Crowdsourcing for Information Retrieval

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

Emine Yilmaz
Microsoft Research Cambridge, UK
eminey@microsoft.com
Koc University Istanbul, Turkey
eyilmaz@ku.edu.tr

Abstract

The 2nd SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR 2011) was held on July 28, 2011 in Beijing, China, in conjunction with the 34th Annual ACM SIGIR Conference[1]. The workshop brought together researchers and practitioners to disseminate recent advances in theory, empirical methods, and novel applications of crowdsourcing for information retrieval (IR). The workshop program included three invited talks, a panel discussion entitled *Beyond the Lab: State-of-the-Art and Open Challenges in Practical Crowdsourcing*, and presentation of nine refereed research papers and one demonstration paper. A *Best Paper Award*, sponored by Microsoft Bing, was awarded to Jun Wang and Bei Yu for their paper entitled *Labeling Images with Queries: A Recall-based Image Retrieval Game Approach*. A *Crowdsourcing Challenge* contest was also announced prior to the workshop, sponsored by CrowdFlower. The contest offered both seed funding and advanced technical support for the winner to use CrowdFlower's services for innovative work. Workshop organizers selected Mark Smucker as the winner based on his proposal entitled: *The Crowd vs. the Lab: A Comparison of Crowd-Sourced and University Laboratory Participant Behavior*. Proceedings of the workshop are available online[2] [15].

1 Introduction

Crowdsourcing's rapid development continues to both offer new ideas and challenges to our traditional methods for designing, training, and evaluating 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.

Much research to date in crowdsourcing for IR has focused on investigating strategies for reducing the time, cost, and effort required for annotation, evaluation, and other manual

[1]The organizers thank Microsoft and Crowdflower for their generous sponsorship of the workshop.
[2]`https://sites.google.com/site/cir2011ws/proceedings`

tasks which underlie and support automated IR systems. "Wisdom of the crowd" aggregation strategies which combine information from multiple annotators suggest intriguing potential to reduce bias and improve accuracy vs. traditional assessment practices using in-house annotators (e.g. [4]). Consider, for example, the well-established Cranfield paradigm for evaluating IR systems [8], 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 [5, 6, 30], 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 [16]. Traditional costs associated with data annotation have driven recent machine learning work toward greater use of unsupervised and semi-supervised methods [11]. The recent emergence of crowdsourcing has made labeled data far easier to acquire (e.g. [25]), driving a potential resurgence in the degree of and methodology for best utilizing labeled data.

Crowdsourcing has also introduced intriguing new possibilities for integrating human computation with automated systems: validating search results in near-real time [29], 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, entitled *Crowdsourcing for Search Evaluation* (CSE 2010)[3] [7, 12], was well-attended with enthusiastic discussion by participants continuing well into the evening. Building on the strong interest and participation of this event, a subsequent workshop entitled *Crowdsourcing for Search and Data Mining* (CSDM)[4] [13, 14] was held soon thereafter in conjunction with the Fourth ACM International Conference on Web Search and Data Mining (WSDM 2011). This workshop explored crowdsourcing topics in IR beyond search evaluation, as well as inviting wider participation from the WSDM community.

Following on the first *Crowdsourcing for Relevance Evaluation* tutorial at ECIR 2010 [1], a 2nd tutorial on crowdsourcing was organized at WSDM 2011 to provide further opportunities for the community to learn more about this emerging area [2]. In conjunction with this 2nd SIGIR Workshop on Crowdsourcing for Information Retrieval, a 3rd Crowdsourcing for IR tutorial was offered with updated and expanded content [3] to complement the workshop. Slides from all three tutorials are available online (see the References section).

This report on the CIR 2011 workshop describes recent advances in the state-of-the-art in using crowdsourcing for IR. We begin by acknowledging the workshop's Program Committee who invested significant time and effort refereeing the submitted research papers. We then outline the workshop's program. The remainder of this report summarizes the workshop's invited keynote talks, panel discussion, and the accepted research papers and demonstration.

