

Wacky Weights in Learned Sparse Representations and the Revenge of Score-at-a-Time Query Evaluation

Joel Mackenzie,¹ Andrew Trotman,² Jimmy Lin³

¹ School of Computing and Information Systems, The University of Melbourne, Australia

² Department of Computer Science, University of Otago, Dunedin, New Zealand

³ David R. Cheriton School of Computer Science, University of Waterloo, Ontario, Canada

ABSTRACT

Recent advances in retrieval models based on learned sparse representations generated by transformers have led us to, once again, consider score-at-a-time query evaluation techniques for the top- k retrieval problem. Previous studies comparing document-at-a-time and score-at-a-time approaches have consistently found that the former approach yields lower mean query latency, although the latter approach has more predictable query latency. In our experiments with four different retrieval models that exploit representational learning with bags of words, we find that transformers generate “wacky weights” that appear to greatly reduce the opportunities for skipping and early exiting optimizations that lie at the core of standard document-at-a-time techniques. As a result, score-at-a-time approaches appear to be more competitive in terms of query evaluation latency than in previous studies. We find that, if an effectiveness loss of up to three percent can be tolerated, a score-at-a-time approach can yield substantial gains in mean query latency while at the same time dramatically reducing tail latency.

1 INTRODUCTION

Despite various investigations of alternatives over the past few decades, document-at-a-time (DAAT) query evaluation algorithms remain the dominant solution for the top- k retrieval problem that lies at the core of systems-focused information retrieval research as well as deployed production systems. The most recent systematic exploration of different query evaluation strategies that we are aware of is the work of Crane et al. [5], who compared the then latest score-at-a-time (SAAT) query evaluation algorithm against the WAND-family of DAAT index-traversal techniques [4, 9]. They concluded that despite advances in score-at-a-time query evaluation, the best approach at the time, known as JASS [23, 39], was still slower than Block-Max WAND (BMW) in terms of mean query latency, although the query latency of JASS was more consistent, with much lower tail latency [7].

In this work, we once again compare DAAT and SAAT query evaluation techniques, but in the context of a new class of retrieval models that rely on learned sparse representations based on transformers. A natural question to ask: Why are we relitigating these comparisons? Recent empirical studies have shown that this new class of retrieval models yields effectiveness that is superior to unsupervised sparse retrieval models such as BM25, and at least on par with popular dense retrieval techniques [17]. Thus, it is important to study the behavior of query evaluation techniques in the context of these models.

Interestingly, we observe that the neural models (typically, transformers) that underlie these learned sparse retrieval models often assign “wacky” weights to terms in bag-of-words representations.

The adjective “wacky” is used in at least two ways: First, the distribution of term weights do not appear to allow standard DAAT optimizations to effectively perform skipping and early exiting, which represent the source of accelerated query evaluation performance with traditional bag-of-words scoring models like BM25. Second, manual examination of the assigned term weights reveals that high scores are frequently assigned to terms in counter-intuitive ways, e.g., large weights placed on stopwords and subwords that lack any meaningful semantic content. These weight assignments are particularly puzzling since the models have been demonstrated to be more effective than existing bag-of-words models such as BM25 on various benchmark datasets.

The contributions of this work are twofold: First, we systematically characterize the wackiness of these learned sparse representations and their effects on query evaluation performance. Although these issues have been observed before [20, 29], our work confirms that these wacky weights are indeed quite pervasive. Second, we are the first to compare DAAT vs. SAAT approaches in this context and show that this wackiness differentially affects DAAT more than SAAT. More precisely, the weight distributions in learned sparse representations appear to greatly reduce opportunities for skipping and early exiting, thus reducing the efficiency of DAAT approaches. In contrast, since SAAT approaches rely less on advantageous weight distributions, slowdown is less pronounced. On the whole, if approximate query evaluation with an effectiveness loss of up to three percent can be tolerated, a score-at-a-time approach can yield substantial gains in mean query latency while at the same time dramatically reducing tail latency.

