References

[1] Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of CIKM*, pages 601–610, 2009.

[2] James O. Berger. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1):1–32.

[3] Ben Carterette. Model-based inference about ir systems. In *Proceedings of ICTIR*, 2011.

[4] Ben Carterette. Multiple testing in statistical analysis of systems-based IR experiments. *ACM TOIS*, 2012.

[5] Ben Carterette, Evangelos Kanoulas, Virgil Pavlu, and Hui Fang. Building reusable test collections through experimental design. In *Proceedings of SIGIR*, 2010.

[6] Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. Simulating simple user behavior for systems effectiveness evaluation. In *Proceedings of CIKM*, 2011.

[7] Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. Incorporating variability in user behavior into systems based evaluation. In *Proceedings of CIKM*, 2012.

[8] Ben Carterette and Mark D. Smucker. Hypothesis testing with incomplete relevance judgments. In *Proceedings of the 16th ACM International Conference on Information and Knowledge Management*, pages 643–652, 2007.

[9] G. V. Cormack and T. R. Lyman. Statistical precision of information retrieval evaluation. In *Proceedings of SIGIR*, pages 533–540, 2006.

[10] John P. A. Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, 294(2):218–228, 2005.

[11] John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8), 2005.

[12] Karen Sparck Jones, editor. *Information Retrieval Experiment*. Buttersworth, 1981.

[13] Karen Sparck Jones and Peter Willett, editors. *Readings in Information Retrieval*. Morgan Kaufmann Publishers, 1997.

[14] Alistair Moffat and Justin Zobel. What does it mean to "measure performance"? In *Proceedings of WISE*, pages 1–12, 2004.

[15] Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

[16] Mark Sanderson and Justin Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of SIGIR*, pages 162–169, 2005.

[17] Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of CIKM*, pages 623–632, 2007.

[18] Mark D. Smucker, James Allan, and Ben Carterette. Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes. In *Proceedings of SIGIR*, pages 630–631, 2009.

[19] Jean Tague. The pragmatics of information retrieval evaluation. In Jones [12], pages 59–102.

[20] Jean Tague-Sutcliffe. The pragmatics of information retrieval evaluation revisited. In Jones and Willett [13], pages 205–216.

[21] Jean Tague-Sutcliffe and James Blustein. A statistical analysis of the TREC-3 data. In *Proceedings of the 3rd Text REtrieval Conference (TREC)*, pages 385–399, 1994.

[22] Ellen Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of SIGIR*, pages 315–323, 1998.

[23] Ellen M. Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of SIGIR*, pages 316–323, 2002.

[24] Ellen M. Voorhees and Donna K. Harman. *TREC: Experiments and evaluation in information retrieval*. The MIT Press, 2005.

[25] William Webber, Alistair Moffat, and Justin Zobel. Statistical power in retrieval experimentation. In *Proceedings of CIKM*, pages 571–580, 2008.

[26] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of SIGIR*, pages 307–314, 1998.