ReNeuIR at SIGIR 2024: The Third Workshop on Reaching Efficiency in Neural Information Retrieval

Maik Fröbe
Friedrich-Schiller-Universität Jena
Jena, Germany
maik.froebe@uni-jena.de

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

Bhaskar Mitra
Microsoft Research
Montréal, Canada
bmitra@microsoft.com

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

Martin Potthast
University of Kassel, hessian.AI, and ScaDS.AI
Kassel, Germany
martin.potthast@uni-kassel.de

ABSTRACT

The Information Retrieval (IR) community has a rich history of empirically measuring novel retrieval methods in terms of effectiveness and efficiency. However, as the search ecosystem is developing rapidly, comparatively little attention has been paid to evaluating efficiency in recent years, which raises the question of the cost-benefit ratio between effectiveness and efficiency. In this regard, it has become difficult to compare and contrast systems in an empirically fair way. Factors including hardware configurations, software versioning, experimental settings, and measurement methods all contribute to the difficulty of meaningfully comparing search systems, especially where efficiency is a key component of the evaluation. Furthermore, efficiency is no longer limited to time and space but has found new, challenging dimensions that stretch to resource, sample, and energy efficiency and have implications for users, researchers, and the environment. Examining algorithms and models through the lens of efficiency and its trade-off with effectiveness requires revisiting and establishing new standards and principles, from defining relevant concepts, to designing measures, to creating guidelines for making sense of the significance of findings. The third iteration of ReNeuIR aims to bring the community together to debate these questions and collaboratively test and improve a benchmarking framework for efficiency derived from the discussions of the first two iterations of this workshop. We provide a first prototype of this framework by organizing a shared task track focused on comparability and reproducibility at the workshop.

CCS CONCEPTS

• **Information systems → Search engine architectures and scalability**.

KEYWORDS

Efficiency, neural IR, sustainable IR, retrieval, ranking, algorithms

ACM Reference Format:
Maik Fröbe, Joel Mackenzie, Bhaskar Mitra, Franco Maria Nardini, and Martin Potthast. 2024. ReNeuIR at SIGIR 2024: The Third Workshop on Reaching Efficiency in Neural Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3626772.3657994

1 MOTIVATION AND THEME

We rely on a suite of algorithmic tools to help us find relevant information in a timely manner. In particular, search and recommendation systems help us to search for news articles, discover new movies or songs, find answers to our urgent questions and retrieve our emails with flight or hotel bookings, to name just a few examples. A key commonality in many of these search applications is the problem of *retrieving and ranking* relevant items from a large collection to provide maximum utility to the user in satisfying their information need.

In the last few decades, search has evolved from simple Boolean matching, to statistical ranking models, then to first-generation Learning to Rank systems [34], arriving at today's deep neural networks and large language models which have significantly advanced the state-of-the-art in ranking [31, 49, 50, 52].

Beyond ranking, deep learning methods have offered a range of other improvements to the field, such as dense retrieval methods [27, 61], document expansion techniques [51], and others. These recent developments mark the beginning of a new era known as Neural Information Retrieval (NIR), and retrieval-augmented generation (RAG) as a new way of combining retrieval technologies with generative AI [30]. Indeed, the transition from inexpensive statistical heuristics to more accurate—yet expensive—deep learning-based models has pervaded many domains of IR research. While this progression has enabled new frontiers in quality, it has done so with orders of magnitude more learnable parameters, requiring greater scales of data and, in turn, computational resources. The growing scale of retrieval approaches from decision forests to deep neural networks has dramatically increased the computational and environmental costs of model training and inference. This has lead the research community to question whether optimizing for effectiveness alone is the way to go or whether we need to trade effectiveness for efficiency and examine the impact of such a trade-off [55, 57, 63].

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657994

In the recent past, finding the delicate balance between efficiency and effectiveness in the context of learning to rank systems motivated a line of research on *learning to efficiently rank* [11], leading to several innovations. Multi-stage rankers, for example, were proposed to separate the light-weight retrieval and ranking of large sets of documents from the more costly re-ranking of top candidates to speed up inference at the expense of quality [1, 4, 17, 18, 33, 41, 58]. From probabilistic data structures [2, 3], to cost-aware training and *post hoc* pruning of decision forests [5, 19, 23, 37, 39, 48], to early-exit strategies and fast inference algorithms [6, 7, 14, 15, 29, 36, 38], the IR community thoroughly considered the practicality and scalability of complex ranking algorithms, arriving at solutions that provide competitive trade-offs on massive scales of data.

