Crowdsourcing for Search and Data Mining

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

Vitor R. Carvalho
Intelius
Bellevue, WA
vcarvalho@intelius.com

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

Abstract

The Crowdsourcing for Search and Data Mining (CSDM 2011) workshop was held on February 9, 2011 in Hong Kong, China, in conjunction with the Fourth ACM International Conference on Web Search and Data Mining (WSDM 2011)[1]. The workshop addressed recent advances in theory and empirical methods, as well as novel applications, in crowdsourcing for search and data mining. Three invited talks were presented, along with eight refereed papers. Workshop proceedings and presentation slides can be found online[2].

1 Introduction

The advent of crowdsourcing is leading us to rethink the ways we design, train, and evaluate information retrieval (IR) systems. Thanks to global growth in Internet connectivity and bandwidth, we can now harness "human computation" in near-real time from a vast and ever-growing, distributed population of online Internet users. Moreover, a rapidly growing array of internet marketplaces, platforms, games, and other internet services has made facilitating such interactions easier than ever before. Such capabilities raise a variety of intriguing new opportunities and challenges for IR research to explore.

To date, research in crowdsourcing for IR has largely focused on investigating strategies for reducing the time, cost, and effort required for annotation, evaluation, and other manual tasks which underlie and support automated IR systems. We have also seen that "wisdom of the crowd" aggregation strategies which combine information from multiple annotators have potential to reduce bias and improve accuracy vs. traditional assessment practices using in-house annotators (e.g. [3]). Consider, for example, the well-established Cranfield paradigm for evaluating IR systems [7], 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 [4, 5, 25], assessment itself remains expensive and slow. Calling upon this distributed, on-demand workforce in place of in-house annotators offers one avenue for addressing this challenge.

Another impacted area is collecting labeled data to train supervised learning systems, such as for learning to rank [15]. Traditional costs associated with data annotation have

[1]The organizers thank Microsoft and CrowdFlower for their generous sponsorship of the workshop.
[2]`http://ir.ischool.utexas.edu/csdm2011`

driven recent machine learning work toward greater use of unsupervised and semi-supervised methods [10]. The recent emergence of crowdsourcing has made labeled data far easier to acquire (e.g. [22]), driving a potential resurgence in use of labeled data.

Crowdsourcing has also introduced intriguing new possibilities for integrating human computation with automated systems: validating search results in near-real time [24], handling difficult cases where automation fails, or exploiting the breadth of backgrounds and geographic dispersion of crowd workers for more diverse and representative assessment.

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. The 1st SIGIR Workshop on Crowdsourcing for IR, *Crowdsourcing for Search Evaluation* (CSE 2010) [6, 14], was well-attended with enthusiastic discussion by participants continuing well into the evening. Building on the strong interest and participation of this event, the CSDM 2011 workshop was organized to generalize crowdsourcing work in IR beyond search evaluation, as well as to invite wider participation from the WSDM community. A complementary tutorial on crowdsourcing was also organized at WSDM 2011 to provide an additional opportunity for the community to learn more about this emerging area [2].

This report on the CSDM 2011 workshop describes advances in the state-of-the-art in using crowdsourcing for search and data mining. Building on the success of both this workshop and the aforementioned events, a 2nd SIGIR Workshop on Crowdsourcing for Information Retrieval[3] will be held on July 28, 2011 in Beijing, China, in conjunction with SIGIR 2011. Another related event, the 2011 Text REtrieval Conference (TREC) Crowdsourcing Track[4], will take place in conjunction with TREC from November 15-18, 2011 in Gaithersburg, MD.

The remainder of this report identifies the CSDM 2011 workshop program committee and briefly summarizes the invited keynote talks and accepted research papers.

2 Program Committee

Omar Alonso, Microsoft Bing
Adam Bradley, Amazon
Ben Carterette, University of Delaware
Charlie Clarke,University of Waterloo
Deepak Ganesan, University of Massachusetts Amherst
Panagiotis G. Ipeirotis, New York University
Gareth Jones, Dublin City University
Jaap Kamps, University of Amsterdam
Gabriella Kazai, Microsoft Research
Mounia Lalmas, University of Glasgow
Martha Larson, Delft University of Technology
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

[3] `https://sites.google.com/site/cir2011ws`
[4] `https://sites.google.com/site/treccrowd2011`

Mark Smucker, University of Waterloo
Siddharth Suri, Yahoo! Research
Fang Xu, Saarland University

3 Workshop Program

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

Invited Talks

- *The Smarter Crowd: Active Learning, Knowledge Corroboration, and Collective IQs* [9]
  Thore Graepel, Microsoft Research
- *Crowdsourcing using Mechanical Turk: Quality Management and Scalability* [12]
  Panagiotis G. Ipeirotis, Stern School of Business, New York University
- *Individual vs. Group Success in Social Networks* [16]
  Winter Mason, Yahoo! Research

