Inf Retrieval (2013) 16:91–100
DOI 10.1007/s10791-013-9222-7

Crowdsourcing for information retrieval: introduction to the special issue

Matthew Lease · Emine Yilmaz

Received: 4 March 2013 / Accepted: 4 March 2013 / Published online: 26 March 2013

Abstract This introduction to the special issue summarizes and contextualizes six novel research contributions at the intersection of information retrieval (IR) and *crowdsourcing* (also overlapping crowdsourcing's closely-related sibling, *human computation*). Several of the papers included in this special issue represent deeper investigations into research topics for which earlier stages of the authors' research were disseminated at crowdsourcing workshops at SIGIR and WSDM conferences, as well as at the NIST TREC conference. Since the first proposed use of crowdsourcing for IR in 2008, interest in this area has quickly accelerated and led to three workshops, an ongoing NIST TREC track, and a great variety of published papers, talks, and tutorials. We briefly summarize the area in order to help situate the contributions appearing in this special issue. We also discuss some broader current trends and issues in crowdsourcing which bear upon its use in IR and other fields.

Keywords Crowdsourcing · Human computation · Search evaluation

1 Introduction

The first computers were people (Grier 2005). Today, Internet-based access to 24/7 online human crowds has led to a renaissance of research in *human computation* (Quinn and Bederson 2011; Law and Ahn 2011) enabled by crowdsourcing (Howe 2006). These new opportunities have brought a disruptive shift to research and practice for how we build and evaluate intelligent systems today. Not only can labeled data for training and evaluation be

M. Lease (✉)
School of Information, University of Texas at Austin, Austin, TX, USA
e-mail: ml@ischool.utexas.edu

E. Yilmaz
Microsoft Research, Cambridge, UK
e-mail: eminey@microsoft.com

E. Yilmaz
Koc University, Istanbul, Turkey
e-mail: eyilmaz@ku.edu.tr

collected faster, cheaper, and easier than ever before, but we now see human computation being integrated into the systems themselves, operating in concert with automation to create a new class of hybrid applications [e.g., validating search results in near-real time (Yan et al. 2010; McCreadie et al. 2012)]. New systems for collective intelligence and wisdom of crowds (Surowiecki 2005) are helping us to engage and aggregate information from groups of people more effectively, as well as with automatic methods. information retrieval (IR) is one of many Information Science sub-fields that has been rapidly changing in recent years as a consequence.

Crowdsourcing's rapid development continues to both offer new ideas and challenges to our traditional methods for designing, training, and evaluating IR systems. Beginning with the first use of crowdsourcing for IR evaluation (Alonso et al. 2008), research to date has predominantly focused on developing strategies for maximizing data quality while reducing the time, cost, and effort required for annotation, evaluation, and other manual tasks which underlie and support automated IR systems. Consider, for example, the well-established Cranfield paradigm for evaluating IR systems (Cleverdon 1997), which depends on human judges manually assessing documents for topical relevance. Although recent advances in stochastic evaluation algorithms have greatly reduced the number of such assessments needed for reliable evaluation (Aslam et al. 2006; Carterette et al. 2006; Yilmaz et al. 2008), assessment itself remains expensive and slow. Calling upon this distributed, on-demand workforce in place of in-house annotators offers one potential avenue for gaining additional traction on this problem and helping to ensure the continuing scalability of IR evaluation practices to support tomorrow's ever-larger information collections.

Following this Introduction, we first describe synergistic activities at the intersection of IR and crowdsourcing (i.e., workshops, tutorial, and TREC Crowdsourcing Track). We then briefly summarize the six research papers accepted to appear in the final issue.

Prior to concluding, a section on Broader Trends and Issues in Crowdsourcing provides broader surrounding context of current trends in crowdsourcing and ramifications for IR research. Many of these issues are discussed in greater detail in a recent 2013 survey paper on "The Future of Crowd Work" (Kittur et al. 2013), to which we refer the interested reader.

