benkler 2006) singles out two user behaviors borne out of access to information networks, which in turn underlie the rise of crowdsourcing. First, individuals are more empowered to operate autonomously, for themselves with less reliance on mass-market goods. At the same time, loose collaborations are easier to organize, allowing the pursuit of individual needs at scales beyond the capabilities of a single person.

a history for it where it is a side-effect of external influences.

[2] http://mystarbucksidea.force.com/

2.2 Taxonomies of Crowdsourcing

The space of crowdsourcing is large, and there have been a number of attempts to organize the sub-concepts within it or to reconcile it alongside other areas of research. Some of the most important questions in differentiating crowdsourcing include:

- Who are the contributors? What are their skills?
- How are contributors motivated? Are they paid or do they volunteer for other incentives?
- Are contributions new, or do they react to existing documents or entities?
- Are contributions presented or used as a whole, or are they combined into a larger contribution?
- What do the contributions look like? Are they subjective or objective?
- Who is asking for the contributions? Who is benefiting?
- Is the collaboration indirect (i.e. contributors work on parts independently) or manifest?
- Is the crowdsourcing central to the system?
- How is quality controlled for?

2.2.1 Motivation

The incentives for contributors to participate in crowdsourcing are complex and not always consistent from contributor to contributor.

Intrinsic / Extrinsic Motivation

Motivation in crowdsourcing follows related work in the motivations of humans in general (e.g. Maslow 1943; Alderfer 1969; Ryan and Deci 2000). While a review of that work is beyond the scope of this work, many views of crowdsourcing motivation adopt the lens of motivation as a mixture of *intrinsic* factors and *extrinsic* factors (Ryan and Deci 2000). With intrinsic factors, fulfillment is internal to the contributor –psychologically motivated – while with extrinsic factors the rewards are external.

Category	Description	Sub-categories
Motivation	How are contributors incentivized?	Primary/Secondary (Organisciak 2010), Contribution/commitment (Kraut and Resnick 2011) Extrinsic/Intrinsic
Centrality	How central is the crowdsourcing to the overall project?	Core / Peripheral (Organisciak 2013)
Beneficiary	Who benefits? What is their relationship to contributors?	Autonomous / sponsored (Zwass 2010) Crowd / individual
Aggregation	How are diverse contributions reconciled into a common product?	Selective /Integrative (Geiger et al. 2011; Schenk and Guittard 2009) Summative / Iterative / Averaged
Type of Work	What is the nature of the work?	Human computation / Creative Generative / Reactive Subjective / Objective
Type of Crowd	What are the dimensions of the crowd and how they are expected to perform?	Unskilled, locally trained, specialized heterogenous / diverse

Table 2.1: Overview of crowdsourcing taxonomy

The spectrum of intrinsic to extrinsic motivators is commonly paralleled in crowdsourcing literature through dichotomy of paid and volunteer crowdsourcing(Rouse 2010; Geiger et al. 2011; Kraut and Resnick 2011; Schenk and Guittard 2009).

Paid and volunteer crowdsourcing are not exclusive, and there are extrinsic motivators beyond money. However, this separation is common because of it accounts for some of the starkest differences between how crowdsourcing is implemented and motivated. There are differing design implications around people being paid and performing work for other reasons: money is a direct currency for obtaining labor, while convincing volunteers to contribute requires a greater sensitivity of their needs and ultimately more complexity in engineering the crowdsourcing system.

It has been shown that intrinsic motivation still plays a part in paid crowdsourcing (Mason and Watts 2010), and some systems mix intrinsically motivated tasks with payment or the chance at remuneration. For example, some contest-based marketplaces are popular among users looking to practice their skills, such 99Designs for designers or Quirky for aspiring inventors.

Some taxonomies make a distinction between forms of payment. Geiger et al. (2011) makes the distinction between fixed remuneration, with a pre-agreed fee, and success-based remuneration, such as contest winnings or bonus.

Specific Motivators

Taxonomies of specific motivators seen in crowdsourcing have been previously attempted, with varying results that touch on similar issues. In Organisciak (2010), I identified a series of primary and secondary motivators from a diverse set of crowdsourcing websites. Below I adopt the categories from that study, as they accommodate related work well.

Primary motivators are those that are considered critical parts of a system's interaction. Systems do not require all of them, but to attract and retain contributions, they need one or more of them. In contrast, *secondary motivators* are system mechanics that were not observed as necessary components of systems, but were elements that encourage increased interaction by people that are already contributors. Kraut and Resnick (2011) parallel the primary/secondary split by differentiating between encouraging contributions and encouraging commitment.

The motivators in Organisciak (2010) were observed from a content analysis of 13 crowdsourcing websites and subsequent user interviews. For sampling, 300 websites most commonly described as 'crowdsourcing' in online bookmarks were classified with a bottom-up ontology, then the 13 final sites were selected through purposive stratified sampling, to represent the breadth of the types of crowdsourcing seen.

Below is a list of primary motivators seen in Organisciak (2010), but also paralleled and supported by the similar broad view social study published by Kraut and Resnick (2011).

- **Money and extrinsic reward**. Paying crowds is not particularly novel, but it is the most reliable approach for collecting contributions.

In the absence of other motivators or where certainty is required, reimbursement will attract contributors. However, it also introduces bottlenecks of scale, and negates some of the benefits of intrinsic motivation. Mason and Watts (2010) note that, while intrinsic motivation still exists on paid crowdsourcing platforms, it is overwhelmed when tasks are too closely tied to reimbursement, resulting in contributions that are done minimally, briskly, and with less enjoyment. Kraut and Resnick (2011) point to psychology research that shows the ability of reward in other settings to subvert intrinsic motivation, leading to less interested contributors.

- **Interest in the Topic**. Sites that cater to people that have a pre-existing interest in their subject matter or outcomes tend to get longer, more consistent engagement. For example, the Australian Newspaper Digitisation Project (now know as a larger project called Trove) found that that amateur genealogists, with pre-existing communities and a willingness to learn new technologies, took "to text correction like ducks to water" Holley (2009). Similarly, Galaxy Zoo found similar success with amateur astronomers. Kraut and Resnick likewise argue that asking people to perform tasks that interest them results in more engagement than asking people at random.
- **Ease of entry and ease of participation**. Low barriers to entry and participation were cited by every single user interviewed for the study. Wikipedia has a low barrier to entry but its interface and demanding community standards have been criticized in recent years for raising the barrier to participation (Angwin and Fowler 2009; Sanger 2009). "Simple requests" generally lead to more productive contributions, according to Kraut and Resnick.
- **Altruism and Meaningful contribution**. People like to help if they believe in what they're helping. Writing about Flickr Commons, Library of Congress noted that they "appear to have tapped into the Web community's altruistic substratum by asking people for help. People wanted to participate and liked being asked to contribute". (Springer et al. 2008). With Galaxy Zoo , people often cite the fact that it is a tangible way to contribute to real science. Kraut and Resnick (2011) argue that appeals to the value of a contributions are more effective for

people that care about the domain.

- **Sincerity**. "People are more likely to comply with requests the more they like the requester," Kraut and Resnick (2011) note. A recurring theme among interview participants in Organisciak (2010) was whether a project seems sincere or exploitative. Since crowd contributions often exist as a parallel to labour, crowds are often weary of anything that smells like them being taken advantage of.
- **Appeal to knowledge and opinions**. One curious source of motivation observed in the study is simply asking the right people. Online visitors presented with a question are often compelled to answer it simply because they know the response, be it part of their knowledge, skills, circumstance, or opinions. The 'appeal' itself can be explicit or implicit. Kraut and Resnick (2011) refer to this sort of appeal as "Ask and Ye Shall Receive", asserting that online communities stand to benefit from easily accessible lists of what work needs to be done. They also assert that direct requests for contribution are better than broadcast.

One motivator overlooked in Organisciak (2010) is *novelty*. Novelty or curiosity is ephemeral and unsustainable, but nonetheless a unique idea can attract contributions for a short amount of time. Kraut and Resnick (2011) also note structure, goals, and deadlines as incentives. Such an effect is strongly felt on Kickstarter, where the tenor of crowdfunding for projects changes relative to the funding end date.

The supplemental secondary motivators observed in the study were:

- **"Cred": External indicators of progress and reputation**. Using games, badges, or leaderboards encourages more contribution among certain people. An important caveat is that this form of performance feedback needs to be perceived as sincere (Kraut and Resnick 2011).
- **Feedback and impression of change**. Showing the contribution in the system or conveying how it fits into the whole. Kraut and Resnick (2011) tie feedback to goals, emphasizing the importance of showing progress relative to personal or site-wide goals.
- **Recommendations and the social**. Prodding by friends, colleagues, and like-minded individuals. Simply seeing that other people have con-

tributed makes a person more likely to contribute (Kraut and Resnick 2011).

- **Window fixing**. Nurturing a well-maintained community where the members are compelled to support it's health.

2.2.2 Centrality

How central, or necessary, is the crowdsourcing to the task at hand? Is it *peripheral*, or *core*?

The work in Organisciak (2013) tried to counterbalance a perceived focus on whole-hog crowdsourcing – the large, highly novel initiatives like Wikipedia – by introducing *incidental crowdsourcing*. Incidental crowdsourcing focused on types of crowdsourcing – like rating, commenting, or tagging – that are peripheral and non-critical. The shift to an incidental mode brings with it its own design tendencies, such as lower bandwidth forms of contribution and fallback strategies for low engagement cases.

2.2.3 Aggregation

Schenk and Guittard (2009) and Geiger et al. (2011) discuss two types of aggregation: *integrative* and *selective*. Integrative aggregation pools contributions into a common product, like a wiki, while selective aggregation tries to choose the best contributions, such as in contests.

This simple separation hides some of the complexity seen in aggregation approaches. Integrative aggregation can be approached in a number of ways. I argue the following finer views on integrative aggregation are useful:

- **Summative**. In summative aggregation, people contribute to an ever-expanding base of information. Contributions are clearly part of a bigger whole, but their individual form is retained. Examples: online reviews
- **Iterative**. In versioned aggregation, multiple contributions are used toward a larger product, but the contributions are permutations of a common work. Examples: wikis
- **Averaged**. In averaged aggregation. Contributions are still pooled, but a consensus-seeking processing tries to reconcile them. Examples:

ratings, multiple-keyed classifications

2.2.4 Director / Beneficiary

Who directs the crowdsourcing activities and who benefits from the contributions?

Considering the director of a crowdsourcing task, Zwass (2010) distinguishes between *autonomous* and *sponsored* forms of crowdsourcing.

Sponsored crowdsourcing is when there is a entity at the top soliciting the contributions: a client of sorts. In contrast *autonomous* crowdsourcing serves the community itself. Autonomous crowdsourcing can be in a centralized location, like a community-written wiki or video-sharing website, or exist loosely, as in blogs. Zwass (2010) explains: “Marketable value is not necessarily consigned to the market—it may be placed in the commons, as is the case with Wikipedia.”

Considering the soliciting party as a case of sponsorship or autonomy is useful, though a further distinction should be made between the collective (the *crowd*) and the individual (the *contributors*). Crowds collaborate toward a shared goal, as with Wikipedia or a type of open-source software development, while individuals are more self-possessed. For example, in citation analysis through web links, as was done with PageRank (Page et al. 1999), the large-scale benefits of the crowds are unrelated to what the individuals creating the links are thinking.

One way to view this relationship between contributor and director is in light of effort against benefit. Do both director and contributor benefit (symbiosis)? Does one benefit at the expense of the other (parasitism)? Or is it a case of commensalism, where both benefit without affecting each other?

2.2.5 Type of Work

The type of work performed by crowds can vary greatly in its complexity and style.

One notable form of crowdsourced work is represented by the concept of human computation, where “the problems fit the general paradigm of computation, and as such might someday be solvable by computers” (Quinn and Bederson 2011). Understanding that crowdsourcing is not solely human computation tasks, the inferred corollary to these types of tasks are those that are expected to be too complex for computers: creative, judgment-based, or requiring critical thinking. Creative crowdsourcing might take the form of artistic human expression, such as online contributors collectively animating a music video (Johnny Cash Project) or the sum of YouTube. Opinion or judgement-based crowdsourcing often does not have a definitive answer, and is seen in areas such as movie reviews or product ratings. More complex critical thinking tasks do not fit the paradigm of computation and are much more complex, such as Wikipedia or protein-folding project FoldIt.

Schenk and Guittard (2009) distinguish between three types of crowdsourcing. First are routine tasks, such as crowdsourcing of OCR text correction with ReCaptcha. The majority of human computation tasks would likely fall within this category of rote tasks. Second are complex tasks, such are open-source software development. Finally, they suggest creative tasks. An examples would be a system like MyStarbucksIdea, where people suggest changes they would like to see at the coffee chain Starbucks. Since Schenk and Guittard (2009) focus on crowdsourcing when there is a client, usually a corporate client, they do not consider the wider space of creative crowdsourcing tasks.

Another view that touches on the nature of the contribution is *creative* versus *reactive*. In the former, new intellectual products are created. With reactive work, the work is a reaction or interpretation of an existing information object: reviews, ratings, encoding. This is the view that this study takes in focusing on descriptive crowdsourcing, where the crowd is teaching us about existing objects.

Subjective vs. Objective Crowdsourcing

Another parallel being drawn in recent years is that of objective or subjective crowdsourcing tasks.

Objective tasks are assumed to have an authoritative truth, even if it is unknown. For example, in transcribing scanned texts, it is assumed that there is a ‘correct’ passage in the work that has been scanned.

In contrast, subjective tasks have a variable concept of correctness, as they are are not expected to be consistent between contributors.

Human computation undertakings are commonly objective tasks, and taxonomic effort of human computation – such as Schenk's split of routine, complex, and creative – does not touch on the subjective/objective separation in a direct way (Schenk and Guittard 2009).

This designation also applies to aggregation. Multiple contributions can be aggregated with an objective assumption, expecting a truth a deviations from it as bad work or data. Other systems try to aggregate a normative opinion or judgment of subjective contributions. This latter assumption is seen often in opinion ratings, such as film or restaurant ratings: just because there is an aggregated rating presented, there is an understanding that some people might disagree and that they are not incorrect for doing so.