Accepted Papers

- *Perspectives on Infrastructure for Crowdsourcing* [1]
  Omar Alonso
- *How Crowdsourcable is Your Task?* (Received **Most Innovative Paper Award**) [8]
  Carsten Eickhoff and Arjen de Vries
- *Youre Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks*
  Christopher Harris [11]
- *Modeling Annotator Accuracies for Supervised Learning* [13]
  Abhimanu Kumar and Matthew Lease
- *Crowdsourcing Blog Track Top News Judgments at TREC* [17]
  Richard McCreadie, Craig Macdonald, and Iadh Ounis
- *Investigating Factors Influencing Crowdsourcing Tasks with High Imaginative Load* [21]
  Raynor Vliegendhart, Martha Larson, Christoph Kofler, Carsten Eickhoff, and Johan Pouwelse
- *Estimating Completion Time for Crowdsourced Tasks Using Survival Analysis Models*
  Jing Wang, Siamak Faridani, and Panagiotis Ipeirotis [23]
- *Crowdsourcing Interactions - A proposal for capturing user interactions through crowdsourcing*
  Guido Zuccon, Teerapong Leelanupab, Stewart Whiting, Joemon Jose, and Leif Azzopardi [26]

3.1 Invited Talks

The workshop opened with an invited talk by Winter Mason, who described how information sharing between multiple individuals responding to an open call impacts the rate at which a solution is identified by the community, as well as who reaps the greatest benefit. To investigate, his team recruited participants via Amazon's Mechanical Turk (MTurk)[5] to play a game modeling this scenario. Gameplay involved exploring the space of a reward function which resembled a rugged landscape and contained a single, global optimum value. Players

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

were paid proportional to points earned during gameplay, but most importantly, could see information about exactly three other players' behaviors: where they had explored and how much they had earned, allowing mimicry. By varying which players were connected in different network structures (hidden to the players), Mason's group studied the impact of different network structures on overall community behavior and outcomes and discovered an interesting tension between local incentives vs. community benefits.

The second invited keynote of the day by Thore Graepel described joint work with Ralf Herbrich, Ulrich Paquet, David Stern, Jurgen Van Gael, Gjergji Kasneci, and Michal Kosinksi. Graepel identified three approaches for better crowdsourcing. By using active learning to select examples for labeling, labeling effort could be more efficiently employed, as demonstrated in the FUSE/MSRC news recommender system `http://projectemporia.com` to categorise news stories in a cost-efficient way. To combat noise in crowd labels, Graepel discussed the importance of consensus methods for label aggregation which account for reliability and expertise of workers as well as the nature and difficulty of the tasks. This method was employed to verify facts in the entity-relationship knowledge base Yago [20]. Finally, Graepel discussed an intriguing approach to measuring collective intelligence of MTurk workers as a function of observable marketplace properties (such as price and required track record for participation).

The final invited talk of the day by Panagiotis G. Ipeirotis discussed the repeated labeling of examples when the labeling is noisy in order to improve label quality for supervised learning [19]. Ipeirotis described how repeated-labeling can improve label quality and model quality, how simple methods and carefully example selection alone can be extremely effective vs. more complicated strategies, and how worker accuracy can be estimated on the fly with systematic bias correction. Results were presented in the context of a real-life application from on-line advertising: using MTurk to determine whether or not certain web pages are objectionable to advertisers. Perhaps most fascinating, Ipeirotis presented recent results suggesting similar characteristics between MTurk worker behavior and that of mice under certain experimental conditions.

3.2 Refereed Papers

The CSDM 2011 Program Committee accepted eight papers for presentation at the workshop.

The first paper presented in the workshop was "*Crowdsourcing Interactions - A proposal for capturing user interactions through crowdsourcing*", by Guido Zuccon, Teerapong Leelanupab, Stewart Whiting, Joemon Jose, and Leif Azzopardi [26]. Guido Zuccon presented the paper, which proposed use of crowdsourcing to collect data on users' interactions with an interactive IR system. Zuccon discussed problematic issues that arose during the design process, together with preliminary findings from implementing their approach on MTurk. Reported results demonstrated the promise of gathering interaction data via crowdsourcing and suggest further research in this vein. In particular the authors suggest that crowdsourcing might be used for evaluating interactive IR systems.

Richard McCreadie presented the second paper "*Crowdsourcing Blog Track Top News Judgments at TREC*" [17], which he co-authored with Craig Macdonald and Iadh Ounis. This paper studies the crowdsourcing of relevance assessments for the 2010 TREC Blog track. The authors built upon their previous work "Crowdsoucing a News Query Classification Dataset" [18], which studied the generation and validation of a news query classification dataset. They used a judging system on top of MTurk to assess the importance of a story,

and proposed a fast manual worker validation approach for quality assurance. Using this system, the authors find that crowdsourcing is a fast, cheap and effective alternative to using specialist assessors or participating groups for this task, and conclude with suggestions for best practices when crowdsourcing.

