EVENT REPORT

Report on the 1st Workshop on Reaching Efficiency in Neural Information Retrieval (ReNeuIR 2022) at SIGIR 2022

Sebastian Bruch
Pinecone
USA
sbruch@acm.org

Claudio Lucchese
Ca' Foscary University of Venice
Italy
claudio.lucchese@unive.it

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

Abstract

As Information Retrieval (IR) researchers, we not only develop algorithmic solutions to hard problems, but we also insist on a proper, multifaceted evaluation of ideas. The IR literature on the fundamental topic of retrieval and ranking, for instance, 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. This is evidenced, for example, by more than a decade of research on efficient training and inference of large decision forest models in Learning to Rank (LtR). As we move towards even more complex, deep learning models in a wide range of applications, questions on efficiency have once again become relevant with renewed urgency. Indeed, efficiency is no longer limited to time- and space-efficiency; instead it has found new, challenging dimensions that stretch to resource-, sample- and energy-efficiency with ramifications for researchers, users, and the environment.

As a step towards bringing together experts from industry and academia and creating a forum for a critical discussion and the promotion of efficiency in the era of Neural Information Retrieval (NIR), we held the ReNeuIR workshop on July 15th, 2022 as a hybrid event—in person in Madrid, Spain along with online attendees—in conjunction with ACM SIGIR 2022. Recognizing the importance of this topic, approximately 80 participants answered our call and attended the workshop over three sessions. The event included a total of two keynotes and eight paper presentations, and 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 benchmark datasets, platforms, and tools.

Date: 15 July, 2022.

Website: https://ReNeuIR.org.

1 Introduction

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 information retrieval—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).

Consider Web search, and in particular text ranking. In text ranking, a user provides a query q as an expression of an intent and information need in the form of keywords or in natural language. The task is to sift through a large collection of documents $\mathcal{D}$, find a subset that is relevant to the query, and order the resulting set in some order that maximizes an application-dependent user satisfaction metric Q.

Over a decade ago, machine learning transformed how we approach the text ranking problem. That wave resulted in a paradigm shift from early statistical methods, heuristics, and hand-crafted rules to determine the relevance of documents to a query, to what would later be called LtR [Liu, 2009]. This leap was perhaps best exemplified by LambdaMART [Burges, 2010] in the Yahoo! Learning-to-Rank Challenge [Chapelle and Chang, 2011].

A decade later, deep neural networks, and in particular, Transformer-based [Vaswani et al., 2017] models advanced the state-of-the-art in ranking dramatically [Lin et al., 2021; Nogueira and Cho, 2020; Nogueira et al., 2019a, 2020]. Learnt representations of queries and documents by deep networks, too, offer a range of opportunities including the development of a new generation of retrieval methods [Karpukhin et al., 2020; Xiong et al., 2021], document expansion techniques [Nogueira et al., 2019b], and others. These recent developments mark the beginning of a new era known as NIR.

This march from inexpensive statistical methods to complex, expensive machine learning models is not unique to the text ranking problem and can be seen across IR research. While this progression enabled reaching new peaks in quality and effectiveness, it has done so with orders of magnitude more learnable parameters using much 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 above motivated a line of research to systematically investigate questions of efficiency and explore the trade-offs between efficiency and effectiveness, leading to several innovations in the early days. In the area of ranking alone, for example, the community widely adopted multi-stage rankers, separating light-weight ranking on large sets of documents from costly re-ranking of top candidates to speed up inference at the expense of quality [Wang et al., 2011; Asadi and Lin, 2013b; Dang et al., 2013; Culpepper et al., 2016; Mackenzie et al., 2018; Liu et al., 2017; Asadi, 2013]. From probabilistic data structures [Asadi and Lin, 2012, 2013a], to cost-aware training and *post hoc* pruning of decision forests [Asadi and Lin, 2013c; Lucchese et al., 2017, 2016a; Dato et al., 2016], to early-exit strategies and fast inference algorithms [Cambazoglu et al., 2010; Asadi et al., 2014; Lucchese et al., 2016b, 2015], the information retrieval community thoroughly considered the practicality and scalability of complex ranking algorithms.

