Crowdsourcing for Search Evaluation

Vitor R. Carvalho
Microsoft Corporation
Redmond, WA
vitor@microsoft.com

Matthew Lease
School of Information
University of Texas at Austin
Austin, TX
ml@ischool.utexas.edu

Emine Yilmaz
Microsoft Research Cambridge
Cambridge, UK
eminey@microsoft.com

Abstract

The Crowdsourcing for Search Evaluation Workshop (CSE 2010) was held on July 23, 2010 in Geneva, Switzerland, in conjunction with the 33rd Annual ACM SIGIR Conference[1]. The workshop addressed the latest advances in theory and empirical methods in crowdsourcing for search evaluation, as well as novel applications of crowdsourcing for evaluating search systems. Three invited talks were presented, along with seven refereed papers. Proceedings from the workshop, along with presentation slides, have been made available online[2].

1 Introduction

While automated Information Retrieval (IR) technologies have enabled people to quickly and easily find desired information, development of these technologies has historically depended on slow, tedious, and expensive data annotation. For example, the Cranfield paradigm for evaluating IR systems [5] 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 [3, 4, 11], assessment nonetheless remains an expensive and slow process.

Crowdsourcing represents a promising new avenue for reducing effort, time, and cost involved in evaluating search systems. The key idea of crowdsourcing is to tap into the vast and growing, distributed workforce available online today due to world-wide growth in internet connectivity and the development of online marketplaces, platforms, and services supporting this new form of online labor. With regard to search evaluation, rather than employing in-house annotators for relevance assessment, one can instead leverage the "wisdom of the crowd" via this distributed workforce. While crowdsourcing poses a variety of new challenges

[1]The organizers thank Microsoft and CrowdFlower for their generous sponsorship of the workshop.
[2]`http://ir.ischool.utexas.edu/cse2010/program.htm`

in interacting with workers and ensuring standards for quality control, a variety studies have shown that the crowd in aggregate can produce superior annotations to in-house assessors in less time and at significantly lower cost (cf. [2]). Crowdsourcing also facilitates intriguing new opportunities to leverage workforce diversity, geographic dispersion, and near real-time response of this on-demand and under-utilized online workforce.

While search evaluation studies using crowdsourcing have been quite encouraging, many questions remain as to how crowdsourcing methods can be most effectively and efficiently employed in practice. The Crowdsourcing for Search Evaluation Workshop reported on advances in the state-of-the-art in using crowdsourcing for evaluating information retrieval systems. The workshop was well-attended with enthusiastic discussion by participants continuing well into the evening beyond the day's scheduled activities. Building on the success of this workshop, a follow-on one-day workshop on Crowdsourcing for Search and Data Mining (CSDM 2011)[3] will be held on February 9, 2011 in Hong Kong, in conjunction with the 4th Annual ACM WSDM Conference.

In the remainder of this report, we first present the workshop program committee and program. We then summarize the invited keynote talks and the accepted research papers.

2 Program Committee

Eugene Agichtein, Emory University
Ben Carterette, University of Delaware
Charlie Clarke,University of Waterloo
Gareth Jones, Dublin City University
Michael Kaisser. University of Edinburgh
Jaap Kamps, University of Amsterdam
Gabriella Kazai, Microsoft Research
Mounia Lalmas, University of Glasgow
Winter Mason, Yahoo! Research
Don Metzler, University of Southern California
Stefano Mizzaro, University of Udine
Gheorghe Muresan, Microsoft Bing
Iadh Ounis, University of Glasgow
Mark Sanderson, University of Sheffield
Mark Smucker, University of Waterloo
Siddharth Suri, Yahoo! Research
Fang Xu, Saarland University

3 Workshop Program

The workshop program included three invited talks and seven refereed paper presentations.

Invited Talks

- *Design of experiments for crowdsourcing search evaluation: challenges and opportunities* Omar Alonso, Microsoft Bing

[3] `http://ir.ischool.utexas.edu/csdm2011`

- *Insights into Mechanical Turk*
  Adam Bradley, Amazon
- *Better Crowdsourcing through Automated Methods for Quality Control*
  Lukas Biewald, CrowdFlower

Accepted Papers, in Order of Presentation

- *Crowdsourcing for Affective Annotation of Video: Development of a Viewer-reported Boredom Corpus* (Runner-up: **Most Innovative Paper Award**)
  Mohammad Soleymani and Martha Larson
- *Crowdsourcing Preference Judgments for Evaluation of Music Similarity Tasks*
  (Winner: **Most Innovative Paper Award**)
  Julian Urbano, Jorge Morato, Monica Marrero and Diego Martin
- *Ensuring quality in crowdsourced search relevance evaluation*
  John Le, Andy Edmonds, Vaughn Hester and Lukas Biewald
- *An Analysis of Assessor Behavior in Crowdsourced Preference Judgments*
  Dongqing Zhu and Ben Carterette
- *Logging the Search Self-Efficacy of Amazon Mechanical Turkers*
  Henry Feild, Rosie Jones, Robert C. Miller, Rajeev Nayak, Elizabeth F. Churchill and Emre Velipasaoglu
- *Crowdsourcing a News Query Classification Dataset* Richard M. C. McCreadie, Craig Macdonald and Iadh Ounis
- *Detecting Uninteresting Content in Text Streams*
  Omar Alonso, Chad Carson, David Gerster, Xiang Ji, Shubha U. Nabar

