Efficient and Effective Tree-based and Neural Learning to Rank

Other titles in Foundations and Trends® in Information Retrieval

Quantum-Inspired Neural Language Representation, Matching and Understanding
Peng Zhang, Hui Gao, Jing Zhang and Dawei Song
ISBN: 978-1-63828-204-4

Pre-training Methods in Information Retrieval
Yixing Fan, Xiaohui Xie, Yinqiong Cai, Jia Chen, Xinyu Ma, Xiangsheng Li, Ruqing Zhang and Jiafeng Guo
ISBN: 978-1-63828-062-0

Fairness in Information Access Systems
Michael D. Ekstrand, Anubrata Das, Robin Burke and Fernando Diaz
ISBN: 978-1-63828-040-8

Deep Learning for Dialogue Systems: Chit-Chat and Beyond
Rui Yan, Juntao Li and Zhou Yu
ISBN: 978-1-63828-022-4

Search Interface Design and Evaluation
Chang Liu, Ying-Hsang Liu, Jingjing Liu and Ralf Bierig
ISBN: 978-1-68083-922-7

Efficient and Effective Tree-based and Neural Learning to Rank

Sebastian Bruch
Pinecone
sbruch@acm.org

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

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

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

S. Bruch *et al.*. *Efficient and Effective Tree-based and Neural Learning to Rank*. Foundations and Trends® in Information Retrieval, vol. 17, no. 1, pp. 1–123, 2023.

ISBN: 978-1-63828-199-3

Foundations and Trends® in Information Retrieval
Volume 17, Issue 1, 2023

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2023, Volume 17, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Contents

Efficient and Effective Tree-based and Neural Learning to Rank

Sebastian Bruch[1], Claudio Lucchese[2] and Franco Maria Nardini[3]

[1] *Pinecone, USA; sbruch@acm.org*
[2] *Ca' Foscari University of Venice, Italy; claudio.lucchese@unive.it*
[3] *ISTI-CNR, Pisa, Italy; francomaria.nardini@isti.cnr.it*

ABSTRACT

As information retrieval researchers, we not only develop algorithmic solutions to hard problems, but we also insist on a proper, multifaceted evaluation of ideas. The 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, while 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 resurfaced 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.

This monograph takes a step towards promoting the study of efficiency in the era of neural information retrieval by

Sebastian Bruch, Claudio Lucchese and Franco Maria Nardini (2023), "Efficient and Effective Tree-based and Neural Learning to Rank", Foundations and Trends® in Information Retrieval: Vol. 17, No. 1, pp 1–123. DOI: 10.1561/1500000071.

offering a comprehensive survey of the literature on efficiency and effectiveness in ranking, and to a limited extent, retrieval. This monograph was inspired by the parallels that exist between the challenges in neural network-based ranking solutions and their predecessors, decision forest-based LtR models, as well as the connections between the solutions the literature to date has to offer. We believe that by understanding the fundamentals underpinning these algorithmic and data structure solutions for containing the contentious relationship between efficiency and effectiveness, one can better identify future directions and more efficiently determine the merits of ideas. We also present what we believe to be important research directions in the forefront of efficiency and effectiveness in retrieval and ranking.

1

Introduction

Search engines are a familiar tool to the reader of this monograph. In fact, you have likely arrived at this copy by typing a few keywords into one and perusing the relevant links and page descriptions in its results page. Indeed, the abundance of data on the web makes search engines an integral tool, without which it would be nearly impossible to discover the right information and satisfy an information need.

We similarly rely on a suite of other algorithmic tools to get what 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 tools have in common is that they are different manifestations of the *retrieval and ranking* problem, which seeks to discover a *set of relevant items* from a large collection and order them according to some *criteria* and with respect to some *context*.

Definition 1.1 (The Document Ranking Problem)**.** Given a query q (*context*) and a set of documents D (*items*), the goal is to order elements of D such that the resulting ranked list maximizes a user satisfaction metric Q (*criteria*).

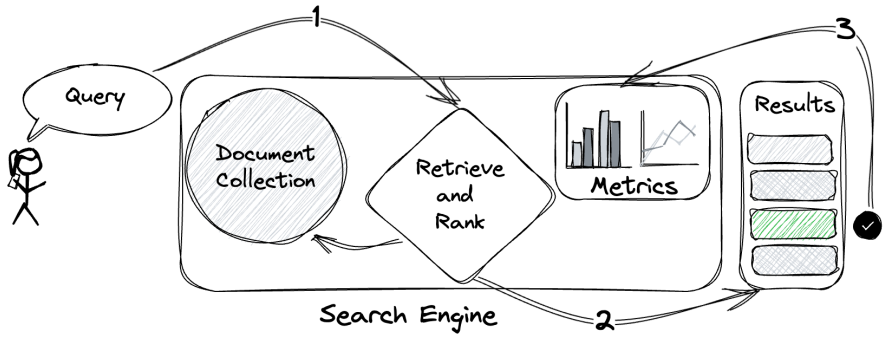

Figure 1.1: The Document Ranking Problem in the context of web search—our running example. The user sends a text query to the search engine (1), which, in turn, *retrieves* the most relevant documents from a large collection, and presents them as a *ranked* list (2). The user then decides if and to what extent the ranked list satisfies their information need, which affects metrics of interest (3).

We take web search as the theme of this monograph and delve into the ranking problem in that context. In document ranking, the query q is an intent expressed (often briefly) as a set of textual keywords or in natural language, the documents D are (possibly long) texts written in natural language, and Q is any utility metric that captures the relevance of an ordered list to q. We have illustrated this setup in Figure 1.1.

Document ranking presents a number of unique questions that are the subject of much research in the field of information retrieval: How do we define Q to quantify the perceived quality of a ranked list and its utility to a user? How do we capture and interpret implicit, noisy, and sometimes circular user preferences, which are represented by clicks? And, more pertinent to this monograph, how do we arrive at a ranked list given a query, a set of documents, a metric, possibly subject to a set of other constraints?

Over a decade ago, machine learning transformed how we approach the document ranking problem and answer the questions above. 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 Learning to Rank (LtR) (Liu, 2009), where the relevance of a document to a query is estimated by a learnt

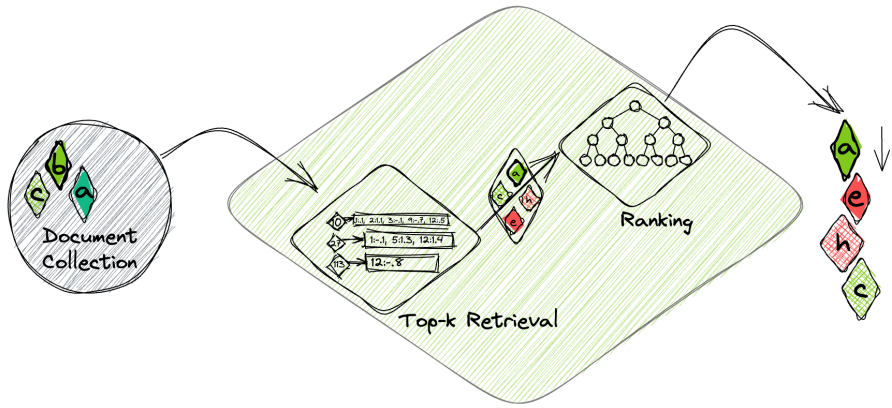

Figure 1.2: Retrieval and ranking algorithms in a modern search system. The *retrieval* algorithm often solves one form of the maximum inner product search (MIPS) problem using, for example, an approximate nearest neighbor (ANN) search or an inverted index-based top-k retrieval algorithm where closeness is determined by lexical matching scores. The *ranking* algorithm may be as simple as an identity function (e.g., in deep learning-based "dense retrieval") or a complex learnt function such as decision forests or deep learning models.

function, hence "learning" to rank. This leap was perhaps best exemplified by LambdaMART (Burges, 2010) in the Yahoo! Learning-to-Rank Challenge (Chapelle and Chang, 2011).