[3] `http://ir.ischool.utexas.edu/cse2010/program.htm`

[4] `http://ir.ischool.utexas.edu/csdm2011/proceedings.html`

2 Program Committee

Omar Alonso, Microsoft Bing
Paul Bennett, Microsoft Research
Adam Bradley, Amazon.com
Ben Carterette, University of Delaware
Charlie Clarke, University of Waterloo
Harry Halpin, University of Edinburgh
Gareth Jones, Dublin City University
Jaap Kamps, University of Amsterdam
Martha Larson, Delft University of Technology
Gabriella Kazai, Microsoft Research
Mounia Lalmas, Yahoo! Research
Edith Law, Carnegie Mellon University
Don Metzler, University of Southern California
Stefano Mizzaro, University of Udine
Stefanie Nowak, Fraunhofer IDMT
Iadh Ounis, University of Glasgow
Mark Sanderson, RMIT University
Mark Smucker, University of Waterloo
Ian Soboroff, National Institute of Standards
Siddharth Suri, Yahoo! Research

3 Workshop Program

The workshop program included three invited talks by noted researchers, a panel session of industry experts addressing practical challenges of achieving reliable crowdsourcing at scale, and presentation of nine refereed research papers and one demonstration paper.

By popular vote of workshop attendees, the workshop's *Best Paper Award*, sponsored by Microsoft Bing, was awarded to Jun Wang and Bei Yu for their paper entitled *Labeling Images with Queries: A Recall-based Image Retrieval Game Approach.*

Invited Talks

Effects of Defensive HIT Design on Crowd Diversity
Gabriella Kazai, Microsoft Research

Experiences and Lessons from Collecting Relevance Judgments
Ian Soboroff, National Institute of Standards and Technology (NIST)

Issues of Quality in Human Computation
Praveen Paritosh, Google

Panel Discussion

Beyond the Lab: State-of-the-Art and Open Challenges in Practical Crowdsourcing
Omar Alonso, Microsoft Bing
Roi Blanco, Yahoo! Research
Praveen Paritosh, Google

Accepted Papers

Genealogical Search Analysis Using Crowd Sourcing
Patrick Schone and Michael Jones

The Crowd vs. the Lab: A Comparison of Crowd-Sourced and University Laboratory Participant Behavior
Mark Smucker and Chandra Prakash Jethani
Winner: ***Crowdsourcing Challenge*** contest

A Comparison of On-Demand Workforce with Trained Judges for Web Search Relevance Evaluation
Maria Stone, Kylee Kim, Suvda Myagmar and Omar Alonso

An Ensemble Framework for Predicting Best Community Answers
Qi Su

Quality Control of Crowdsourcing through Workers Experience
Li Tai, Zhang Chuang, Xia Tao, Wu Ming and Xie Jingjing

Semi-Supervised Consensus Labeling for Crowdsourcing
Wei Tang and Mathew Lease

Crowdsourced Evaluation of Personalization and Diversification Techniques in Web Search
David Vallet

How Much Spam Can You Take? An Analysis of Crowdsourcing Results to Increase Accuracy
Jeroen Vuurens, Arjen P. De Vries and Carsten Eickhoff

Labeling Images with Queries: A Recall-based Image Retrieval Game Approach
Jun Wang and Bei Yu
Winner: ***Best Paper Award***

Accepted Demonstration Paper

GEAnn - Games for Engaging Annotations
Carsten Eickhoff, Christopher G. Harris, Padmini Srinivasan and Arjen P. de Vries

3.1 Invited Talks

Effects of Defensive HIT Design on Crowd Diversity
Gabriella Kazai, Microsoft Research

The first invited talk by Gabriella Kazai focused on the design of the design of crowdsourcing tasks (HITs). Gabriella described a detailed investigation of two HIT designs in the context of a relevance labeling task and showed that there is a clear segmentation of the crowd based on the HIT design. She showed that a design rich in quality control features attracts more conscientious workers, mostly from the US, while a simpler design attracts younger and less serious workers, mostly from Asia. She recommended a two stage method for tasks where quality control is a requirement: (1) initial HITs are run as a recruitment campaign to determine task sensitivity to spam, pricing, personality profiles, and task-related features of the workers with impact on quality and (2) final HIT design is created to target (pre-filter) the specific segment of the crowd.

Experiences and Lessons from Collecting Relevance Judgments
Ian Soboroff, National Institute of Standards and Technology (NIST)

In his talk, Ian Soboroff described the current methodology used by TREC for obtaining relevance judgments, problems they will face if they were to completely switch to crowdsourcing instead of laboratory based judgments. He mentioned that the best role for TREC is to help establish best practices for the IR community, by leveraging the TREC community and possibly TREC resources to determine what IR crowdsourcing for lab experiments should do.

