ReNeuIR at SIGIR 2023: The Second Workshop on Reaching Efficiency in Neural Information Retrieval

Sebastian Bruch
Pinecone
New York, United States
sbruch@acm.org

Joel Mackenzie
The University of Queensland
Brisbane, Australia
joel.mackenzie@uq.edu.au

Maria Maistro
University of Copenhagen
Copenhagen, Denmark
mm@di.ku.dk

Franco Maria Nardini
ISTI-CNR
Pisa, Italy
francomaria.nardini@isti.cnr.it

ABSTRACT

Multifaceted, empirical evaluation of algorithmic ideas is one of the central pillars of Information Retrieval (IR) research. The IR community has a rich history of studying the effectiveness of indexes, retrieval algorithms, and complex machine learning rankers and, at the same time, quantifying their computational costs, from creation and training to application and inference. As the community moves towards even more complex deep learning models, questions on efficiency have once again become relevant with renewed urgency. Indeed, efficiency is no longer limited to time and space; instead it has found new, challenging dimensions that stretch to resource-, sample- and energy-efficiency with ramifications for researchers, users, and the environment alike. Examining algorithms and models through the lens of holistic efficiency requires the establishment of standards and principles, from defining relevant concepts, to designing metrics, to creating guidelines for making sense of the significance of new findings. The second iteration of the ReNeuIR workshop aims to bring the community together to debate these questions, with the express purpose of moving towards a common benchmarking framework for efficiency.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Retrieval models and ranking**.

KEYWORDS

efficiency, neural IR, sustainable IR, retrieval, ranking, algorithms

ACM Reference Format:
Sebastian Bruch, Joel Mackenzie, Maria Maistro, and Franco Maria Nardini. 2023. ReNeuIR at SIGIR 2023: The Second Workshop on Reaching Efficiency in Neural Information Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3539618.3591922

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591922

1 MOTIVATION AND THEME

We rely on a suite of algorithmic tools to get the information that is pertinent to us, such as discovering news articles, movies, or songs (recommendation systems), getting answers to natural language questions (question answering and conversational agents), finding images depicting a given description (image search), and many more. What all of these applications have in common is that they are different manifestations of the *retrieval and ranking* problem—a fundamental question in IR—which seeks to discover a *set of relevant items* (research articles) from a large collection (the Web) and order them according to some *criteria* (relevance) and with respect to some *context* (the user and their query).

Machine learning transformed how we approach the ranking problem. In particular, we have witnessed a paradigm shift from early statistical methods, heuristics, and hand-crafted rules, to what is now known as Learning to Rank [24]. This leap was perhaps best exemplified by LambdaMART [9] in the Yahoo! Learning-to-Rank Challenge [11], and almost a decade later, by deep neural networks and foundational language models which advanced the state-of-the-art in ranking dramatically [21, 37, 38, 40].

It is not just ranking that benefited from the advent of deep learning. Learnt representations of queries and documents by deep networks, too, offered a range of new opportunities including the development of a generation of "dense" retrieval methods [19, 48], document expansion techniques [39], and others. These recent developments mark the beginning of a new era known as Neural Information Retrieval (NIR).

This transition from inexpensive statistical methods to complex, expensive machine learning models is not unique to the text ranking problem and can be seen across many domains of IR research. While this progression enabled new peaks in quality and effectiveness to be attained, it has done so with orders of magnitude more learnable parameters, thereby requiring greater amounts of data and computational resources. The growth in scale from decision forests to deep neural networks, in particular, drastically increases the computational and economic costs of model training and inference, for example, leaving the research community wondering if we must lose quality to find a less costly solution, and trade off effectiveness for *efficiency*.