2 Synergistic activities

While IR studies using crowdsourcing have been quite encouraging, many questions remain as to how crowdsourcing methods can be most effectively and efficiently employed in practice. Such open questions motivated the organization of three crowdsourcing workshops at top IR conferences to foster community and expertise-sharing in this emerging area, as well as promote early research:

1. SIGIR 2010: *Crowdsourcing for Search Evaluation* (CSE 2010)[1] (Carvalho et al. 2010; Lease et al. 2010)
2. WSDM 2011: *Crowdsourcing for Search and Data Mining* (CSDM 2011)[2] (Lease et al. 2011a, b)
3. SIGIR 2011: *Crowdsourcing for Information Retrieval* (CIR 2011)[3] (Lease and Yilmaz 2011b; Lease et al. 2011c)

[1] ir.ischool.utexas.edu/cse2010/program.htm.

[2] ir.ischool.utexas.edu/csdm2011/proceedings.html.

[3] sites.google.com/site/cir2011ws/proceedings.

Following Omar Alonso's *Crowdsourcing for Relevance Evaluation* tutorial at ECIR 2010 (Alonso 2010), Alonso and Lease offered updated tutorials at WSDM 2011 (Alonso and Lease 2011a), SIGIR 2011 (Alonso and Lease 2011b), and SIGIR 2012 (Lease and Alonso 2012). Slides from all three tutorials are available online (see References).

Since its inception in 2011, the NIST TREC Crowdsourcing Track is now in its third year[4] (Lease and Kazai 2011a; Smucker et al. 2013). Following in the long tradition of TREC, this track offers a shared task which challenges participants to innovate how IR evaluation can be accomplished at scale with crowds, preserving quality of relevance assessments for reliable evaluation of IR systems, while at the same time lowering cost, time, and effort traditionally required by assessment. The track invites participation from any and all interested groups.

Industrial sponsors As models of crowdsourcing based on *crowd work* typically involve offering of financial incentives, several commercial crowdsourcing service providers have offered free use of their services for researchers participating in the above workshops and the ongoing TREC Crowdsourcing Track. We greatly appreciate the generosity of Amazon (Mechanical Turk)[5] (Chen et al. 2011), CrowdFlower (formerly Dolores Labs)[6] (Snow et al. 2008; Le et al. 2010; Oleson et al. 2011), and MobileWorks[7] (Narula et al. 2011; Kulkarni et al. 2012) for their continuing sponsorship. We also thank Microsoft for their sponsorship of best paper awards at the three workshops.

3 Papers accepted to the special issue

This special issue includes six excellent research papers investigating topics at the intersection of IR and crowdsourcing. We briefly summarize these papers below.

In his paper entitled, "Implementing crowdsourcing-based relevance experimentation: an industrial perspective" (Alonso 2013), Omar Alonso emphasizes the importance of approaching quality assurance as an end-to-end process which requires all of the following: clear instructions, a well-designed user interface, content quality, inter-rater agreement metrics, and worker feedback analysis. Moreover, Alonso stresses that designing and implementing experiments that require thousands or millions of labels is fundamentally different than conducting small scale experiments, and enabling a framework for continuous crowdsourcing experiments requires even more rigorous design. Alonso's paper builds upon his earlier CSDM 2011 workshop paper (Alonso 2011), as well as a long line of related work (Alonso et al. 2008; Alonso and Mizzaro 2009, 2012). Using examples based on TREC, INEX, and Wikipedia data sets, Alonso illustrates many traits of successful crowdsourcing experiments impacting quality of final results, with an invaluable industrial perspective complementing this special issue's other papers by academic researchers.

Carsten Eickhoff and Arjen P. de Vries investigate a less-traveled road to quality assurance in their paper, "Increasing cheat robustness of crowdsourcing tasks" (Eickhoff and de Vries 2013). Their work builds on earlier work at CIR and CSDM workshops (Eickhoff and de Vries 2011; Vliegendhart et al. 2011; Vuurens et al. 2011) and is related

[4] sites.google.com/site/treccrowd.

[5] http://www.mturk.com.