2.2.6 Type of Crowd

Vukovic and Batolini (Vukovic and Bartolini 2010) define two extremes of crowd types: *internal* and *external*. Internal crowds are composed solely of contributors from the organization that is crowdsourcing, if it is thus centralized. External crowds are members outside of the institution. Vukovic and Batolini also note that *mixed* crowds are observable.

Necessary Skills

A potential separation between crowd methods is the skills required to perform the work. *Unskilled*, *locally training*, and *specialized* are all seen among crowdsourcing systems. Where unskilled labour encourages contributions from anybody at anytime, systems that use methods for authority control leave certain tasks to long-term, involved contributors. For example, on Stack Overflow, a user's administrative ability grows more open as they contribute more to the management of the system, a way of ensuring that those users have learned the proper management of the site.

2.2.7 Quality Control

The classification of human computation systems by Quinn and Bederson (2011) includes quality control as a primary dimension. Here they consider how the system protects against poor contributions, such as reputation systems, input or output agreement, multi-contribution redundancy, a crowd review workflow, expert review, and designs that disincentive poor quality or obstruct the ability to do so.

2.2.8 Common Design Patterns

A number of design patterns have been established and repeated in crowdsourcing, some organically and some, like the ESP Game, carefully engineered. These include:

Microtasking: the concept of splitting a large task into many smaller parts to be worked on by different people was an important tide change in the history of open-source software (Raymond 1999), and the same models have been emulated in crowdsourcing. With so-called 'microtasks', the overhead to participation is low, and the pressure or dependence on any one contributor is low.

Gamification: Gamification is predicated on a reframing of what would traditionally be labour into a game-like or leisurely tasks. Gamification follows in the philosophy, as Twain wrote, "that work consists of whatever a body is obliged to do, and that play consists of whatever a body is not obliged to do." (Twain 1920) The ethics of gamification have been argued for as an extension of contributors' desire to perform meaningful work. Shirky, for example, argues that people have a 'cognitive surplus' to give during their leisure time, a desire to spend their free time doing useful, creative or stimulating tasks. Gamification is an extension of serious games – games meant to do more than simply entertain (Abt 1987; Michael and S. L. Chen 2005; Ritterfeld, Cody, and Vorderer 2010). In areas of crowdsourcing and human computation, Games with a Purpose (L. v. Ahn 2006) is an extension of serious games in the context of distributed, collaborative crowds. Harris and Srinivasan (Harris and Srinivasan 2012) consider the applicability of applying games with a purpose to various facets of information retrieval, concluding it is a feasible approach for tasks such as term resolution, document classification, and relevance judgment. Eikhoff et al. (Eickhoff, Harris, et al. 2012) have investigated the gamification of relevance judgements further, augmenting the financial incentive on paid crowdsourcing platforms.

Opinion Ratings: A standard and highly familiar activity online is soliciting qualitative judgments from visitors. These ratings have different granularities, most commonly 5-level (e.g. 1 to 5 stars) or binary (e.g. thumbs up/thumbs down). Unary judgments have grown in popularity as ways of showing support with minimal effort. Their popularity seems to stem from when social network Friendfeed implement a unary voting button labelled, succinctly, "I like this" (Taylor 2007) and subsequently when similar wording was adopted by Facebook after acquiring Friendfeed.

Platforms: There is a cottage industry of services that offer the infrastructure for requesters to crowdsourcing, using in domain-specific ways. For example, Kickstarter and Indiegogo ease crowdfunding, 99Designs enables contest-based design tasks, and Mechanical Turk offers the tools and people for microtasks.

Contests: In the contest design pattern, a requester offers a bounty to the best solution to a problem or task of their choosing, such as in design (e.g. 99Designs), coding (e.g. TopCoder), and research and development (e.g. Innocentive). Here the "crowdsourcing" is simply using internet to connect to many potentially talented individuals, though contests have been integrated into more collaborative workflows. For example, with the collaborative product incubator Quirky, the community votes on the best ideas to develop into products, discussing how to improve the ideas openly. One reason for this may be that, in addition to the large portion of future profits that an idea originator may earn if it is voted into development, the rest of the community also receive points for supporting the best ideas.

Wisdom of crowds: Wisdom of the crowds is a design pattern which emphasizes the effectiveness of human judgment in aggregate (Surowiecki 2004), provided the participants are rationally organized. This is embodied by multiple-keying for tasks which are expected to have a real answer, such as classifying galaxies, or averaging opinions for subjective tasks to derive a normative judgment.

2.2.9 Other Taxonomies

Geiger et al. (2011) identify crowdsourcing processes by four defining characteristics: the pre-selection process for contributors, the accessibility of peer contributions, the aggregation of contributions, and the form of remuneration for contributors. While these are all valid ways of viewing crowdsourcing, I believe more qualitative or naturalistic separations are also necessary in order to understand crowdsourcing websites, such as motivation or centrality.

Quinn and Bederson (2011) provide a taxonomy of human computation along six facets: motivation, quality control, aggregation, human skill, process order, and task-request cardinality. Their taxonomy is thorough and relevant to crowdsourcing in general.

2.3 Top Research in Crowdsourcing

The lit review of the proposed dissertation will include a treatment of the most notable research in crowdsourcing: a "greatest hits" of sorts. Here I offer a brief overview of some of the research that will be discussed.

The ESP Game(L. v. Ahn 2006) demonstrated an effective and fun way to support research data through games.

Soylent(Bernstein et al. 2010) integrated paid crowdsourcing into word-processing tools. In the process, the study termed the Find-Fix-Verify design pattern.

Mason and Watts (2010) found that increasing wages on paid crowdsourcing sites did not improve quality of results, just quantity of contributions, due to an anchoring effect whereby perceived value of the task also grew with payment. They also identified that intrinsic motivation is still at play in paid crowdsourcing, elegantly showing that it grows weaker the more closely the task is bound to extrinsic reward.

Kraut and Resnick (2011) mined social literature for one of the most insightful looks at how online communities function effectively

Yochai Benkler's Wealth of Networks (2006), discussed earlier, which took a political economy view to commons-based forms of production, and the affordances that technology provides for sidestepping corporate production in this way.

2.4 Crowdsourcing in the Wild

There is a great deal of crowdsourcing “in the wild”, including notable successes and failures. Some of the successful project provide archetypal blueprints of crowdsourcing dimensions, and will be good to understand moving forward. Below I describe a few projects to aid in understanding the possibilities of crowdsourcing; more detail will be provided in the proposed dissertation literature review.

Wikipedia is a collaboratively written encyclopedia, where the majority of contributors are volunteers. Wikipedia, formed in 2001 and now containing 4,579,708 articles (as August 13, 2014 (*Wikipedia* 2014)), has an open editing policy that allows anonymous contributions and only restricts who can edit a page for few special cases where vandalism is likely. The policy also ensures that readers are latent editors (Shirky 2009), helping police, correct, and improve poor quality content.

Threadless is a community of artists that design and vote on T-shirt designs. Winning designs are licensed by Threadless to print and sell, providing a commission to the designer and additional profit for subsequent shirt reprintings.

The **Netflix Prize** was a competition run by film rental (and now streaming) company Netflix, offering a million dollar bounty to the person or team that could improve film recommendation by 10% over the root-mean-squared-error performance of Netflix’s own system. Claiming the prize required the winner to publish their results but did not require transfer of intellectual property, only a license for Netflix. A 2008 New York Times article about the prize noted that the community of participants were notably open in sharing their insights(Thompson 2008).

Kickstarter is a microfunding platform that enables patronage of artists and creators in their project through small but plentiful contributions. A project creator on Kickstarter proposes a project and offers tiers of rewards for backers that contribute varying amounts. When researched in Organisciak (2010), the balance between the altruistic support-based motivation and opportunistic reward-based incentives seemed to weigh slightly more toward the former, though I expect this has changed in recent years as more products have been offered on the site. Regardless, the model of small contributions from many has been seen in many other so-called crowdfunding contexts, including charity, politics (Fung 2014), and small business (Cortese 2011; Cortese 2013).

2.4.1 Academic

Zooniverse is a series of crowdsourcing projects that started with Galaxy Zoo. Galaxy Zoo allowed the general public to classify images of galaxies from the Sloan Digital Sky Survey, many being seen for the first time, at a pace much quicker than any one human could perform. Another popular project, Old Weather, transcribes weather logs from old ship's journals. In Snapshot Serengeti, participants classify animals photographed in camera traps. Many of the Zooniverse projects follow a similar pattern: encoding of curious, novel, or interesting images while contributing to real research.

FoldIt is a game where users try to develop the most efficient folding of a protein (Khatib et al. 2011). Folds are scored and placed on a leaderboard, adding a competitive edge. FoldIt shows that, when well matched to competitive impulses, complex problems can be tackled through semi-anonymous online workers.

ReCaptcha (**von˙ahn˙recaptcha:˙2008**) cleverly took a system intended to distinguish humans from bots – obfuscated text transcription with Captchas – and combined it with a problem that by definition only humans can do: fixing scanned text that computational techniques failed at. With ReCaptcha, online visitors prove they are human and help digitize scanned archives at the same time.

2.5 Research in Information Science

In information retrieval, the focus on crowdsourcing has been predominantly in the use of paid crowds for generating evaluation datasets, though there have been efforts to use crowds to improve document representation or even query-specific ranking.

The benefit of paid crowds for relevance judgments is that it allows for on-demand evaluation datasets (Alonso, Rose, and Stewart 2008). This has been a costly and exhausting process in the past, making it difficult to perform IR research on more novel datasets than the judged sets available from TREC.

Relevance judgments benefit from the agreement among multiple humans, since the concept of 'relevance' is not clear-cut but rather normative. The ability to attract a breadth of rater types also positions paid crowdsourcing as an effective means to collecting evaluation data.

For three years, TREC has run a crowdsourcing track that emphasizes the collection of high quality relevance judgments through paid crowds (Lease and Kazai 2011; Smucker, Kazai, and Lease 2012). While much of the focus was on identifying and accounting for lower quality workers, there were also some efforts which built novel interfaces to try to streamline contributions or increase reliability. For example, the Glasgow team encourage fast turnaround, reducing rating click counts, pre-loading pages, and floating the assessment question (McCreadie, Macdonald, Santos, et al. 2011). Earlier, the same team crowdsourced judgments for the TREC Blog track with a design that color coded completed tasks based on whether they matched other raters and a gold standard (McCreadie, Macdonald, Santos, et al. 2011).

Grady and Lease also 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, including in Library and Information Science work. 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).

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.

Using crowdsourcing in the machine, as evidence for search engine algorithms rather than evaluation, is less common. PageRank is one such effort, utilizing the linking habits of web page authors as a proxy for authoritativeness and quality(Page et al. 1999).

One of the best explored spaces of retrieval over or incorporating crowdsourced information is in folksonomies. Folksonomies refer to free-text labelling (i.e. 'tagging') by non-professionals. A popular resource for folksonomies over general web documents is the older incarnation of bookmarking website del.icio.us. In folksonomies such as on del.icio.us, over 50% of tags contribute information that was not contained in the document; for music tags (on the website Last.fm), over 98% of tags provide text information not previously help in the record(Bischoff et al. 2008). Information retrieval can benefit for this extra information, and a comparison of web query logs to folksonomies from del.icio.us, Flickr, and Last.fm shows that 58.43-71.22% of queries overlap at least partially with tags in those systems (ibid).

Studying ways to retrieval saved bookmarks on del.icio.us, (Hotho et al. 2006) present *FolkRank*, a manner to adjust authority of authors and importance of tags in order to find important resources. While their approach has limited success as a generalized retrieval approach, they find that it holds value in identifying communities of interest within the community.

(Zhou et al. 2008) present a generalized framework for dealing with social annotations within the language modeling approach. Their model categorizes users by expertise domain and builds domain topics from related annotations. These are linearly smoothed with document and query language models. In the context of del.icio.us, their approach improves over traditional unigram models over the document text. Their model parallels some of the approaches proposed for this dissertation's study over Pinterest. The proposed study makes similar assumptions on the meaning of an individual annotation, but differs in that it makes no assumptions about the expertise about a user and relies on the interpretations from user curated groupings rather than inferred topics.

Finally, Harris and Srinivasan (2012) provide a comprehensive overview of ways that crowdsourcing and games with a purpose can be incorporated in the information retrieval workflow. While crowdsourcing is noted as highly feasible for evaluation, it is also noted as an approach which can help in building document collections, identifying information needs, and query refinement.

2.6 Summary

Crowdsourcing is a phenomenon with a wide umbrella and a broad range of parameterizations. For information science, it is potentially very valuable for its ability to efficiently gather extra-textual information about existing objects. The next two chapters propose two studies for learning more about this quality.

Chapter 3

DESIGNING RELIABLE TASKS FOR DESCRIPTIVE CROWDSOURCING

3.1 Introduction

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 the first of two research chapters, the proposed dissertation will investigate how the design of crowdsourcing tasks for collecting useful metadata for information retrieval metadata affects the quality of the content.

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 (e.g. E. 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; Organisciak, Efron, et al. 2012; Jeff 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.

In keeping with goals of the proposed dissertation to explore and develop methods for controlling intercoder reliability, this sub-study will compare 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 designs 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 research by Komarov, Reinecke, and Gajos (2013) found, the effects seen in traditional user studies are still present in online crowd markets. Their finding suggests that non-crowdsourcing research in human-computer interaction is informative for our purposes. This proposal chapter explores some possible design decisions and argues why they should be studied.

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 expands beyond a scope than can reasonably be answered, but it should nonetheless be addressed as thoroughly as possible.

3.1.1 Overview of proposed research design

In this chapter, I will evaluate two interfaces for encouraging less deviation between human contributors by providing training and feedback mechanisms, respectively. These interfaces, motivated by efforts in my earlier work, are intended to slow down workers and make them aware of how their perception of the task deviates from the standard. They will be compared to a baseline basic interface, as well as an alternative system that encourages quicker responses.