Issues of Quality in Human Computation
Praveen Paritosh, Google

Praveen Paritosh presented some fundamental issues regarding computing the quality of work in crowdsourced human computation settings. The key to most measures of quality are intrinsically tied to the participants in the system, and do not have properties like comparability across time and across tasks. The talk by Praveen described some approaches to ameliorate this problem. Praveen also described new capabilities provided by some crowdsourcing platforms (e.g. training the judges before crowdsourcing, crowdsourcing the management of judges, etc.), why they are needed and how they are used.

3.2 Panel Discussion

A panel session, entitled *Beyond the Lab: State-of-the-Art and Open Challenges in Practical Crowdsourcing*, was organized with three industry experts addressing practical challenges of achieving reliable crowdsourcing at scale. Panelists included Omar Alonso (Microsoft Bing), Roi Blanco (Yahoo! Research), and Praveen Paritosh (Google). The session's main goal was to develop a broader understanding of how search companies use crowdsourcing in practice.

Panelists were asked to respond to the following questions:

- For what situations or kinds of problems have you or your company considered using crowdsourcing? Why did you decide to use it or not to use it? When you did use it, how did you use it, and what challenges did you encounter?
- What privacy issues have you encountered? What kinds of data have been at issue (e.g. customer data, corporate records / intellectual property, etc.) or might you want to crowdsource if you could (e.g. annotating rare queries which are often personal)?
- While search companies are reputed to used very detailed judging guidelines (e.g. leaked google quality raters handbook [18]), the most popular micro-task model assumes minimal training and instructions. Have you adopted the micro task model and bridged this gap somehow, or adopted a different model?
- How have you achieved practical, reliable, scalable, affordable effectiveness in "real world" crowdsourcing (i.e. when it matters to the company's success and your job)?
- What crowdsourcing problems should be worked on by (or left to) (a) academics (b) industry users (search engine companies) (c) industry providers (e.g. Amazon, Crowdflower)? Where are good opportunities for academics and industry to collaborate?

In general, all companies seem to be investigating ways of using crowdsourcing for commercial purposes but many important practical issues continue to preclude broader adoption

of crowdsourcing methods at scale in a production envrionment. One of the notable obstacles to wider adoption by commercial entities continues to be ensuring privacy of user data. Unfortunately, developing effective methods for protecting customer privacy remains an open research problem with some highly visible recent failures. Moreover, companies have greater cause for concern with the Federal Trade Commission's recent move toward more aggressively acting to protect consumers from data breeches caused by commercial entities [28].

3.3 Refereed Papers

The workshop's Program Committee reviewed and accepted nine research papers for presentation and inclusion in workshop proceedings. An additional demonstration paper was also accepted. We briefly summarize these contributed papers below; we refer the interested reader to the papers themselves for additional details.

Genealogical Search Analysis Using Crowd Sourcing
Patrick Schone and Michael Jones [17]

The authors describe the creation and analysis of what they believe to be the largest genealogical evaluation set ever developed. The evaluation was made possible through crowdsourcing efforts of 2277 genealogical patrons over a period of about two months. This evaluation resulted in the annotation of almost 145,000 search results from 3781 genealogical queries issued against a collection of billions of digitized historical records. The authors described some of the interesting analysis and discoveries from this new evaluation corpus and proposed a metric which can serve to quantify systems of this kind.

For the interested reader, we note a distinct IR collection related to message boards in the genealogy domain was also released by a different research group [10].

The Crowd vs. the Lab: A Comparison of Crowd-Sourced and University Laboratory Participant Behavior
Mark Smucker and Chandra Prakash Jethani [19]

The authors investigate differences in remuneration and environment between crowdsourced workers and more traditional laboratory study participants. The authors find that if crowdsourced participants are to be used for IR user studies, we need to know if and to what extent their behavior on IR tasks differs from the accepted standard of laboratory participants. The authors conducted an experiment to measure the relevance judging behavior of both crowd-sourced and laboratory participants, and found that while only 30% of the crowd-sourced workers qualified for inclusion in the final group of participants, 100% of the laboratory participants qualified. They showed that both groups had similar true positive rates, but the crowd-sourced participants had a significantly higher false positive rate and judged documents nearly twice as fast as the laboratory participants.

A Comparison of On-Demand Workforce with Trained Judges for Web Search Relevance Evaluation
Maria Stone, Kylee Kim, Suvda Myagmar and Omar Alonso [20]

Stone et al. focused on comparing crowdsourced judges with expert judges for evaluation of search engine effectiveness. The standard approach used to understand how well crowdsourcing works relative to expert judges is to generate a set of labels that has been pre-judged by the expert judges (gold standard data), and then determine how closely crowd-

sourced data approximate these expert labels. This approach inherently assumes that expert judges are better and are better able to assign appropriate relevance labels. Instead, Stone et al. proposed an independent test of how well two groups will do, without relying on a set that has been pre-judged by the expert judges. They chose to rely on a ranker of one of the major search engines to generate the gold standard data. If the ranker is doing its job, then removing the top result should damage most search results pages for most queries. They studied how well expert judges and Mechanical Turk workers are able to label full, unaltered pages and pages with top result removed.