[6] http://www.crowdflower.com.

[7] http://www.mobileworks.com.

to a recent journal paper in an IEEE Computing Special Issue on Crowdsourcing (Vuurens and de Vries 2012). They investigate the following four research questions:

1. How does the concrete task type influence the number of observed cheaters?
2. Does interface design affect the share of cheaters?
3. Can we reduce fraudulent tendencies by explicitly filtering the crowd?
4. Is there a connection between the size of HIT batches and observed cheater rates?

Along the way, they identify three "dysfunctional worker types" and two common "cheating strategies", and they discuss limitations of relying on MTurk acceptance rates for filtering out such problems. Interestingly, they flag problematic workers per task, rather than per-batch or globally, arguing that work on other tasks may be of higher quality, and so each task should be assessed on its own merits. Moreover, they follow Kittur et al. (2008) in suggesting traditional principles of optimizing interface design may not yield the best results with a diverse crowd. In particular, they suggest that poor work might be best prevented by careful HIT design rather than filtering our poor quality results after the fact, arguing that complex and creative tasks attract fewer cheaters. In distinguishing financially-motivated versus entertainment-driven workers, they connect to their currently ongoing work developing crowdsourcing games for collecting relevance judgments (Eickhoff et al. 2011, 2012).

In "An analysis of human factors and label accuracy in crowdsourcing relevance judgments" (Kazai et al. 2013), Gabriella Kazai, Jaap Kamps and Natasa Milic-Frayling investigate the following two research questions:

1. How do different conditions of pay, required effort, and selection of workers based on proven reliability affect the quality of the crowdsourced relevance labels?
2. How do various human factors, such as motivation, expertise, level of interest, perceived task difficulty and satisfaction with the offered pay, that characterize a group of workers under a specific task condition, relate to the resulting label quality?

This paper builds on a variety of related work by the authors (Kazai 2011; Kazai et al. 2011a, b, 2012). In this work, workers were asked to perform relevance judging of book pages taken from the INEX Book Search task, as well as complete a questionnaire which inquired regarding their primary motivation, topic familiarity, interest in the task, and satisfaction with offered pay. Workers were also asked to rate each task's difficulty. This data allows Kazai et al. to correlate task performance with each questionnaire response type, yielding a variety of interesting findings. For example, those working for fun tend to under-perform those motivated by financial gain, those who claim expertise tend to perform worse than those who do not, and those who are better paid rate task difficulty as easier. The authors main general conclusion is that varied "task conditions do indeed attract a different crowd, and these differences are affecting the quality of the work."

In their paper entitled, "Identifying top news using crowdsourcing" (McCreadie et al. 2013), Richard McCreadie, Craig Macdonald and Iadh Ounis report on use of crowdsourcing in the TREC Blog Track (Ounis and Soboroff 2011). This paper builds on the authors' earlier work at the CSE 2010 and CSDM 2011 workshops (Richard et al. 2010; McCreadie et al. 2011). Their aim is "to determine how feasible and effective crowdsourcing is at relevance assessment for a modern TREC task." The Blog Track's two tasks involved two types of assessment activities: (1) labeling newsworthiness and news category; and (2) assessing topical relevance as well as various facets—factual versus opinionated, sentiment, succinctness versus depth, and predictions versus aftermath. Interestingly, instead of inserting some percentage of questions with known answers into the work queue in order to assess quality, the authors instead sample a percentage of submitted judgments and manually verify them, using color coding to accelerate the verification step. The authors also show, through post-hoc sampling, that by collecting a third relevance judgment only upon disagreement of the first two, only 20 % of the judgments require such a tie-breaker, enabling greater savings. Of further interest, a majority of their judgments were collected from only three workers, suggesting their tasks achieved high retention value for maintaining consistency of judgments. They conclude their paper with a brief list of recommended best practices.