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 [24, 28, 32, 35, 42–46, 50, 51, 56, 59, 60, 62], alongside a series of novel approaches [22, 25, 26, 47, 53].

Despite these efforts, efficiency has typically been measured via mean space or time efficiency, primarily in the context of online inference. But as Scells et al. [55] show through a comparison of a range of models, complex neural models are energy-hungry, especially during training. Moreover, these aspects of model evaluation are often ignored or under-reported—perhaps as an indirect result of effectiveness-driven competitions and leaderboards in IR [54].

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, with the aim of fostering the IR community to discuss and debate these themes in the context of modern neural search systems. Our primary goal for the third ReNeuIR workshop is to implement the ideas and discussions of the previous two iterations into executing an efficiency-oriented shared task that enables the development of new evaluation measures that paint a holistic picture of neural models in IR by considering both efficiency and effectiveness. 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 define future directions collectively. 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 [9, 10]. The program included two keynote talks, three sessions of paper presentations, and a lively discussion tailored to identify gaps in existing research and brainstorm future directions. Within ReNeuIR 2022 we discussed—as a community—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 for the development of efficiency-centered evaluation metrics, datasets, platforms, and tools.

The second edition of ReNeuIR took place at ACM SIGIR 2023 as a hybrid event—in person in Taipei, Taiwan, with support for online attendees [12].Highlights from the program include a keynote talk, a session of paper presentations, and a joint poster session with two other SIGIR 2023 workshops, i.e., “Retrieval-enhanced Machine Learning” (REML) and “Generative Information Retrieval” (GenIR). The last session of ReNeuIR 2023 has been devoted to discussing an efficiency-oriented shared task to facilitate the development of a fair *efficiency-first* execution and measurement framework.

We believe the first two editions of the workshop helped to identify and unify a community of researchers who are active in the space of efficiency in IR, thereby raising awareness of ongoing work and existing gaps. The third edition of ReNeuIR will serve as a community-building exercise and a forum to keep the existing community abreast of the progress made over the elapsed year. Moreover, our specific objective for the third ReNeuIR workshop is to implement the ideas and discussions of the previous two iterations into the execution of an efficiency-oriented shared task that allows the development of new evaluation measures towards a more complete evaluation of neural models in IR that considers both efficiency and effectiveness. ReNeuIR 2024 will report on the results of the shared task, allowing the community to discuss these results and identify new challenges in this research area.

3 TOPICS

To promote the themes discussed in the preceding section and enable a critical analysis and debate of each point, we solicit contributions on the following topics, including but not limited to: 1) Novel NIR models that reach competitive quality but are designed to provide fast training or fast inference; 2) Efficient NIR models for decentralized IR tasks such as conversational search; 3) Strategies to speed up training or inference of NIR models; 4) Sample-efficient training of NIR models; 5) Efficiency-driven distillation, pruning, quantization, retraining, and transfer learning; 6) Empirical investigation of the complexity of existing NIR models through an analysis of quality, interpretability, robustness, and environmental impact; and, 7) Evaluation protocols for efficiency in NIR.

4 EFFICIENCY-ORIENTED IR SHARED TASK

The shared task aims to collect and measure NIR systems to foster the development of new IR measures that incorporate efficiency and effectiveness. The methodology of the shared task was developed in the first two iterations of the ReNeuIR workshop [9, 10, 12]. Therefore, a range of different options and their trade-offs were discussed [13] that were based on the assumptions that (1) the task uses a unified hardware/execution environment; (2) effectiveness evaluation is handled via existing evaluation tools and relevance judgments, and (3) submitted systems follow a standardized workflow. The rest of this section describes the methodology we will use in the first iteration of the shared task.

Retrieval pipelines are expected to (1) index, (2) retrieve, (3) and re-rank different workloads sampled from ir_datasets [40]. We monitor their execution with Scaphandre[1] and make all run files publicly available together with Scaphandre traces of their execution in a Kubernetes cluster using the ir_metadata [8] standard. To enable a comparable and reproducible evaluation of retrieval pipeline efficiency, we employ TIRA/TIREx [20, 21].[2]