Since the focus is on design for crowdsourcing in information retrieval, I will 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.

While a more common use of paid crowdsourcing is for evaluation dataset creation, this dissertation looks as crowdsourcing for document metadata. A task where more information in encoding about the document is a more appropriate task to study in this context, even though evaluation can be considered loosely relevant.

3.2 Motivation

Why is task design important to descriptive crowdsourcing? Task design has been noted on multiple occasions as an intuitively important consideration, but has only been studied in limited formal circumstances (e.g. Alonso and Baeza-Yates 2011; Grady and Lease 2010).

Crowdsourcing is increasingly being used for information retrieval. Particularly, the on-demand nature of paid crowdsourcing is being embraced for uses such as relevance judgments (Alonso, Rose, and Stewart 2008), to describing queries (E. Chen and Jain 2013), and to annotating entities (Finin et al. 2010).

Particularly, my past doctoral-level research motivates this chapter, suggesting the importance of the problem and making this study a logical extension of my doctoral work.

- In Organisciak, Efron, 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.
- During research on Organisciak, Teevan, et al. (2013), we found that asking people to reflect on their response changed the nature of their response, with less internal consistency.
- In a sample study comparing the space of incidental crowdsourcing across two systems (Organisciak 2013), I 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.
- In recently-completed research (submitted paper pending), I looked at low grader consistency in the ground truth for the Audio Music Similarity (AMS) task in the Music Information Retrieval Exchange (MIREX). One of the results found that redesigning the task to attach finer instruction to the rating improved the quality of judgments by crowdsourced judges.

Better workers read codebooks

In Organisciak, Efron, et al. (2012), we consider a variety of methods for removing error from relevance judgments collected on Mechanical Turk. One of the study's findings hints at the importance of task design.

In relevance judgments, workers are shown a *query*, a *description* of what is relevant to the query, and a set of *results* to classify as relevant or not relevant to the query. In our tasks, we showed ten results per tasks, using full item records from an online catalogue.

Looking at the amount of time that a worker spends on classifying each result, we found that overall neither slower workers nor faster performing workers are more consistently reliable (Wilcoxon rank sum $p = 0.064$, but $p = 0.154$ when excluding extreme outliers). However, looking at that value blocked by the order of results – i.e. how reliable are workers taking x time on the y^{th} result of a task – we found that the time spent on correct classifications for the first judgment in a task is significantly higher than the time spent on incorrect classifications (Wilcoxon Rank Sum one-sided $p = 0.01$). For all other relevance judgments in a task, the amount of time spent was insignificant.

Why are slower workers better for the first judgment in a task set, but not for subsequent items? Remembering that the time spent on the first relevance judgment is confounded with the start of the task and reading the description of what is relevant, we found that the effect is actually related to closer reading of instructions. Those tasks where a worker spent more time on the first judgment were correlated to better performance overall not just on that first judgment.

It does not matter how much time a worker spends on a task, as long as they spend enough time figuring it out at the beginning. How do you encourage this form of behaviour? Organisciak, Efron, et al. (2012) did not explore the design of tasks, but the findings suggest two things.

First, it encourages the assumption of honest-but-biased workers. If a worker interprets the codebook differently then they will disagree with the consensus and be considered a bad performer. Nothing here suggests a malicious worker, only a misguided one.

The second implication is that a better designed task can help. While much research tends to focus on the honesty and quality of paid crowdsourcing workers, part of the responsibility toward good quality crowdsourcing data is the requesters. In a task such as relevance judgments, encouraging better work might require workers to be more aware of their performance or to reassess their understanding of the task when it is necessary. Some possibilities to research might be,

- A training task, where workers are walked through the codebook in relation to actual tasks, and told why specific results are relevant or not relevant.
- Online feedback, showing workers their estimated performance based on agreement with other workers.
- Starting tasks with a known item, and alerting workers when they fail that task.

The importance of designing tasks was also seen in recent unpublished work on collecting human audio similarity judgments. I investigated ways to address low intra-coder consistency in the ground truth for the Audio Music Similarity (AMS) task of the Music Information Retrieval EXchange (MIREX), including a change in the collection design. In the original task, similarity judgments were collected on a three-point "broad" scale – not similar, somewhat similar, and very similar – and on a 101-point open-ended "fine" scale. In comparing the open-ended similarity scores across multiple years, the scores for song pairs did not appear strongly consistent, with a root-mean-squared-error (RMSE) of 16.58. Changing the design to provide more particular naming to ranges of the 101-point scale (i.e. a change to the codebook and how it is communicated) improved agreement on song pair similarity to an RMSE of 11.44 (an improvement of 31%). Paired with other strategies, this enabled an improvement to $RMSE = 5.40$, a 66.1% improvement.

Dynamic of attention affect the style of contributions

In Organisciak (2013), I looked at a facet of crowdsourcing that I called *incidental crowdsourcing* (IC).

Incidental crowdsourcing refers to crowdsourcing in the periphery. These are elements that are unobtrusive and non-critical to the user or the system. They generally:

- describe existing information,
- collect contributions in low-granularity mechanisms,
- and favor interface choices over statements.

An example of an incidental crowdsourcing mechanic might be a 'thumbs-up' voting button on a video-sharing website or tagging functionality on an image-sharing website that does not force itself upon users.

By not taking users hostage or asking users for overly attentive or time-consuming contributions, incidental crowdsourcing contributions result in greater numbers of contributions, even if individual contributions are less frequent.

Contributors to IC also tend toward different patterns of contribution. This was clear in a small example study comparing a crowdsourcing element on two systems: one that was designed more ‘incidental’ while another put up more hurdles for contributing quality (Organisciak 2013). In the systems, I compared the ratings of identical product on two application marketplaces for the Android smartphone operating system: Amazon’s Appstore and Google Play. Users on both systems are able to rate applications on a five-point scale; however, while Google’s store allows them to offer this rating quickly and without any other guidelines, Amazon’s store requires an accompanying review with a minimum word length, and also has a codebook that reviewers must abide by.

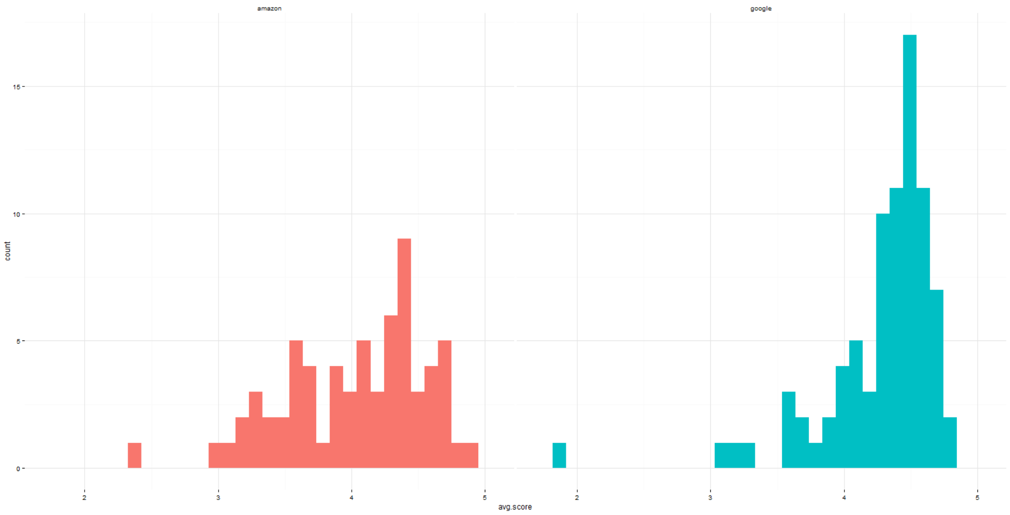


Figure 3.1: The difference in average ratings for the same items between Amazon Appstore and Google Play

Though other potential factors may also affect differences between the systems, this cursory comparison suggests that the attentiveness and introspection required of a crowdsourcing contribution affects what that contribution will look like.

Since most incidental crowdsourcing deals with reaction to existing documents rather than creation of outright new works, the future directions suggested by this work (in addition the work itself) are pertinent to the proposed dissertation’s focus on crowdsourcing for information retrieval indexing.

What are you *feeling*? Introspection changes rating habits

Additional evidence for this was encountered in a peripheral finding while conducting research for Organisciak, Teevan, et al. (2013). [^Note that the result here was not the primary focus of the study and is thus unpublished. Despite being unpublished, this work was purformed for Microsoft Research and should not be considered a contribution of this document.] In this case, we were again looking at opinion ratings on a five-point scale, this time on Mechanical Turk. Workers were asked to provide their personal opinions on whether they liked the style of each of 100 salt shakers. In one group of workers, this is all they were asked. In another group, we also asked workers to explain why they gave that rating. Alongside their rating of the item, they would provide a short explanation, such as 'I like the colors'. In most other regards the ratings were collected in the same way: on Mechanical Turk during the US work day, in an interface designed identically – besides the additional text input boxes for the latter formulation – and with the same restrictions on workers. The payment for the task was scaled to account for the extra time necessary for completion.

The figure below shows the distribution of workers by their average rating. In the first set, where workers rated quickly, the workers' average ratings were across the board: there were some unambiguously negative raters, some notably positive raters, and everything in-between. While the average worker could be expect to give a rating of 2.52 stars, the standard deviation was 0.69. In contrast, the workers that were asked to explain their rating tempered their opinions more. The distribution of average ratings of individuals was normal ($p = 0.9644$,Shapiro-Wilk) about a mean of 2.68 with a much lower standard deviation of 0.47.

It is clear that workers in the set with more introspection performed work differently. However, their contributions were also not markedly different in reliability when compared against an internal consistency test, averaging a difference of 0.33 stars when asked to rate the same item, versus an average difference of 0.39 stars.

These findings are suggestive of two issues related to this dissertation. First, they support the premise that different designs for ostensibly the same task change the contours of the resulting data. They also show that simply scaling a task to take more time does not provide corresponding returns in

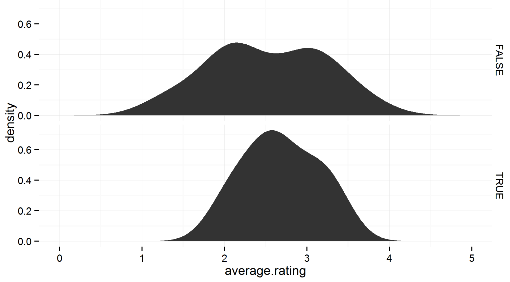


Figure 3.2: A comparison of raters' average ratings of salt shaker preferences, when only asked for a preference rating (above) and when also asked to explain the reason for their rating (below)

reliability improvements.

3.3 Scope

In looking at the design of contribution tasks, I will be focusing on paid crowdsourcing. For the scope of this study, it would be intractable to look at at the design of both paid and volunteer crowdsourcing task, so I will pursue the part that is more common for information retrieval researchers to have control over. Paying workers is only one part of crowdsourcing, and one that arguably tethers the scalability of a task by anchoring it to financial means. However, it is easier to control for because it sidesteps much of the complexities of motivation. Information retrieval researchers are using the predictability of paid crowd markets like Amazon's Mechanical Turk to generate on-demand data, making design for those systems important.

3.4 Approach

In this chapter, I will investigate the effect of different parameterizations of a microblogging encoding task.

3.4.1 Task

The metadata encoding task being controlled for is a microblog annotation task.

Microblogging messages, in this case from Twitter, are notably truncated. Since Twitter messages are limited to 140 characters, messages often are brief, missing context, and heavily abbreviated. This creates problems for parsing the topic of an individual message. On Twitter, 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 (E. Chen and Jain 2013).

Microblogging users partially address this problem by occasionally encoding metadata into messages in a structured way. For example, *hashtags* are used to describe the topic, theme, or context of a tweet (Efron 2011). Unfortunately, these are inconsistently applied and with differing purposes.

Due to the sparse information and novel needs of microblog IR, crowdsourcing has been used in this area, both for augmenting tweets and for creating datasets to train classifiers specific to the corpus. Twitter, the company, uses Mechanical Turk in this way to understand the context of trending queries. When there is a spike in search volume, human classifiers are able to differentiate between possible interpretations of the query (E. Chen and Jain 2013). For example, when "Big Bird" began trending after a political candidate made a polarizing comment about public television funding, it was easily identified as a political query.

The task will be 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.

Data

This study's dataset will be collected from Twitter through their Search API. Since we are evaluating the effectiveness of user annotation of Twitter messages, *tweets*, there is no need for a population of tweets, such as the dataset used in the TREC microblogging retrieval track.

The dataset will compile two types of data. The first data type will be random tweets from the stream, with named entity recognition used to identify entities that are potentially, but not necessarily, the topic of the tweet. The second data type collected will be all tweets about prominent topics, such as President Obama or musician Justin Bieber, again without prejudice over whether they are the topic of the message or not. The actual topic will be chosen later. The size of the dataset will be determined after consideration of typical time per task, statistical power, and their trade-off with cost.

For evaluation, a gold standard set will be encoded by myself, as a reliability coder. A ground truth dataset will be built after all experiments are run, by pooling all responses for a consensus. For additional rigour, the pooled dataset will be compared against my ground truth, so that points of disagreement can be re-evaluation.

3.4.2 Task Flow

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.

3.4.3 Gedanken Experiments

Within this framework, a number of factors are observable that may potentially affect how our microblog encoding task is completed. First are the parameterizations of the task within its existing structure – changes to the goal, instructions, item, action, and even the task itself. Ways that these can change from task to task include:

- **Task**
  - Payment.
  - Bonuses.
  - Number of tasks available.
- **Goal**
- **Instructions**
  - Clarity.
  - Restrictive vs. interpretable.
  - Length.
- **Item**
  - Number of items in a task.
- **Action**
  - Complexity of action. e.g. granularity.

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, such as the instructions, though it is hard to imagine that doing would would have a positive effect on the reliability or variance of the data.

The possibilities are endless for adding parts to the basic task. To inspire useful ones, it is helpful to consider one final, 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, such as experience, skill, time spent per task, and attentiveness. Which of these can be influenced by external factors?