This transformation of the document ranking problem culminated in a framework that comprises of two distinct algorithms, depicted in Figure 1.2: *top-k retrieval*, which finds a *subset* of k documents that are more relevant to a query, followed by *ranking* which orders the documents in the top-k set. In LtR, the ranking stage uses an often expensive function that was trained using supervised or online learning methods, while the retrieval algorithm solves a form of the maximum inner product search (MIPS) problem. As we will describe later, in "dense retrieval," retrieval is often (but not always) an approximate nearest neighbor search while ranking is the identity function.

1.1 The Importance of Efficiency

Any solution that addresses the ranking problem, including LtR, by definition seeks to maximize a user satisfaction metric, Q. But in many real-world applications achieving the highest **effectiveness** is only one of many requirements. We may indeed desire to impose additional constraints on the ranked list, such as a requirement that ranked lists fairly represent underrepresented categories; that they guarantee privacy when the set D consists of documents private to a user; or that they counter biases and ensure trust. Each of these additional constraints is an important objective to optimize in its own right.

An objective that is equally as important as effectiveness in many applications is the **efficiency** of the retrieval and ranking systems. For example, it is often imperative to find the right documents and finalize a ranked list within a small time budget to meet demand and ensure a timely delivery of information. In fact, a perfectly-ordered ranked list may be of little value or have a low perceived quality if delivered too late or with substantial delay.[1]

The question of efficiency gained increasing significance with the rise of LtR whose training and serving require large amounts of computational power. Indeed, the success of LambdaMART and subsequent decision forest-based descendents (Ganjisaffar *et al.*, 2011; Dato *et al.*, 2016; Bruch, 2021; Lucchese *et al.*, 2018b) in improving the quality of rankings came at the expense of the efficiency of training and inference. The training of such models is expensive because we must often (and repeatedly) learn ensembles of hundreds to thousands of deep decision trees sequentially with gradient boosting (Friedman, 2001), with each node in every tree requiring a search in the feature space (Breiman *et al.*, 1984). To become accurate, these large models need to be trained on vast amounts of data, often represented as complex features that are in turn costly to compute. Inference, too, is computationally intensive because estimating the relevance of a single document to a query requires the traversal of paths, from roots to leaves, of every decision tree in the model.

[1] Kohavi *et al.* (2013), reporting on an experiment conducted at Bing, a web search engine, estimated that *"every 100msec improves revenue by 0.6%."*

1.2 Efficiency Considerations Beyond Latency

A decade later, deep neural networks, and in particular, Transformer-based (Vaswani *et al.*, 2017) pre-trained language models advanced the state-of-the-art in ranking dramatically (Lin *et al.*, 2021; Nogueira and Cho, 2020; Nogueira *et al.*, 2019a; Nogueira *et al.*, 2020). Learnt representations of queries and documents by deep networks, too, offer a range of opportunities including the development of a new generation of "dense" 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 Neural Information Retrieval (NIR).

NIR is a leap forward, reaching new highs in quality. Whatever the reason behind its success may be, NIR achieves a greater effectiveness than the previous wave of machine learning models like decision forests on many information retrieval tasks, but with orders of magnitude more learnable parameters and much greater amounts of data. The new scale drastically increases the computational and economic costs of model training and inference. GPT-3 (Brown *et al.*, 2020), for example, required 285,000 CPU cores and 10,000 GPUs to train, with an estimated economic cost of $4.6M.[2] Although it may be argued that the high cost of training deep models is amortized because large language models can, through a process known as "fine-tuning," be recycled and reused for a variety of applications with a substantially smaller effort, it is still a significant price to pay upfront. Furthermore, not all large neural models can be easily recycled—in fact, that is one of the properties Scells *et al.* (2022) call out in their article. What is more, once trained, the use of such large models in production similarly requires a nontrivial amount of tensor multiplications and other complex operations.

Due to their alarming computational requirements, NIR models underline several dimensions of efficiency that have thus far been less obvious. Crucially, "efficiency" is no longer characterized by low latency, but is instead a concept that amalgamates space-, sample-, and energy-efficiency, among other emerging factors, as summarized in Table 1.1.

[2] https://lambdalabs.com/blog/demystifying-gpt-3/.

Table 1.1: Taxonomy of a multi-faceted view of ranking efficiency and the stages in which they manifest.

Dimension	Definition	Scope
Query	Time elapsed between the arrival of a query and the presentation of ranked list of documents	Inference
Sample	Number of training examples required to learn a ranking function	Training
Space	Total storage used to serve a ranking model	Training; Inference
Training	Time required to train a ranking model	Training
Energy	Amount of energy required to train a model or evaluate a learnt model on a query-document pair	Training; Inference

In other words, the inefficiency of an algorithm cannot and should not be understood solely in terms of negative user experience due to greater latencies, but instead, we must acknowledge that ineffciency has adverse implications for resource-constrained researchers and practitioners, and more importantly, for the environment (in the form of emissions and carbon footprint) (Scells *et al.*, 2022; Strubell *et al.*, 2019; Xu *et al.*, 2021). We must therefore acknowledge that, due to environmental factors, attempting to address the efficiency problem by relying on advances in hardware systems or by utilizing more resources is not a sustainable long-term solution. Instead, combating this multi-faceted issue of efficiency necessitates a careful study and design of efficient algorithms and data structures, as highlighted by deliberations at recent academic workshops (e.g., the Workshop on Reaching Efficiency in Neural Information Retrieval (Bruch *et al.*, 2022b; Bruch *et al.*, 2023)).

1.3 Efficient and Effective Ranking

Accuracy by way of ever-increasing complexity presents a challenge: how do we then optimize for both effectiveness and efficiency? Must

we lose accuracy to find a more efficient solution, inevitably trading off effectiveness for efficiency and vice versa? These and other similar questions give rise to a research topic that extends the document ranking problem as follows:

Definition 1.2 (The Efficient Document Ranking Problem)**.** Given a query q and a set of documents D, the goal is to order elements of D *efficiently* such that the resulting ranked list maximizes a user satisfaction metric Q.

The problem above spawned a line of research in the information retrieval community to systematically investigate questions of efficiency and explore the trade-offs between efficiency and effectiveness in ranking models, leading to several innovations. The community widely adopted multi-stage, *cascade* 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, 2013a; 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; Asadi and Lin, 2013b), to cost-aware training and *post hoc* pruning of decision forests (Asadi and Lin, 2013c; Lucchese *et al.*, 2017b; Lucchese *et al.*, 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; Lucchese *et al.*, 2015b), the information retrieval community thoroughly considered the practicality and scalability of complex ranking algorithms.