As complex neural network-based models come to dominate the research on ranking, it is unsurprising that there is renewed interest in the question above, with many of the proposals put forward to date to tame efficiency being reincarnations of past ideas [Nogueira et al., 2019a; Matsubara et al., 2020; Soldaini and Moschitti, 2020; Xin et al., 2020, 2021; Gordon et al., 2020; McCarley et al., 2021; Lin et al., 2020; Liu et al., 2021; Zhuang and Zuccon, 2022; Nogueira et al., 2019b; Mallia et al., 2022; Lassance and Clinchant, 2022] and a few novel ideas [Jiao et al., 2020; Sanh et al., 2020; Gao et al., 2020; Mitra et al., 2021; Hofstätter et al., 2020].

Despite these efforts, efficiency has always been taken to mean space- or time-efficiency, primarily in the context of inference. But as Scells et al. [2022] show through a comparison of a range of models from decision tree-based to Transformer-based rankers, complex neural models are energy-hungry, especially during training. This increased energy consumption along with the need for larger and larger datasets present new challenges to the IR community, especially considering the environmental impact of this research.

We organized the 1^{st} Workshop on Reaching Efficiency in Neural Information Retrieval (ReNeuIR 2022), held jointly with ACM SIGIR 2022 as a hybrid event, as a community building exercise and to draw attention to the challenges before us, identify gaps in existing research, and find new research directions. We argue that as IR researchers, we must develop better theoretical frameworks and practical tools to conduct a holistic evaluation of NIR systems, and standardize the interpretation and understanding of the Pareto front in the space of effectiveness and efficiency.

2 The Workshop

The workshop's call for paper included short and full papers—up to 9 pages long—and accepted both original works as well as re-submissions. We received a total of 8 valid submissions, each of which was reviewed by at least three members of the program committee in a single-blind peer-review process. In the end, the program committee accepted 5 papers. The workshop chairs later identified and invited the authors of 3 additional works from the ACM SIGIR 2022 proceedings that topically complemented the set of accepted papers, to also present their work at the workshop. We took care to select works that gave every research group who is active in this area a chance to present their work and perspective. In the end, the workshop included 8 paper presentations with authors affiliated with 8 research institutes and industry labs from 6 countries.[1] In addition to paper presentations, the ReNeuIR program included two keynote talks by Hamed Zamani and Bhaskar Mitra, as well as a discussion led by a panel of experts from academia and industry including Bhaskar Mitra, Nils Reimers, Grace Hui Yang, and Guido Zuccon.

As with the main SIGIR conference, the full-day workshop was held as a hybrid event with in-person and online attendees. Over three sessions of paper presentations—20 minutes each including Q&A—two keynote talks—each 40 minutes long, including Q&A—and an hour-long panel discussion in the end, we estimate that about 80 participants were present in the event's

[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.

physical and virtual rooms combined. Our speakers, too, 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 details of the program was made available online prior to the event and was later updated with links to papers and slides from the workshop. [2]

2.1 Program Committee

ReNeuIR was made possible by seven researchers who volunteered their time to review the submissions. We have listed the members of the program committee below. We thank each member for their time and commitment to the workshop.

- **Claudia Hauff**, Delft University of Technology
- **Amir Ingber**, Pinecone
- **Sean MacAvaney**, University of Glasgow
- **Bhaskar Mitra**, Microsoft
- **Tommaso di Noia**, Polytechnic University of Bari
- **Hamed Zamani**, University of Massachusetts Amherst
- **Min Zhang**, Tsinghua University

2.2 Keynotes

ReNeuIR featured two invited keynote speakers. We present the title and summary of each talk below along with a short bio of each speaker.

2.2.1 Efficient Neural Models for Representing, Indexing, and Retrieving Documents

The first keynote talk was delivered by **Hamed Zamani**. Hamed is an Assistant Professor in the Manning College of Information and Computer Sciences (CICS) at the University of Massachusetts Amherst (UMass), where he also serves as the Associate Director of the Center for Intelligent Information Retrieval (CIIR). Prior to UMass, he was a Researcher at Microsoft. In 2019, he received his Ph.D. from UMass under supervision of W. Bruce Croft and received the CICS Outstanding Dissertation Award. His research focuses on developing and evaluating statistical and machine learning models with application to (interactive) information access systems including search engines, recommender systems, and question answering. He is an active member of the information retrieval community and has published over 75 peer-reviewed articles. He is mostly known for his work on neural information retrieval and conversational search. He is a recipient of NSF CAREER Award and his papers have received awards from ICTIR 2019 and CIKM 2020. He has organized multiple workshops at SIGIR, WSDM, and RecSys and has served as the Program Committee Co-Chair for SIGIR 2022 - Short Paper Track.