The challenge of balancing both efficiency and effectiveness motivated a line of research on *learning to efficiently rank*, leading to several innovations. In the area of ranking alone, for example, the

community widely adopted multi-stage rankers, separating lightweight ranking on large sets of documents from costly re-ranking of top candidates to speed up inference at the expense of quality [1, 4, 12, 13, 23, 30, 45]. From probabilistic data structures [2, 3], to cost-aware training and *post hoc* pruning of decision forests [5, 14, 28, 29], to early-exit strategies and fast inference algorithms [6, 10, 26, 27], the IR community thoroughly considered the practicality and scalability of complex ranking algorithms.

As complex neural network-based models come to dominate the research on ranking, there is renewed interest in this research area, with many recent proposals appearing as reincarnations of past ideas [16, 20, 22, 25, 31–35, 38, 39, 44, 46, 47, 49], alongside a series of novel approaches [15, 17, 18, 36, 41].

Despite these efforts, efficiency has always been taken to mean space- or time-efficiency, primarily in the context of online inference. But as Scells et al. [43] show through a comparison of a range of models from decision tree-based to language model-based rankers, complex neural models are energy-hungry, especially during training. What is more, these aspects of model evaluation are often conveniently ignored or under-reported—perhaps as an indirect result of effectiveness-driven competitions and leaderboards in IR [42].

Following this evolution of ideas, we propose this workshop as a forum for the discussion of efficient and effective models in NIR such as ranking and dense retrieval. In particular, we wish to promote the following notions and encourage the IR community to raise and debate questions on these themes:

- **Justification**: We believe it is important to justify the ever-growing model complexity through empirical analysis.
- **Training and inference efficiency**: We encourage the development of models that require less data or computational resources for training and fine-tuning, and that offer similarly fast inference. We also ask if there are meaningful simplifications of the existing training processes or model architectures that lead to comparable quality.
- **Evaluation and reporting**: We draw attention to the lessons learnt from past IR studies and encourage a multi-faceted evaluation of NIR models from quality to efficiency to environmental impact, and the design of reusable models, benchmarks, and standardized metrics.

Our specific objective for the second ReNeuIR workshop is to explore ideas for benchmarking frameworks and metrics that, in addition to effectiveness, are efficiency-aware and that paint a holistic picture of neural models in IR.

Finally, we believe that the ACM SIGIR conference is an appropriate venue for our proposed workshop. This gathering of IR researchers—who increasingly use and develop neural network-based models in their work—would help identify specific questions and challenges within this space, allowing us to collectively define future directions. We hope our forum will foster collaboration across interested groups.

2 RELATED WORKSHOPS

The ReNeuIR workshop debuted at ACM SIGIR 2022 as a hybrid event—in person in Madrid, Spain, with support for online attendees [8]. Recognizing the importance of the workshop, approximately 80 participants answered our call and attended the full-day event. The program included two keynote talks by Prof. Hamed Zamani of the University of Massachusetts Amherst, and Dr. Bhaskar Mitra of Microsoft Research—each 40 minutes long, including Q&A. Over three sessions of paper presentations—20 minutes each including Q&A—we learnt about the most recent works by authors affiliated with 8 research institutes and industry labs from 6 countries.[1] Our speakers presented their work both in person and online with the morning slots scheduled for remote Asia and Oceania speakers and the afternoon sessions catered for those presenting from the Americas.

The day concluded with a lively discussion where participants helped identify gaps in existing research and brainstormed future research directions. We had consensus in recognizing that efficiency is not simply latency; that a holistic, concrete definition of efficiency is needed to guide researchers and reviewers alike; and that more research is necessary in the development of efficiency-centered evaluation metrics and standard datasets, platforms, and tools. We refer the reader to the official workshop report for details [7].

We believe the first instance of the workshop proved instrumental in identifying and bringing together a community of researchers who are active in the space of efficiency in IR, thereby raising awareness of ongoing work and existing gaps. We therefore hope that, like in 2022, the second installment of ReNeuIR will first and foremost serve as a community building exercise, and as a forum to keep the existing community abreast of the progress made over the elapsed year. Furthermore, having identified under-researched areas in the neural efficiency space, we hope to prepare a program that is tailored to research that attempts to bridge those gaps.

