REFERENCES

[1] Alfred V Aho and Margaret J Corasick. 1975. Efficient string matching: an aid to bibliographic search. *Commun. ACM* 18, 6 (1975), 333–340.

[2] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, 3490–3496.

[3] Azzah Al-Maskari, Mark Sanderson, Paul Clough, and Eija Airio. 2008. The good and the bad system: does the test collection predict users' effectiveness?. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 59–66.

[4] Peter Bailey, Nick Craswell, Ryen W White, Liwei Chen, Ashwin Satyanarayana, and Seyed MM Tahaghoghi. 2010. Evaluating search systems using result page context. In *Proceedings of the third symposium on Information interaction in context*.

[5] Ben Carterette, James Allan, and Ramesh Sitaraman. 2006. Minimal Test Collections for Retrieval Evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Seattle, Washington, USA) *(SIGIR '06)*. Association for Computing Machinery, New York, NY, USA, 268–275. https://doi.org/10.1145/1148170.1148219

[6] Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. 2008. Here or There: Preference Judgments for Relevance. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval* (Glasgow, UK) *(ECIR'08)*. Springer-Verlag, Berlin, Heidelberg, 16–27.

[7] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*.

[8] Zhuyun Dai and Jamie Callan. 2019. Context-Aware Sentence/Passage Term Importance Estimation For First Stage Retrieval. In *arXiv:1910.10687*.

[9] Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[10] Hui Fang, Tao Tao, and Chengxiang Zhai. 2011. Diagnostic evaluation of information retrieval models. *ACM Transactions on Information Systems (TOIS)* 29, 2 (2011), 1–42.

[11] Marco Ferrante, Nicola Ferro, and Norbert Fuhr. 2021. Towards Meaningful Statements in IR Evaluation. Mapping Evaluation Measures to Interval Scales. *arXiv preprint arXiv:2101.02668* (2021).

[12] N. Fuhr. 2018. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum* 51 (2018), 32–41.

[13] Ning Gao, Mossaab Bagdouri, and Douglas W Oard. 2016. Pearson rank: a head-weighted gap-sensitive score-based correlation coefficient. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 941–944.

[14] Ning Gao and Douglas Oard. 2015. A head-weighted gap-sensitive correlation coefficient. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 799–802.

[15] Peter B Golbus, Imed Zitouni, Jin Young Kim, Ahmed Hassan, and Fernando Diaz. 2014. Contextual and dimensional relevance judgments for reusable SERP-level evaluation. In *Proceedings of the 23rd international conference on World wide web*.

[16] John Guiver, Stefano Mizzaro, and Stephen Robertson. 2009. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Transactions on Information Systems (TOIS)* 27, 4 (2009), 1–26.

[17] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-Hoc Retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM 2016)*. Indianapolis, Indiana, 55–64.

[18] Sebastian Hofstätter, Markus Zlabinger, and Allan Hanbury. 2020. Interpretable & Time-Budget-Constrained Contextualization for Re-Ranking. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*.

[19] Sebastian Hofstätter, Markus Zlabinger, and Allan Hanbury. 2019. Neural-IR-Explorer: A Content-Focused Tool to Explore Neural Re-Ranking Results. arXiv:1912.04713 [cs.IR]

[20] Sebastian Hofstätter, Markus Zlabinger, and Allan Hanbury. 2019. TU Wien @ TREC Deep Learning '19 – Simple Contextualization for Re-ranking. arXiv:1912.01385 [cs.IR]

[21] Mehdi Hosseini, Ingemar J Cox, Natasa Milic-Frayling, Vishwa Vinay, and Trevor Sweeting. 2011. Selecting a subset of queries for acquisition of further relevance judgements. In *Conference on the theory of information retrieval*. Springer.

[22] Kai Hui, Andrew Yates, Klaus Berberich, and Gerard De Melo. 2018. Co-PACRR: A context-aware neural IR model for ad-hoc retrieval. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 279–287.

[23] Gabriella Kazai and Homer Sung. 2014. Dissimilarity based query selection for efficient preference based IR evaluation. In *European conference on information retrieval*. Springer, 172–183.

[24] Mucahid Kutlu, Tamer Elsayed, Maram Hasanain, and Matthew Lease. 2018. When rank order isn't enough: New statistical-significance-aware correlation measures. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 397–406.

[25] Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. 2017. Learning to effectively select topics for information retrieval test collections. *arXiv:1701.07810* (2017).

[26] Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2020. PARADE: Passage representation aggregation for document reranking. *arXiv preprint arXiv:2008.09093* (2020).

[27] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained transformers for text ranking: Bert and beyond. *arXiv:2010.06467* (2020).

[28] Sean MacAvaney. 2020. OpenNIR: A Complete Neural Ad-Hoc Ranking Pipeline. In *Proceedings of the Thirteenth ACM International Conference on Web Search and Data Mining*. 845–848. https://doi.org/10.1145/3336191.3371864

[29] Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. 2020. ABNIRML: Analyzing the Behavior of Neural IR Models. *arXiv:2011.00696* (2020).

[30] Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Expansion via Prediction of Importance with Contextualization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)* (Virtual Event, China). 1573–1576.

[31] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized Embeddings for Document Ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1101–1104. https://doi.org/10.1145/3331184.3331317

[32] Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. 2021. Simplified Data Wrangling with ir_datasets. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://doi.org/10.1145/3404835.3463254

[33] C. MacDonald and N. Tonellotto. 2020. Declarative Experimentation in Information Retrieval using PyTerrier. *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval* (2020).

[34] Bhaskar Mitra, Corby Rosset, David Hawking, Nick Craswell, Fernando Diaz, and Emine Yilmaz. 2019. Incorporating query term independence assumption for efficient retrieval and ranking using deep neural networks. *arXiv:1907.03693* (2019).

[35] Daniël Rennings, Felipe Moraes, and Claudia Hauff. 2019. An axiomatic approach to diagnosing neural IR models. In *European Conference on Information Retrieval*. Springer, 489–503.

[36] T. Sakai and N. Kando. 2008. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval* 11 (2008), 447–470.

[37] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do user preferences and evaluation measures line up?. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 555–562.

[38] Paul Thomas and David Hawking. 2006. Evaluation by comparing result sets in context. In *Proceedings of the 15th ACM international conference on Information and knowledge management*. 94–101.

[39] Christophe Van Gysel and Maarten de Rijke. 2018. Pytrec_eval: An Extremely Fast Python Interface to trec_eval. In *SIGIR*. ACM.

[40] Sebastiano Vigna. 2015. A weighted correlation index for rankings with ties. In *Proceedings of the 24th international conference on World Wide Web*. 1166–1176.

[41] E. Voorhees. 2019. The Evolution of Cranfield. In *Information Retrieval Evaluation in a Changing World*.

[42] Ellen M. Voorhees. 2004. Overview of the TREC 2004 Robust Track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*. Gaithersburg, Maryland, 52–69.

[43] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)* 28, 4 (2010), 1–38.

[44] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-End Neural Ad-Hoc Ranking with Kernel Pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*. Tokyo, Japan, 55–64.

[45] Andrew Yates, Kevin Martin Jose, Xinyu Zhang, and Jimmy Lin. 2020. Flexible IR pipelines with Capreolus. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3181–3188.

[46] Emine Yilmaz, Javed A Aslam, and Stephen Robertson. 2008. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 587–594.