In addition to volumes of publications, the output of this research effort included standardized algorithms and reusable software packages (Ke *et al.*, 2017; Lucchese *et al.*, 2015b). Perhaps more crucially, the community developed an understanding that quality is not the be-all and end-all of information retrieval research and that model complexity must be managed (through more efficient training and inference) and justified (e.g., by contextualizing quality gains in terms of the amount of computational resources required).

As complex neural network-based models come to dominate the research on document ranking, it is unsurprising that there is renewed interest in the question above, not just in the information retrieval com-

munity but also in related branches such as natural language processing. Interestingly, many of the proposals put forward to date to contain efficiency are reincarnations of past ideas, such as stage-wise ranking with BERT-based models (Nogueira *et al.*, 2019a; Matsubara *et al.*, 2020), early-exit strategies in Transformers (Soldaini and Moschitti, 2020; Xin *et al.*, 2020; Xin *et al.*, 2021), neural connection pruning (Gordon *et al.*, 2020; McCarley *et al.*, 2021; Lin *et al.*, 2020b; Liu *et al.*, 2021), precomputation of representations (MacAvaney *et al.*, 2020b), and enhancing indexes (Zhuang and Zuccon, 2022; Nogueira *et al.*, 2019b; Mallia *et al.*, 2022; Lassance and Clinchant, 2022). Other novel but general ideas such as knowledge distillation (Jiao *et al.*, 2020; Sanh *et al.*, 2020; Gao *et al.*, 2020) have also proved effective in reducing the size of deep models. Yet other innovative ideas developed specifically for ranking include efforts to reinvent Transformers from the ground-up (Mitra *et al.*, 2021; Hofstätter *et al.*, 2020).

1.4 About this Monograph

Given the resurgence of the question of efficiency and the trade-offs between efficiency and effectiveness in ranking, and the apparent overlap between the neural and pre-neural ideas to address this question, we believe it is necessary to present a comprehensive review of this literature with a particular focus on the document ranking problem. We have thus prepared this monograph in four parts in the hope that it serves as one such resource.

The first part introduces the document ranking problem and reviews a machine learning formulation of it in the context of web search in depth. We also describe the architecture of a modern search engine to illustrate an application of ranking that is of primary interest to this work. As we explain the ingredients of a search engine and all that is involved in the training and serving of a ranking model within this framework, we highlight the costs to efficiency and call out the levers that trade off effectiveness for efficiency.

While the first part of this monograph concerns an abstract, general setup, the two subsequent parts get more specific and examine two popular families of ranking algorithms through the lens of efficiency.

One presents a treatment of a branch of LtR that is based on forests of decision trees, while another turns to neural networks and deep learning methods for retrieval and ranking. Each family presents its own unique challenges and requires its own set of solutions to explore the Pareto front on the efficiency-effectiveness optimization landscape.

As the reader will notice, the approaches developed for the two families of ranking algorithms appear to be—and in many ways, are—independent. But the ideas behind them overlap too. We attempt, in the last part of the monograph, to identify the common threads that can help translate ideas from one space to another. We also discuss emerging research directions, made urgent by the rise of deep neural networks in information retrieval, and explore open challenges within this space.

References

Ai, Q., X. Wang, S. Bruch, N. Golbandi, M. Bendersky, and M. Najork. (2019). "Learning Groupwise Multivariate Scoring Functions Using Deep Neural Networks". In: *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. Santa Clara, CA, USA. 85–92.

Akkalyoncu Yilmaz, Z., S. Wang, W. Yang, H. Zhang, and J. Lin. (2019). "Applying BERT to Document Retrieval with Birch". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*.

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

Asadi, N. and J. Lin. (2012). "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*. Maui, Hawaii, USA. 2419–2422.

Asadi, N. and J. Lin. (2013a). "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*. Dublin, Ireland. 997–1000.

Asadi, N. and J. Lin. (2013b). "Fast Candidate Generation for Real-Time Tweet Search with Bloom Filter Chains". *ACM Transactions on Information Systems*. 31(3).

Asadi, N. and J. Lin. (2013c). "Training Efficient Tree-Based Models for Document Ranking". In: *Proceedings of the 35th European Conference on Advances in Information Retrieval*. Moscow, Russia. 146–157.

Asadi, N., J. Lin, and A. P. de Vries. (2014). "Runtime Optimizations for Tree-Based Machine Learning Models." *IEEE Transactions on Knowledge and Data Engineering*. 26(9): 2281–2292.

Ba, J. and R. Caruana. (2014). "Do Deep Nets Really Need to be Deep?" In: *Advances in neural information processing systems*. 2654–2662.

Bendersky, M., W. B. Croft, and Y. Diao. (2011). "Quality-Biased Ranking of Web Documents". In: *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*. Hong Kong, China. 95–104.

Bennett, P. N., K. Svore, and S. T. Dumais. (2010). "Classification-enhanced Ranking". In: *Proceedings of the 19th International Conference on World Wide Web*. 111–120.

Blondel, M., O. Teboul, Q. Berthet, and J. Djolonga. (2020). "Fast Differentiable Sorting and Ranking". In: *Proceedings of the 37th International Conference on Machine Learning*.

Borisov, A., I. Markov, M. de Rijke, and P. Serdyukov. (2016). "A Neural Click Model for Web Search". In: *Proceedings of the 25th International Conference on World Wide Web*. 531–541.

Breiman, L., J. Friedman, C. J. Stone, and R. Olshen. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Broder, A. Z., D. Carmel, M. Herscovici, A. Soffer, and J. Zien. (2003). "Efficient Query Evaluation Using a Two-Level Retrieval Process". In: *Proceedings of the 12th International Conference on Information and Knowledge Management*. New Orleans, LA, USA. 426–434.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. (2020). "Language Models are Few-Shot Learners". arXiv: 2005.14165 [cs.CL].

Bruch, S. (2021). "An Alternative Cross Entropy Loss for Learning-to-Rank". In: *Proceedings of the Web Conference 2021*. Ljubljana, Slovenia. 118–126.

Bruch, S., S. Gai, and A. Ingber. (2022a). "An Analysis of Fusion Functions for Hybrid Retrieval". arXiv: 2210.11934 [cs.IR].

Bruch, S., S. Han, M. Bendersky, and M. Najork. (2020). "A Stochastic Treatment of Learning to Rank Scoring Functions". In: *Proceedings of the 13th International Conference on Web Search and Data Mining*. 61–69.

Bruch, S., C. Lucchese, and F. M. Nardini. (2022b). "ReNeuIR: Reaching Efficiency in Neural Information Retrieval". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Madrid, Spain. 3462–3465.

Bruch, S., C. Lucchese, and F. M. Nardini. (2023). "Report on the 1st Workshop on Reaching Efficiency in Neural Information Retrieval (ReNeuIR 2022) at SIGIR 2022". *SIGIR Forum*. 56(2).

Bruch, S., X. Wang, M. Bendersky, and M. Najork. (2019a). "An Analysis of the Softmax Cross Entropy Loss for Learning-to-Rank with Binary Relevance". In: *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. Santa Clara, CA, USA. 75–78.