3 TOPICS

With the objective of promoting the themes discussed in the preceding section and enabling a critical analysis and debate of each point, we solicit contributions on the following topics, including but not limited to:

- Novel NIR models that reach competitive quality but are designed to provide fast training or fast inference;
- Efficient NIR models for decentralized IR tasks such as conversational search;
- Strategies to speed up training or inference of existing NIR models;
- Sample-efficient training of NIR models;
- Efficiency-driven distillation, pruning, quantization, retraining, and transfer learning;
- Empirical investigation of the complexity of existing NIR models through an analysis of quality, interpretability, robustness, and environmental impact; and,
- Evaluation protocols for efficiency in NIR.

[1] Affiliations included: New York University, USA; University of Queensland, Australia; University of Pisa, Italy; Delft University of Technology, the Netherlands; Georgetown University, USA; Sorbonne Université, France; Amazon Music, Germany; and, Naver Labs Europe, France.

4 ORGANIZATION

Sebastian Bruch[2] completed his Ph.D. in Computer Science at the University of Maryland, College Park in 2013. His research since has centered around probabilistic data structures, streaming algorithms, and the application of machine learning to information retrieval with a particular focus on efficiency. He has published in and served on the program committees and senior program committees of premier IR and data mining conferences like SIGIR, WSDM, SIGKDD, and the Web Conference. He currently works at Pinecone in the United States as a staff research scientist.

Joel Mackenzie is a lecturer at the University of Queensland in Brisbane, Australia. He received his Ph.D. in Computer Science from RMIT University in 2019. His research focuses on efficient representations and algorithms for large-scale data analysis and retrieval. He is also interested in empirical experimentation, measurement, reproducibility, and user behavior analysis. He has co-authored over 30 papers in and acted as a program committee member for conferences and journals such as SIGIR, WSDM, WWW, CIKM, ECIR, EMNLP, TKDE, TOIS, and IPM. He was the Program Co-Chair for ADCS 2021 and 2022, an area chair for COLING 2022, and the Proceedings Chair for CHIIR 2021 and WSDM 2019.

Maria Maistro is a tenure track assistant professor at the Department of Computer Science, University of Copenhagen (DIKU). Prior to this, she was a Marie Curie fellow at DIKU. She conducts research in IR, on evaluation, reproducibility, click log analysis, and recommender systems, publishing papers at conferences as SIGIR, ECIR, RecSys, and CIKM, as well as journal articles in IRJ and TOIS. She has served as guest editor, member of programme committees, and reviewer for highly ranked conferences and journals in IR. She has co-organized and co-chaired several international scientific events among which: co-organizer of the TREC Health Misinformation Track (2019-2022), co-chair of ECIR reproducibility track in 2021, short paper track in 2023, workshop track in 2024, evaluation lab co-chair of CLEF 2021, programme co-chair at the Malawi Data Science Bootcamp 2021, co-chair of SIGIR 2022 tutorial track, and general co-chair of ACM and DDSA RecSys Summer School 2023.

Franco Maria Nardini[2] is a senior researcher with ISTI-CNR in Pisa, Italy. He received a Ph.D. in Information Engineering from the University of Pisa in 2011. His research interests are focused on Web Information Retrieval (IR), Machine Learning (ML), and Data Mining (DM). He authored more than 70 papers in peer-reviewed international journals, conferences, and other venues. In the past, he has been Tutorial Co-Chair of ACM WSDM 2021, Demo Papers Co-Chair of ECIR 2021, Program Committee Chair of the Italian Information Retrieval Workshop (IIR) in 2016, and General Chair of the International Workshop on Tourism Facilities (co-located with IEEE/WIC/ACM Web Intelligence) in 2012. He also participated in the organization of ACM SIGIR 2016 held in Pisa, Italy. Moreover, he participated in and coordinated activities in European and Italian national projects. He is a co-recipient of the ECIR 2022 Industry Impact Award, the ACM SIGIR 2015 Best Paper Award, and the ECIR 2014 Best Demo Paper Award. He is a member of the program committee of several top-level conferences in IR, ML, and DM, like SIGIR, ECIR, SIGKDD, CIKM, WSDM, IJCAI, and ECML-PKDD.