- *Experience.* Experience is a product of sustained interaction with the current type of task. It can affecting indirectly by focusing on methods to extend the length of a user's interaction, such as bonus payments for staying around.
- *Skill.* Skill is developed over time and is mostly affected by factors internal to the worker. To the extent that we could affect it, most functionality would encourage greater experience. Teaching workers by reinforcing their successes and failures might also have an effect.
- *Self-confidence and decisiveness.* Contributors or workers that second-guess themselves more often may be less internally consistent.
- *Attentiveness and fatigue.* Environmental distractions or fatigue can change how consistently a task is completed. The microtasking design pattern in paid crowdsourcing is meant to negate some of the fatigue seen in traditional classification labour, but there is no way to anticipate other outside factors, such has how many tasks from other requesters were completed. It is possible to affect attentiveness and fatigue within a task, however, with higher- or lower-effort tasks.
- *Perceived importance of task.* The perceived importance of a task might affect some other factors, such as attentiveness or self-confidence.
- *Time spent on each task.* The time spend on a task does not always translate to an indicator or quality, but might encourage greater numbers of contributions or more decisive contributions when controlled.

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.

Consider this study's Twitter encoding task. How would the contribution change if:

- Tasks were 100 items long? 200? 1000? Only 1?
- Instructions were written very tersely? Verbosely, with many examples?
- 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?

- Contributors were asked to volunteer their time? Were paid 1c per task? Were paid 10c per task? Were paid by the hour?
- Contributors were paid bonuses for performance against a ground truth or internal consistency? For continued task completion? For difficulty of their classification?
- Contributors were shown their performance (or estimated performance)? What if they were ranked against other workers? What if they gained levels or earned badges for performance?
- Contributors had tasks/time quotas to meet for bonuses? What if they were forced into these quotas (with tasks automatically moving forward)? What if a timer ticked away until their task disappear?
- 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-term engagement from individuals do not appear to be a promising direction. Worker experience was previously measured (Organisciak, Efron, 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, notably by Mason and Watts (2010). With regards to designs that mislead workers about their performance, there are ethical and trust issues that limit such an approach, in addition to the warning by Kraut and Resnick (2011) that feedback is only effective when contributors believe it is sincere.

3.4.4 Proposed designs

So what tweaks will this study measure?

As outlined in the overview of my own doctoral research, a few directions look like promising continuations of my research.

- It is still unclear whether simple encoding tasks benefit more from workers' gut instincts or careful consideration. Designs that can change a worker's attentiveness address an interesting problem and may bring potential improvements.

- Having previously found that reading instructions slowly is important for properly performing work, it should be seen whether a task can push a worker into internalizing the codebook rather than interpreting it.
- Understanding 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, I propose the three interfaces to study for crowdsourced data collection: a training interface, a feedback interface, and a time-limited interface.

Basic interface

The basic interface will resemble an archetypal task, following conventions seen in Mechanical Turk usage. It will show workers an tasks with a goal, description, and ten items to perform actions on. The goal of the interface will be to "identify the topic of a tweet." For each item, a multiple-choice question will be posed, with the proper noun phrases provided as options, as well as a free-text "Other" category and an "unknown" option. The description will explain what a "topic" is, and make clear the difference between a topic and simply a mention. An example will be included with the description, but as a pop up window behind a "See Example" link that needs to be clicked. This is done to conform to the convention that instructions should not be too long, in order not to push the actual action items 'below the fold'. Amazon's own advice for designing good tasks states that the task should not require scrolling to start (*Requester Best Practices* 2011).

Training interface

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

The training tasks will be hand-designed, based on a random sample of items.

Feedback Interface

In the feedback interface, a worker is shown feedback about their estimated performance on past tasks. The first that they complete is identical to the basic interface. Starting with the second task, however, the top of the interface will tell users:

- Their estimated performance, in terms of agreement with other workers.
- A visualization of where they fall in the distribution of all workers, from best performing to worst,

Since the interaction of this interface truly begins on the second task, evaluation of this interface will also focus on users returning after their first task.

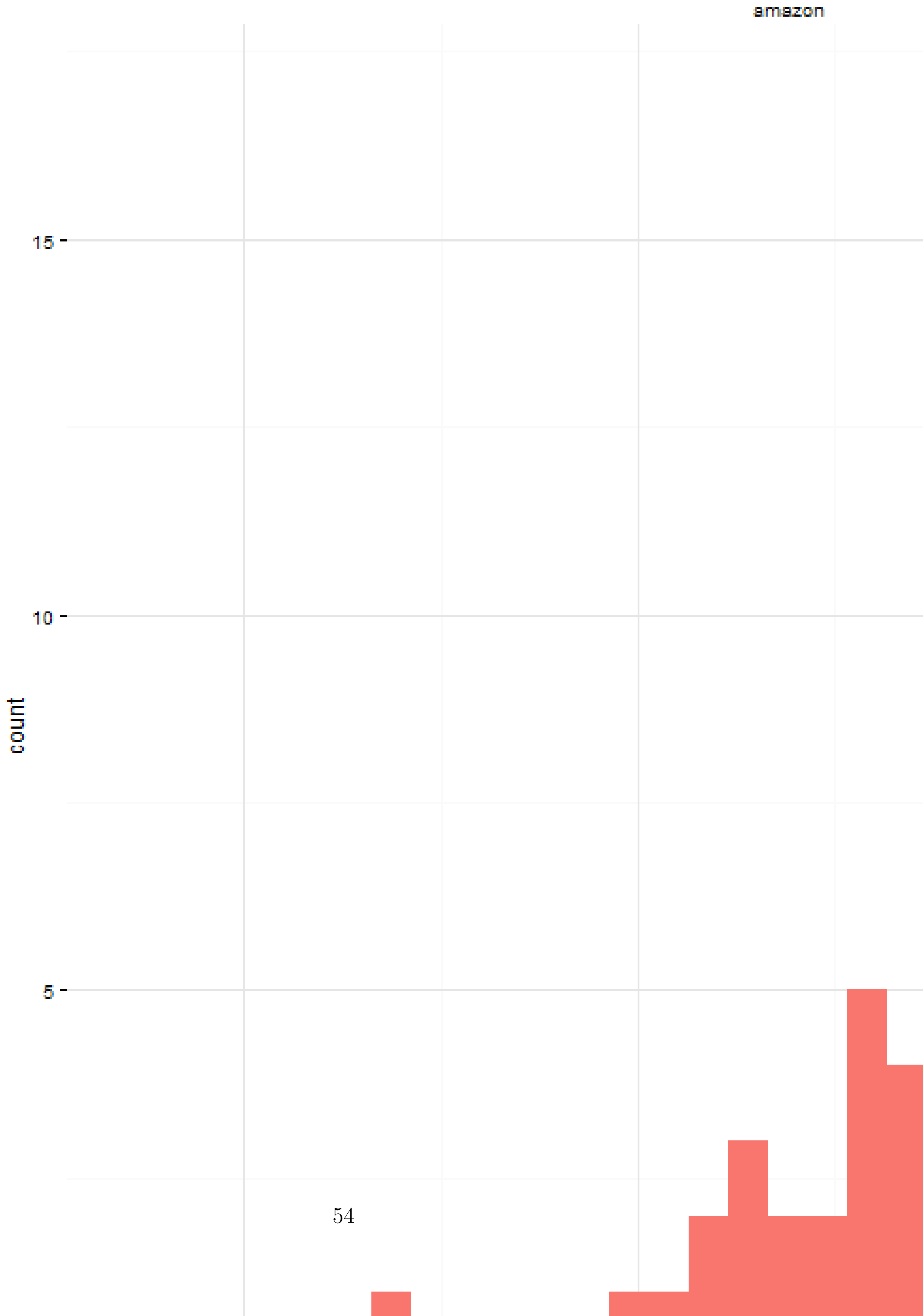
McCreadie, Macdonald, and Ounis (2011) attempted a similar approach, where contributors were shown a sidebar color-coding all their contributions based on their agreement with other raters and the authors. Showing this information with such granularity encourages workers to go back to reconsider debated answers, whereas this study's take tries to encourage more care and competition moving forward.

Time-Limited Interface

As hinted at during my past work, 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 timer to complete all tasks.

It is important not to distress the worker when trying to push them into a visceral form of task completion, as this might have the opposite effect.


amazon
count
15
10
5

Instead, this design should encourage flow, where a user moves seamlessly through the tasks without over-thinking their answers. To avoid the potentially distress of thinking about what is to come, this interface will not show a list of tasks to complete (e.g. "complete these ten tasks in a minute"). Instead, tasks will be shown one at a time (e.g."See how many tasks you can complete in a minute"), with bonuses paid for each complete task and increased for correct answers.

3.4.5 Evaluation

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. For example, workers cannot be forced to perform multiple tasks, simply encouraged to do so. Also, the actual user pools testing the different interfaces are not necessarily the same individuals. Thus, it is important that the users are similarly representational: it would be problematic if one interface was used mainly by Indian residents while another was performed mainly by American residents (the second and first larges nationalities on Mechanical Turk, respectively).

For this reason, each interface will be evaluated with temporal and geographic restrictions. Workers will be restricted to American workers, and tasks will each be posted during the American work day: between 10am and 1pm Pacific Time during weekdays.

Baseline

The baseline for evaluation is the performance of workers on the basic contribution interface.

Measurement

Primary questions

- Quantitative
- What is the mean agreement between workers encoding the same tweet?
  - Null hypothesis: That the agreement distributions for interface X is equal to or less than the distribution for the basic interface.
- What is the internal consistency of workers when they are asked to encode the exact same tweet?
  - Null hypothesis: That the consistency distributions for interface X is equal to or less than the distribution for the basic interface.
- Qualitative
- User feedback: all tasks will include optional feedback forms. This will include an free-text field for any communication that workers may want to pass on, and a Likert scale question on how interesting the task was.

Secondary questions

- What is the mean time a worker spends on their first task set? What is the amount of time workers spend on ten task sets?
- What is the mean number of task sets that workers perform?
- What is the expected cost per contribution?

3.4.6 Implementation

The experiments will be performed on Amazon's Mechanical Turk, using an API that allows external pages to be hosted within the Mechanical Turk interface. Funds for workers will be provided out of pocket, though I will seek dissertation research funding where available. The Graduate College at Illinois lists grant funding opportunities that may be worth pursuing [1].

The systems themselves will be developed using JavaScript for the front end, built on top of the Backbone.js model-view library with require.js used to modularize the code. Backbone.js is a strong choice for binding data to arbitrary views, offering flexibility for our comparative interfaces. On the back-end, the stack will be run on a Node.js server with MongoDB for data storage. These options are not critical, but they are fast for concurrent activities, reducing my server needs.

[1] http://nres.illinois.edu/graduate/funding-your-research

The software will be developed by myself, but is not drastically different than past systems that I have built and should not present any difficulties beyond time commitment. Honoring the notion of dissertation work as a public contribution, all development will be performed in a reusable manner, and released with an open-source Apache license[2].

3.5 Conclusion

Crowdsourcing is a promising approach for collecting descriptive document metadata. We have seen that the contribution changes as the instrument does, but there has not been much study into the effect of design on the reliability and consistency of contributions. This chapter of the dissertation allows these issues to be explored, stimulated by a study comparing the data from various designs of a microblog message encoding task.

[2] https://github.com/organisciak/crowdy

Chapter 4

MODELING USER-CONTRIBUTED DOCUMENT METADATA

4.1 Introduction

While on-demand collection of document metadata through crowdsourcing can be invaluable in controlled circumstances, there are cases where either the data is already collected, or where there is a limit to the amount of control a system designer can exert over contributors before discouraging them.

In the second of two research chapters, the proposed dissertation will investigate the issue of interpreting human-contributed metadata in an information retrieval context, viewed through the lens of an ranked retrieval study on the online community *Pinterest*. Whereas the previous chapter looks at issues related to *collecting* document information from a crowd, this chapter will consider the ways that this form of information is *used*.

The focus of this chapter will be ranked retrieval over content in the online community *Pinterest*, a system for users to publicly save web images in curated lists. Pinterest is a heavily user-generated website where the majority of crowd contributions fit into the descriptive or curatorial mode that this dissertation is concerned with. It is an appropriate focus for this study because Pinterest generates new information about existing information object, and does so in a very loosely constrained manner. Studying retrieval in the context of Pinterest allows broader exploration of interpreting abstract contributions in concrete ways more broadly.

The fundamental difficulty in incorporating user-contributed evidence into an information retrieval model is that often it is subjective in nature. While this is the type of contribution that sets human contributors apart, it also presents difficulties for interpreting it in a computational way. This chapter treats the description of 'pins' – community created visual bookmarks on Pinterest – as mixtures of human interpretations on their aboutness. Documents are treated as generative language models, where the actual text of a pin is smoothed against language from co-occurring pins in member lists and language from other users' pins of the source material.

The research performed in Organisciak, Efron, et al. (2012) sheds light on a particular problem of interpreting human contributions: consensus-building in ground truth tasks. We studied time, experience, and agreement as indicators of the quality of contributions for a paid relevance feedback task. Answering these questions provided valuable insights into how to treat workers and their data when using paid crowdsourcing for building ground truth datasets. However, modeling a crowd contribution is an issue that extends beyond paid ground truth generation. Various types of crowdsourcing data have been used for understanding information retrieval documents, including page links(Page et al. 1999), microblogging discussion of the documents(Dong et al. 2010), social tags(Lamere 2008), and implicit relevance feedback(Agichtein, Brill, and Dumais 2006). On contribution-heavy Pinterest, this study will look at how curated lists can expand a system's language model of the typically text-sparse Pinterest documents. Equally interesting, it will provide us an opportunity to look at where crowdsourcing information fails, perhaps due to unexpected patterns of contribution or misuse of the system.

In this chapter, I will review how various forms of crowd contributions have been used for information retrieval and other modelling uses, focusing on both volunteer and paid contributions and grounded by a study on Pinterest information retrieval.