In striking contrast with other papers in this special issue, Robert Munro's investigation of crowdsourcing techniques to improve support for humanitarian crises provides a compelling account of crowdsourcing's impact when applied beyond the confines of traditional IR tasks. In his paper entitled, "Crowdsourcing and the crisis-affected community: Lessons learned and looking forward from Mission 4636" (Munro 2013), Munro reports on the findings of Mission 4636, a real-time humanitarian crowdsourcing initiative that processed 80,000 text messages (SMS) sent from within Haiti following the 2010 earthquake. Contrary to all previous papers, studies, and media reports about Mission 4636, which have typically chosen to exclude empirical analyses and the involvement of the Haitian population, Munro reports that the greatest volume, speed and accuracy in information processing was actually achieved by Haitian nationals, the Haitian diaspora, and those working closest with them. Moreover, no new technologies were found to play a significant role. Consequently, Munro recommends that future humanitarian deployments of crowdsourcing focus on information processing specifically within the populations they serve, utilizing crowdsourcing to engage those possessing crucial local knowledge, wherever those people happen to be in the world.

Last but not least, in their paper "Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems" (Zuccon et al. 2013), authors Guido Zuccon, Teerapong Leelanupab, Stewart Whiting, Emine Yilmaz, Joemon M. Jose, and Leif Azzopardi propose an experimental methodology that can be an alternative to laboratory-based user studies in information retrieval experiments. This paper builds on earlier research presented at the CSDM 2011 workshop (Zuccon et al. 2011). The methodology they propose uses a crowdsourcing platform as a means of engaging study participants. They show that their crowdsourcing based approach can capture user interactions and searching behaviors at a lower cost, with more data, and within a shorter period than traditional laboratory-based user studies. They also show the characteristic differences of the crowdsourcing based approach with respect to traditional experimental and evaluation procedures, comparing crowdsourcing-based evaluation with laboratory-based evaluation of information retrieval systems.

4 Broader trends and issues in crowdsourcing

Crowdsourcing continues to rapidly evolve, with increasing impact on both industrial and research practice. For many, crowdsourcing may be interesting only so far as it enables them to better advance their research programs and gain new traction on old problems. For others, crowdsourcing has become a phenomenon to study in its own right, with interdisciplinary issues spanning engineering, psychology, sociology, economics, policy, and ethics (at least).

One of the editors of this Special Issue maintains a simple Webpage which tracks related academic research activities, such as conferences, workshops, journals, etc.[8] The

[8] ir.ischool.utexas.edu/crowd.

CHI community maintains an active shared blog which provides a focal point for promoting community and dissemination of related crowdsourcing research and activities.[9] In the KDD and AAAI areas, 4 years of a Human Computation workshop are now giving rise to a new AAAI Conference on Human Computation.[10] This year will also mark the fourth industrial CrowdConf Conference.[11] A broader Collective Intelligence conference held in 2012[12] may occur again. A variety of Special Issues have been published or are forthcoming.

Crowdsourcing and human computation can often seem like magic. We invoke a computer function like we always have, but now that function suddenly produces output outperforming state-of-the-art AI. This sense of magic stems somewhat from our general use of such black box abstractions, which let us focus our attention on integrating heterogeneous software modules while encapsulating details of the modules internal characteristics. With crowdsourcing, however, there is a very real danger here: we may forget there are real people behind the abstraction, we may impact their lives in ways that do not penetrate the abstraction, and we may not realize or fully appreciate those impacts we are having. Terminology such as "Human Computation", "Human Processing Units (HPUs)" (Davis et al. 2010), and "Remote Person Calls (RPCs)" offer conceptually useful (and amusing) ways for us to think about the computation of crowdsourcing, but the same terminology also helps perpetuate the invisibility of a global workforce that is by its very distributed nature difficult to put a face on. To help us see these people in a way that statistics on crowd demographics do not seem to make as visceral to us, Andy Baio created a collage of worker faces (Baio 2008). Leila Chirayath Janah, who founded the non-profit SamaSource platform,[13] regularly gives talks showing people living in African refugee camps performing online crowd work as one of their only opportunities to earn income and exert some measure of control in otherwise chaotic living environments. As a form of design activism, Lilly Irani and collaborators built Turkopticon, combatting the invisibility of crowd workers and raising collective awareness of worker concerns (Silberman et al. 2010; Irani and Silberman 2013). Despite such efforts, we still know relatively little about the lives and working conditions of many of the people powering today's crowdsourcing applications.