3.1 Invited Talks

The first invited talk by Omar Alonso discussed effective design of experiments for crowdsourcing search evaluation. Alonso presented a detailed workflow of employing crowdsourcing in evaluation starting with defining the evaluation exercise and collecting the data to be used and testing their quality. Designing the experiment was the next crucial step described which requires a careful user experience (UX) design, careful selection of workers, policy and process for awarding payment and determining its amount, and testing and analyzing the design itself. Iteration over this process was recommended to produce the best possible design of the experiment. Great tips regarding running the actual experiment were also given, with the major point made being the need for splitting the experiment into batches and improving the experiment over the next batch on the basis of the results of the current batch. Quality control was a major point of Alsonso's presentation, along with the benefits of allowing users to give any feedback regarding the justification of their decisions. Alonso concluded his talk laying a number of challenges and opportunities for improving crowdsourcing. He pointed out, for instance, the strengths and limitations of the current crowdsourcing platforms, the need for data analysis tools, browsing and searching features and integration with the databases technology. The talk offered the attendees a variety of practical advice for increasing the quality of crowdsourcing based on Alonso's extensive experience.

The second invited speaker of the day was Adam Bradley, lead architect for Amazon's Mechanical Turk (MTurk) marketplace[4] and responsible for its scalability, operational excellence, and marketplace quality. Bradley began by presenting MTurk's service from an insider's perspective, providing an overview of how the system works and revealing unique statistics on the its utilization by various demographics. Bradley then discussed some of the recently deployed features in MTurk and how these changes were received by the users and MTurk workers. Bradley then concluded by opening the microphones for direct feature requests, feedback, complaints and ideas for the next releases of MTurk – an effort to bring researchers and users of the system closer to its developers.

Lukas Biewald, founder and CEO of CrowdFlower[5], gave the final invited talk of the day. Biewald's presentation spanned useful applications of crowdsourcing (such as using crowdsourcing to translate text messages during the Haiti crisis), to the characteristics of different crowdsourcing workforces, and a variety of different platforms that could be used for crowdsourcing (e.g. iPhone applications). Biewald presented an extensive break down of MTurk workers on the basis of country/continent – with India making up 46.85% increasing its percentage compared to past survey figures and US making up 42.7% – gender, age, education level, household income level, motivation – with money being the most significant motivation for everyone but money aside, people from India are there to learn and people from the US are there to have fun – working hours, number of tasks workers undertake and task completion time – showing a large diversity across workers – and so on. Regarding the crowdsourcing worker's motivation, Biewald pointed out an increasing workforce funded by virtual currency from online social games. A break down of that workforce on the same basis of that of MTurk identified important differences between the two pools of workers. This workforce also showed many similarities to Amazon's MTurk workforce from two years ago. Biewald further discussed the accuracy of the data generated by crowdsourcing showing a number of examples of inter-worker agreement and how it increases with more judgments.

3.2 Refereed Papers

The CSE 2010 Program Committee accepted seven papers for presentation at the workshop.

The first paper presented was "*Crowdsourcing for Affective Annotation of Video: Development of a Viewer-reported Boredom Corpus*", by Soleymani and Larson [9]. This paper was the runner-up for the **Most Innovative Paper Award**, carrying a $100 cash prize sponsored Microsoft Bing. Soleymani and Larson [9] focused on utilizing crowdsourcing for the problem of predictions of viewer affective response to video in order to enhance the performance of multimedia retrieval and recommendation systems. They reported on the development of a new corpus to be used to evaluate algorithms for prediction of viewer-reported boredom. When preparing the corpus, they made use of crowdsourcing in order to address two shortcomings of previous affective video corpora: small number of annotators and gap between annotators and target viewer group. The authors described the design of the MTurk setup that was used to generate the affective annotations for the corpus, as well as some specific issues that arose and how they were resolved. They also presented an analysis of the annotations collected and a list of recommended practices for the collection of self-reported affective annotations using crowdsourcing techniques.