Bruch, S., M. Zoghi, M. Bendersky, and M. Najork. (2019b). "Revisiting Approximate Metric Optimization in the Age of Deep Neural Networks". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Paris, France. 1241–1244.

Buckley, C. and E. Voorhees. (2005). "Retrieval System Evaluation". In: *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press. Chap. 3.

Burges, C., T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. (2005). "Learning to Rank using Gradient Descent". In: *Proceedings of the 22nd international conference on Machine learning*. ACM. 89–96.

Burges, C. J. (2010). "From RankNet to LambdaRank to LambdaMART: An Overview". *Tech. rep.* No. MSR-TR-2010-82.

Busolin, F., C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, and S. Trani. (2021). "Learning Early Exit Strategies for Additive Ranking Ensembles". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada*. 2217–2221.

Cambazoglu, B. B., H. Zaragoza, O. Chapelle, J. Chen, C. Liao, Z. Zheng, and J. Degenhardt. (2010). "Early Exit Optimizations for Additive Machine Learned Ranking Systems". In: *Proceedings of the 3rd International Conference on Web Search and Web Data Mining (WSDM)*. ACM. 411–420.

Cao, Z., T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. (2007). "Learning to Rank: from Pairwise Approach to Listwise Approach". In: *Proceedings of the 24th International Conference on Machine learning*. ACM. 129–136.

Capannini, G., C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, and N. Tonellotto. (2016). "Quality Versus Efficiency in Document Scoring with Learning-to-rank Models". *Information Processing & Management*. 52(6): 1161–1177.

Carterette, B., P. Bennett, D. Chickering, and S. Dumais. (2008). "Here or there". *Advances in Information Retrieval*: 16–27.

Chapelle, O. and Y. Chang. (2011). "Yahoo! Learning to Rank Challenge Overview". In: *Proceedings of the Learning to Rank Challenge*. 1–24.

Chapelle, O., T. Joachims, F. Radlinski, and Y. Yue. (2012). "Large-scale Validation and Analysis of Interleaved Search Evaluation". *ACM Transactions on Information Systems*. 30(1): 6.

Chapelle, O., D. Metlzer, Y. Zhang, and P. Grinspan. (2009). "Expected Reciprocal Rank for Graded Relevance". In: *Proceedings of the 18th ACM conference on Information and knowledge management*. 621–630.

Chapelle, O., P. Shivaswamy, S. Vadrevu, K. Weinberger, Y. Zhang, and B. Tseng. (2011). "Boosted Multi-task Learning". *Machine learning*. 85(1-2): 149–173.

Chen, M., Z. Xu, K. Weinberger, O. Chapelle, and D. Kedem. (2012). "Classifier cascade for minimizing feature evaluation cost". In: *Artificial Intelligence and Statistics*. 218–226.

Chen, R.-C., L. Gallagher, R. Blanco, and J. S. Culpepper. (2017). "Efficient Cost-Aware Cascade Ranking in Multi-Stage Retrieval". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Shinjuku, Tokyo, Japan. 445–454.

Chen, T., M. Zhang, J. Lu, M. Bendersky, and M. Najork. (2022). "Out-of-Domain Semantics to the Rescue! Zero-Shot Hybrid Retrieval Models". In: *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*. Stavanger, Norway. 95–110.

Chen, T. and C. Guestrin. (2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA. 785–794.

Chuklin, A., I. Markov, and M. de Rijke. (2015). *Click Models for Web Search*. Morgan & Claypool.

Cohen, D., J. Foley, H. Zamani, J. Allan, and W. B. Croft. (2018). "Universal Approximation Functions for Fast Learning to Rank: Replacing Expensive Regression Forests with Simple Feed-forward Networks". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM. 1017–1020.

Cormack, G. V., M. D. Smucker, and C. L. Clarke. (2011). "Efficient and Effective Spam Filtering and Re-ranking for Large Web Datasets". *Information Retrieval*. 14: 441–465.

Culpepper, J. S., C. L. Clarke, and J. Lin. (2016). "Dynamic Cutoff Prediction in Multi-stage Retrieval Systems". In: *Proceedings of the 21st Australasian Document Computing Symposium*. ACM. 17–24.

Cuturi, M., O. Teboul, and J.-P. Vert. (2019). "Differentiable Ranking and Sorting using Optimal Transport". In: *Advances in Neural Information Processing Systems*. Vol. 32.

Dai, Z. and J. Callan. (2019). "Deeper Text Understanding for IR with Contextual Neural Language Modeling". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Paris, France. 985–988.

Dai, Z., C. Xiong, J. Callan, and Z. Liu. (2018). "Convolutional Neural Networks for Soft-Matching N-Grams in Ad-Hoc Search". In: *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. Marina Del Rey, CA, USA. 126–134.

Dang, V., M. Bendersky, and W. B. Croft. (2013). "Two-Stage learning to rank for information retrieval". In: *Advances in Information Retrieval*. Springer. 423–434.

Dato, D., C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. (2016). "Fast Ranking with Additive Ensembles of Oblivious and Non-Oblivious Regression Trees". *ACM Transactions on Information Systems*. 35(2): 15:1–15:31.

Dehghani, M., H. Zamani, A. Severyn, J. Kamps, and W. B. Croft. (2017). "Neural Ranking Models with Weak Supervision". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Tokyo, Japan. 65–74.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

Diaz, F., B. Mitra, M. D. Ekstrand, A. J. Biega, and B. Carterette. (2020). "Evaluating Stochastic Rankings with Expected Exposure". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. Virtual Event, Ireland. 275–284.

Ding, S. and T. Suel. (2011). "Faster Top-k Document Retrieval Using Block-Max Indexes". In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Beijing, China. 993–1002.

Dredze, M., R. Gevaryahu, and A. Elias-Bachrach. (2007). "Learning Fast Classifiers for Image Spam." In: *CEAS*. 2007–487.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, *et al.* (2004). "Least Angle Regression". *The Annals of Statistics*. 32(2): 407–499.

Formal, T., C. Lassance, B. Piwowarski, and S. Clinchant. (2022). "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*. Madrid, Spain. 2353–2359.

Freund, Y., R. Iyer, R. E. Schapire, and Y. Singer. (2003). "An Efficient Boosting Algorithm for Combining Preferences". *Journal of Machine Learning Research*. 4(Nov): 933–969.

Friedman, J. H. (2001). "Greedy Function Approximation: a Gradient Boosting Machine". *Annals of Statistics*: 1189–1232.

Gallagher, L., R.-C. Chen, R. Blanco, and J. S. Culpepper. (2019). "Joint Optimization of Cascade Ranking Models". In: *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. Melbourne VIC, Australia. 15–23.

Ganjisaffar, Y., R. Caruana, and C. V. Lopes. (2011). "Bagging Gradient-boosted Trees for High Precision, Low Variance Ranking Models". In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 85–94.

Gao, L. and J. Callan. (2021a). "Condenser: a Pre-training Architecture for Dense Retrieval". arXiv: 2104.08253 [cs.CL].

Gao, L. and J. Callan. (2021b). "Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval". arXiv: 2108.05540 [cs.IR].