As seen with earlier outsourcing, global market forces are increasingly moving computer work to regions of the world where it can be completed more quickly and affordably. Crowd work savings arise from increase in labor supply, lower cost of living in other geographic regions, and the ability to decompose work into very fine-granularity units which can be efficiently and affordably distributed. While early demographic studies suggested that crowd work was typically performed for supplemental rather than primary income, subsequent studies have indicated crowd work is increasingly become a source of primary income, especially in developing economies (Ross et al. 2010; Ipeirotis 2010). Relatively low wages, depersonalized work, and asymmetric power relationships have led some conscientious researchers to express concern that we may be building a future of crowd-powered computing on the backs of exploited workers in digital sweat shops (Mitchell 2010). At the same time, crowdsourcing is conversely being seen as "The New

[9] crowdresearch.org/blog.

[10] http://www.humancomputation.com/2013.

[11] http://www.crowdconf.com.

[12] http://www.ci2012.org.

[13] samasource.org.

Sewing Machine" (Paritosh et al. 2011), creating new opportunities for income and social mobility in regions of the world where local economies are stagnant and local governmental structures may preclude traditional outsourcing. Just as consumers can choose to buy fair trade goods or invest in social choice funds, some crowdsourcing services now promise worker protections and living wages (SamaSource, MobileWorks, and CloudFactory[14]).

It would be valuable for our community to reflect upon and discuss ethical questions surrounding this crowdsourcing enterprise we are engaged in, to be intentional about our activities and their consequences. Of course, such questions are not easy. For example, is it really better to pay nothing at all (i.e., crowdsourcing games) than to pay people something, even if it is low (Fort et al. 2011). If we want to build search engines which automatically filter out search results containing adult or violent content, and our state-of-the-art AI methods prove insufficient, what are the implications of asking crowd workers to filter such content (Harmanci 2012)?

There are tremendously exciting applications of crowdsourcing emerging with great potential to transform and advance our society for the betterment of all. We scientists can play a significant, positive role in helping shape this future. Clearly there are many fascinating technical challenges and opportunities to explore, but we also have a responsibility to remain equally aware and vigilant about mitigating any potentially hidden costs of our "advances".

While such a detailed discussion of socio-technical issues may seem unnecessary to some readers for a Special Issue focused on the utility of crowdsourcing, it is precisely this exclusive focus on utility that should give us pause (a focus that widespread across technical research on crowdsourcing today). Our impact extends beyond our technical contributions, and we are affecting the real human lives powering our human computation systems today.

5 Conclusion

It has been a tremendous honor to read all of the papers submitted for this Special Issue, as well as to interact with the authors involved and see their works further improve into the six papers appearing here. Overall, it is an exciting time of change and growth in the crowdsourcing space at large, and we are witnessing increasing progress and creativity in related applications to IR. There are a variety of ongoing opportunities for the interested reader to participate in related workshops and conferences, as well as in the ongoing TREC Crowdsourcing Track. For those who have primarily followed the technical aspects of crowdsourcing research, the last section on Broader Trends and Issues in Crowdsourcing is intended to provide a valuable holistic, contextual grounding of crowdsourcing as a social-technical enterprise, one that is ultimately powered by a world of people just like us.

Acknowledgments We thank the authors who contributed the papers to make this special issue possible, as well as the global crowd of individual around the world who have contributed and continue to contribute to crowdsourcing projects, such as those described here. Thank you for making crowdsourcing possible. Matthew Lease was supported in part by an NSF CAREER award, a DARPA Young Faculty Award N66001-12-1-4256, a John P. Commons Fellowship, and a Temple Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors alone and do not express the views of any of the funding agencies.

[14] cloudfactory.com.

