A Proposed Efficiency Benchmark for Modern Information Retrieval Systems

Presented at ReNeuIR 2023 (at SIGIR 2023, Tapei)

Sebastian Bruch,[1] Joel Mackenzie,[2] Maria Maistro,[3] Franco Maria Nardini[4]

Pinecone,[1] The University of Queensland,[2] University of Copenhagen,[3] ISTI-CNR[4]

July 19, 2023

Abstract

The Information Retrieval community has a rich history of empirically measuring novel retrieval methods in terms of both effectiveness and efficiency. However, as the search ecosystem continues to rapidly develop, it is becoming difficult to empirically compare and contrast systems in a fair way. Factors including hardware configurations, software versioning, experimental settings, and measurement methods all contribute to the difficulty of meaningfully comparing search systems, especially where efficiency is a key component of the evaluation. Fortunately, this is not a new problem; many others have already considered these difficulties and proposed various solutions to solve them. In this proposal, we briefly highlight some of the key issues that the community faces with the state of benchmarking — especially in terms of what it means to be efficient — and propose a new campaign for benchmarking the current state-of-the-art search systems in a meaningful way. We hope to receive useful and actionable feedback on how the community can collectively move forward in this challenging yet important direction.

1 Introduction

This document outlines a rough proposal for an efficiency-focused benchmark for modern Information Retrieval (IR) systems. While number of prior efforts have been made in the broad direction of comparing systems in a unified way, none have gained sufficient traction to be widely adopted (or continued) in the field [4, 10, 11].[1] On the other hand, effectiveness leaderboards within IR — such as the MS MARCO passage and document ranking tasks [5, 2]) — and efficiency efforts outside of IR — like the *ANN* and the *BigANN* benchmarks [1, 19] — *have* achieved critical mass. Our aim is to solicit community feedback and to try and converge on a benchmark task that will motivate wide community participation in order to gain a clearer picture of the current state of modern IR systems.

2 Desiderata

While effectiveness measurement is fairly standardized (with some exceptions), it is often difficult to compare efficiency characteristics in a fair and repeatable manner; differences in hardware, software, measurement norms, treatment of data, and experimental settings mean that results for the *same system and experiment* can often differ from paper to paper, and from group to group. Furthermore, "efficiency" can mean a number of things [21, 3]. Traditionally, efficiency in IR has typically been measured via the online processing latency (or perhaps throughput) and the total space occupancy of the given system. More recently, however, other efficiency measures of interest have included the total computational resources — measured in CPU cycles or energy consumption — as well as subsequent CO2 emissions [18, 20] and even estimated water consumption [22]. Furthermore, the offline (training and indexing) cost of most systems is typically considered less important than the online serving efficiency.

[1] We refer the reader to the work of Fröbe et al. [8] who provide a detailed overview of these efforts.

The main goal of the proposed task is to facilitate a fair, reproducible comparison between the efficiency and effectiveness characteristics of both traditional and modern retrieval and ranking mechanisms to provide a more detailed picture of their inherent trade-offs. It is our view that both efficiency and effectiveness, together, are important factors that provide a snapshot of the total operating spectrum of interest (from highly efficient and less effective, to highly effective but less efficient, and everything in between); any system on a Pareto frontier is presumably a system of interest in trading between those dimensions. However, while most current benchmarks are *effectiveness-first* (with efficiency a secondary factor), what we are proposing here is an *efficiency-first* benchmark.

3 Tasks and Measurement

Typically, the "right" measurements and metrics depend on the task at hand. To begin with, we propose the use of a small to moderately sized collection to lower the barrier to entry. Two such exemplars are the `MSMARCO-v1` collection (with around 3.2 million documents and 8.8 million passages) [2] and the `ISTELLA22` collection (8.4 million multilingual documents with the majority being Italian and then English) [6]. Indeed, much recent research has been conducted on `MSMARCO-v1`, so a wide range of competitive systems could easily be submitted. In both cases, suitable metrics can include those evaluating candidate generation algorithms such as deep recall-based metrics, as well as shallow final-stage ranking metrics like NDCG@10. Differentiating between early and late stages of retrieval may provide further insights into system performance, as systems are often tailored towards a specific part of the end-to-end search pipeline.

Efficiency measurement then becomes the main challenge. Both latency and computational footprint should be measured, to account for systems that heavily parallelize their execution to accelerate latency. Space consumption can be readily measured via peak memory occupancy, index size, or both. Furthermore, it would be beneficial to measure power consumption and emissions if possible, and a number of open-source tools are readily available for instrumenting experiments in this way. A relatively standard approach to facilitate such measurement involves supplying a large log of queries to a system, and allowing them to be executed, with efficiency measurement being conducted across the lifetime of the execution. Indexing cost could be measured in a similar way. Scalability — in terms of measuring query throughput on a real-world query load — would be very useful to measure, but is likely too difficult to simulate in the first instance of the benchmark due to engineering constraints.