Hamed reviewed in detail how deep learning has transformed IR research, and how their millions or billions of parameters present unique challenges to efficiency. Hamed then discussed recent

[2] https://ReNeuIR.org

methods that aim to improve the efficiency of NIR models by designing an efficient representation of long documents, efficient indexing, and efficient retrieval of documents. The talk concluded with a note on the importance of research on efficient NIR by drawing attention to a recent study on the role of retrieval systems in improving machine learning [Zamani et al., 2022].

2.2.2 Efficient Machine Learning and Machine Learning for Efficiency in Information Retrieval

Our second speaker was **Bhaskar Mitra**, who delivered the keynote talk remotely. Bhaskar is a Principal Researcher at Microsoft Research based in Montreal, Canada. He received a Ph.D. in Computer Science from University College London under the supervision of Dr. Emine Yilmaz. His research interests are at the intersection of information retrieval, deep learning, and FATE (Fairness, Accountability, Transparency, and Ethics). He joined Microsoft in 2006 and in his 15+ years at Microsoft, he shipped several search quality improvements for Bing and conducted research with strong focus on both academic and product impact. He co-organized the Neural IR (Neu-IR) Workshop in 2016 and 2017 which was the first to attempt to bring together a community of IR researchers interested in deep learning methods, and since then have influenced the research vision for the community through development of the MS MARCO benchmark, co-founding the TREC Deep Learning Track, co-authoring a book on the topic of neural information retrieval, serving as a guest editor for the special issue of the Information Retrieval Journal, co-organizing multiple tutorials on the same topic, and through his own research.

Bhaskar's talk centered on the fact that emerging machine learning methods for IR have made significant improvements in accuracy but at the cost of increasing model complexity, leading to a rise in computational and environmental costs. Bhaskar notes that "In web search, these costs are further compounded by the necessity to train on large-scale datasets, consume long documents as inputs, and retrieve relevant documents from Web-scale collections within milliseconds in response to high volume query traffic." He further observes that "A typical playbook for developing deep learning models for IR involves largely ignoring efficiency concerns during model development and then later scaling these methods by either finding faster approximations of the same models or employing heuristics to reduce the input space over which these models operate. Domain knowledge about the specific IR task and deeper understanding of system design and data structures in whose context these models are deployed can significantly help with not only model simplification but also to inform data-structure specific machine learning model design. Alternatively, predictive machine learning can also be employed specifically to improve efficiency in large scale IR settings." In this talk, Bhaskar reviewed several case studies on improving the efficiency of NIR models, including how machine learning methods themselves can help improve retrieval efficiency. Bhaskar highlighted potential future directions for efficiency-sensitive benchmarking of NIR solutions.

2.3 Papers

As noted earlier, the program included 8 paper presentations. In the subsections that follow, we categorize them by topic and present their titles and authors along with a summary of their contributions.

2.3.1 Time and Space Efficiency in Sparse Retrieval

One theme that was prominent among the works presented at ReNeuIR was on the important topic of *sparse representations* and their use in retrieval and ranking, with an emphasis on time- and space-efficiency. The crux of the idea is to learn *sparse* representations in a space that has as many dimensions as there are terms in the vocabulary, where each coordinate encodes the "importance" of the corresponding term in the context of a query or document. Such a representation enables us to leverage traditional inverted index-based algorithms for efficient retrieval.

In the papers presented at ReNeuIR, we learnt about: TILDEv2 [Zhuang and Zuccon, 2022] that offers fast inference on commodity hardware (CPU) by relying on an inexpensive query tokenizer and thereby removing the need for costly inference of a Transformer-based model; an adaptation of inverted indexes to sparse representations and a novel query processing algorithm for efficient sparse retrieval [Mallia et al., 2022]; methods to make SPLADE models more effective and efficient [Lassance and Clinchant, 2022; Formal et al., 2022]; and, a general indexing framework for sparse, interaction-based neural retrieval systems [Dong et al., 2022]. We review the five papers that fall into this category in the remainder of this section in the order in which they were presented.