References

Alonso, O. (2010). Tutorial: Crowdsourcing for relevance evaluation. In *Proceedings of the 32nd European conference on IR research (ECIR)*. Slides available online at http://ir.ischool.utexas.edu/cse2010/materials/alonso-ecir2010-tutorial.pdf.

Alonso, O. (2011). Perspectives on infrastructure for crowdsourcing. In M. Lease, V. Carvalho & E. Yilmaz (Eds.), *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. 7–10.

Alonso, O. (2013). Implementing crowdsourcing-based relevance experimentation: An industrial perspective. *Information Retrieval, 16*(2). doi:10.1007/s10791-012-9204-1.

Alonso, O., & Lease, M. (2011a). Crowdsourcing 101: Putting the WSDM of crowds to work for you. In *Proceedings of the fourth ACM international conference on web search and data mining (WSDM)*, pp. 1–2. Slides available online at http://ir.ischool.utexas.edu/wsdm2011_tutorial.pdf.

Alonso, O., & Lease, M. (2011b). Crowdsourcing for information retrieval: Principles, methods, and applications. In *Tutorial at the 34th annual ACM SIGIR conference*, p. 1299, Beijing, China. Slides available online at http://www.slideshare.net/mattlease/crowdsourcing-for-information-retrieval-principles-methods-and-applications.

Alonso, O., & Mizzaro, S. (2009). Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR workshop on the future of IR evaluation*, pp. 15–16.

Alonso, O., & Mizzaro, S. (2012). Using crowdsourcing for trec relevance assessment. *Information Processing & Management.*

Alonso, O., Rose, D. E., & Stewart, B. (2008). Crowdsourcing for relevance evaluation. *ACM SIGIR Forum, 42*(2), 9–15.

Aslam, J. A., Pavlu, V., & Yilmaz, E. (2006). A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 541–548.

Baio, A. (2008). The faces of mechanical turk. November 20. waxy.org/2008/11/the_faces_of_mechanical_turk.

Carterette, B., Allan, J., & Sitaraman, R. (2006). Minimal test collections for retrieval evaluation. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, pp. 268–275.

Carvalho, V., Lease, M., & Yilmaz, E. (2010). Crowdsourcing for search evaluation. *ACM SIGIR Forum, 44*(2), 17–22.

Chen, J. J., Menezes, N. J., Bradley, A. D., & North, T. (2011). Opportunities for crowdsourcing research on amazon mechanical turk. In *CHI workshop on crowdsourcing and human computation*.

Cleverdon, C. (1997). The cranfield tests on index language devices. *Readings in Information Retrieval*, 47–59.

Davis, J., Arderiu, J., Lin, H., Nevins, Z., Schuon, S., Gallo, O., et al. (2010). The HPU. In *Computer vision and pattern recognition workshops (CVPRW)*, pp. 9–16.

Eickhoff, C., Harris, C. G., Srinivasan, P., & de Vries, A. P. (2011). GEAnn—games for engaging annotations. In M. Lease, V. Hester, A. Sorokin & E. Yilmaz (Eds.), *Proceedings of the ACM SIGIR 2011 workshop on crowdsourcing for information retrieval (CIR 2011)*, p. 63, Beijing, China.

Eickhoff, C., Harris, C. G., de Vries, A. P., & Srinivasan, P. (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*, pp. 871–880. ACM.

Eickhoff, C., & de Vries, A. (2011). How crowdsourcable is your task? In M. Lease, V. Carvalho, & E. Yilmaz (Eds.), 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. 11–14, Hong Kong, China. Received Most Innovative Paper Award.

Eickhoff, C., & de Vries, A. P. (2013). Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval, 16*(2). doi:10.1007/s10791-011-9181-9.

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics, 37*(2), 413–420.

Grier, D. A. (2005). *When computers were human*, Vol. 316. Princeton: Princeton University Press.

Harmanci, R. (2012). *The googler who looked at the worst of the internet*. BuzzFeed August 21. http://www.buzzfeed.com/reyhan/tech-confessional-the-googler-who-looks-at-the-wo.