Gao, L., Z. Dai, and J. Callan. (2020). "Understanding BERT Rankers Under Distillation". In: *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. Virtual Event, Norway. 149–152.

Geng, X., T.-Y. Liu, T. Qin, and H. Li. (2007). "Feature Selection for Ranking". In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Amsterdam, The Netherlands. 407–414.

Gigli, A., C. Lucchese, F. M. Nardini, and R. Perego. (2016). "Fast Feature Selection for Learning to Rank". In: *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. Newark, Delaware, USA. 167–170.

Gil-Costa, V., F. Loor, R. Molina, F. M. Nardini, R. Perego, and S. Trani. (2022). "Ensemble Model Compression for Fast and Energy-Efficient Ranking on FPGAs". In: *Advances in Information Retrieval*. Springer. 260–273.

Gomes, B. (2017). "Our Latest Quality Improvements for Search". URL: https://www.blog.google/products/search/our-latest-quality-improvements-search/.

Gordon, M., K. Duh, and N. Andrews. (2020). "Compressing BERT: Studying the Effects of Weight Pruning on Transfer Learning". In: *Proceedings of the 5th Workshop on Representation Learning for NLP*. 143–155.

Guo, J., Y. Fan, Q. Ai, and W. B. Croft. (2016). "A Deep Relevance Matching Model for Ad-Hoc Retrieval". In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. Indianapolis, Indiana, USA. 55–64.

Guo, J., Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, and X. Cheng. (2020). "A Deep Look into Neural Ranking Models for Information Retrieval". *Information Processing & Management*. 57(6).

He, X., J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, *et al.* (2014). "Practical Lessons from Predicting Clicks on Ads at Facebook". In: *Proceedings of the 8th International Workshop on Data Mining for Online Advertising*. 1–9.

Henzinger, M. R. *et al.* (2000). "Link Analysis in Web Information Retrieval". *IEEE Data Engineering Bulletin*. 23(3): 3–8.

Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. (2012). "Improving Neural Networks by Preventing Co-adaptation of Feature Detectors". arXiv: 1207.0580 [cs.NE].

Hoerl, A. E. and R. W. Kennard. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems". *Technometrics*. 12(1): 55–67.

Hofmann, K., A. Schuth, S. Whiteson, and M. de Rijke. (2013a). "Reusing Historical Interaction Data for Faster Online Learning to Rank for IR". In: *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*. 183–192.

Hofmann, K., S. Whiteson, and M. de Rijke. (2013b). "Balancing Exploration and Exploitation in Listwise and Pairwise Online Learning to Rank for Information Retrieval". *Information Retrieval*. 16(1): 63–90.

Hofstätter, S., H. Zamani, B. Mitra, N. Craswell, and A. Hanbury. (2020). "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*. Virtual Event, China. 2021–2024.

Huang, P.-S., X. He, J. Gao, L. Deng, A. Acero, and L. Heck. (2013). "Learning Deep Structured Semantic Models for Web Search using Clickthrough Data". In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. ACM. 2333–2338.

Jagerman, R., Z. Qin, X. Wang, M. Bendersky, and M. Najork. (2022). "On Optimizing Top-K Metrics for Neural Ranking Models". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Madrid, Spain. 2303–2307.

Järvelin, K. and J. Kekäläinen. (2000). "IR Evaluation Methods for Retrieving Highly Relevant Documents". In: *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 41–48.

Jiang, D., K. W.-T. Leung, and W. Ng. (2016). "Query Intent Mining with Multiple Dimensions of Web Search Data". *Proceedings of the 25th International Conference on World Wide Web*. 19(3): 475–497.

Jiao, X., Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu. (2020). "TinyBERT: Distilling BERT for Natural Language Understanding". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Jin, X., T. Yang, and X. Tang. (2016). "A Comparison of Cache Blocking Methods for Fast Execution of Ensemble-based Score Computation". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Pisa, Italy. 629–638.

Joachims, T. (2002). "Optimizing Search Engines using Clickthrough Data". In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 133–142.

Joachims, T., L. Granka, B. Pan, H. Hembrooke, and G. Gay. (2005). "Accurately Interpreting Clickthrough Data as Implicit Feedback". In: *Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 154–161.

Joachims, T., A. Swaminathan, and T. Schnabel. (2017). "Unbiased Learning-to-rank with Biased Feedback". In: *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*. 781–789.

Johnson, J., M. Douze, and H. Jégou. (2021). "Billion-Scale Similarity Search with GPUs". *IEEE Transactions on Big Data*. 7: 535–547.

Jones, K. S., S. Walker, and S. E. Robertson. (2000). "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments: Part 2". *Information processing & management*. 36(6): 809–840.

Karpukhin, V., B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. (2020). "Dense Passage Retrieval for Open-Domain Question Answering". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6769–6781.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. (2017). "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, California, USA. 3149–3157.

Khattab, O. and M. Zaharia. (2020). "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. Virtual Event, China. 39–48.

Kohavi, R., A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann. (2013). "Online Controlled Experiments at Large Scale". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1168–1176.

Kusner, M. J., W. Chen, Q. Zhou, Z. E. Xu, K. Q. Weinberger, and Y. Chen. (2014). "Feature-Cost Sensitive Learning with Submodular Trees of Classifiers". In: *AAAI*. 1939–1945.

Kveton, B., C. Szepesvari, Z. Wen, and A. Ashkan. (2015). "Cascading Bandits: Learning to Rank in the Cascade Model". In: *International Conference on Machine Learning*. 767–776.

Lassance, C. and S. Clinchant. (2022). "An Efficiency Study for SPLADE Models". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Madrid, Spain. 2220–2226.

Lettich, F., C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. (2019). "Parallel Traversal of Large Ensembles of Decision Trees". *IEEE Transactions on Parallel and Distributed Systems*. 30(9): 2075–2089.

Li, C., A. Yates, S. MacAvaney, B. He, and Y. Sun. (2020). "PARADE: Passage Representation Aggregation for Document Reranking". arXiv: 2008.09093 [cs.IR].

Lin, J., R. Nogueira, and A. Yates. (2021). "Pretrained Transformers for Text Ranking: BERT and Beyond". arXiv: 2010.06467 [cs.IR].

Lin, S.-C., J.-H. Yang, and J. Lin. (2020a). "Distilling Dense Representations for Ranking using Tightly-coupled Teachers". arXiv: 2010.11386 [cs.IR].

Lin, Z., J. Liu, Z. Yang, N. Hua, and D. Roth. (2020b). "Pruning Redundant Mappings in Transformer Models via Spectral-Normalized Identity Prior". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Lindgren, E., S. Reddi, R. Guo, and S. Kumar. (2021). "Efficient Training of Retrieval Models using Negative Cache". *Advances in Neural Information Processing Systems*. 34: 4134–4146.

Ling, X., W. Deng, C. Gu, H. Zhou, C. Li, and F. Sun. (2017). "Model Ensemble for Click Prediction in Bing Search Ads". In: *Proceedings of the 26th International Conference on World Wide Web Companion*. 689–698.

Liu, S., F. Xiao, W. Ou, and L. Si. (2017). "Cascade Ranking for Operational E-commerce Search". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 1557–1565.