4 Possible Approaches

Next, we outline some assumptions of our benchmark, and then briefly highlight a series of possible approaches and their inherent trade-offs.

4.1 Assumptions

We make some basic assumptions regarding the benchmark.

1. We assume that some centralized and unified hardware/execution environment is available that can run all submitted systems. The specifics of this hardware may depend on the approach taken, but unified hardware is important for achieving a valid benchmark. It is also important to consider CPU, GPU, and memory constraints that may limit participants with what they can realistically submit.

2. We assume that effectiveness evaluation can be handled easily via existing evaluation tools and libraries.

3. We assume that all participant systems will have efficiency measured through the implementation of specified functions (such as `index()`, `search()`, and `evaluate()`) that will incorporate standardized measurement. This may be realized through the implementation of a containerized solution (such as a Docker image), similar to the "jig" tooling used in OSIRRC [4]. It is important to standardize measurement to ensure sound experimental outcomes.

4.2 Approach 1: Maximal Freedom

The first possible approach, and the least restrictive, is to allow participants to access a raw collection, build their systems, and submit them for measurement. A very small degree of standardization may be applied (through the specific `index()`, `search()`, and `evaluate()` functions), allowing participants to be free in their design and implementation (subject to the available hardware). However, this approach means that any differences in the participants local environment (where they build and test their system) and the benchmark execution environment may result in a number of problems from software or hardware incompatibilities, possibly voiding measurements. One possible remedy here is to provide each participant with access to a standardized system (such as an AWS instance), but this may not scale well in terms of cost.

4.3 Approach 2: Constrained Inputs

The first approach also suffers from a lack of standardization on how the input data is processed, which may also impact both efficiency and effectiveness comparisons. The second approach is a slightly more constrained version of the first, which aims to remedy this issue. The idea is to also constrain the input of each system, to allow for simpler and more fair comparisons. Constraining the input may involve using existing tooling such as ir_datasets [13] with a specific set of pre-processing rules to ensure all systems ingest the same plain/raw data.

4.4 Approach 3: Pre-Indexing

Taking the prior approach one step further, participants could provide their system *and* a pre-built index to offload computation from the shared hardware. The downside of this approach is that instrumenting the efficiency of the indexing stage would not be feasible, as participants would be expected to build indexes on their own hardware. Similarly, building and shipping large indexes may increase the barrier to entry for some groups.

4.5 Approach 4: Coupling with TIREx

Recently, [8] proposed the *Information Retrieval Experiment Platform* (TIREx) which aims to provide a standardized environment for reproducible and scalable research. In particular, TIREx solves many of the aforementioned problems regarding standardization by providing a unified environment and approach for running end-to-end experiments based on TIRA [9, 7]. TIREx provides a way of executing Docker images on data, including integration with existing tooling such as ir_datasets [13], ir_measures [14], and PyTerrier [16, 15]. The main barrier with this approach is how efficiency measurement can be incorporated into TIREx, and the possible lack of hardware configuration available. However, a recent NLP-focused *evaluation-as-a-service* framework [12] has successfully incorporated efficiency measurement, so this may not be a barrier.

A secondary version of this approach is to structure our own measurement tooling around the TIREx framework (in terms of the images and the expected API), but handle the efficiency measurement on other hardware. Further discussion with the TIREx maintainers is required to better understand whether each of these approaches are feasible, and the possible limitations therein.

5 Limitations and Future Work

The main aim of this proposal is to start small with a feasible benchmark/challenge that will interest — and be accessible to — the community. As such, there are many limitations and ideas that might be worth considering in future iterations; some of these are described here.

5.1 Hardware and Cost

One very real problem with a unified hardware environment is ensuring that there is sufficient time and resources available to run all of the submitted systems. It is unclear whether the computational burden of running, say, 50 unique systems from end-to-end will be too large for a single environment, especially if they are quite heavy systems. It is possible to employ a uniform environment that can be readily scaled (like providing a common Amazon EC2 instance to participants [11]) but it is unclear who should (or will!) pay for those instances. Similarly, if different instances are required to support varying levels of memory, CPU, and GPU capacity, moving beyond pure efficiency measurement to a more uniform *cost-per-query* approach may be necessary Santhanam et al. [17].

5.2 Multi-Dimensional Measurement

As discussed previously, efficiency has many dimensions. Deciding on what to measure (and in which context to measure it) is still an open question. One possible solution is to have a number of distinct challenges that limit resources in different ways, allowing *in-class* comparisons to be made meaningfully.

5.3 Tasks and Data

With a number of possible tasks and collections available, our focus was to select existing datasets that are large enough to be interesting from an efficiency standpoint, yet small enough to reduce the barrier of participation. However, it would be ideal to expand the benchmark to other (and possibly larger) collections in the future. In particular, many modern IR systems perform well on simple unstructured passage retrieval tasks; however, learning-to-rank systems are still strong competitors on collections containing rich structural features [6].