Guido Zuccon presented his joint work with Shengyao Zhuang—both of the University of Queensland—entitled "Fast Passage Re-ranking with Contextualized Exact Term Matching and Efficient Passage Expansion." Their work builds on TILDE [Zhuang and Zuccon, 2021] which, for every passage, stores in the index the "importance" scores for the entire BERT vocabulary; in other words, every passage has a posting in the postings list of every term in the BERT vocabulary. While the index may be inflated, this representation allows one to circumvent inference of the underlying BERT model at query time, leading to a substantial cut in query latency, albeit at the cost of ranking quality. In this work, however, Shengyao and Guido show that efficiency and effectiveness need not always compete: By expanding each passage and storing in the index only the part of the BERT vocabulary that is present in the expanded passage, they reduce TILDE's memory footprint by 99% and, at the same time, improve quality.

Joel Mackenzie (University of Queensland) was invited to present "Faster Learned Sparse Retrieval with Guided Traversal" [Mallia et al., 2022], a joint work with Antonio Mallia (NYU), Torsten Suel (NYU), and Nicola Tonellotto (University of Pisa), that was published in ACM SIGIR 2022. Their work is motivated by the observation that sparse retrieval is typically slower than more traditional statistics-based methods such as BM25, despite tht fact that they both use inverted indices. By examining scores from a state-of-the-art sparse retrieval model, they show that many more terms end up with a high impact score than vanilla BM25, which suggests that score distributions produced by sparse retrieval models may be less friendly to dynamic pruning algorithms developed for inverted index traversal over the past decades. Their work then presents a heuristic that uses BM25 scores to decide which documents to process, but uses the sparse retrieval model to compute the document scores. This work is notable in that it shows the utility of traditional, statistical-based methods of retrieval in achieving better efficiency in sparse retrieval.

In the second session, **Thibault Formal** (Naver Labs Europe) was invited to present "From Distillation to Hard Negative Sampling: Making Sparse Neural IR Models More Effective" [Formal et al., 2022], a joint work with Carlos Lassance (Naver Labs Europe), Benjamin Piwowarski

(Sorbonne Université), and Stéphane Clinchant (Naver Labs Europe). Thibault showed us that by taking the latest innovations (such as distillation and better negative sampling strategies) from the literature on the training of dense retrieval models and adopting them to the sparse retrieval setting, we can improve the effectiveness of SPLADE. What is notable in their observations is that many of the gains from the dense retrieval techniques proved to be additive, leading to better sparse retrieval models. In their work, Thibault et al. also conducted experiments that studied the trade-offs between effectiveness and time-efficiency, which served as a segue into the next paper.

Carlos Lassance also presented another orthogonal but related work on "An Efficiency Study for SPLADE Models" [Lassance and Clinchant, 2022], a joint work with Stéphane Clinchant that was also published as a Short Paper in the proceedings of ACM SIGIR 2022. Having noticed that enforcing sparsity through existing regularization does not make SPLADE sufficiently time-efficient (using only a single CPU core) compared to a traditional retrieval measure such as BM25, they propose a suite of techniques to address that gap. As an important step, they modify SPLADE so that it uses separate encoders to project queries and documents to the embedding space. By imposing further sparsity on the query representations via an L1 regularization and the use of other techniques, they show that SPLADE can reach time-efficiency that is on a par BM25-based retrieval with little degredation in quality.

Finally, **Grace Hui Yang** presented a joint work with Sibo Dong and Justin Goldstein—all of the University of Georgetown—entitled "SEINE: SEgment-based Indexing for NEural information retrieval" [Dong et al., 2022]. This work is notable in that it promotes *reusability*—a theme we will return to in a later paper presentation by Guido and Harry: They note that unlike traditional term-level inverted indices, an index formed by a (dense or sparse) representation-based model cannot be easily reused by another retrieval method. They then propose a general indexing framework that is flexible-enough to support a variety of neural retrieval models, and show that through a better structuring of the index they can speed up query processing during inference.