Liu, T.-Y. (2009). "Learning to Rank for Information Retrieval". *Foundations and Trends in Information Retrieval*. 3(3): 225–331.

Liu, W., P. Zhou, Z. Wang, Z. Zhao, H. Deng, and Q. Ju. (2020). "FastBERT: a Self-distilling BERT with Adaptive Inference Time". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6035–6044.

Liu, Z., F. Li, G. Li, and J. Cheng. (2021). "EBERT: Efficient BERT Inference with Dynamic Structured Pruning". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 4814–4823.

Long, B., O. Chapelle, Y. Zhang, Y. Chang, Z. Zheng, and B. Tseng. (2010). "Active Learning for Ranking through Expected Loss Optimization". In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 267–274.

Luan, Y., J. Eisenstein, K. Toutanova, and M. Collins. (2021). "Sparse, Dense, and Attentional Representations for Text Retrieval". *Transactions of the Association for Computational Linguistics*. 9: 329–345.

Lucchese, C., C. I. Muntean, F. M. Nardini, R. Perego, and S. Trani. (2017a). "RankEval: An Evaluation and Analysis Framework for Learning-to-Rank Solutions". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Shinjuku, Tokyo, Japan. 1281–1284.

Lucchese, C., C. I. Muntean, F. M. Nardini, R. Perego, and S. Trani. (2020a). "RankEval: Evaluation and Investigation of Ranking Models". *SoftwareX*. 12: 100614.

Lucchese, C., F. M. Nardini, S. Orlando, R. Perego, F. Silvestri, and S. Trani. (2016a). "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*. Pisa, Italy. 949–952.

Lucchese, C., F. M. Nardini, S. Orlando, R. Perego, F. Silvestri, and S. Trani. (2018a). "X-CLEaVER: Learning Ranking Ensembles by Growing and Pruning Trees". *ACM Transactions on Intelligent Systems and Technology*. 9(6).

Lucchese, C., F. M. Nardini, S. Orlando, R. Perego, and N. Tonellotto. (2015a). "Speeding up Document Ranking with Rank-based Features". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. 895–898.

Lucchese, C., F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. (2015b). "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*. 73–82.

Lucchese, C., F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. (2016b). "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*. Pisa, Italy. 833–836.

Lucchese, C., F. M. Nardini, S. Orlando, R. Perego, and S. Trani. (2017b). "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*. Shinjuku, Tokyo, Japan. 1077–1080.

Lucchese, C., F. M. Nardini, S. Orlando, R. Perego, and S. Trani. (2020b). "Query-Level Early Exit for Additive Learning-to-Rank Ensembles". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20*. Virtual Event, China: ACM. 2033–2036.

Lucchese, C., F. M. Nardini, R. Perego, S. Orlando, and S. Trani. (2018b). "Selective Gradient Boosting for Effective Learning to Rank". In: *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. Ann Arbor, MI, USA. 155–164.

Lucchese, C., S. Orlando, R. Perego, F. Silvestri, and G. Tolomei. (2013). "Discovering Tasks from Search Engine Query Logs". *ACM Transactions on Information Systems*. 31(3): 14.

Ma, X., K. Sun, R. Pradeep, and J. Lin. (2021a). "A Replication Study of Dense Passage Retriever". arXiv: 2104.05740 [cs.IR].

Ma, Z., K. Ethayarajh, T. Thrush, S. Jain, L. Y. Wu, R. Jia, C. Potts, A. Williams, and D. Kiela. (2021b). "Dynaboard: An Evaluation-As-A-Service Platform for Holistic Next-Generation Benchmarking". In: *Neural Information Processing Systems*.

MacAvaney, S., S. Feldman, N. Goharian, D. Downey, and A. Cohan. (2020a). "ABNIRML: Analyzing the Behavior of Neural IR Models". *arXiv*. abs/2011.00696. URL: https://arxiv.org/abs/2011.00696.

MacAvaney, S., C. Macdonald, and I. Ounis. (2022). "Streamlining Evaluation with ir-measures". In: *Advances in Information Retrieval*. Springer International Publishing.

MacAvaney, S., F. M. Nardini, R. Perego, N. Tonellotto, N. Goharian, and O. Frieder. (2020b). "Efficient Document Re-Ranking for Transformers by Precomputing Term Representations". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Virtual Event, China. 49–58.

MacAvaney, S., F. M. Nardini, R. Perego, N. Tonellotto, N. Goharian, and O. Frieder. (2020c). "Expansion via Prediction of Importance with Contextualization". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1573–1576.

MacAvaney, S., A. Yates, A. Cohan, and N. Goharian. (2019). "CEDR: Contextualized Embeddings for Document Ranking". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Paris, France. 1101–1104.

Macdonald, C., R. L. Santos, and I. Ounis. (2012). "On the Usefulness of Query Features for Learning to Rank". In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. 2559–2562.

Macdonald, C., R. L. Santos, and I. Ounis. (2013). "The Whens and Hows of Learning to Rank for Web Search". *Information Retrieval*. 16(5): 584–628.

Mackenzie, J., J. S. Culpepper, R. Blanco, M. Crane, C. L. Clarke, and J. Lin. (2018). "Query Driven Algorithm Selection in Early Stage Retrieval". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM. 396–404.

Mackenzie, J., M. Petri, and A. Moffat. (2021). "Anytime Ranking on Document-Ordered Indexes". *ACM Transactions on Information Systems*. 40(1).

Malkov, Y. A. and D. A. Yashunin. (2016). "Efficient and Robust Approximate Nearest Neighbor Search using Hierarchical Navigable Small World graphs". arXiv: 1603.09320 [cs.DS].

Mallia, A., J. Mackenzie, T. Suel, and N. Tonellotto. (2022). "Faster Learned Sparse Retrieval with Guided Traversal". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Madrid, Spain. 1901–1905.

Matsubara, Y., T. Vu, and A. Moschitti. (2020). "Reranking for Efficient Transformer-Based Answer Selection". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1577–1580.

McCarley, J. S., R. Chakravarti, and A. Sil. (2021). "Structured Pruning of a BERT-based Question Answering Model". arXiv: 1910.06360 [cs.CL].

Metzler, D. and W. B. Croft. (2007). "Linear Feature-based Models for Information Retrieval". *Information Retrieval*. 10(3): 257–274.

Mikolov, T., K. Chen, G. S. Corrado, and J. Dean. (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv: 1301.3781 [cs.CL].

Mitra, B. and N. Craswell. (2017). "Neural Models for Information Retrieval". arXiv: 1705.01509 [cs.IR].

Mitra, B., F. Diaz, and N. Craswell. (2017). "Learning to Match using Local and Distributed Representations of Text for Web Search". In: *Proceedings of the 26th International Conference on World Wide Web*. 1291–1299.

Mitra, B., S. Hofstätter, H. Zamani, and N. Craswell. (2021). "Improving Transformer-Kernel Ranking Model Using Conformer and Query Term Independence". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1697–1702.

Mitra, B., E. Nalisnick, N. Craswell, and R. Caruana. (2016). "A Dual Embedding Space Model for Document Ranking". arXiv: 1602.01137 [cs.IR].

Moffat, A. and J. Zobel. (2008). "Rank-biased Precision for Measurement of Retrieval Effectiveness". *ACM Transactions on Information Systems*. 27(1): 2.