4.2 Scope

The goals of this chapter will stay unchanged, focusing on crowdsourcing additional metadata for improved information retrieval indexing, with an underlying assumption of honest-but-biased workers. Again, the use cases being looked at are those with objective goals, with the intention of producing an output with minimal divergence from either the norms of a community or the instructions of a task designer. While user-dependent personalization approaches are an equally promising research direction, they are not part of the questions being investigated here.

This study focuses on volunteered data, mainly because it is a more novel space. My past research has already looked at issues around interpreting a contribution in paid-crowdsourcing, post-collection. It is also generally a well-explored space in research (notably including: Ipeirotis, Provost, and Wang 2010; Raykar et al. 2009; Eickhoff and Vries 2012; Sheng, Provost, and Ipeirotis 2008; Welinder and Perona 2010; Whitehill et al. 2009; Snow et al. 2008; Novotney and Callison-Burch 2010). Others have looked at similar situations of consensus-making among experts (B. Wallace et al. 2011). Secondly, there is an appeal to pursuing realistic contexts. In a production information retrieval context, it is more common to design around user needs, leaving the system designer to use the data for retrieval as it comes in.

4.3 Motivation

How do you model a collection of loosely-structured and subjective human contributions into a normative crowd opinion, one that can be used to describe objects in a corpus?

It is common to see differences in the habits of contributors that are trying to achieve the same thing. Many tasks that assume an objective, correct contribution are nonetheless are subject to interpretation, especially in volunteered contributions where detailed codebooks are deterrents to casual contributions. Likewise, even when a task is subjective, where a contribution is understood to be related to each individual's tastes and opinion, there are still problems of reliability beyond differences of opinion. One person's definition of a 3-star opinion judgment or their threshold for what is needed to click an approving 'thumbs up' button might be different from another person's, even if their underlying opinions are identical.

Speaking about tagging in the context of information retrieval, Zhou et al. (2008) warn that "a tag represents an abstract of the document from a *single* perspective of a *single* user."

Subjective crowdsourced contributions can be treated either through a personalized approach and a consensus-seeking approach.

For personalization, a system assumes that each opinion is representative of a specific type of user, and that the data can be used to model the individual users. Collaborative filtering usually makes this assumption. Doing so keeps closer to a user's tastes in highly divisive domains, separating Black Sabbath from Black-Eyed Peas, or Doctor Zhivago from Doctor Doolittle. This was the assumption made in Organisciak, Teevan, et al. (2013).

Another approach to subjectivity is to seek consensus. This approach is not as nuanced to the differences of opinions between humans, but has a few benefits. First, it is easier to communicate to users a single description or opinion. If five out of six users call a film 'funny' but the last person states the opposite, it's easier to tell a user at a glance that the film is 'funny' then trying to communicate the nuance. Secondly, it accommodates users that have never been seen before: recommending the option that has the maximum likelihood of satisfy any user is safe for new users. This is the *Beatles* option: even if you admit, rightly, that the context is highly subjective, you still need to consider the global best-guess.

In a context like Pinterest, users might search for terms that do not have a single right answer, but rather one that is interpreted or negotiated. When a user searches for 'rustic wedding' or 'cute dress', it is difficult to infer what their interpretation of 'rustic' is or what they find to be 'cute' what they consider 'rustic' or 'cute' without knowing anything about the user.

It is this latter approach to representing subjective aspects of document, to seek consensus among many individuals, which is the focus of this study. When using crowdsourced information to inform a general understanding of a document, how do you model the variety of interpretations and what are the pitfalls to avoid?

4.4 Approach

This chapter's focus will be a study measuring the value of crowdsourced information in improving information retrieval ranking against the data from Pinterest.

Pinterest is an online community for saving visual bookmarks called 'pins' to curated lists called 'boards'. On their about page, Pinterest features three primary purposes: saving (as pins), organizing (into boards), and discovery (*About Pinterest* 2014).

RQ1: How can crowdsourced contributions be incorporated into an generative language retrieval model, and to what effect?

RQ2: Can crowdsourced information be improved as evidence by adding contributor-dependent normalization or smoothing techniques?

RQ3: How do you account for novelty, allowing for new items without any crowdsourcing contributions?

For *RQ1*, this study treats descriptive user contributions as evidence for estimates of $P(Q|D)$. As a secondary question in addressing *RQ1*, user contributed quality judgments (i.e. "faves", reshares, sharing to external social networks) are considered as evidence for a document's prior probability, $P(D)$.

The primary contributions is in improving a document's language model by smoothing it against the language model of the curated lists, 'boards', that it belongs to. Documents, are treated in two ways: as an individual user's 'pin' – their visual bookmark of a page on the internet alongside their title, description, and the board they add it to – and as the meta-document, a collection of all user's pins that same source image.

The variability of human interpretations is embraced in modeling a document, rather than pushed against. Not everybody sees the same features in the same document, so user-contributed document metadata is treated as a mixture of interpretations.

Why study Pinterest? Pinterest is a novel website for studying ways to incorporate crowdsourced information into web retrieval.

- The organizational form of Pinterest, grouping documents into curated lists called 'boards', is a interface pattern that is relevant to many forms of information repository. Social OPACs, for example, allow library patrons to collect books into similarly uncontrolled lists.
- Pinterest contains very little information about the source web document. It is feasible to crawl the full text of the source, but as it stands, a Pinterest 'pin' alone offers a record of a *human's interpretation* of the source. It is simple, and helps us avoid confounding the focus on crowd contributions.
- Since the primary form of Pinterest document is a human reaction to a web document, the user contributions on the site may have possible future use for web retrieval.

Finally, Pinterest is an interesting but understudied website. Demographically there is a female skew, interesting precisely it counter-balance the typical male-heavy community demographic.

4.4.1 Data

This study will be performed on Pinterest, a website of curated images.

Pinterest is built entirely on crowd contributions. On Pinterest, the document unit is a ‘pin’: an image, associated with a web URL and page title, and a required text description provided by the user. Though the most common type of pin is saved from an external website, it is also possible to upload personal content. The ‘descriptions’ are required but free-text, meaning they do not necessarily *describe* the image.

Pins are sorted into curated lists, referred to as ‘boards’. Like pins, classification into board is not controlled. While adding a pin to a board is an act of classification, the classes are user-defined and can be created for various reason, such as quality judgments (e.g. “Neat stuff”), thematically descriptive (e.g. “dream wedding”), or miscellany of various sorts (e.g. “inspiration”, “funny”). Boards are user-specific, created by a user with a title, description, category, and optional map.

Animals	Architecture	Art	Cars & Motorcycles
Celebrities	Design	DIY	Crafts
Education	Film	Music & Books	Food & Drink
Gardening	Geek	Hair & Beauty	Health & Fitness
History	Holidays & Events	Home Decor	Humor
Illustrations & Posters	Kids	Men’s Fashion	Outdoors
Photography	Products	Quotes	Science & Nature
Sports	Tattoos	Technology	Travel
Weddings	Women’s Fashion	Other	

Table 4.1: Categories for curated lists (‘boards’) on Pinterest

There is also the concept of a ‘repin’, which involves saving a new pin from an existing pin, using the same source URL and image, but applying a new description and saving to a new board. A document's repin count can be interpreted as a measure of a document's internal influence among the Pinterest community. Additional community-specific social features include commenting on pins and 'liking', which is a unary voting mechanism.

The explicit forms of descriptive crowdsourcing that are seen on Pinterest are:

- Describing pins: description field, choice of board membership
- Describing boards: title, description, category, fields
- Social contribution: commenting on pins, repinning, 'liking', Facebook integration

4.4.2 Design: A Language-modeling approach to curated lists

Adopting a language modeling approach for this study, documents are ranked by estimating the probability of each document's language model generating the query, and that document's prior probability of being relevant. Estimating document d for query q,

$P(d|q) \propto P(q|d)P(d)$,

where q is a set of terms t.

Given a basic case of the unigram model (Ponte and Croft 1998; Song and Croft 1999) a document's prior probability of generating the query, $P(d)$, is assumed to be constant across all documents. This is the approach used for the baseline system, *provided in detail in the baseline section below.*

This work approaches $P(q|d)$ as an estimate that may be improved by user-contributed description of the document and, of secondary focus, $P(d)$ as an estimate that may be improved by quality judgments.

P(q|d)

Most basically, $P(q|d)$ starts with a maximum likelihood estimate of all the query terms occurring in the user's pin: $P_{ml}(t_i|d)$, where $P(q|d) = \prod_{i=1}^{|q|} P(t_i|d)$.

In this study's approach, we assume that co-occurring pins in lists and other users' pins of the same source content represent additional interpretations of the pin's aboutness. This will inform a model for boards, $P(t_i|b)$, which can be used as a fallback model, as well as a language model of other user's saves of the same document, $P(t_i|D)$. Thus, $P_{ml}(t|d)$ provides an estimate on seen words, but smoothing against $P(t_i|b)$ and $P(t_i|D)$ adds information on unseen but potentially likely words, and smoothing against the collection model provides general insight on term probabilities across the Pinterest corpus. These will be incorporated into a document's language model used the cluster approach seen in Liu and Croft (2004).

By smoothing a pin's language model with models provided by other documents and users, probability mass is dispersed among different interpretations of the aboutness of the document. This treatment of multiple subjective interpretations loosens the assumptions of language modeling. However, functionally it is the same as treating a language model as an objective but latent generative model with different probabilities assigned for different term occurrences.

There are some added complexities that will need to be considered while this study is being conducted. One is the proper weights to apply to interpolation: what type of smoothing is necessary between different models? This study will evaluate educated guesses based on work performed by Zhai and Lafferty (2001), and decide on whether an genetic parameter-learning algorithm is necessary.

Two other issues that may need to be considered are occasions when there is an absence of other information, perhaps when a pin is alone in its board and nobody else has saved the same source item. In instances like this, it is important not to overly rely on a single user's interpretation, since taken alone it is a biases description.

As described earlier in the literature review, this approaches closest parallel is in research on retrieval over folksonomies (e.g. Zhou et al. 2008; Bao et al. 2007; Hotho et al. 2006; Bischoff et al. 2008).

4.4.3 Data Collection

Three types of information will be collected from Pinterest:

- pins
- boards

- users

This will be collected according to the listings of all pins, boards, and users in Pinterest's provided sitemap. Based on an initial survey of the sitemaps, I expect approximately 107.5 million users on the website, 207.5 million pins, and 571.95 million boards.

This is a very large amount of data, and a bottleneck that is likely not necessary for this study. Instead, a smaller sample will be collected with the following sampling strategy:

1. A sample of boards is randomly selected
2. All pins that belong to the board sample are collected
3. All pins that save the same source images are collected
4. User data for the creators of the sampled boards is collected
5. A second sample of boards is collected, with all the boards that the sampled pins belong to

The exact size of the sample will be determined once I start collecting data. As a general rule, I would like to collect as much data as possible, while staying within a manageable file size and crawling timeline.

4.4.4 Evaluation

Evaluators

Given that the number of registered users on Pinterest is very high, approximately 107.5 million, it should be feasible to perform a more naturalistic evaluation, recruiting real users as judges for real queries. For evaluation, I will recruit Pinterest users locally to perform relevance judging.

Judgement Design

Users will be asked to judge the relevance of 130 documents per retrieval model on a graded scale. Documents will be shown in randomized order. The large number of results judged per query is influenced by the visual format of Pinterest. Pinterest's visual interface is quicker to browse than text results, and I expect that the common focus on ten ranked results is too small in a realistic setting. Search results on Pinterest's IR system load 65 results initially, though an 'infinite scroll' keeps loading results as a user scrolls down the page. Note that the judging interface format may change following the results of the first half of this dissertation.

Because of the nuances of the document space, binary relevance is a low bar to achieve on Pinterest. Many of the user information needs on Pinterest revolve around taste, and an appropriate evaluation should be sensitive not only to whether a document is a match to the query, but *how good* of a match it is. This is why evaluation will be performed with graded relevance. The primary metric for relevance will subsequently be Normalized Discounted Cumulative Gain (NDCG).

One concern with NDCG is that it needs to be estimated when there do not exist judgments for all results. This is because, for the normalization, one needs to consider the rank of a document relative to the ideal ranking of all results. This study will use the approach used at TREC, of pooling the top results for each algorithm's output, and assuming the result of documents are non-relevant. Since I expect a long tail of somewhat relevant queries on Pinterest, during the performance of the study I will also consider the necessity of NDCG estimation based on random sampling, such as infNDCG (Yilmaz, Kanoulas, and Aslam 2008).

Evaluation Queries

The queries being evaluated will be a mix of evaluators' search history – dependent on what they feel comfortable providing – and on popular queries collected through the Pinterest query input auto-complete feature.

When an evaluator is recruited that has used Pinterest, they will be provided a script that collects their Pinterest search history. They will be asked to cull the list down to queries that they feel comfortable sharing, and to resist the desire to revisit the results until after evaluation.

Additional evaluation queries will be sampled from auto-complete suggestions on Pinterest. When a user starts to type in a query, five suggestions appear. For example, typing 'r' will suggest 'recipes', 'red hair', 'rings', 'relationship quotes', and 'rustic wedding'. These appear to be the five-most probable queries starting with the provided string. For an insight of what types of queries are in the sampling frame and more generally what topics are popular among Pinterest users, Table 4.2 lists the auto-complete suggestions when each letter of the alphabet is entered into the search box.

It should be noted, however, that a sample frame of just the most popular terms is too general. To shift the sample away from the head of the distribution, the sampling frame will also include 500 queries derived from auto-complete suggestions based on two character strings: specifically, the one hundred most common two-character pairs occurring at the start of the English language.[1]

For each query, a description of what constitute the different levels of relevance will be written by myself, and the relevance of the first one hundred results will be rated by paid workers on a graded relevance scale.

It is likely that Pinterest's own retrieval model incorporates additional implicit feedback from users in the form of click-through data. This is a useful indicator of a item's quality, itself a form of crowd-contributed retrieval evidence, but is well-studied and too removed from the scope of this study to undertake.