[1] https://github.com/hubblo-org/scaphandre

[2] Overview of the infrastructure: https://webis.de/facilities.html

Within TIRA, we run the uploaded templates on different datasets derived from the MS MARCO passage dataset [16]. In doing so, we vary the number of queries and passages to be indexed to cover different workloads while monitoring their execution with Scaphandre. The execution happens within the TIRA sandbox, which ensures reproducibility (e.g., all dependencies must be installed in the Docker image), and keeps the test data (i.e., test queries and documents) secret, allowing different document and query distributions to be measured. This also reveals cases where the efficiency of the systems is highly dependent on the observed distribution. The reuse of MS MARCO lowers the barrier to entry, as a wide range of retrieval systems already exist for this benchmark, allowing participants to focus on efficiency. With this in mind, for the first iteration of the shared task in 2024, we focus on submitting pre-trained systems that batch-process the entire dataset. That is, we monitor and evaluate the inference parts of retrieval systems and focus on macro-benchmarking, as this type of batch processing is the standard in ad hoc retrieval research. This focus on macro-benchmarking of the inference part of retrieval pipelines distinguishes our efforts from related projects, e.g., the MLCommons AlgoPerf competition, which focuses on the training efficiency of ML algorithms,[3] or the micro-benchmarks as implemented by the nightly Lucene builds.[4]

To simplify participation, we have published public baselines along with submission instructions.[5] We prepare the software required so that participants can compare the efficiency measurements we perform in TIRA with measurements on their own hardware. We strongly encourage participants to monitor and submit traces of their software on their own hardware, with our focus on Docker helping to run the exact same retrieval system in a cloud setup. Therefore, Scaphandre is used for the efficiency measurements in TIRA as well as on the hardware of the participants, as Scaphandre supports Kubernetes (used in TIRA) as well as Docker (expected on the hardware of participants). The targeted audience of the shared task is researchers with efficient, highly optimized implementations of retrieval systems but also researchers using standard implementations that aim to achieve efficiency via conceptual changes (e.g., skipping computations).

5 ORGANIZATION

Maik Fröbe is a PhD student at the Webis group with research interests in information retrieval. He has co-organized the Touché shared task since 2020 and the SCAI shared task since 2021 and was the main organizer of task 5 at SemEval-2023. He is an active developer of TIRA [20]/TIREx [21], which improved the reproducibility of a number of shared tasks and has an archive of more than 500 research prototypes and is increasingly used in teaching initiatives.

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 40 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, the Proceedings Chair for SIGIR-AP 2023, CHIIR 2021, WSDM 2019, and the Demo chair for SIGIR 2024.

Bhaskar Mitra is a Principal Researcher at Microsoft Research. His research focuses on AI-mediated information and knowledge access and questions of fairness and ethics in the context of these socio-technical systems. Before joining Microsoft Research, he worked on search technologies at Bing for 15+ years. He is serving as the ACM SIGIR Community Relations Coordinator and on the NIST TREC program committee. He co-organized several workshops (Neu-IR @ SIGIR 2016-2017, HIPstIR 2019, and Search Futures @ ECIR 2024), shared evaluation tasks (TREC Deep Learning Track 2019-2023, TREC Tip-of-the-Tongue Track 2023-2024, and MS MARCO ranking leaderboards), and tutorials (WSDM 2017-2018, SIGIR 2017, ECIR 2018, and AFIRM 2019-2020). He also served as a guest editor for the special issue of the Information Retrieval Journal on neural information retrieval in 2017, a Virtual Organization Co-Chair for SIGIR 2021, a DEI Co-Chair for SIGIR 2021, the D&I Scholarship Chair for CIKM 2021, a PhD Symposium Co-Chair for CIKM 2022, a Tutorials Co-Chair for ECIR 2023, a Senior Area Co-Chair for EMNLP 2023, and the Industry Chair for FIRE 2023.

Franco Maria Nardini is a Senior Researcher with ISTI-CNR in Pisa, Italy. His research interests focus on Web Information Retrieval, Machine Learning, and Data Mining. He authored over 100 papers in peer-reviewed international journals, conferences, and other venues. He has been Program Committee Co-Chair of SPIRE 2023, Tutorial Co-Chair of ACM WSDM 2021, Demo Papers Co-Chair of ECIR 2021, Program Committee Co-Chair, and General Co-Chair of ReNeuIR at SIGIR (2022, 2023). 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 editorial board of ACM TOIS and a PC member of SIGIR, ECIR, SIGKDD, CIKM, WSDM, IJCAI, and ECML-PKDD.