An Ensemble Framework for Predicting Best Community Answers
Qi Su [21]

Enlightened by the decomposable characteristics of information credibility, the authors proposed a multi-dimensional learning approach to predict the credibility of answers in the context of community question answering. The author encoded several factors with regard to the content aspect of credibility, including mislead/deviation, certainty and informativeness and proposed a two-level scheme. First, supervised ranking serves as the first-order learner to model three credibility aspects separately. Then, an ensemble framework combines the predictions from the first-order learners to generate final rankings. The author showed that an effective ensemble strategy targeting different credibility aspects contributes much better to the predication performance for information credibility than all the first-order predictors and their feature-level combination.

Quality Control of Crowdsourcing through Workers Experience
Li Tai, Zhang Chuang, Xia Tao, Wu Ming and Xie Jingjing [22]

The authors focused on quality control mechanisms for crowdsourcing. They proposed a new quality control method through worker's experience in the work which has been divided in to several stages. They permitted the workers at each stage to work on a number of HITs in proportion to their estimated accuracy in previous stages. To test the method, they conducted two experiments on CrowdFlower, and created a simulation model based on Gaussian distribution and worker quantitative distribution in some existing crowdsourcing result data. They showed that the accuracy of result has increased from 76% to 85% in the first experiment, and in the simulation the accuracy of result increased from 79.75% to 91.5% in simulation program.

Semi-Supervised Consensus Labeling for Crowdsourcing
Wei Tang and Mathew Lease [23]

The authors consider the consensus task of inferring a single label for each item given multiple crowdsourced labels. While simple majority vote computes consensus by equally weighting each worker's vote, weighted voting assigns greater weight to more accurate workers, where accuracy is estimated by inner-annotator agreement (unsupervised) and/or agreement with known expert labels (supervised). The authors investigated the annotation cost vs. consensus accuracy benefit from increasing the amount of expert supervision, proposing a semi-supervised approach which infers consensus labels using both labeled and unlabeled examples. Using both synthetic data and relevance judgments from Mechanical Turk, the authors showed significant benefit can be derived from a relatively modest amount of supervision, and that consensus accuracy from full supervision with a large amount of labeled

data is matched by their semi-supervised approach requiring much less supervision.

Crowdsourced Evaluation of Personalization and Diversification Techniques in Web Search
David Vallet [24]

Vallet presented two complementary crowdsourcing-based evaluation methodologies to evaluate personalization and/or diversification techniques. The author showed that using the approaches he presented results in results that are consistent with previous experimental results coming from users. These results suggest that these evaluation methodologies could allow performing inexpensive user-centered evaluations in the future. The authors also made a test collection for personalized and diversification approaches available.

How Much Spam Can You Take? An Analysis of Crowdsourcing Results to Increase Accuracy
Jeroen Vuurens, Arjen P. De Vries and Carsten Eickhoff [26]

The authors propose a way to both reduce spam and increase accuracy in crowdsourcing. Using simulations to compare performance between different algorithms and inspecting accuracy and costs for different experimental settings, they show that gold sets and majority voting are less spam-resistant than many believe and can easily be outperformed.

Labeling Images with Queries: A Recall-based Image Retrieval Game Approach
Jun Wang and Bei Yu [27]

Jun Wang discussed how human computation game-based image labeling has proven effective in collecting image descriptions for use in improving image searching, browsing, and accessibility. ESP and Phetch are two successful image description games: the ESP game is designed for collecting keyword descriptions while the Phetch game for sentence descriptions. In their paper, the authors proposed and implemented an image retrieval game for collecting query-like descriptions. The idea is that a player is presented with a target image for a short time and then needs to construct a query based on his recall of the image, with the goal of making the target image ranked as high as possible by an image search engine. The authors showed that the image retrieval game enabled them to collect queries that are comparable to the real queries reported in existing studies, and it can also be used to collect informative tags while avoiding undesired ones such as trivial color words.