Carsten Eickhoff presented his joint paper with Arjen de Vries entitled "*How Crowdsourcable is Your Task?* [8]. This paper received the *Most Innovative Paper Award*, carrying a $300 cash prize sponsored by Microsoft Bing. Eickhoff discussed how malicious workers may try to maximise their financial gains by producing generic answers rather than actually working on the task, and how identifying these individuals challenges both crowdsourcing providers and requesters alike. He went on to discuss measures they identified to increase robustness of the crowdsourcing process to such worker fraud. The authors propose non-repetitive and interactive tasks in order to *a priori* discourage fraudsters looking for easily automatable jobs.

The paper titled "*Youre Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks.*" [11] was presented by its single author, Christopher Harris. Harris described experiments conducted on MTurk to conduct resume reviews in order to reduce manual labor involved with hiring. As part of the work, Harris discussed several incentive models and experimental results of their use and show the effectiveness of certain incentive schemes when the task is designed appropriately.

In "*Estimating Completion Time for Crowdsourced Tasks Using Survival Analysis Models*", Vliegendhartet al. [21] addressed crowdsourcing for high "imaginative load" a term they introduced to designate a task that requires workers to answer questions from a hypothetical point of view that is beyond their daily experiences. Their finding was that workers are able to deliver high quality responses to such tasks, but that it is important that the HIT title allows workers to formulate accurate expectations of the task. Also important was the inclusion of free-text justification questions. Carsten Eickhoff presented the paper.

Omar Alonso presented a position paper titled "*Perspectives on Infrastructure for Crowdsourcing*" [1] which laid out a number of challenges and opportunities for improving crowdsourcing systems. 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 to think about platforms and future directions from three views: the end-user, the developer/experimenter, and the system (back end).

The final paper presented at the workshop was "*Modeling Annotator Accuracies for Supervised Learning*", given by Matthew Lease on behalf of the first author, Abhimanu Kumar [13]. Building upon earlier work by Sheng et al. [19] studying the interaction between annotation noise, consensus methods, and classifier accuracy, Kumar et al. studied the effects of modeling vs. ignoring worker accuracy in the consensus method. While choice of consensus method (among those considered) showed relatively little impact on resultant classifier accuracy, failure to model worker accuracy in whichever consensus method was chosen did significantly degrade classifier accuracy. Consensus methods ignoring worker accuracy (like the oft-used majority vote) may thus perform relatively poorly with increasingly noisy workers.

References

[1] O. Alonso. Perspectives on Infrastructure for Crowdsourcing. 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)*, pages 7–10, 2011.

[2] O. Alonso and M. Lease. 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 '11, pages 1–2, 2011. Slides available online.

[3] 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.

[4] J. A. Aslam, V. Pavlu, and E. Yilmaz. 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*, pages 541–548, 2006.

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

[6] V. Carvalho, M. Lease, and E. Yilmaz. Crowdsourcing for search evaluation. *ACM SIGIR Forum*, 44(2):17–22, December 2010.

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

[8] C. Eickhoff and A. de Vries. How Crowdsourcable is Your Task? In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 11–14, 2011.

[9] T. Graepel. The Smarter Crowd: Active Learning, Knowledge Corroboration, and Collective IQs. 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)*, pages 3–4, 2011.

[10] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12, 2009.

[11] C. Harris. Youre Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks. 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)*, pages 15–18, 2011.

[12] P. Ipeirotis. Crowdsourcing using Mechanical Turk: Quality Management and Scalability. 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)*, page 6, 2011.

[13] A. Kumar and M. Lease. Modeling Annotator Accuracies for Supervised Learning. 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)*, pages 19–22, 2011.

[14] M. Lease, V. Carvalho, and E. Yilmaz, editors. *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*. Geneva, Switzerland, July 2010. Available online at `http://ir.ischool.utexas.edu/cse2010`.

[15] T. Liu. Learning to Rank for Information Retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

[16] W. Mason. Individual vs. Group Success in Social Networks. 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)*, page 5, 2011.

[17] R. McCreadie, C. Macdonald, and I. Ounis. 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)*, pages 23–26, 2011.

[18] 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.

[19] V. Sheng, F. Provost, and P. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622, 2008.

[20] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW 2007)*, 2007.

[21] R. Vliegendhart, M. Larson, C. Kofler, C. Eickhoff, and J. Pouwelse. Investigating Factors Influencing Crowdsourcing Tasks with High Imaginative Load. 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*, pages 27–30, 2011.

[22] L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. of SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004.

[23] J. Wang, S. Faridani, and P. Ipeirotis. Estimating Completion Time for Crowdsourced Tasks Using Survival Analysis Models. 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)*, pages 31–34, 2011.

[24] T. Yan, V. Kumar, and D. Ganesan. 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*, pages 77–90. ACM, 2010.

[25] E. Yilmaz, E. Kanoulas, and J. A. Aslam. 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*, pages 603–610, 2008.

[26] G. Zuccon, T. Leelanupab, S. Whiting, J. Jose, and L. Azzopardi. Crowdsourcing Interactions - A proposal for capturing user interactions through crowdsourcing. 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)*, pages 35–38, 2011.