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine 14*(6), 1–4.

Ipeirotis, P. (2010). Demographics of mechanical turk. Tech. Rep. CeDER-10-01, New York University.

Irani, L., & Silberman, M. (2013). Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceeding of the ACM SIGCHI conference on human factors in computing systems*.

Kazai, G. (2011). In search of quality in crowdsourcing for search engine evaluation. In *European conference on information retrieval (ECIR)*, pp. 165–176.

Kazai, G., Kamps, J., Koolen, M., & Milic-Frayling, N. (2011a). Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *SIGIR*, pp. 205–214.

Kazai, G., Kamps, J., & Milic-Frayling, N. (2011b). Worker types and personality traits in crowdsourcing relevance labels. In *CIKM*, pp. 1941–1944.

Kazai, G., Kamps, J., & Milic-Frayling, N. (2012). The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *CIKM*, pp. 2583–2586.

Kazai, G., Kamps, J., & Milic-Frayling, N. (2013). An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval, 16*(2). doi:10.1007/s10791-012-9205-0.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the ACM annual SIGCHI conference on Human factors in computing systems*, pp. 453–456.

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., et al. (2013). The future of crowd work. In *Proceedings of the ACM conference on computer supported cooperative work (CSCW)*, pp. 1301–1318.

Kulkarni, A., Gutheim, P., Narula, P., Rolnitzky, D., Parikh, T., & Hartmann, B. (2012). Mobileworks: Designing for quality in a managed crowdsourcin architecture. *IEEE Internet Computing, 16*(5), 28.

Law, E., & Ahn, L. (2011). Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 5*(3), 1–121.

Le, J., Edmonds, A., Hester, V., & Biewald, L. (2010). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, pp. 21–26.

Lease, M., & Alonso, O. (2012). Crowdsourcing for search evaluation and social-algorithmic search. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval, SIGIR '12*. Slides available online at: http://www.slideshare.net/mattlease/crowdsourcing-for-search-evaluation-and-socialalgorithmic-search.

Lease, M., Carvalho, V., & Yilmaz, E. (eds.). (2010). *Proceedings of the ACM SIGIR 2010 workshop on crowdsourcing for search evaluation (CSE 2010)*. Geneva, Switzerland. Available online at http://ir.ischool.utexas.edu/cse2010.

Lease, M., Carvalho, V., & Yilmaz, E. (2011a). Crowdsourcing for search and data mining. *ACM SIGIR Forum, 45*(1), 18–24.

Lease, M., Carvalho, V., & Yilmaz, E. (eds.). (2011b). *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)*, Hong Kong, China. Available online at ir.ischool.utexas.edu/csdm2011/proceedings.html.

Lease, M., & Kazai, G. (2011a). Overview of the TREC 2011 crowdsourcing track (conference notebook). In *20th text retrieval conference (TREC)*.

Lease, M., & Yilmaz, E. (2011b). Crowdsourcing for information retrieval. *ACM SIGIR Forum, 45*(2), 66–75.

Lease, M., Yilmaz, E., Sorokin, A., & Hester, V. (eds.). (2011c). *Proceedings of the 2nd ACM SIGIR 2011 workshop on crowdsourcing for information retrieval (CIR)*. Beijing, China. Available online at http://sites.google.com/site/cir2011ws/proceedings.

McCreadie, R., Macdonald, C., & Ounis, I. (2011). Crowdsourcing blog track top news judgments at TREC. In M. Lease, V. Carvalho & E. Yilmaz (Eds.), *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, Hong Kong, China.

McCreadie, R., Macdonald, C., & Ounis, I. (2012). Crowdterrier: Automatic crowdsourced relevance assessments with terrier. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 1005–1005. ACM.

McCreadie, R., Macdonald, C., & Ounis, I. (2013). Identifying top news using crowdsourcing. *Information Retrieval, 16*(2). doi:10.1007/s10791-012-9186-z.

Mitchell, S. (2010). Inside the online sweatshops. PC Pro Magazine. August 6.http://www.pcpro.co.uk/features/360127/inside-the-online-sweatshops.

