
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. McCreddie, 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.