[4] `https://www.mturk.com`

[5] `http://crowdflower.com`

Julin Urbano presented the second paper, "*Crowdsourcing Preference Judgments for Evaluation of Music Similarity Tasks*" [10]. This paper received the **Most Innovative Paper Award**, carrying a $400 cash prize sponsored Microsoft Bing. In this work, the authors focused on the problem of evaluating the quality of music similarity tasks, where musical pieces similar to a query should be retrieved. Traditionally, ground truths based on partially ordered lists were developed to cope with problems regarding relevance judgment, but they require such man-power to be generated that more affordable alternatives were needed. However, in house evaluations keep using these partially ordered lists because they are still more suitable for similarity tasks. The authors of this paper proposed a cheaper alternative to generate these lists by using crowdsourcing to gather music preference judgments. They showed that their method produces lists very similar to the original ones, while dealing with some defects of the original methodology. With this study, they concluded that crowdsourcing is a perfectly viable alternative to evaluate music systems without the need for experts.

John Le presented the paper titled"*Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution*" [7]. One approach to ensure quality of worker judgments in crowdsourcing is to include an initial training period and subsequent sporadic insertion of predefined gold standard data (training data). Workers are notified or rejected when they err on the training data, and trust and quality ratings are adjusted accordingly. In this paper, the authors assessed how this type of dynamic learning environment can affect the workers' results in a search relevance evaluation task completed on Amazon MTurk. They analyzed how the distribution of training set answers impacts training of workers and aggregate quality of worker results and concluded that in a relevance categorization task, a uniform distribution of labels across training data labels produces optimal peaks in 1) individual worker precision and 2) majority voting aggregate result accuracy.

The paper titled "*An Analysis of Assessor Behavior in Crowdsourced Preference Judgments*" by Dongqing Zhu and Ben Carterette [12] described a pilot study using MTurk to collect preference judgments between pairs of full-page layouts including both search results and image results. The authors analyzed the behavior of assessors that participated in their study to identify some patterns that may be broadly indicative of unreliable assessments.

In their paper titled "*Logging the Search Self-Efficacy of Amazon Mechanical Turkers*", Feild et al. [6] addressed the relationship between searcher self-efficacy assessments and their strategies for conducting complex searches. They described a platform for logging actions of MTurk workers and a questionnaire assessing search self-efficacy of Turk workers. They also described the design of an experiment to use workers to evaluate search assistance tools.

The paper titled "*Crowdsourcing a News Query Classification Dataset*" by McCreadie et al. [8] studied the generation and validation of a news query classification dataset comprised of labels crowdsourced from MTurk. The authors mainly focused on two challenges when crowdsourcing news query classification labels: 1) how to overcome the workers lack of information about the news stories from the time of each query and 2) how to ensure quality of the resulting labels. They showed that a workers lack of information about news stories can be addressed through the integration of news-related content into the labeling interface and that this improves the quality of the resulting labels.

Omar Alonso presented the paper titled "*Detecting Uninteresting Content in Text Streams*" [1]. The paper focused on the problem of identifying uninteresting content in text streams from micro-blogging services such as Twitter. The authors used crowdsourcing to estimate the fraction of the Twitter stream that is categorically not interesting, and derived a single, highly effective feature that separates "uninteresting" from "possibly interesting" tweets.

References

[1] O. Alonso, C. Carson, D. Gerster, X. Ji, and S. U. Nabar. Detecting uninteresting content in text streams. In *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, pages 39–42, Geneva, Switzerland, July 2010.

[2] O. Alonso and S. Mizzaro. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 15–16, 2009.

[3] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 541–548, New York, NY, USA, 2006. ACM.

[4] B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 268–275, New York, NY, USA, 2006. ACM.

[5] C. Cleverdon. The cranfield tests on index language devices. *Readings in Information Retrieval*, pages 47–59, 1997.

[6] H. Feild, R. Jones, R. C. Miller, R. Nayak, E. F. Churchill, and E. Velipasaoglu. Logging the search self-efficacy of amazon mechanical turkers. In *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, pages 27–30, Geneva, Switzerland, July 2010.

[7] J. Le, A. Edmonds, V. Hester, and L. Biewald. Ensuring quality in crowdsourced search relevance evaluation. In *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, pages 17–20, Geneva, Switzerland, July 2010.

[8] R. M. C. McCreadie, C. Macdonald, and I. Ounis. Crowdsourcing a news query classification dataset. In *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, pages 31 – 38, Geneva, Switzerland, July 2010.

[9] M. Soleymani and M. Larson. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. In *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, pages 4–8, Geneva, Switzerland, July 2010.

[10] J. Urbano, J. Morato, M. Marrero, and D. Martin. Crowdsourcing preference judgments for evaluation of music similarity tasks. In *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, pages 9–16, Geneva, Switzerland, July 2010.

[11] E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating ap and ndcg. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 603–610, New York, NY, USA, 2008. ACM.

[12] D. Zhu and B. Carterette. An analysis of assessor behavior in crowdsourced preference judgments. In *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, pages 21–26, Geneva, Switzerland, July 2010.