2 BACKGROUND AND RELATED WORK

The standard formulation of document retrieval using bag-of-words representations can be distilled into the following scoring function:

$$S_{d,q} = \sum_{t \in d \cap q} w_{d,t} \cdot w_{q,t} \quad (1)$$

where $w_{d,t}$ is the weight of term t in document d and $w_{q,t}$ represents the weight of term t in the query q . Document weights are typically computed via a function of term statistics such as tf, idf, doclength, etc. Query weights are often set to one, which simplifies query-document scores to the sum of document weights of terms that are found in the query. This formulation covers nearly all major families of retrieval models (probabilistic, vector space, language modeling, divergence from randomness, etc.) and is equivalent to the inner product of two weighted vectors of dimension $|V|$, where V is the vocabulary of the collection.

How to efficiently generate a top- k ranking of documents from an arbitrarily large collection $C = \{d_i\}_{i=0}^N$ in terms of $S_{d,q}$ has been,

quite literally, the subject of many decades of intense study; see Tonellotto et al. [38] for a survey. In the IR context, this is often referred to as the query evaluation problem, and nearly all modern algorithms exploit an inverted index where postings lists are systematically visited (i.e., traversed) in order to generate the top- k documents, stored in data structures such as min-heaps, accumulator tables, or a combination thereof.

Document-at-a-time (DAAT) and score-at-a-time (SAAT) are apt descriptions of the different approaches that query evaluation algorithms can take. At a high level, DAAT approaches work on postings lists that are monotonically sorted by document identifier, and achieve low latency query evaluation by “avoiding work”. Through the use of auxiliary data structures and per-term score upper-bounds, algorithms can cleverly work out when one or more documents cannot possibly be in the top- k and thereby “skip” them (often en masse). Presently, Variable Block-Max WAND (VBMW) [30] is generally acknowledged to represent the state of the art, although the best choice of DAAT traversal algorithm can depend on properties of the collection, query stream, ranker, and specific optimizations enabled [15, 25, 32, 35].

In contrast, SAAT approaches depend on term weights that have first been quantized into integers (called impact scores) and organized in postings lists grouped by descending impact scores [1]. Query evaluation proceeds by considering the score contributions of terms, from highest to lowest. Instead of “avoiding work”, the intuition behind SAAT is to organize computations to maximally take advantage of modern processors architectures (e.g., avoiding pipeline stalls, cache misses, etc.). The most recent take on SAAT query evaluation is the technique known as JASS [23, 39], which has the additional advantage of being an “anytime algorithm”: since term contributions are considered in decreasing order of importance (i.e., contributions to the final score), query evaluation can terminate at any time to yield an approximate ranking.

The experiments of Crane et al. [5] showed, on a number of standard IR test collections, that BMW was faster than JASS (i.e., lower mean query latency), but is susceptible to higher tail latency. That is, for some relatively small fraction of queries, ill-behaved term weights rendered BMW much slower than JASS. This behavior had been previously noted by Petri et al. [35], who examined the efficiency of both WAND and BMW under different bag-of-words ranking models. Subsequent work demonstrated that both DAAT and SAAT techniques can benefit from additional optimizations such as document identifier reassignment [27, 32].

The performance benefits of the various optimization techniques described above depend naturally on the scoring function. For example, the ability to skip blocks of postings depends on the relationship between the various term weights, and how they are distributed across the index (for block-based pruning approaches). Previous evaluations have been based on “traditional” scoring function such as BM25 or language models [35], but recently the field has seen the emergence of models where term weights are *learned* (so-called learned sparse representations [17, 18]). These models still rely on bag-of-words representations, and thus top- k retrieval can still be captured by Eq. (1), and all the foregoing discussions about DAAT vs. SAAT approaches still apply. However, term weights are now supplied by neural networks (today, transformer models), and learned in a supervised manner from large amounts of training data.