[1] Using the frequencies calculated by Norvig (2014), these are: TH, OF, AN, IN, TO, CO, RE, BE, FO, PR, WH, HA, MA, WI, HE, IS, NO, WA, ON, DE, ST, SE, AS, IT, CA, HI, SO, WE, AR, DI, MO, AL, SU, PA, FR, ME, OR, SH, LI, CH, WO, PO, EX, BY, AT, FI, PE, BU, LA, NE, UN, LE, SA, TR, HO, YO, LO, DO, FA, SI, GR, EN, AC, MI, TE, BO, BA, GO, SP, OU, PL, EV, AB, TA, RA, US, BR, CL, DA, GE, TI, FE, AD, MU, IM, AP, RO, NA, SC, PU, EA, CR, VI, CE, OT, AM, AG, UP, RI, VE.

appetizers, art, ab workout, animals, apartment decorating, appetizers, art, ab workout, animals, apartment decorating, christmas, christmas decorations, chicken recipes, crockpot recipes, christmas crafts, diy, dinner recipes, dresses, desserts, disney, easter, engagement rings, elf on the shelf ideas, eye makeup, easter crafts, food, fashion, funny, funny quotes, fall, garden, gift ideas, gluten free, girls bedroom, gardening, hair styles, hair, healthy recipes, halloween costumes, halloween, inspirational quotes, interior design, ikea, i love you, italy, jewelry, jennifer lawrence, jello shots, jeans, jokes, kitchen, kitchen ideas, kids crafts, kitchen decor, kids, love quotes, love, living room, long hair, lingerie, makeup, medium hair styles for women, mothers day, mothers day gifts, master bedroom, nail art, nails, nail designs, nail art designs, nail art for short nails, ombre hair, organization, organization tips, outfits, organizing, prom dresses, pregnancy, prom hair, paleo, puppies, quotes, quinoa, quinoa recipes, quilts, quotes about change, recipes, red hair, rings, relationship quotes, rustic wedding, spring fashion, shoes, short hair styles for women, short hair, sexy, tattoos, thanksgiving, tattoo ideas, thanksgiving recipes, travel, updos, updo hairstyles, ugly christmas sweater, updos for medium length hair, valentines ideas, valentines day gifts for him, valentines day, vintage, valentines crafts, wedding, wedding dresses, wedding hair, wedding rings, wedding ideas, xmas, x, xmas crafts, xmas decorations, x rated, yoga, yoga poses, yoga pants, yellow, yoga workout, zucchini recipes, zucchini, zac efron, zara, zucchini bread

Table 4.2: Popular queries on Pinterest, showing the 5 search input auto-complete suggestions for each letter of the alphabet. Though Pinterest requires users to be logged-in, this list does not appear to be personalized: the same list was derived when I asked other users to run the collection code.

Baseline

The baseline for the system will be a unigram language model, with the query likelihood based on the terms of a document's title and user description, and smoothed against the collection likelihood with linear smoothing (i.e. *Jelinek-Mercer*).

Fundamentally, the language modelling approach assumes each document as a generative model, and estimates the probability of any given document d generating a query q.

The unigram model makes a couple of simplifying assumptions. First, it assumes that the query-independent probability of a document being relevant, $P(d)$, is uniform, meaning it does not affect ranking and can effectively be ignored. Also removing the document-independent query constant, it means that

$P(d|q) \propto (P(q|d))$,

leaving our estimate in the hands of P(q|d). The second assumption for the unigram model is that the probability of a document generating a query is the joint probability of each term in the query occurring in the document. Thus,

$P(q|d) = \prod_{i=1}^{n} P(q_i|d)$.

This assumes a conditional independence between terms. However, taking a maximum likelihood estimate for $P(q_i|d)$—dividing term q_i's occurrences in document d by the sum of all terms—suffers from some problems. First, it heavily punishes documents where a word has not occurred, over-emphasizing seen words in the document's language model. Secondly, it fails to account for the general likelihood of a word occurring in the language. Thus, we smooth between the query term in the document's language model and its probability in the collections language model ($P(q_i|C)$). Smoothing not only de-emphasizes seen words, but also reduces the discriminatory power of common words, akin to the TF-IDF intuition(Zhai and Lafferty 2001).

The baseline in this study will use linear smoothing between the two forms of evidence. With $\lambda = 0.75$ chosen based on a reading of Zhai and Lafferty (2001)'s evaluation of linear smoothing performance, the baseline will use the following estimation of $P(q_i|d)$:

$P(q_i|d) = (1 - \lambda)P(q_i|d) + \lambda P(w|C)$.

System

The testing system will be built on top of Apache Lucene. Lucene provides high performance for large collections. The baseline system will use the LMJelinekMercerSimilarity similarity scorer, and subsequent changes will be custom made.

4.4.5 Risks and fallbacks

This study contains some risks on non-results or failures.

Evaluation problems: If recruitment for evaluators proves difficult or if the methodology for evaluation is too fatiguing, this study can move to a more reliable albeit less naturalistic evaluation through Mechanical Turk.

Another evaluation concern is that, with the raw number of materials on Pinterest, there will be more than enough relevant results for the more popular topics related to head queries, even for the baseline system. The use of NDCG with graded relevance will offer more sensitivity to potential issues.

4.4.6 Related question: stylistic congealment in crowdsourcing

How does the stylistic variance of contributions change through the lifespan of a system? We hypothesize that as crowdsourcing systems grow in size, their contributions also grow less variant in their approach. This question is related to the research questions of the proposed dissertation and this chapter, and will be useful to consider it.

- Directionality of influence
- *If* structure is being solidified: why is it being systematized?
- Contribution attrition
- Simplicity vs. Hand-holding

If our hypothesis holds true, it would suggest that system designers (or self-governed crowds) steer their system toward more algorithmically tractable forms of contributions over time. Such a shift might be afforded by crowds large enough to allow the attrition associated with more complexity, or by the increase of well-experienced participants that are comfortable adapting to tighter guidelines.

One notable example of this type of shift is happening is in Wikipedia. Wikipedia does not set explicit rules, but conventions are established and policed by the community. For example, after a recommendation in 2006, editors began insisting on well-sourced information underlying all significant statement on the website (Wales 2006). The structure has also enabled efforts such as DBPedia, structured data representing the entities in Wikipedia pages. It has also informed information retrieval approaches to understanding entities beyond a query, such as with Google's Knowledge Graph and Microsoft's Satori.

4.5 Conclusion

Descriptive crowdsourcing provides metadata about information objects that has the potential to increase our computation models of them. However, crowds are human, biased in hard to predict ways. This chapter considers the issues involved in modelling crowd contributions, centered around an information retrieval study on Pinterest.

Chapter 5

CONCLUSION

5.1 Workplan

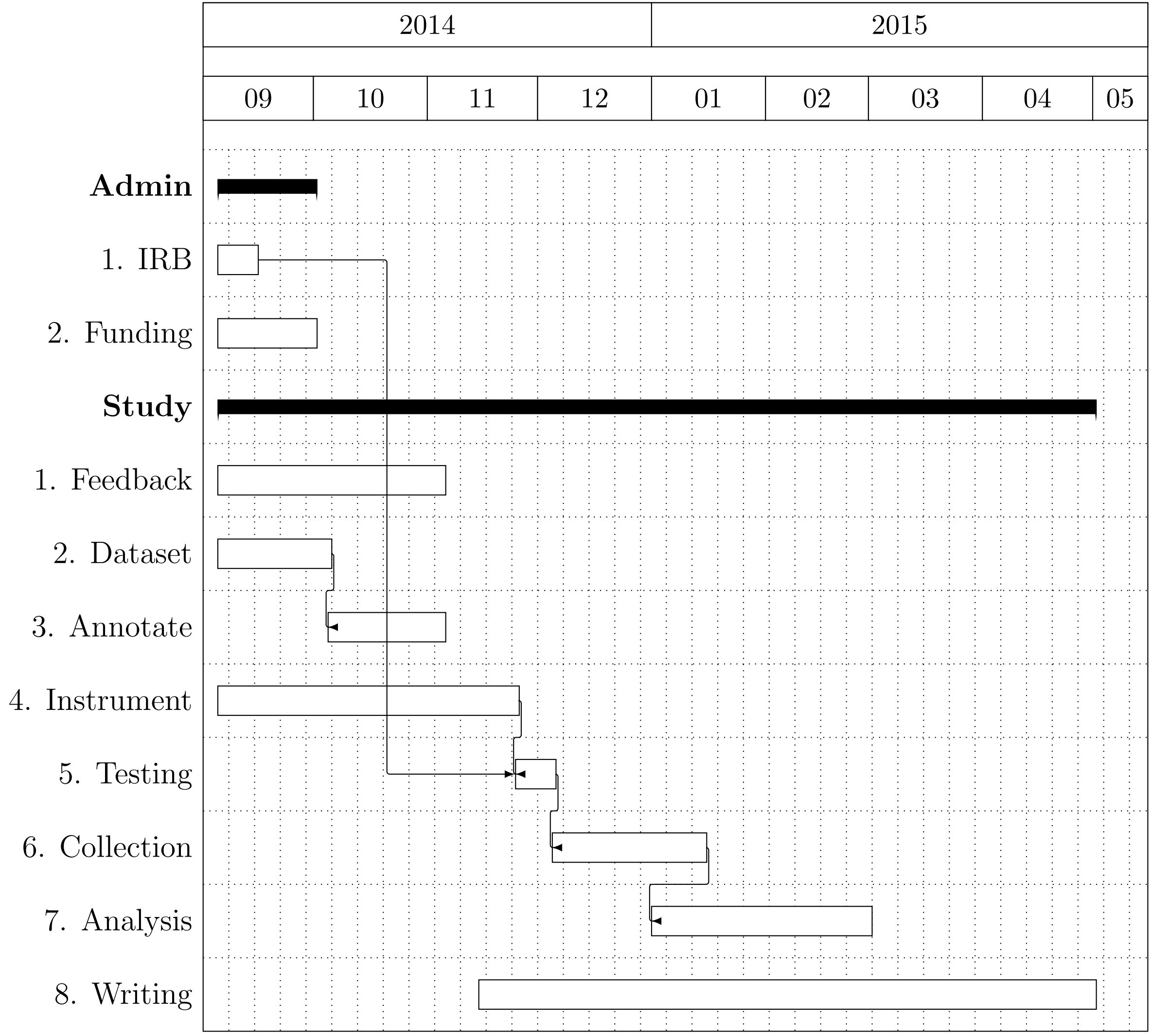
This dissertation will be performed over the course of 7 months, starting immediately in September 2014. Both parts of the dissertation – the study on the effect of task design and the study on retrieval over crowd content on Pinterest – will be pursued concurrently, with early priority on the former because it helps in designing the instruments for the latter.

Throughout this research, I will stay in weekly contact with my advisor, and hope to consult with the committee with increasing frequency as the research progresses.

5.1.1 Designing Crowdsourcing Tasks

Admin 1. Prepare and submit IRB Application. In studying the behaviors of workers, they are being treated as human subjects, requiring IRB approval. 2. Search for dissertation funding opportunities to alleviate cost.

Study Tasks 1. Incorporate dissertation committee feedback into study design 2. Find or collect Twitter dataset 3. Annotate gold standard set 4. Build measurement instruments 1. Basic Design * Prepare Infrastructure for collecting data. The collection interface will be built in JavaScript using the Backbone library, with Node.js and Express powering the back-end. 2. Feedback Design 3. Training Design 4. Time-Limited Design 5. Testing. It is important to understand a crowdsourcing design as a user before subjecting others to it. 6. Data Collection. The Payment amount will be determined based on statistics of time spent. 7. Analysis. How variable is the data? How accurate is it, relative to the gold and silver standards? How quickly do users perform the task, and how much do they enjoy it? What do the free text feedback forms say? How can the trade-off between cost/speed and quality (reliability/consistency/accuracy) be fairly compared. 8. Writing.



5.1.2 Modeling Crowd-contributed Document Metadata in Retrieval

Administrative

1. Secure server and disk space.

Study

1. Incorporate dissertation committee feedback into study design
2. Dataset building

    - Build crawler

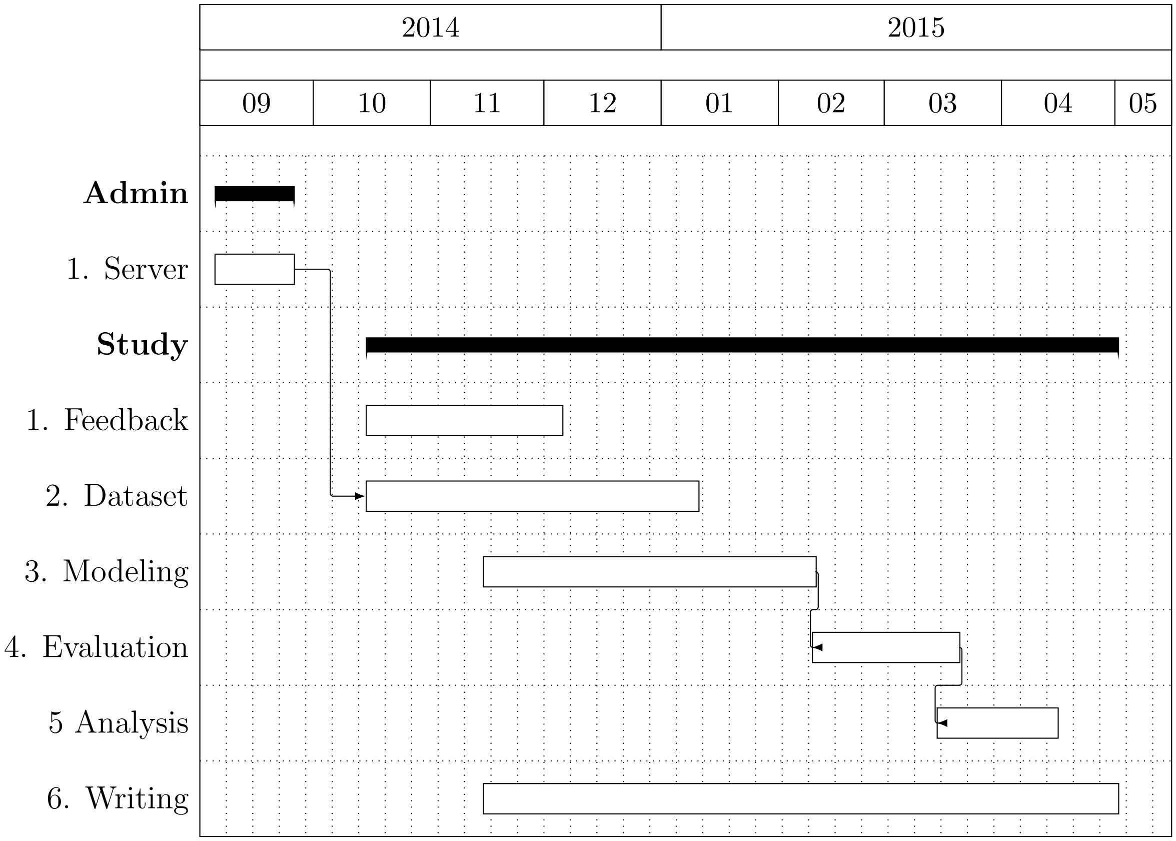
3. System development

    - Prepare baseline system
    - Implement board language model and broader item language model
    - Which mixtures of features will be tested? What is most promising from testing data, informal empirical impressions? Why?