[2]Involved in the organization of ReNeuIR at SIGIR 2022.

REFERENCES

[1] N. Asadi. *Multi-Stage Search Architectures for Streaming Documents*. University of Maryland, 2013. Ph.D. Dissertation.

[2] N. Asadi and J. Lin. Fast candidate generation for two-phase document ranking: Postings list intersection with Bloom filters. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 2419–2422, 2012.

[3] N. Asadi and J. Lin. Fast candidate generation for real-time tweet search with Bloom filter chains. *ACM Trans. Inf. Syst.*, 31(3):13.1–13.36, 2013.

[4] N. Asadi and J. Lin. Effectiveness/efficiency tradeoffs for candidate generation in multi-stage retrieval architectures. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 997–1000, 2013.

[5] N. Asadi and J. Lin. Training efficient tree-based models for document ranking. In *Proceedings of the 35th European Conference on Information Retrieval*, pages 146–157, 2013.

[6] N. Asadi, J. Lin, and A. P. de Vries. Runtime optimizations for tree-based machine learning models. *IEEE Transactions on Knowledge and Data Engineering*, 26(9): 2281–2292, 2014.

[7] S. Bruch, C. Lucchese, and F. M. Nardini. Report on the 1st workshop on reaching efficiency in neural information retrieval (ReNeuIR 2022) at SIGIR 2022. *SIGIR Forum*, 56(2), 2022.

[8] S. Bruch, C. Lucchese, and F. M. Nardini. ReNeuIR: Reaching efficiency in neural information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3462–3465, 2022.

[9] C. J. Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010.

[10] B. B. Cambazoglu, H. Zaragoza, O. Chapelle, J. Chen, C. Liao, Z. Zheng, and J. Degenhardt. Early exit optimizations for additive machine learned ranking systems. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 411–420, 2010.

[11] O. Chapelle and Y. Chang. Yahoo! learning to rank challenge overview. In *Proceedings of the Learning to Rank Challenge*, Proceedings of Machine Learning Research, pages 1–24, 2011.

[12] J. S. Culpepper, C. L. Clarke, and J. Lin. Dynamic cutoff prediction in multi-stage retrieval systems. In *Proceedings of the 21st Australasian Document Computing Symposium*, pages 17–24, 2016.

[13] V. Dang, M. Bendersky, and W. B. Croft. Two-stage learning to rank for information retrieval. In *Proceedings of the 35th European Conference on Information Retrieval*, pages 423–434. 2013.

[14] D. Dato, C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. Fast ranking with additive ensembles of oblivious and non-oblivious regression trees. *ACM Trans. Inf. Syst.*, 35(2):15.1–15.31, 2016.

[15] L. Gao, Z. Dai, and J. Callan. Understanding BERT rankers under distillation. In *Proceedings of the 2020 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 149–152, 2020.

[16] M. Gordon, K. Duh, and N. Andrews. Compressing BERT: Studying the effects of weight pruning on transfer learning. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 143–155, July 2020.

[17] S. Hofstätter, H. Zamani, B. Mitra, N. Craswell, and A. Hanbury. Local Self-Attention over Long Text for Efficient Document Retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2021–2024, 2020.

[18] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

[19] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

[20] C. Lassance and S. Clinchant. An efficiency study for SPLADE models. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2220–2226, 2022.

[21] J. Lin, R. Nogueira, and A. Yates. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Morgan & Claypool Publishers, 2021.