The first example of this class of models using transformers is DeepCT [6], which used a regression model to learn term weights. Mackenzie et al. [24] showed that these weights affect the behavior of DAAT query evaluation approaches, although, with appropriate mitigation, it is possible to accelerate query evaluation without harming effectiveness. In other words, the distribution of term weights assigned by the neural model is qualitatively different from BM25. Additional evidence for this finding comes from the work of Mallia et al. [29] in the context of their proposed model called DeepImpact: query latency is noticeably longer with term weights assigned by a transformer model, compared to the BM25 score over the same sets of terms in each document. To further compound this issue, some learned models also assign weights to *queries*, thereby changing the relative importance between the terms and potentially hindering efficiency further. These interesting observations suggest that we should take another detailed look at query evaluation algorithms in the context of learned sparse retrieval models, and thus the “previously settled” DAAT vs. SAAT debate should be reopened.

Note that our work only examines learned *sparse* representations for retrieval, where documents are represented by bags of words, i.e., the basis of the vector representation is the vocabulary space. There is, of course, another large class of learned *dense* representations, exemplified by models such as DPR [14] and ANCE [44]; see Lin et al. [22] for a survey. Although transformer-based models are involved in both classes of retrieval models, dense retrieval techniques require a completely different “software stack”. For example, top- k retrieval is formulated as nearest neighbor search and implemented with approximate techniques such as HNSW [28]. These techniques do not use inverted indexes and thus our discussions of DAAT and SAAT query evaluation algorithms are not applicable. However, for the interested reader, a number of papers have attempted to draw conceptual connections between dense and sparse learned representations [17, 18, 20].

3 EXPERIMENTAL SETUP

Our experiments used the popular MS MARCO passage corpus, comprising 8.8M passages, and all models were evaluated on the 6980 queries in the development set [3]. While it would have been desirable to explore multiple test collections, we are limited by the trained models that are publicly available for download, and training models from scratch is beyond the scope of this work. Unfortunately, most models today have only been trained and evaluated on the MS MARCO data; evaluations on more test collections will be explored in future work.

3.1 Retrieval Models

As points of comparison, we adopted the following baselines:

BM25 simply performs retrieval using the ubiquitous BM25 scoring function [37] over bag-of-words representations of the passages in the corpus. We set $k_1 = 0.82$ and $b = 0.68$, based on tuning on a selection of training instances on the MS MARCO passage ranking test collection [19].

BM25 w/ doc2query-T5 [33, 34] (BM25-T5 for short) augments passages in the corpus with query predictions generated by the T5 [36] neural sequence-to-sequence model. The expanded passages are scored using BM25 at retrieval time, with the same BM25

formulation and parameters above. Thus, while neural models are involved in corpus preparation, the assignment of term weights does not involve any neural networks.

We examined the following retrieval models that leverage sparse learned representations using transformers:

DeepImpact [29] uses doc2query-T5 to identify dimensions in the passage’s bag-of-words representation that should have non-zero weights (i.e., expansion terms) and learns a term weighting model based on a pairwise loss between relevant and non-relevant passages with respect to a query.

uniCOIL + doc2query-T5 [18] (uniCOIL-T5 for short) is a simplified variant of COIL [12] that assigns scalar weights to terms (as opposed to vector weights in the original COIL formulation). This model additionally benefits from doc2query-T5 expansions.

uniCOIL + TILDE [46] (uniCOIL-TILDE for short) can be best characterized as replacing the doc2query-T5 expansion component with an alternative model based on TILDE [47] that has lower inference costs but appears to be just as effective.

SPLADEv2 [10] represents an improvement over SPLADEv1 [11], which itself builds on SparTerm [2]. For this family of sparse retrieval models, the expansion component can be best characterized as being based on masked language modeling. SPLADEv2 further improves effectiveness via distillation techniques.

Note that for all the models, our experiments are based on our own implementation of code and generation of data, with Anserini as a starting point (see details below). In all applicable cases, we started with checkpoints of the neural models provided by the authors. Thus, our results are quite close, but not exactly the same, as figures reported in the respective authors’ original papers. For the sparse learned retrieval models, the corpus and queries are both pretokenized. The corpus already includes term weights for each term, and the same for the queries. Thus, none of these experiments involved neural inference, which eliminates a source of non-determinism in neural models.