4. Evaluation

    - Develop front-end for evaluation. The evaluation interface can build upon early impressions from the other chapter.
    - Generate a test collection for better sensitivity to methods during development.
    - Recruit evaluators
    - Sample evaluation queries
    - Collect relevance judgments

5. Analysis. Which approaches worked? Which didn't? What types of documents are favoured? What's being neglected? What types of queries do the retrieval models fail on?
6. Writing



5.1.3 Other Tasks

- Write richer literature review
- Publish code source online
- Tertiary studies
- Develop taxonomy of crowdsourcing, from this proposal's literature review, into a paper.
- Pursue study on the variability of crowdsourcing site design over time, testing the hypothesis that style congeals over time into a more rigid and computationally tractable form.

5.2 Publication Targets

There are a number of conferences that represent the target audience of this dissertation, generally mixed-domain areas that combine computer science or information science with sensitivity to social or human-centered issues.

Research from the chapter on design of crowdsourcing tasks would be appropriate at HCOMP, CSCW, CHI, or ASIS&T. The research from the chapter on retrieval over crowdsourced information on Pinterest would be appropriate at JCDL, CIKM, or potentially SIGIR.

5.3 Risks

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.

It is possible that the studies within the proposed dissertation may have null results. While this may be deflating, there is nonetheless value in learning from failure. The research proposed in this dissertation is novel and will contribute to a better understanding of how information system designers can interact with human contributors.

BIBLIOGRAPHY

About Pinterest (2014). Pinterest. URL: http://about.pinterest.com/en (visited on 08/15/2014).

Abt, Clark C. (1987). *Serious Games*. University Press of America. 200 pp. ISBN: 9780819161482.

Agichtein, Eugene, Eric Brill, and Susan Dumais (2006). “Improving Web Search Ranking by Incorporating User Behavior Information”. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’06. New York, NY, USA: ACM, pp. 19–26. ISBN: 1-59593-369-7. DOI: 10.1145/1148170.1148177. URL: http://doi.acm.org/10.1145/1148170.1148177 (visited on 08/11/2014).

Ahn, L. von (2006). “Games with a purpose”. In: *Computer* 39.6, pp. 96–98. ISSN: 0018-9162. URL: http://scholar.google.ca.login.ezproxy.library.ualberta.ca/scholar?hl=en&lr=&cluster=7220788619130524050 (visited on 04/17/2009).

Ahn, Luis von and Laura Dabbish (2004). “Labeling images with a computer game”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. Vienna, Austria: ACM, pp. 319–326. ISBN: 1-58113-702-8. URL: http://dl.acm.org/citation.cfm?id=985733 (visited on 11/03/2008).

Alderfer, Clayton P. (1969). “An empirical test of a new theory of human needs”. In: *Organizational behavior and human performance* 4.2, pp. 142–175. URL: http://www.sciencedirect.com/science/article/pii/003050736990004X (visited on 08/16/2014).

Alonso, Omar and Ricardo Baeza-Yates (2011). “Design and Implementation of Relevance Assessments Using Crowdsourcing”. In: *Advances in Information Retrieval*. Ed. by Paul Clough et al. Lecture Notes in Computer Science 6611. Springer Berlin Heidelberg, pp. 153–164. ISBN: 978-3-642-

20160-8, 978-3-642-20161-5. URL: http://link.springer.com.proxy2.library.illinois.edu/chapter/10.1007/978-3-642-20161-5_16 (visited on 01/27/2014).

Alonso, Omar, Daniel E. Rose, and Benjamin Stewart (2008). "Crowdsourcing for relevance evaluation". In: *SIGIR Forum* 42.2, pp. 9–15. ISSN: 0163-5840. DOI: 10.1145/1480506.1480508. URL: http://doi.acm.org/10.1145/1480506.1480508 (visited on 09/05/2012).

Angwin, Julia and Geoffrey A. Fowler (2009). "Volunteers log off as Wikipedia ages". In: *Wall Street Journal* 23. URL: https://secure.strategyone.net/mtr/Canon/2009/2009_11/228771_11252009/BIOE3434479_1_H.pdf (visited on 08/15/2014).

Bao, Shenghua et al. (2007). "Optimizing Web Search Using Social Annotations". In: *Proceedings of the 16th International Conference on World Wide Web*. WWW '07. New York, NY, USA: ACM, pp. 501–510. ISBN: 978-1-59593-654-7. DOI: 10.1145/1242572.1242640. URL: http://doi.acm.org/10.1145/1242572.1242640 (visited on 08/25/2014).

Bell, Robert M., Yehuda Koren, and Chris Volinsky (2008). "The BellKor 2008 Solution to the Netflix Prize". In: *Statistics Research Department at AT&T Research*. URL: ftp://140.118.199.9:2100/public9/2010_ALL/2010_TRU/2010Fall_Matlab-LABfiles/00_VECTOR-at-MATH/%C3%A5%C2%90%C2%91%C3%A9%C2%87%C2%8F%C3%A7%C2%9A%C2%84%C3%A5%C2%88%C2%86%C3%A8%C2%A7%C2%A3/Public9/zTEMP/ZTemp2011/zMSIC-NSC101-DM/40--Dec27/Netflix-Articles-AWARD/(netflix)-ProgressPrize2008_BellKor.pdf (visited on 01/12/2014).

Benkler, Yochai (2006). *Wealth of Networks*. New Haven: Yale University Press. URL: http://cyber.law.harvard.edu/wealth_of_networks/Download_PDFs_of_the_book (visited on 10/20/2008).

Bernstein, Michael S. et al. (2010). "Soylent: a word processor with a crowd inside." In: *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. UIST '10. New York, NY: ACM Press, pp. 313–322. ISBN: 9781450302715. DOI: 10.1145/1866029.1866078. URL: http://dl.acm.org/citation.cfm?id=1866078 (visited on 09/12/2011).

bias, adj., n., and adv. (2014). In: *OED Online*. Oxford University Press. URL: http://www.oed.com.proxy2.library.illinois.edu/view/Entry/18564 (visited on 01/09/2014).

Bischoff, Kerstin et al. (2008). "Can All Tags Be Used for Search?" In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM '08. New York, NY, USA: ACM, pp. 193–202. ISBN: 978-1-59593-991-3. DOI: 10.1145/1458082.1458112. URL: http://doi.acm.org/10.1145/1458082.1458112 (visited on 08/25/2014).

Causer, Tim, Justin Tonra, and Valerie Wallace (2012). "Transcription maximized; expense minimized? Crowdsourcing and editing The Collected Works of Jeremy Bentham". In: *Literary and Linguistic Computing* 27.2, pp. 119–137. ISSN: 0268-1145, 1477-4615. DOI: 10.1093/llc/fqs004. URL: http://llc.oxfordjournals.org/content/27/2/119 (visited on 08/26/2013).

Chen, Edwin and Alpa Jain (2013). *Improving Twitter search with real-time human computation*. Twitter Engineering Blog. URL: https://blog.twitter.com/2013/improving-twitter-search-real-time-human-computation (visited on 12/09/2013).

Cortese, Amy (2011). "A Proposal to Allow Small Private Companies to Get Investors Online". In: *The New York Times*. ISSN: 0362-4331. URL: http://www.nytimes.com/2011/09/26/opinion/a-proposal-to-allow-small-private-companies-to-get-investors-online.html (visited on 08/21/2014).

— (2013). "Crowdfunding for Small Business Is Still an Unclear Path". In: *The New York Times*. ISSN: 0362-4331. URL: http://www.nytimes.com/2013/01/06/business/crowdfunding-for-small-business-is-still-an-unclear-path.html (visited on 08/21/2014).

Dong, Anlei et al. (2010). "Time is of the Essence: Improving Recency Ranking Using Twitter Data". In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. New York, NY, USA: ACM, pp. 331–340. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772725. URL: http://doi.acm.org/10.1145/1772690.1772725 (visited on 07/21/2014).

Efron, Miles (2011). "Information search and retrieval in microblogs". In: *Journal of the American Society for Information Science and Technology* 62.6, pp. 996–1008. ISSN: 1532-2890. DOI: 10.1002/asi.21512. URL: http://onlinelibrary.wiley.com/doi/10.1002/asi.21512/abstract (visited on 06/14/2013).

Eickhoff, Carsten, Christopher G. Harris, et al. (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.* SIGIR '12. New York, NY, USA: ACM, pp. 871–880. ISBN: 978-1-4503-1472-5. DOI: 10.1145/2348283.2348400. URL: http://doi.acm.org.proxy2.library.illinois.edu/10.1145/2348283.2348400 (visited on 01/27/2014).

Eickhoff, Carsten and Arjen P. Vries (2012). "Increasing cheat robustness of crowdsourcing tasks". In: *Information Retrieval.* ISSN: 1386-4564, 1573-7659. DOI: 10.1007/s10791-011-9181-9. URL: http://www.springerlink.com/content/70807017421n1462/ (visited on 02/19/2012).

Finin, Tim et al. (2010). "Annotating Named Entities in Twitter Data with Crowdsourcing". In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.* CSLDAMT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 80–88. URL: http://dl.acm.org/citation.cfm?id=1866696.1866709 (visited on 12/07/2013).

Fung, Brian (2014). "Larry Lessig's super PAC to end super PACs raised $2.5 million in just 2 days. Here's what comes next." In: *The Washington Post.* ISSN: 0190-8286. URL: http://www.washingtonpost.com/blogs/the-switch/wp/2014/07/07/larry-lessigs-super-pac-to-end-super-pacs-raised-2-5-million-in-2-days/ (visited on 08/21/2014).

Galton, Francis (1907). "Vox populi". In: *Nature* 75, pp. 450–451. URL: http://adsabs.harvard.edu/abs/1907Natur..75..450G (visited on 10/22/2013).

Geiger, David et al. (2011). "Managing the crowd: towards a taxonomy of crowdsourcing processes". In: *Proceedings of the seventeenth Americas conference on information systems, Detroit, Michigan*, pp. 1–15. URL: http://schader.bwl.uni-mannheim.de/fileadmin/files/schader/files/publikationen/Geiger_et_al._-_2011_-_Managing_the_Crowd_Towards_a_Taxonomy_of_Crowdsourcing_Processes.pdf (visited on 05/24/2013).

Grady, Catherine 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.* CSLDAMT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 172–179. URL: http://dl.acm.org.proxy2.library.illinois.edu/citation.cfm?id=1866696.1866723 (visited on 02/19/2012).

Harris, Christopher G. and Padmini Srinivasan (2012). "Applying human computation mechanisms to information retrieval". In: *Proceedings of the American Society for Information Science and Technology* 49.1, pp. 1–10. ISSN: 1550-8390. DOI: 10.1002/meet.14504901050. URL: http://onlinelibrary.wiley.com.proxy2.library.illinois.edu/doi/10.1002/meet.14504901050/abstract (visited on 03/25/2014).

Hippel, Eric von (1988). *The Sources of Innovation*. SSRN Scholarly Paper ID 1496218. Rochester, NY: Social Science Research Network. URL: http://papers.ssrn.com.proxy2.library.illinois.edu/abstract=1496218 (visited on 06/11/2014).

— (2006). "Democratizing Innovation". In: URL: http://econpapers.repec.org/bookchap/mtptitles/0262720477.htm (visited on 06/12/2014).

Hofmann, Thomas (2004). "Latent semantic models for collaborative filtering". In: *ACM Trans. Inf. Syst.* 22.1, pp. 89–115. ISSN: 1046-8188. DOI: 10.1145/963770.963774. URL: http://doi.acm.org/10.1145/963770.963774 (visited on 09/18/2013).

Holley, Rose (2009). *Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers*. National Library of Australia. URL: http://www-prod.nla.gov.au/openpublish/index.php/nlasp/article/viewArticle/1406.

Hotho, Andreas et al. (2006). *Information retrieval in folksonomies: Search and ranking*. Springer. URL: http://link.springer.com/chapter/10.1007/11762256_31 (visited on 08/25/2014).

Howe, J. (2006a). *Crowdsourcing: A Definition*. URL: http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html.

— (2006b). "The rise of crowdsourcing". In: *Wired Magazine* 14.6.

Howe, Jeff (2006). *Birth of a Meme*. Crowdsourcing. URL: http://www.crowdsourcing.com/cs/2006/05/birth_of_a_meme.html (visited on 04/26/2014).

— (2008). *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. 1st ed. Crown Business. 320 pp. ISBN: 0307396207.

Ipeirotis, Panagiotis G., Foster Provost, and Jing Wang (2010). "Quality management on Amazon Mechanical Turk". In: *Proceedings of the ACM*

SIGKDD Workshop on Human Computation. HCOMP '10. New York, NY, USA: ACM, pp. 64–67. ISBN: 978-1-4503-0222-7. DOI: 10.1145/1837885.1837906. URL: http://doi.acm.org.proxy2.library.illinois.edu/10.1145/1837885.1837906 (visited on 02/19/2012).