Mohan, A., Z. Chen, and K. Weinberger. (2011). "Web-search Ranking with Initialized Gradient Boosted Regression Trees". In: *Proceedings of the learning to rank challenge*. 77–89.

Molina, R., F. Loor, V. Gil-Costa, F. M. Nardini, R. Perego, and S. Trani. (2021). "Efficient Traversal of Decision Tree Ensembles with FPGAs". *Journal of Parallel and Distributed Computing*. 155: 38–49.

Nardini, F. M., C. Rulli, S. Trani, and R. Venturini. (2022). "Distilled Neural Networks for Efficient Learning to Rank". *IEEE Transactions on Knowledge and Data Engineering*: 1–1.

Nguyen, T., M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. (2016). "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset". Nov.

Nogueira, R. and K. Cho. (2020). "Passage Re-ranking with BERT". arXiv: 1901.04085 [cs.IR].

Nogueira, R., Z. Jiang, R. Pradeep, and J. Lin. (2020). "Document Ranking with a Pretrained Sequence-to-Sequence Model". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 708–718.

Nogueira, R. and J. Lin. (2019). "From doc2query to docTTTTTquery".

Nogueira, R., W. Yang, K. Cho, and J. Lin. (2019a). "Multi-stage document ranking with BERT". arXiv: 1910.14424 [cs.IR].

Nogueira, R., W. Yang, J. Lin, and K. Cho. (2019b). "Document Expansion by Query Prediction". arXiv: 1904.08375 [cs.IR].

Onal, K. D., Y. Zhang, I. S. Altingovde, M. M. Rahman, P. Karagoz, A. Braylan, B. Dang, H.-L. Chang, H. Kim, Q. Mcnamara, A. Angert, E. Banner, V. Khetan, T. Mcdonnell, A. T. Nguyen, D. Xu, B. C. Wallace, M. Rijke, and M. Lease. (2018). "Neural Information Retrieval: At the End of the Early Years". *Information Retrieval*. 21(2–3): 111–182.

Oosterhuis, H. (2021). "Computationally Efficient Optimization of Plackett-Luce Ranking Models for Relevance and Fairness". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Oosterhuis, H., R. Jagerman, and M. de Rijke. (2020). "Unbiased Learning to Rank: Counterfactual and Online Approaches". In: *Companion Proceedings of the Web Conference 2020*. Taipei, Taiwan. 299–300.

Oosterhuis, H. and M. de Rijke. (2017). "Balancing Speed and Quality in Online Learning to Rank for Information Retrieval". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM. 277–286.

Pang, L., J. Xu, Q. Ai, Y. Lan, X. Cheng, and J. Wen. (2020). "SetRank: Learning a Permutation-Invariant Ranking Model for Information Retrieval". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Petri, M., A. Moffat, J. Mackenzie, J. S. Culpepper, and D. Beck. (2019). "Accelerated Query Processing Via Similarity Score Prediction". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Paris, France. 485–494.

Ponte, J. M. and W. B. Croft. (1998). "A Language Modeling Approach to Information Retrieval". In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 275–281.

Pradeep, R., R. Nogueira, and J. Lin. (2021). "The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models". arXiv: 2101.05667 [cs.IR].

Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. (2018). "CatBoost: Unbiased Boosting with Categorical Features". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Montréal, Canada. 6639–6649.

Qin, T., T.-Y. Liu, and H. Li. (2010). "A General Approximation Framework for Direct Optimization of Information Retrieval Measures". *Information Retrieval*. 13(4): 375–397.

Qu, Y., Y. Ding, J. Liu, K. Liu, R. Ren, W. X. Zhao, D. Dong, H. Wu, and H. Wang. (2021). "RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5835–5847.

Radlinski, F. and T. Joachims. (2005). "Query Chains: Learning to Rank from Implicit Feedback". In: *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. ACM. 239–248.

Radlinski, F., R. Kleinberg, and T. Joachims. (2008). "Learning Diverse Rankings with Multi-armed Bandits". In: *Proceedings of the 25th International Conference on Machine Learning*. 784–791.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". *Journal of Machine Learning Research*. 21(140): 1–67.

Rasolofo, Y. and J. Savoy. (2003). "Term Proximity Scoring for Keyword-based Retrieval Systems". *Advances in Information Retrieval*: 79–79.

Robertson, S., H. Zaragoza, and M. Taylor. (2004). "Simple BM25 Extension to Multiple Weighted Fields". In: *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*. 42–49.

Salton, G. and C. Buckley. (1988). "Term-weighting approaches in automatic text retrieval". *Information Processing & Management*. 24(5): 513–523.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. (2020). "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter". arXiv: 1910.01108 [cs.CL].

Santhanam, K., O. Khattab, J. Saad-Falcon, C. Potts, and M. Zaharia. (2022a). "ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3715–3734.

Santhanam, K., J. Saad-Falcon, M. Franz, O. Khattab, A. Sil, R. Florian, M. A. Sultan, S. Roukos, M. Zaharia, and C. Potts. (2022b). "Moving Beyond Downstream Task Accuracy for Information Retrieval Benchmarking". arXiv: 2212.01340 [cs.IR].

Scells, H., S. Zhuang, and G. Zuccon. (2022). "Reduce, Reuse, Recycle: Green Information Retrieval Research". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Madrid, Spain. 2825–2837.

Schwartz, R., G. Stanovsky, S. Swayamdipta, J. Dodge, and N. A. Smith. (2020). "The Right Tool for the Job: Matching Model and Instance Complexities". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6640–6651.

Severyn, A. and A. Moschitti. (2015). "Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. 373–382.

Shen, Y., X. He, J. Gao, L. Deng, and G. Mesnil. (2014). "Learning Semantic Representations using Convolutional Neural Networks for Web Search". In: *Proceedings of the 23rd International Conference on World Wide Web*. ACM. 373–374.

Soldaini, L. and A. Moschitti. (2020). "The Cascade Transformer: an Application for Efficient Answer Sentence Selection". In: *ACL*.

Sorokina, D. and E. Cantú-Paz. (2016). "Amazon Search: The Joy of Ranking Products". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–460.

Sparck Jones, K. (1972). "A Statistical Interpretation of Term Specificity and its Application in Retrieval". *Journal of documentation*. 28(1): 11–21.

Strubell, E., A. Ganesh, and A. McCallum. (2019). "Energy and Policy Considerations for Deep Learning in NLP". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy. 3645–3650.

Swezey, R., A. Grover, B. Charron, and S. Ermon. (2021). "PiRank: Scalable Learning To Rank via Differentiable Sorting". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. 21644–21654.

Szummer, M. and E. Yilmaz. (2011). "Semi-supervised Learning to Rank with Preference Regularization". In: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. 269–278.

Tang, J. and K. Wang. (2018). "Ranking Distillation: Learning Compact Ranking Models With High Performance for Recommender System". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2289–2298.

Tang, X., X. Jin, and T. Yang. (2014). "Cache-conscious Runtime Optimization for Ranking Ensembles". In: *Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 1123–1126.

Tax, N., S. Bockting, and D. Hiemstra. (2015). "A Cross-benchmark Comparison of 87 Learning to Rank Methods". *Information Processing & Management*. 51(6): 757–772.

