where S is

$$\mathbb{V}\left[\overline{J}_N\right]\left(\frac{1}{(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)^2}+\frac{(\overline{j}-1+\widehat{m}_{J_N})^2}{(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)^4}+2\frac{\overline{j}-1+\widehat{m}_{J_N}}{(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)^3}\right)$$

Then expand S as

$$S=\frac{\mathbb{V}\left[\overline{J}_N\right]}{(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)^4}\big((\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)^2+(\overline{j}-1+\widehat{m}_{J_N})^2-2(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)(\overline{j}-1+\widehat{m}_{J_N})\big)$$

or equivalently,

$$S=\frac{\mathbb{V}\left[\overline{J}_N\right]}{(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)^4}S_1$$

where

$$S_1=(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)^2-(\overline{j}-1+\widehat{m}_{J_N})^2-2(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)(\overline{j}-1+\widehat{m}_{J_N})$$

Rewrite S_1 with $z=\widehat{m}_{J_N}-1$ to get

$$S_1=(\widehat{m}_{J_R}+z)^2+(\overline{j}+z)^2-2(\widehat{m}_{J_R}+z)(\overline{j}+z)$$

which simplifies to

$$S_1=\big((\widehat{m}_{J_R}+z)-(\overline{j}+z)\big)^2=(\widehat{m}_{J_R}-\overline{j})^2$$

Putting these pieces together gives Equation (7).

B. REFERENCES

[1] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9, 2008.

[2] P. Bailey, N. Craswell, I. Soboroff, A. P. D. Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *SIGIR 2008*, pages 667–674, 2008.

[3] G. V. Cormack and T. R. Lynam. Statistical precision of information retrieval evaluation. In *SIGIR 2006*, page 533, 2006.

[4] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

[5] L. A. Granka, T. Joachims, and G. Gay. Eye-Tracking Analysis of User Behavior in WWW Search. In *SIGIR 2004*, pages 478–479, 2004.

[6] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *TOIS*, 20(4):422–446, 2002.

[7] M. Joglekar, H. Garcia-Molina, and A. Parameswaran. Evaluating the crowd with confidence. In *KDD 2013*, page 686, 2013.

[8] G. Kazai and M. Lalmas. eXtended Cumulated Gain Measures for the Evaluation of Content-oriented XML Retrieval. *TOIS*, 24(4):503–542, 2006.

[9] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *TOIS*, 27(1):1–27, 2008.

[10] R. J. Passonneau and B. Carpenter. The Benefits of a Model of Annotation. *Transactions of the Association for Comptuational Linguistics*, 2:311–326, 2014.

[11] B. Piwowarski, A. Trotman, and M. Lalmas. Sound and complete relevance assessment for XML retrieval. *TOIS*, 27:1:1–1:37, 2008.

[12] M. Sanderson, F. Scholer, and A. Turpin. Relatively relevant: Assessor shift in document judgements. In *ADCS 2010*, pages 60–67, 2010.

[13] F. Scholer, A. Turpin, and M. Sanderson. Quantifying Test Collection Quality Based on the Consistency of Relevance Judgements. In *SIGIR 2011*, pages 1063–1072, 2011.

[14] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM 2007*, pages 623–632, 2007.

[15] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

[16] W. Tang and M. Lease. Semi-supervised consensus labeling for crowdsourcing. In *SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval*, pages 36–41, 2011.

[17] J. Vuurens, A. de Vries, and C. Eickhoff. How Much Spam Can You Take? An Analysis of Crowdsourcing Results to Increase Accuracy. In *ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, pages 21–26, 2011.

[18] Y. Wang, L. Wang, Y. Li, D. He, W. Chen, and T.-Y. Liu. A Theoretical Analysis of NDCG Ranking Measures. In *26th Annual Conference on Learning Theory*, pages 1–30, 2013.

[19] W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *TOIS*, 28(4):1–38, 2010.

[20] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration. *VLDB*, 5(6):550–561, 2012.