2.3.2 Sample Efficiency

The second category of papers can be roughly described as an investigation of sample-efficiency in LtR. **Alexander Buchholz** presented "Low-variance estimation in the Plackett-Luce model via quasi-Monte Carlo sampling," [Buchholz et al., 2022], a joint work with Jan Malte Lichtenberg, Giuseppe Di Benedetto, Yannik Stein, Vito Bellini, and Matteo Ruffini—all of Amazon Music. When optimizing the Plackett-Luce model to learn a ranking function, we often need to estimate an expectation, that does not have a closed form, by sampling from its underlying distribution. This work shows that using quasi Monte Carlo sampling—instead of the vanilla Monte Carlo method—leads to a low-variance estimation of the expectation term. This, in turn, makes the learning algorithm more sample-efficient, another important category of efficiency.

2.3.3 Resource Efficiency

In their study entitled "Moving stuff around: a study on efficiency of moving documents into memory for Neural IR models", [Câmara and Hauf, 2022] **Arthur Câmara** and Claudia Hauff, both of the Delft University of Technology, investigate how better data handling and transfer between disk, memory (RAM), and video memory (VRAM) affects efficiency of training complex NIR models. In particular, they study the impact on speed and memory footprint of three ways of moving documents between storage and different layers of memory, and how those methods scale with multiple GPUs. Interestingly, they show that loading all documents into memory is not always the fastest or most scalable method. Additionally, they consider popular techniques for improving data loading times and how they lead to further reductions in model training time.

2.3.4 Energy Efficiency and Environmental Impact

Finally, in a creative format, **Guido Zuccon** and **Harry Scells** jointly presented their ACM SIGIR 2022 Perspective Paper entitled "Reduce, Reuse, Recycle: Green Information Retrieval Research," [Scells et al., 2022] a work co-authored by Shengyao Zhuang. This work, which was the third and last paper invited to the workshop, defines its own category. The authors extensively study a multitude of retrieval and ranking models, ranging from bag-of-words to decision forests and NIR, from the lens of effectiveness, training and inference time-efficiency, as well as *energy* consumption. They present methods to calculate the energy consumption of models in popular programming and computing frameworks, and ultimately quantify the environmental impact of model training and inference in terms of emissions. In their paper, Harry, Shengyao, and Guido also provide a framework to guide what they refer to as "Green Information Retrieval" research based on waste management principles: 'reduce, reuse, recycle.'

They conclude their work by noting that as a community "we must be mindful of the potential costs that our research may have. The ways that we measure and address the environmental impact of our research are just one of the many brushstrokes that coalesce into a larger landscape that portray our impacts on society at large." They observe that the IR community "is at a turning point in terms of the types of deep learning models used, the scale of those models, and how those models are trained. While the investigation into and development of such models are valuable research goals, we believe that it is important to be mindful of the costs and environmental impacts of these techniques." [Scells et al., 2022]

2.4 Panel Discussion

ReNeuIR concluded with a lively discussion led by our panelists and great engagement from the audience. ReNeuIR attendees learnt about and organized the challenges that we face together, and used this opportunity to brainstorm high-level research directions for the coming year.

We explored as a group what efficiency actually means ("low latency" and "efficiency" are not interchangeable) and realized, because it encompasses a multitude of factors, that its definition itself is not well-understood. This observation encourages us to work together and, as a stepping stone in shaping future research directions, define the notion of efficiency more comprehensively and concretely. That has become one of the objectives of this newly-formed community.

Through debate, we also observed that it is difficult to judge the merits of a work that improves certain dimensions of efficiency, but not others, and at the cost of effectiveness. This multidimensional view of efficiency requires a new way of thinking about evaluation and interpreting empirical results. This highlights the importance of research on designing efficiency-oriented evaluation metrics and providing clear guidelines to the research community to aid in understanding empirical findings.

It also became clear that, even if we had the right tools to measure efficiency and the right mindset to interpret our measurements, it is unclear if the existing datasets and benchmarking platforms are adequate for an evaluation of efficiency of arbitrary algorithms and models. The need to prepare appropriate datasets and standardized platforms became another research objective of this community, especially for our experts from industry.