Taylor, M., J. Guiver, S. Robertson, and T. Minka. (2008). "SoftRank: Optimizing Non-Smooth Rank Metrics". In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*. Palo Alto, California, USA. 77–86.

Thakur, N., N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych. (2021). "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models". In: *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

The Guardian. (2017). "Google tells Army of 'Quality Raters' to Flag Holocaust denial". URL: https://www.theguardian.com/technology/2017/mar/15/google-quality-raters-flag-holocaust-denial-fake-news.

Tonellotto, N. and C. Macdonald. (2021). "Query Embedding Pruning for Dense Retrieval". In: *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*. Virtual Event, Queensland, Australia. 3453–3457.

Tseng, P. *et al.* (1988). "Coordinate Ascent for Maximizing Nondifferentiable Concave Functions".

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. (2017). "Attention is All You Need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, California, USA. 6000–6010.

Vinayak, R. K. and R. Gilad-Bachrach. (2015). "DART: Dropouts meet Multiple Additive Regression Trees". In: *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*. Ed. by G. Lebanon and S. V. N. Vishwanathan. Vol. 38. *Proceedings of Machine Learning Research*. San Diego, California, USA: PMLR. 489–497.

Wang, L., J. Lin, and D. Metzler. (2011). "A Cascade Ranking Model for Efficient Ranked Retrieval". In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Beijing, China. 105–114.

Wang, L., J. J. Lin, and D. Metzler. (2010a). "Learning to Efficiently Rank". In: *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 138–145.

Wang, L., D. Metzler, and J. Lin. (2010b). "Ranking Under Temporal Constraints". In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. Toronto, ON, Canada. 79–88.

Wang, M., X. Xu, Q. Yue, and Y. Wang. (2021a). "A Comprehensive Survey and Experimental Comparison of Graph-Based Approximate Nearest Neighbor Search". *Proc. VLDB Endow.* 14(11): 1964–1978.

Wang, S., S. Zhuang, and G. Zuccon. (2021b). "BERT-Based Dense Retrievers Require Interpolation with BM25 for Effective Passage Retrieval". In: *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. Virtual Event, Canada. 317–324.

Xia, F., T.-Y. Liu, J. Wang, W. Zhang, and H. Li. (2008). "Listwise Approach to Learning to Rank: Theory and Algorithm". In: *Proceedings of the 25th International Conference on Machine Learning*. Helsinki, Finland. 1192–1199.

Xie, Y., H. Dai, M. Chen, B. Dai, T. Zhao, H. Zha, W. Wei, and T. Pfister. (2020). "Differentiable Top-k with Optimal Transport". In: *Advances in Neural Information Processing Systems*. Vol. 33. 20520–20531.

Xin, J., R. Tang, J. Lee, Y. Yu, and J. Lin. (2020). "DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Xin, J., R. Tang, Y. Yu, and J. Lin. (2021). "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*. 91–104.

Xiong, C., Z. Dai, J. Callan, Z. Liu, and R. Power. (2017). "End-to-End Neural Ad-Hoc Ranking with Kernel Pooling". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Shinjuku, Tokyo, Japan. 55–64.

Xiong, L., C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. Bennett, J. Ahmed, and A. Overwijk. (2021). "Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval". In: *International Conference on Learning Representations*.

Xu, C. and J. McAuley. (2022). "A Survey on Model Compression and Acceleration for Pretrained Language Models". arXiv: 2202.07105 [cs.CL].

Xu, J., W. Zhou, Z. Fu, H. Zhou, and L. Li. (2021). "A Survey on Green Deep Learning". arXiv: 2111.05193 [cs.CL].

Xu, J. and H. Li. (2007). "AdaRank: a Boosting Algorithm for Information Retrieval". In: *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 391–398.

Xu, Z., M. J. Kusner, K. Q. Weinberger, and M. Chen. (2013). "Cost-sensitive Tree of Classifiers". In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. Atlanta, GA, USA. I-133–I-141.

Ye, T., H. Zhou, W. Y. Zou, B. Gao, and R. Zhang. (2018). "RapidScorer: Fast Tree Ensemble Evaluation by Maximizing Compactness in Data Level Parallelization". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 941–950.

Yilmaz, E. and S. Robertson. (2009). "Deep versus Shallow Judgments in Learning to Rank". In: *Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval*. 662–663.

Yin, D., Y. Hu, J. Tang, T. D. Jr., M. Zhou, H. Ouyang, J. Chen, C. Kang, H. Deng, C. Nobata, J.-M. Langlois, and Y. Chang. (2016). "Ranking Relevance in Yahoo Search". In: *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Yue, Y., J. Broder, R. Kleinberg, and T. Joachims. (2012). "The k-armed Dueling Bandits Problem". *Journal of Computer and System Sciences*. 78(5): 1538–1556.

Yue, Y. and T. Joachims. (2009). "Interactively Optimizing Information Retrieval Systems as a Dueling Bandits Problem". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 1201–1208.

Zamani, H., M. Bendersky, D. Metzler, H. Zhuang, and M. Najork. (2022). "Stochastic Retrieval-Conditioned Reranking". In: *Proceedings of the 2022 ACM SIGIR International Conference on the Theory of Information Retrieval*. Madrid, Spain.

Zhan, J., J. Mao, Y. Liu, J. Guo, M. Zhang, and S. Ma. (2021). "Jointly Optimizing Query Encoder and Product Quantization to Improve Retrieval Performance". In: *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*. Virtual Event, Queensland, Australia. 2487–2496.

Zhan, J., J. Mao, Y. Liu, J. Guo, M. Zhang, and S. Ma. (2022). "Learning Discrete Representations via Constrained Clustering for Effective and Efficient Dense Retrieval". In: *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*. Virtual Event, AZ, USA. 1328–1336.

Zhan, J., J. Mao, Y. Liu, M. Zhang, and S. Ma. (2020). "RepBERT: Contextualized Text Embeddings for First-Stage Retrieval". arXiv: 2006.15498 [cs.IR].

Zhang, Y., C. Hu, Y. Liu, H. Fang, and J. Lin. (2021). "Learning to Rank in the Age of Muppets: Effectiveness–Efficiency Tradeoffs in Multi-Stage Ranking". In: *Proceedings of the 2nd Workshop on Simple and Efficient Natural Language Processing*. 64–73.

Zhao, W. X., J. Liu, R. Ren, and J.-R. Wen. (2022). "Dense Text Retrieval based on Pretrained Language Models: A Survey". arXiv: 2211.14876 [cs.IR].

Zheng, Z., H. Zha, T. Zhang, O. Chapelle, K. Chen, and G. Sun. (2008). "A General Boosting Method and its Application to Learning Ranking Functions for Web Search". In: *Advances in Neural Information Processing Systems*. 1697–1704.

Zhuang, H., Z. Qin, S. Han, X. Wang, M. Bendersky, and M. Najork. (2021). "Ensemble Distillation for BERT-Based Ranking Models". In: *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. Virtual Event, Canada. 131–136.

Zhuang, S. and G. Zuccon. (2021). "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*. Virtual Event, Canada. 1483–1492.

Zhuang, S. and G. Zuccon. (2022). "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*.