3.4 Demonstration Paper

GEAnn - Games for Engaging Annotations
Carsten Eickhoff, Christopher G. Harris, Padmini Srinivasan and Arjen P. de Vries [9]

Based on their previous experience using crowdsourcing as well as insights from psychology, Eickoff et al. proposed and showed the use of a game in crowdsourcing scenarios in order to attract and retain a larger share of entertainment seekers to relevance assessment tasks.

4 Conclusion

The 2nd SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR 2011) brought together researchers and industry practioners to exchange state-of-the-art methodology and best-practices, as well as identify open challenges remaining to be addressed when utilizing

crowdsourcing for IR. We anticipate a variety of exciting future events (tutorials, workshops, conference sessions, etc.) further exploring these issues in the years to come. In the short-term, we look forward to additional findings and lessons learned to be disseminated from the 2011 Text REtrieval Conference (TREC)'s Crowdsourcing Track[5], which took place in conjunction with TREC from November 15-18, 2011 in Gaithersburg, MD USA.

References

[1] O. Alonso. Tutorial: Crowdsourcing for Relevance Evaluation. In *Proceedings of the 32nd European Conference on IR Research (ECIR)*, 2010. Slides available online at `http://ir.ischool.utexas.edu/cse2010/materials/alonso-ecir2010-tutorial.pdf`.

[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)*, pages 1–2, 2011. Slides available online at `http://ir.ischool.utexas.edu/wsdm2011_tutorial.pdf`.

[3] O. Alonso and M. Lease. Crowdsourcing for Information Retrieval: Principles, Methods, and Applications. In *Tutorial at the 34th Annual ACM SIGIR Conference*, page 1299, Beijing, China, July 2011. Slides available online at `http://www.slideshare.net/mattlease/crowdsourcing-for-information-retrieval-principles-methods-and-applications`.

[4] 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.

[5] 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.

[6] 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.

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

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

[9] C. Eickhoff, C. G. Harris, P. Srinivasan, and A. P. de Vries. GEAnn - Games for Engaging Annotations. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, 2011.

[10] J. L. Elsas. Ancestry.com online forum test collection. Technical Report CMU-LTI-017, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2011.

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

[12] 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`.

[13] M. Lease, V. Carvalho, and E. Yilmaz. Crowdsourcing for search and data mining. *ACM SIGIR Forum*, 45(1):18–24, June 2011.

[5] `https://sites.google.com/site/treccrowd2011`

[14] M. Lease, V. Carvalho, and E. Yilmaz, editors. *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, February 2011. Available online at `http://ir.ischool.utexas.edu/cse2010/program.html`.

[15] M. Lease, E. Yilmaz, A. Sorokin, and V. Hester, editors. *Proceedings of the 2nd ACM SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval (CIR)*. Beijing, China, July 2011. Available online at `https://sites.google.com/site/cir2011ws/proceedings`.

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

[17] P. Schone and M. Jones. Genealogical Search Analysis Using Crowd Sourcing. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, 2011.

[18] B. Schwartz. The Google Quality Raters Handbook, 2008. March 14. `http://searchengineland.com/the-google-quality-raters-handbook-13575`.

[19] M. Smucker and C. P. Jethani. The Crowd vs. the Lab: A Comparison of Crowd-Sourced and University Laboratory Participant Behavior. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, 2011.

[20] M. Stone, K. Kim, S. Myagmar, and O. Alonso. A Comparison of On-Demand Workforce with Trained Judges for Web Search Relevance Evaluation. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, 2011.

[21] Q. Su. An Ensemble Framework for Predicting Best Community Answers. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, 2011.

[22] L. Tai, Z. Chuang, X. Tao, W. Ming, and X. Jingjing. Quality Control of Crowdsourcing through Workers Experience. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, 2011.

[23] W. Tang and M. Lease. Semi-Supervised Consensus Labeling for Crowdsourcing. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, 2011.

[24] D. Vallet. Crowdsourced Evaluation of Personalization and Diversification Techniques in Web Search. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, 2011.

[25] 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.

[26] J. Vuurens, A. P. D. Vries, and C. Eickhoff. How Much Spam Can You Take? An Analysis of Crowdsourcing Results to Increase Accuracy. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, 2011.

[27] J. Wang and B. Yu. Labeling Images with Queries: A Recall-based Image Retrieval Game Approach. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, 2011.

[28] S. Wolfson and M. Lease. Look Before You Leap: Legal Pitfalls of Crowdsourcing. In *Proceedings of the 74th Annual Meeting of the American Society for Information Science and Technology (ASIS&T)*, 2011.

[29] 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 (MOBISYS)*, pages 77–90. ACM, 2010.

[30] 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.