3 Summary

As stated in our proposal [Bruch et al., 2022], we organized this workshop with the express purpose of forming a focus group within the IR community to examine the current state of efficiency in IR research. In that way, ReNeuIR was successful. This event drew experts with a variety of perspectives on efficiency—with some even challenging our preconceived notion of what it means for a system to be efficient—and served as an opportunity for everyone in this sub-community to learn about the others' ideas and ongoing works.

In the discussion before the closing of the workshop, we debated the open questions and challenges we face, and brainstormed how we may collectively approach them. For instance, we broadly agreed that the definition of efficiency itself needs to be revisited, and so defining it more concretely became one of our objectives for the year ahead. We understood that assessing the impact of novel methods is made difficult by the various, often competing effectiveness and efficiency factors (e.g., as a reviewer, how do we understand and evaluate the novelty of a method that leads to a loss in quality but an improvement in one dimension of efficiency?). We have also recognized the paucity of standard evaluation protocols and benchmark datasets for efficiency-oriented studies. In helping us identify and make precise some of these questions and set a research agenda for the future, too, ReNeuIR was successful.

We would be remiss, however, if we did not mention the challenge we faced in gathering a more diverse group of speakers and presenters. We believe this was partly due to the hybrid nature of the event and the difficulty of accommodating schedules across time zones. We also think, in the absence of an established community, it proved difficult to identify the research groups who are actively pursuing efficiency-centered work and who have a vested interest in the topic; our recruitment can best be described as a shot in the dark. Having identified the root causes, however, and with the lessons learnt and the newly-formed network of researchers, we believe we have all the tools to take steps towards creating a more diverse event in the future, and to encourage more researchers to investigate efficiency in NIR and, more generally, in IR. We look forward to the year ahead.

Acknowledgments

We are grateful to the SIGIR Workshop Chairs, Julián Urbano and Johanne Trippas, for facilitating this hybrid event and for helping us navigate its administration and logistics. We also thank the members of the technical staff who directed and coordinated sound, video, presentations, and the online platform on the day of the event and helped make the event run smoothly despite the many challenges. Finally, we extend our thanks to Pinecone for their sponsorship of the event.

References

Nima Asadi. *Multi-Stage Search Architectures for Streaming Documents*. University of Maryland, 2013.

Nima Asadi and Jimmy 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*, page 2419–2422, 2012.

Nima Asadi and Jimmy Lin. Fast candidate generation for real-time tweet search with bloom filter chains. *ACM Trans. Inf. Syst.*, 31(3), aug 2013a.

Nima Asadi and Jimmy 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*, page 997–1000, 2013b.

Nima Asadi and Jimmy Lin. Training efficient tree-based models for document ranking. In *European Conference on Information Retrieval*, pages 146–157. Springer, 2013c.

Nima Asadi, Jimmy Lin, and Arjen P. de Vries. Runtime optimizations for tree-based machine learning models. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2281–2292, 2014.

Sebastian Bruch, Claudio Lucchese, and Franco Maria Nardini. Reneuir: Reaching efficiency in neural information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 3462–3465, 2022.

Alexander Buchholz, Jan Malte Lichtenberg, Giuseppe Di Benedetto, Yannik Stein, Vito Bellini, and Matteo Ruffini. Low-variance estimation in the plackett-luce model via quasi-monte carlo sampling. In *Workshop on Reaching Efficiency in Neural Information Retrieval, the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

Christopher J.C. Burges. From ranknet to lambdarank to lambdamart: An overview. Technical Report MSR-TR-2010-82, June 2010.

Arthur Câmara and Claudia Hauf. Moving stuff around: A study on the efficiency of moving documents into memory for neural ir models. In *Workshop on Reaching Efficiency in Neural Information Retrieval, the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

Berkant Barla Cambazoglu, Hugo Zaragoza, Olivier Chapelle, Jiang Chen, Ciya Liao, Zhaohui Zheng, and Jon Degenhardt. Early exit optimizations for additive machine learned ranking systems. In *Proceedings of the Third International Conference on Web Search and Web Data Mining (WSDM)*, pages 411–420. ACM, 2010.

Olivier Chapelle and Yi Chang. Yahoo! learning to rank challenge overview. In *Proceedings of the Learning to Rank Challenge*, pages 1–24, 2011.

J Shane Culpepper, Charles LA Clarke, and Jimmy Lin. Dynamic cutoff prediction in multi-stage retrieval systems. In *Proceedings of the 21st Australasian Document Computing Symposium*, pages 17–24. ACM, 2016.