[22] Z. Lin, J. Liu, Z. Yang, N. Hua, and D. Roth. Pruning redundant mappings in transformer models via spectral-normalized identity prior. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[23] S. Liu, F. Xiao, W. Ou, and L. Si. Cascade ranking for operational e-commerce search. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1557–1565, 2017.

[24] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[25] Z. Liu, F. Li, G. Li, and J. Cheng. EBERT: Efficient BERT inference with dynamic structured pruning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4814–4823, 2021.

[26] C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. Quickscorer: A fast algorithm to rank documents with additive ensembles of regression trees. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 73–82, 2015.

[27] C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. Exploiting CPU SIMD extensions to speed-up document scoring with tree ensembles. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 833–836, 2016.

[28] C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, F. Silvestri, and S. Trani. Post-learning optimization of tree ensembles for efficient ranking. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 949–952, 2016.

[29] C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, and S. Trani. X-DART: blending dropout and pruning for efficient learning to rank. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1077–1080, 2017.

[30] J. Mackenzie, J. S. Culpepper, R. Blanco, M. Crane, C. L. Clarke, and J. Lin. Query driven algorithm selection in early stage retrieval. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 396–404, 2018.

[31] J. Mackenzie, A. Mallia, A. Moffat, and M. Petri. Accelerating learned sparse indexes via term impact decomposition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022.

[32] J. Mackenzie, A. Trotman, and J. Lin. Efficient document-at-a-time and score-at-a-time query evaluation for learned sparse representations. *ACM Trans. Inf. Syst.*, 2022. Just Accepted.

[33] A. Mallia, J. Mackenzie, T. Suel, and N. Tonellotto. Faster learned sparse retrieval with guided traversal. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1901–1905, 2022.

[34] Y. Matsubara, T. Vu, and A. Moschitti. *Reranking for Efficient Transformer-Based Answer Selection*, pages 1577–1580. 2020.

[35] J. S. McCarley, R. Chakravarti, and A. Sil. Structured pruning of a BERT-based question answering model. *arXiv:1910.06360*, 2021.

[36] B. Mitra, S. Hofstätter, H. Zamani, and N. Craswell. *Improving Transformer-Kernel Ranking Model Using Conformer and Query Term Independence*, pages 1697–1702. 2021.

[37] R. Nogueira and K. Cho. Passage re-ranking with BERT. *arXiv:1901.04085*, 2020.

[38] R. Nogueira, W. Yang, K. Cho, and J. Lin. Multi-stage document ranking with BERT. *arXiv:1910.14424*, 2019.

[39] R. Nogueira, W. Yang, J. Lin, and K. Cho. Document expansion by query prediction. *arXiv:1904.08375*, 2019.

[40] R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, 2020.

[41] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108*, 2020.

[42] K. Santhanam, J. Saad-Falcon, M. Franz, O. Khattab, A. Sil, R. Florian, M. A. Sultan, S. Roukos, M. Zaharia, and C. Potts. Moving beyond downstream task accuracy for information retrieval benchmarking. *arXiv:2212.01340*, 2022.

[43] H. Scells, S. Zhuang, and G. Zuccon. Reduce, Reuse, Recycle: Green information retrieval research. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2825–2837, 2022.

[44] L. Soldaini and A. Moschitti. The cascade transformer: an application for efficient answer sentence selection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5697–5708, 2020.

[45] L. Wang, J. Lin, and D. Metzler. A cascade ranking model for efficient ranked retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 105–114, 2011.

[46] J. Xin, R. Tang, J. Lee, Y. Yu, and J. Lin. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[47] J. Xin, R. Tang, Y. Yu, and J. Lin. BERxiT: Early exiting for BERT with better fine-tuning and extension to regression. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 91–104, 2021.

[48] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. Bennett, J. Ahmed, and A. Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021.

[49] S. Zhuang and G. Zuccon. Fast passage re-ranking with contextualized exact term matching and efficient passage expansion. In *Workshop on Reaching Efficiency in Neural Information Retrieval, the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.