Martin Potthast is professor at the University of Kassel, Germany. In his research on information retrieval and natural language processing, he has put a special focus on the reproducibility of experimental evaluations. With TIRA, he has introduced and presented the first working prototype of a cloud-based evaluation framework that implements the evaluation-as-a-service paradigm, enabling the submission of working software instead of merely software runs while minimzing organizer overhead. Martin is co-initiator of the PAN network for digital text forensics, hosted at the CLEF conference, where he has organized annual shared tasks since 2009. Since 2012, TIRA has been used as the exclusive submission system. Besides PAN, Martin has co-initiated and co-organized several shared task events at various editions of WSDM, MediaEval, SemEval, CoNLL, and INLG, and most recently the annual Touché-Lab at CLEF on argument retrieval. In his research, Martin has published regularly at major IR and ACL conferences over the years.

ACKNOWLEDGMENTS

This work has been partially supported by the OpenWebSearch.eu project (funded by the EU; GA 101070014) and by the "Extreme Food Risk Analytics" (EFRA) project (funded by the EU; GA 101093026).

[3] https://mlcommons.org/2023/11/mlc-algoperf-training-algorithms-competition/
[4] http://home.apache.org/~mikemccand/lucenebench/
[5] https://github.com/reneuir/reneuir-code

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 *Proc. CIKM*, 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 *Proc. SIGIR*, pages 997–1000, 2013.

[5] N. Asadi and J. Lin. Training efficient tree-based models for document ranking. In *Proc. ECIR*, pages 146–157, 2013.

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

[7] L. Beretta, F. M. Nardini, R. Trani, and R. Venturini. An optimal algorithm for finding champions in tournament graphs. *IEEE Trans. Knowl. Data Eng*, 35(10): 10197–10209, 2023.

[8] T. Breuer, J. Keller, and P. Schaer. ir_metadata: An extensible metadata schema for IR experiments. In *Proc. SIGIR*, pages 3078–3089, 2022.

[9] 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.

[10] S. Bruch, C. Lucchese, and F. M. Nardini. ReNeuIR: Reaching efficiency in neural information retrieval. In *Proc. SIGIR*, pages 3462–3465, 2022.

[11] S. Bruch, C. Lucchese, and F. M. Nardini. Efficient and effective tree-based and neural learning to rank. *Found. Trnd. Inf. Retr.*, 17(1):1–123, 2023.

[12] S. Bruch, J. Mackenzie, M. Maistro, and F. M. Nardini. Reneuir at SIGIR 2023: The second workshop on reaching efficiency in neural information retrieval. In *Proc. SIGIR*, pages 3456–3459, 2023.

[13] S. Bruch, J. Mackenzie, M. Maistro, and F. M. Nardini. A proposed efficiency benchmark for modern information retrieval systems. 2023. URL https://reneuir.org/assets/pdfs/ReNeuIR_2023_benchmark_proposal.pdf.

[14] F. Busolin, C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, and S. Trani. Early exit strategies for learning-to-rank cascades. *IEEE Access*, 11:126691–126704, 2023.

[15] 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 *Proc. WSDM*, pages 411–420, 2010.

[16] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, and J. Lin. MS MARCO: Benchmarking ranking models in the large-data regime. In *Proc. SIGIR*, pages 1566–1576, 2021.

[17] J. S. Culpepper, C. L. Clarke, and J. Lin. Dynamic cutoff prediction in multi-stage retrieval systems. In *Proc. ADCS*, pages 17–24, 2016.

[18] V. Dang, M. Bendersky, and W. B. Croft. Two-stage learning to rank for information retrieval. In *Proc. ECIR*, pages 423–434. 2013.

[19] 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.

[20] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, and M. Potthast. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Proc. ECIR*, pages 236–241.

[21] M. Fröbe, J. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, and M. Potthast. The Information Retrieval Experiment Platform. In *Proc. SIGIR*, pages 2826–2836, July 2023.

[22] L. Gao, Z. Dai, and J. Callan. Understanding BERT rankers under distillation. In *Proc. ICTIR*, pages 149–152, 2020.

[23] A. Gigli, C. Lucchese, F. M. Nardini, and R. Perego. Fast feature selection for learning to rank. In *Proc. ICTIR*, page 167–170, 2016.

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

[25] S. Hofstätter, H. Zamani, B. Mitra, N. Craswell, and A. Hanbury. Local Self-Attention over Long Text for Efficient Document Retrieval. In *Proc. SIGIR*, pages 2021–2024, 2020.

[26] 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 *Proc. EMNLP Findings*.

[27] 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 *Proc. EMNLP*, 2020.

[28] C. Lassance and S. Clinchant. An efficiency study for SPLADE models. In *Proc. SIGIR*, pages 2220–2226, 2022.