Van Dang, Michael Bendersky, and W Bruce Croft. Two-stage learning to rank for information retrieval. In *Advances in Information Retrieval*, pages 423–434. Springer, 2013.

Domenico Dato, Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Nicola Tonellotto, and Rossano Venturini. Fast ranking with additive ensembles of oblivious and non-oblivious regression trees. *ACM Trans. Inf. Syst.*, 35(2):15:1–15:31, December 2016. ISSN 1046-8188.

Sibo Dong, Justin Goldstein, and Grace Hui Yang. Seine: Segment-based indexing for neural information retrieval. In *Workshop on Reaching Efficiency in Neural Information Retrieval, the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. From distillation to hard negative sampling: Making sparse neural ir models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2353–2359, 2022.

Luyu Gao, Zhuyun Dai, and Jamie Callan. Understanding bert rankers under distillation. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, page 149–152, 2020.

Mitchell Gordon, Kevin Duh, and Nicholas 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.

Sebastian Hofstätter, Hamed Zamani, Bhaskar Mitra, Nick Craswell, and Allan Hanbury. Local Self-Attention over Long Text for Efficient Document Retrieval. In *Proc. of SIGIR*, 2020.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, November 2020.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, November 2020.

Carlos Lassance and Stéphane Clinchant. An efficiency study for splade models. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2220–2226, 2022.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. Pretrained transformers for text ranking: Bert and beyond, 2021.

Zi Lin, Jeremiah Liu, Zi Yang, Nan Hua, and Dan Roth. Pruning redundant mappings in transformer models via spectral-normalized identity prior. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, November 2020.

Shichen Liu, Fei Xiao, Wenwu Ou, and Luo 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. ACM, 2017.

Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

Zejian Liu, Fanrong Li, Gang Li, and Jian Cheng. EBERT: Efficient BERT inference with dynamic structured pruning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4814–4823, August 2021.

Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Nicola Tonellotto, and Rossano 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 (SIGIR)*, pages 73–82, 2015. ISBN 978-1-4503-3621-5.

Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Salvatore 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*, SIGIR '16, pages 949–952, 2016a. ISBN 978-1-4503-4069-4.

Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Nicola Tonellotto, and Rossano 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*, SIGIR '16, pages 833–836, 2016b. ISBN 978-1-4503-4069-4.

Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, and Salvatore 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*, SIGIR '17, pages 1077–1080, 2017. ISBN 978-1-4503-5022-8.

Joel Mackenzie, J Shane Culpepper, Roi Blanco, Matt Crane, Charles LA Clarke, and Jimmy 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. ACM, 2018.

Antonio Mallia, Joel Mackenzie, Torsten Suel, and Nicola Tonellotto. Faster learned sparse retrieval with guided traversal. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1901–1905, 2022.

Yoshitomo Matsubara, Thuy Vu, and Alessandro Moschitti. *Reranking for Efficient Transformer-Based Answer Selection*, page 1577–1580. 2020.

J. S. McCarley, Rishav Chakravarti, and Avirup Sil. Structured pruning of a bert-based question answering model, 2021.

Bhaskar Mitra, Sebastian Hofstätter, Hamed Zamani, and Nick Craswell. *Improving Transformer-Kernel Ranking Model Using Conformer and Query Term Independence*, page 1697–1702. 2021.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert, 2020.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*, 2019a.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*, 2019b.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, November 2020.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

Harrisen Scells, Shengyao Zhuang, and Guido Zuccon. Reduce, reuse, recycle: Green information retrieval research. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2825–2837, 2022.

Luca Soldaini and Alessandro Moschitti. The cascade transformer: an application for efficient answer sentence selection. In *ACL*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, 2017.

Lidan Wang, Jimmy Lin, and Donald 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. ACM, 2011.

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy 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, April 2021.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, April 2021.

Hamed Zamani, Fernando Diaz, Mostafa Dehghani, Donald Metzler, and Michael Bendersky. Retrieval-enhanced machine learning. page 2875–2886, 2022.

Shengyao Zhuang and Guido Zuccon. Tilde: Term independent likelihood model for passage re-ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1483–1492, 2021.

Shengyao Zhuang and Guido 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.