5.4 Implementation Details

A final shortcoming is that different implementations will result in different empirical performance due to aspects such as the programming language, engineering effort, specific optimizations, and so on. While it is not possible to handle these differences, we suggest that each participant can provide nuanced information on these details, thereby allowing a clearer picture of the trade-offs to be made.

6 Break-Out Session and Suggestions

We have provided a sketch of some possible approaches that could be used to benchmark IR systems, each with their own trade-offs. Please reflect on these approaches and provide your feedback. We have included some questions below to help you get started.

- Are you interested in participating in this sort of benchmark?
- Are there any barriers prohibiting your participation? If so, what are they?
- Would you like to be involved in organizing or running this sort of benchmark?
- Would you be likely to submit a system to a 2024 version of this benchmark challenge?
- What do you consider the most important efficiency dimensions to measure?
- Do you have any alternative suggestions or thoughts on these approaches?

Please take some notes and share them with the track organizers at the conclusion of the break-out session, or record them here: `https://forms.gle/yDkmRZMJsB7LJXY76`.

References

[1] M. Aumüller, E. Bernhardsson, and A. Faithfull. ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Information Systems*, 87:101374, 2020.

[2] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv:1611.09268v3*, 2018.

[3] S. Bruch, C. Lucchese, and F. M. Nardini. Efficient and effective tree-based and neural learning to rank. *Foundations and Trends in Information Retrieval*, 17(1):1–123, 2023.

[4] R. Clancy, N. Ferro, C. Hauff, J. Lin, T. Sakai, and Z. Z. Wu. Overview of the 2019 Open-Source IR Replicability Challenge (OSIRRC 2019). In *Proc. OSIRRC at SIGIR 2019*, pages 1–7, 2019.

[5] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, and J. Lin. MS MARCO: Benchmarking ranking models in the large-data regime. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1566–1576, 2021.

[6] D. Dato, S. MacAvaney, F. M. Nardini, R. Perego, and N. Tonellotto. The Istella22 dataset: Bridging traditional and neural learning to rank evaluation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3099–3107, 2022.

[7] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, and M. Potthast. Continuous integration for reproducible shared tasks with TIRA.io. In *Proceedings of the 45th European Conference on IR Research*, pages 236–241, 2023.

[8] M. Fröbe, J. H. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, and M. Potthast. The information retrieval experiment platform. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

[9] T. Gollub, B. Stein, S. Burrows, and D. Hoppe. TIRA: Configuring, executing, and disseminating information retrieval experiments. In *23rd International Workshop on Database and Expert Systems Applications*, pages 151–155, 2012.

[10] S. Hofstätter and A. Hanbury. Let's measure run time! Extending the IR replicability infrastructure to include performance aspects. In *Proc. OSIRRC at SIGIR 2019*, pages 12–16, 2019.

[11] J. Lin, M. Crane, A. Trotman, J. Callan, I. Chattopadhyaya, J. Foley, G. Ingersoll, C. Macdonald, and S. Vigna. Toward reproducible baselines: The open-source IR reproducibility challenge. In *Proceedings of the 38th European Conference on Information Retrieval*, pages 408–420, 2016.

[12] Z. Ma, K. Ethayarajh, T. Thrush, S. Jain, L. Y. Wu, R. Jia, C. Potts, A. Williams, and D. Kiela. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. In *Advances in Neural Information Processing Systems*, 2021.

[13] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, and N. Goharian. Simplified data wrangling with ir_datasets. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2429–2436, 2021.

[14] S. MacAvaney, C. Macdonald, and I. Ounis. Streamlining evaluation with ir-measures. In *Proceedings of the 44th European Conference on IR Research*, pages 305–310, 2022.

[15] C. Macdonald and N. Tonellotto. Declarative experimentation in information retrieval using pyterrier. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, 2020.

[16] C. Macdonald, N. Tonellotto, S. MacAvaney, and I. Ounis. Pyterrier: Declarative experimentation in Python from BM25 to dense retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4526–4533, 2021.

[17] K. Santhanam, J. Saad-Falcon, M. Franz, O. Khattab, A. Sil, R. Florian, M. A. Sultan, S. Roukos, M. Zaharia, and C. Potts. Moving beyond downstream task accuracy for information retrieval benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11613–11628, 2023.

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

[19] H. V. Simhadri, G. Williams, M. Aumüller, M. Douze, A. Babenko, D. Baranchuk, Q. Chen, L. Hosseini, R. Krishnaswamny, G. Srinivasa, S. J. Subramanya, and J. Wang. Results of the NeurIPS'21 challenge on billion-scale approximate nearest neighbor search. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages 177–189, 2022.

[20] E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, 2019.

[21] N. Tonellotto, C. Macdonald, and I. Ounis. Efficient query processing for scalable web search. *Foundations and Trends in Information Retrieval*, 12(4–5):319–500, 2018.

[22] G. Zuccon, H. Scells, and S. Zhuang. Beyond CO2 emissions: The overlooked impact of water consumption of information retrieval models. In *Proceedings of the 13th International Conference on the Theory of Information Retrieval*, page To appear, 2023.