Munro, R. (2013). Crowdsourcing and the crisis-affected community lessons learned and looking forward from mission 4636. *Information Retrieval, 16*(2). doi:10.1007/s10791-012-9203-2.

Narula, P., Gutheim, P., Rolnitzky, D., Kulkarni, A., & Hartmann, B. (2011). Mobileworks: A mobile crowdsourcing platform for workers at the bottom of the pyramid. In *AAAI human computation workshop*.

Oleson, D., Sorokin, A., Laughlin, G., Hester, V., Le, J., & Biewald, L. (2011). Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *AAAI workshop on human computation*.

Ounis, I., & Soboroff, I. (2011). Overview of the TREC 2010 blog track. In *Proceedings of the 19th Text REtrieval Conference (TREC)*.

Paritosh, P., Ipeirotis, P., Cooper, M., & Suri, S. (2011). The computer is the new sewing machine: Benefits and perils of crowdsourcing. In *Proceedings of the 20th international conference companion on world wide web*, pp. 325–326. ACM.

Quinn, A. J., & Bederson, B. B. (2011). Human computation: A survey and taxonomy of a growing field. In *2011 Annual ACM SIGCHI conference on Human factors in computing systems*, pp. 1403–1412.

Richard, M. C., McCreadie, C. M., & Ounis, I. (2010). Crowdsourcing a news query classification dataset. In M. Lease, V. Carvalho & E. Yilmaz (Eds.), *Proceedings of the ACM SIGIR 2010 workshop on crowdsourcing for search evaluation (CSE 2010)*, pp. 31–38. Geneva, Switzerland.

Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers?: Shifting demographics in mechanical turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, pp. 2863–2872. ACM.

Silberman, M., Irani, L., & Ross, J. (2010). Ethics and tactics of professional crowdwork. *XRDS: Crossroads, The ACM Magazine for Students, 17*(2), 39–43.

Smucker, M., Kazai, G., & Lease, M. (2013). Overview of the TREC 2012 crowdsourcing Track. In *Proceedings of the 21st NIST text retrieval conference (TREC)*.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A.Y. (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, pp. 254–263. Association for Computational Linguistics.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Vliegendhart, R., Larson, M., Kofler, C., Eickhoff, C., & Pouwelse, J. (2011). Investigating factors influencing crowdsourcing tasks with high imaginative load. In M. Lease, V. Carvalho & E. Yilmaz (Eds.), *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. 27–30.

Vuurens, J., Vries, A. P. D., & Eickhoff, C. (2011). How much spam can you take? An analysis of crowdsourcing results to increase accuracy. In M. Lease, V. Hester, A. Sorokin & E. Yilmaz (Eds.), *Proceedings of the ACM SIGIR 2011 workshop on crowdsourcing for information retrieval (CIR 2011)*, pp. 48–55.

Vuurens, J.B., & de Vries, A. P. (2012). Obtaining high-quality relevance judgments using crowdsourcing. *IEEE Internet Computing 16*(5), 20–27. http://doi.ieeecomputersociety.org/10.1109/MIC.2012.71.

Yan, T., Kumar, V., & Ganesan, D. (2010). CrowdSearch: Exploiting crowds for accurate real-time image search on mobile phones. In *Proceedings of the 8th international conference on mobile systems, applications, and services (MOBISYS)*, pp. 77–90. ACM.

Yilmaz, E., Kanoulas, E., & Aslam, J. A. (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*, pp. 603–610.

Zuccon, G., Leelanupab, T., Whiting, S., Jose, J., & Azzopardi, L. (2011). Crowdsourcing interactions—a proposal for capturing user interactions through crowdsourcing. In M. Lease, V. Carvalho & E. Yilmaz (Eds.), *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. 35–38.

Zuccon, G., Leelanupab, T., Whiting, S., Yilmaz, E., Jose, J. M., & Azzopardi, L. (2013). Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information Retrieval, 16*(2). doi:10.1007/s10791-012-9206-z.