Khatib, Firas et al. (2011). "Algorithm discovery by protein folding game players". In: *Proceedings of the National Academy of Sciences* 108.47, pp. 18949–18953. URL: http://www.pnas.org/content/108/47/18949.short (visited on 08/25/2014).

Komarov, Steven, Katharina Reinecke, and Krzysztof Z. Gajos (2013). "Crowdsourcing Performance Evaluations of User Interfaces". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. New York, NY, USA: ACM, pp. 207–216. ISBN: 978-1-4503-1899-0. DOI: 10.1145/2470654.2470684. URL: http://doi.acm.org/10.1145/2470654.2470684 (visited on 08/17/2014).

Kraut, Robert E. and Paul Resnick (2011). *Building Successful Online Communities*. Cambridge, MA: MIT Press.

Lakhani, Karim R. and Eric von Hippel (2003). "How open source software works: "free" user-to-user assistance". In: *Research Policy* 32.6, pp. 923–943. DOI: 10.1016/S0048-7333(02)00095-1. URL: http://www.sciencedirect.com/science/article/B6V77-479TM54-1/2/5672b73de696a2d8a1d68e6d5747a2cb (visited on 10/20/2008).

Lamere, Paul (2008). "Social tagging and music information retrieval". In: *Journal of New Music Research* 37.2, pp. 101–114. URL: http://www.tandfonline.com.proxy2.library.illinois.edu/doi/abs/10.1080/09298210802479284 (visited on 04/04/2014).

Law, Edith and Luis von Ahn (2011). "Human Computation". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5.3, pp. 1–121. ISSN: 1939-4608, 1939-4616. DOI: 10.2200/S00371ED1V01Y201107AIM013. URL: http://www.morganclaypool.com.proxy2.library.illinois.edu/doi/abs/10.2200/S00371ED1V01Y201107AIM013 (visited on 09/18/2013).

Le Bon, Gustav (1896). *The Crowd: A Study of the Popular Mind*. URL: http://socserv.socsci.mcmaster.ca/~econ/ugcm/3ll3/lebon/Crowds.pdf (visited on 10/20/2008).

Lease, Matthew and Gabriella Kazai (2011). "Overview of the TREC 2011 Crowdsourcing Track (Conference Notebook)". In: *Text Retrieval Conference Notebook*.

Lists (2011). Bibliocommons. URL: http://help.bibliocommons.com/en-ca/045faq/060faq_lists (visited on 10/11/2011).

Liu, Xiaoyong and W. Bruce Croft (2004). "Cluster-based retrieval using language models". In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.* SIGIR '04. New York, NY, USA: ACM, pp. 186–193. ISBN: 1-58113-881-4. DOI: 10.1145/1008992.1009026. URL: http://doi.acm.org/10.1145/1008992.1009026 (visited on 02/04/2013).

Mackay, Charles (1852). *Memoirs of Extraordinary Popular Delusions and the Madness of Crowds.*

Maslow, A.H. (1943). "A theory of human motivation". In: *Psychological Review* 50.4, pp. 370–396. ISSN: 1939-1471(Electronic);0033-295X(Print). DOI: 10.1037/h0054346.

Mason, Winter and Duncan J. Watts (2010). "Financial incentives and the "performance of crowds"". In: *SIGKDD Explor. Newsl.* 11.2, pp. 100–108. ISSN: 1931-0145. DOI: 10.1145/1809400.1809422. URL: http://doi.acm.org/10.1145/1809400.1809422 (visited on 06/13/2012).

McCreadie, Richard, Craig Macdonald, and Iadh Ounis (2011). "Crowdsourcing blog track top news judgments at TREC". 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)*, pp. 23–26. URL: http://ir.ischool.utexas.edu/csdm2011/csdm2011_proceedings.pdf#page=23 (visited on 08/25/2014).

McCreadie, Richard, Craig Macdonald, Rodrygo LT Santos, et al. (2011). "University of Glasgow at TREC 2011: Experiments with Terrier in Crowdsourcing, Microblog, and Web Tracks." In: *TREC.* URL: http://homepages.dcc.ufmg.br/~rodrygo/wp-content/papercite-data/pdf/mccreadie2011trec.pdf (visited on 08/25/2014).

Michael, David R. and Sandra L. Chen (2005). *Serious Games: Games That Educate, Train, and Inform.* Muska & Lipman/Premier-Trade. ISBN: 1592006221.

Moyle, M., J. Tonra, and V. Wallace (2010). "Manuscript transcription by crowdsourcing: Transcribe Bentham". In: *LIBER Quarterly* 20.3. URL: http://liber.library.uu.nl/publish/issues/2010-3_4/index.html?000514 (visited on 02/01/2012).

Neuendorf, Kimberly A. (2002). *The Content Analysis Guidebook.* Thousand Oaks, CA, USA: Sage Publications. 301 pp.

Norvig, Peter (2014). *English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLDCU*. URL: http://norvig.com/mayzner.html (visited on 05/20/2014).

Novotney, S. and C. Callison-Burch (2010). “Cheap, fast and good enough: Automatic speech recognition with non-expert transcription”. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. HLT ’10. Stroudsburg, PA, USA, pp. 207–215.

Organisciak, Peter (2010). “Why Bother? Examining the motivations of users in large-scale crowd-powered online initiatives”. Thesis. Edmonton, Alberta: University of Alberta. 167 pp. URL: http://hdl.handle.net/10048/1370.

— (2013). “Incidental Crowdsourcing: Crowdsourcing in the Periphery”. In: Digital Humanities 2013. Lincoln, Nebraska. URL: http://dh2013.unl.edu/abstracts/ab-273.html.

Organisciak, Peter, Miles Efron, et al. (2012). “Evaluating rater quality and rating difficulty in online annotation activities”. In: *Proceedings of the American Society for Information Science and Technology* 49.1, pp. 1–10. ISSN: 1550-8390. DOI: 10.1002/meet.14504901166. URL: http://onlinelibrary.wiley.com/doi/10.1002/meet.14504901166/abstract (visited on 11/23/2013).

Organisciak, Peter, Jaime Teevan, et al. (2013). “Personalized Human Computation”. In: HCOMP 2013. Palm Spring, CA.

Page, Lawrence et al. (1999). *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab. URL: http://ilpubs.stanford.edu:8090/422/.

Ponte, Jay M. and W. Bruce Croft (1998). “A Language Modeling Approach to Information Retrieval”. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’98. New York, NY, USA: ACM, pp. 275–281. ISBN: 1-58113-015-5. DOI: 10.1145/290941.291008. URL: http://doi.acm.org/10.1145/290941.291008 (visited on 07/24/2014).

Quinn, Alexander J. and Benjamin B. Bederson (2011). “Human computation”. In: ACM Press, p. 1403. ISBN: 9781450302289. DOI: 10.1145/1978942.1979148. URL: http://dl.acm.org.proxy2.library.illinois.edu/citation.cfm?id=1979148 (visited on 10/19/2011).

Raykar, Vikas C. et al. (2009). "Supervised learning from multiple experts: whom to trust when everyone lies a bit". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. New York, NY, USA: ACM, pp. 889–896. ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553488. URL: http://doi.acm.org.proxy2.library.illinois.edu/10.1145/1553374.1553488 (visited on 02/19/2012).

Raymond, Eric S. (1999). *The Cathedral and the Bazaar*. O'Reilly Media. 241 pp.

Requester Best Practices (2011). URL: http://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf (visited on 08/08/2014).

Ritterfeld, Ute, Michael Cody, and Peter Vorderer (2010). *Serious Games: Mechanisms and Effects*. Routledge. 553 pp. ISBN: 9781135848910.

Rouse, Anne (2010). "A Preliminary Taxonomy of Crowdsourcing". In: *ACIS 2010 Proceedings*. URL: http://aisel.aisnet.org/acis2010/76.

Ryan, Richard M. and Edward L. Deci (2000). "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions". In: *Contemporary Educational Psychology* 25.1, pp. 54–67. ISSN: 0361-476X. DOI: 10.1006/ceps.1999.1020. URL: http://www.sciencedirect.com/science/article/pii/S0361476X99910202 (visited on 03/19/2014).

Sanger, Lawrence M. (2009). "The Fate of Expertise after Wikipedia". In: *Episteme* 6.1, pp. 52–73. DOI: 10.3366/E1742360008000543.

Schenk, Eric and Claude Guittard (2009). "Crowdsourcing: What can be Outsourced to the Crowd, and Why?" In: *Workshop on Open Source Innovation, Strasbourg, France*. URL: http://raptor1.bizlab.mtsu.edu/s-drive/DMORRELL/Mgmt%204990/Crowdsourcing/Schenk%20and%20Guittard.pdf (visited on 01/30/2014).

Sheng, Victor S., Foster Provost, and Panagiotis G. Ipeirotis (2008). "Get another label? improving data quality and data mining using multiple, noisy labelers". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '08. New York, NY, USA: ACM, pp. 614–622. ISBN: 978-1-60558-193-4. DOI: 10.1145/1401890.1401965. URL: http://doi.acm.org.proxy2.library.illinois.edu/10.1145/1401890.1401965 (visited on 02/19/2012).

Shirky, C. (2009). *Here comes everybody*. Penguin Books.

Smucker, Mark D., Gabriella Kazai, and Matthew Lease (2012). *Overview of the trec 2012 crowdsourcing track*. DTIC Document. URL: http://

oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA578659 (visited on 08/25/2014).

Snow, R. et al. (2008). "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 254–263. URL: http://dl.acm.org.proxy2.library.illinois.edu/citation.cfm?id=1613715.1613751.

Song, Fei and W. Bruce Croft (1999). "A General Language Model for Information Retrieval". In: *Proceedings of the Eighth International Conference on Information and Knowledge Management*. CIKM '99. New York, NY, USA: ACM, pp. 316–321. ISBN: 1-58113-146-1. DOI: 10.1145/319950.320022. URL: http://doi.acm.org/10.1145/319950.320022 (visited on 07/24/2014).

Spiteri, Louise F. (2011). "Social discovery tools: Cataloguing meets user convenience". In: *Proceedings from North American Symposium on Knowledge Organization*. Vol. 3. URL: http://journals.lib.washington.edu/index.php/nasko/article/view/12790.

Springer, Michelle et al. (2008). "For the common good: The Library of Congress Flickr pilot project". In: URL: http://www.loc.gov/rr/print/flickr_report_final.pdf (visited on 08/26/2013).

Surowiecki, James (2004). *The Wisdom of Crowds*. Doubleday.

Taylor, Bret (2007). *FriendFeed Blog: I like it, I like it*. friendblog. URL: http://blog.friendfeed.com/2007/10/i-like-it-i-like-it.html (visited on 08/09/2014).

Thompson, Clive (2008). "If You Liked This, You're Sure to Love That". In: *The New York Times*. ISSN: 0362-4331. URL: http://www.nytimes.com/2008/11/23/magazine/23Netflix-t.html (visited on 08/13/2014).

Twain, Mark (1920). *The Adventures of Tom Sawyer*. Harper & brothers. 330 pp.

Vukovic, Maja and Claudio Bartolini (2010). "Towards a Research Agenda for Enterprise Crowdsourcing". In: *Leveraging Applications of Formal Methods, Verification, and Validation*. Ed. by Tiziana Margaria and Bernhard Steffen. Lecture Notes in Computer Science 6415. Springer Berlin Heidelberg, pp. 425–434. ISBN: 978-3-642-16557-3, 978-3-642-16558-0. URL:

http://link.springer.com.proxy2.library.illinois.edu/chapter/10.1007/978-3-642-16558-0_36 (visited on 01/30/2014).

Wales, Jimmy (2006). *Insist on Sources*. WikiEN-l. URL: http://lists.wikimedia.org/pipermail/wikien-l/2006-July/050773.html.

Wallace, B. et al. (2011). "Who should label what? Instance allocation in multiple expert active learning". In: *Proceedings of the SIAM International Conference on Data Mining (SDM)*.

Welinder, P. and P. Perona (2010). "Online crowdsourcing: Rating annotators and obtaining cost-effective labels". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 25–32. ISBN: 978-1-4244-7029-7. DOI: 10.1109/CVPRW.2010.5543189.

What is reCAPTCHA? (2008). Recaptcha. URL: http://recaptcha.net/learnmore.html (visited on 09/27/2008).

Whitehill, J. et al. (2009). "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise". In: *Advances in Neural Information Processing Systems* 22, pp. 2035–2043.

Wikipedia (2014). *Wikipedia:Size of Wikipedia*. In: *Wikipedia, the free encyclopedia*. Page Version ID: 615924147. URL: http://en.wikipedia.org/w/index.php?title=Wikipedia:Size_of_Wikipedia&oldid=615924147 (visited on 08/13/2014).

Yilmaz, Emine, Evangelos Kanoulas, and Javed A. Aslam (2008). "A Simple and Efficient Sampling Method for Estimating AP and NDCG". In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. New York, NY, USA: ACM, pp. 603–610. ISBN: 978-1-60558-164-4. DOI: 10.1145/1390334.1390437. URL: http://doi.acm.org/10.1145/1390334.1390437 (visited on 08/16/2014).

Zhai, Chengxiang and John Lafferty (2001). "A study of smoothing methods for language models applied to Ad Hoc information retrieval". In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '01. New York, NY, USA: ACM, pp. 334–342. ISBN: 1-58113-331-6. DOI: 10.1145/383952.384019. URL: http://doi.acm.org/10.1145/383952.384019 (visited on 09/20/2012).

Zhou, Ding et al. (2008). "Exploring Social Annotations for Information Retrieval". In: *Proceedings of the 17th International Conference on World Wide Web*. WWW '08. New York, NY, USA: ACM, pp. 715–724. ISBN: 978-1-60558-085-2. DOI: 10.1145/1367497.1367594. URL: http://doi.acm.org/10.1145/1367497.1367594 (visited on 08/25/2014).

Zwass, Vladamir (2010). "Co-Creation: Toward a Taxonomy and an Integrated Research Perspective". In: *International Journal of Electronic Commerce* 15.1, pp. 11–48. DOI: 10.2753/JEC1086-4415150101.