[29] F. Lettich, C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. Parallel traversal of large ensembles of decision trees. *IEEE Trans. on Par. Dist. Sys.*, 30(9):2075–2089, 2019.

[30] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. NeurIPS*, 2020.

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

[32] Z. Lin, J. Liu, Z. Yang, N. Hua, and D. Roth. Pruning redundant mappings in transformer models via spectral-normalized identity prior. In *Proc. EMNLP Findings*, 2020.

[33] S. Liu, F. Xiao, W. Ou, and L. Si. Cascade ranking for operational e-commerce search. In *Proc. SIGKDD*, pages 1557–1565, 2017.

[34] T.-Y. Liu. Learning to rank for information retrieval. *Found. Trnd. Inf. Retr*, 3(3): 225–331, 2009.

[35] Z. Liu, F. Li, G. Li, and J. Cheng. EBERT: Efficient BERT inference with dynamic structured pruning. In *Proc. ACL-IJCNLP Findings*, pages 4814–4823, 2021.

[36] 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 *Proc. SIGIR*, pages 73–82, 2015.

[37] C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, F. Silvestri, and S. Trani. Post-learning optimization of tree ensembles for efficient ranking. In *Proce. SIGIR*, pages 949–952, 2016.

[38] 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 *Proc. SIGIR*, pages 833–836, 2016.

[39] C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, and S. Trani. X-DART: blending dropout and pruning for efficient learning to rank. In *Proc. SIGIR*, pages 1077–1080, 2017.

[40] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, and N. Goharian. Simplified data wrangling with ir_datasets. In *Proc. SIGIR*, pages 2429–2436, 2021.

[41] J. Mackenzie, J. S. Culpepper, R. Blanco, M. Crane, C. L. Clarke, and J. Lin. Query driven algorithm selection in early stage retrieval. In *Proc. WSDM*, pages 396–404, 2018.

[42] J. Mackenzie, A. Mallia, A. Moffat, and M. Petri. Accelerating learned sparse indexes via term impact decomposition. In *Proc. EMNLP Findings*, 2022.

[43] 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.*, 41(4), 2023.

[44] A. Mallia, J. Mackenzie, T. Suel, and N. Tonellotto. Faster learned sparse retrieval with guided traversal. In *Proc. SIGIR*, pages 1901–1905, 2022.

[45] Y. Matsubara, T. Vu, and A. Moschitti. Reranking for efficient transformer-based answer selection. In *Proc. SIGIR*, pages 1577–1580, 2020.

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

[47] B. Mitra, S. Hofstätter, H. Zamani, and N. Craswell. Improving transformer-kernel ranking model using conformer and query term independence. In *Proc. SIGIR*, pages 1697–1702, 2021.

[48] F. M. Nardini, C. Rulli, S. Trani, and R. Venturini. Distilled neural networks for efficient learning to rank. *IEEE Trans. Knowl. Data Eng.*, 35(5):4695–4712, 2023.

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

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

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

[52] R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin. Document ranking with a pretrained sequence-to-sequence model. In *Proc. EMNLP Findings*, pages 708–718, 2020.

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

[54] 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.

[55] H. Scells, S. Zhuang, and G. Zuccon. Reduce, Reuse, Recycle: Green information retrieval research. In *Proc. SIGIR*, pages 2825–2837, 2022.

[56] L. Soldaini and A. Moschitti. The cascade transformer: an application for efficient answer sentence selection. In *Proc. ACL*, pages 5697–5708, 2020.

[57] E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in NLP. In *Proc. ACL*, pages 3645–3650, 2019.

[58] L. Wang, J. Lin, and D. Metzler. A cascade ranking model for efficient ranked retrieval. In *Proc. SIGIR*, pages 105–114, 2011.

[59] J. Xin, R. Tang, J. Lee, Y. Yu, and J. Lin. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proc. ACL*, 2020.

[60] J. Xin, R. Tang, Y. Yu, and J. Lin. BERxiT: Early exiting for BERT with better fine-tuning and extension to regression. In *Proc. EACL*, pages 91–104, 2021.

[61] 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 *Proc. ICLR*, 2021.

[62] S. Zhuang and G. Zuccon. Fast passage re-ranking with contextualized exact term matching and efficient passage expansion. In *Proc. Workshop on ReNeuIR at SIGIR*, 2022.

[63] G. Zuccon, H. Scells, and S. Zhuang. Beyond CO2 emissions: The overlooked impact of water consumption of information retrieval models. In *Proc. ICTIR*, pages 283–289, 2023.