3.2 Systems

We conducted experiments with three different retrieval systems:

Anserini [45] is an IR toolkit built on the open-source Lucene search library written in Java. The version of Anserini used in our experiments is based on Lucene 8.3.0, which uses BM25 for query evaluation [13]. We made no special modifications to the default query evaluation settings.

PISA [31] is an efficiency-focused IR system written in C++. In our experiments, PISA executes MaxScore [41] processing over SIMD-BP128 encoded postings lists [16]. While VBMW is currently the state-of-the-art approach for DAAT traversal, recent work has demonstrated that MaxScore is a better choice for larger values of k (such as $k = 1000$) and for long queries.

JASS [23, 39] is, as far as we are aware, the only actively maintained open-source SAAT search engine available today. It, too, was implemented in C++ for efficiency and uses a SIMD-enhanced Elias gamma encoding scheme for compressing the postings [40]. Query evaluation proceeds by processing impact-ordered postings using simple integer arithmetic and storing the results in an accumulator

array managed with a unique page-table like algorithm. JASS controls the tradeoff between query latency and effectiveness through a parameter ρ , which limits the total number of postings processed per query. Although JASS usually employs 16-bit accumulators (allowing maximum per document scores of 2^{16}), 32-bit accumulators were necessary in these experiments in order to avoid overflows, as the learned sparse impacts and weights often result in scores exceeding $2^{16} = 65,536$. A preliminary benchmark suggests that moving from 16 to 32-bit accumulators adds an almost 50% overhead to the exact (rank-safe) BoW BM25 run in Table 1, row (3a).

In all cases, our implementations used the “pseudo-document” trick to assign custom impact scores for each document term. That is, if term X was assigned a (quantized) integer weight of ten by the transformer, we simply repeat the term ten times in a dynamically created “fake” document. This allows us to use all the existing systems without any modifications, by simply swapping in “sum of term frequency” as the scoring function.

Finally, all experiments were conducted in memory using a single thread on a Linux machine with two 3.50 GHz Intel Xeon Gold 6144 CPUs and 512 GiB of RAM. Indexes were built with Anserini and exported to the Common Index File Format [21] before being imported into PISA and JASS. In addition, both PISA and JASS made use of document reordering [8, 26] which has been shown to improve compression and accelerate query processing in both DAAT and SAAT [27]. In our experiments, all systems retrieved the top $k = 1000$ documents.

4 RESULTS

4.1 Main Findings

Our main results are presented in Table 1, where each combination of retrieval model and system is characterized in terms of output quality, time, and space. Output quality is measured in terms of mean reciprocal rank at cutoff 10 (RR@10), the official metric of the test collection; time is measured in terms of query latency in milliseconds, and space in terms of index size measured in megabytes. While PISA and JASS were designed specifically as platforms for research in query evaluation algorithms and thus engineered for speed, none of the systems had small index sizes explicitly as a design goal (beyond incidental effects on query evaluation performance that result from index compression techniques).

Each block of the table is devoted to a system, and rows represent each retrieval model. Note that JASS results occupy two blocks; the top one represents *exact* (i.e., rank-safe) processing (where all postings are processed) and the bottom one represents *approximate* processing. For the latter, we used a maximum of $\rho = 1$ million postings visited, based on heuristics provided by the JASS authors [23, 39].

There are a few takeaways from these results: First, our experiments confirm what previous researchers have already discovered [24, 29]—that retrieval models based on learned sparse representations substantially alter the behavior of query evaluation algorithms. However, our results more thoroughly and comprehensively capture the effects, illustrating that these issues pervade sparse learned models more broadly. For DAAT approaches, both Anserini (Lucene) and PISA, we see that, in general, models with higher effectiveness are slower. In Lucene, the difference between

Method	Quality		Time	Space
	RR@10	Latency (ms)	Index Size (MB)	
Anserini (Lucene): DAA T				
(1a) BM25	0.187	40.1	661	
(1b) BM25-T5	0.277	62.8	1036	
(1c) DeepImpact	0.325	244.1	1417	
(1d) uniCOIL-T5	0.352	222.3	1313	
(1e) uniCOIL-TILDE	0.350	194.6	2067	
(1f) SPLADEv2	0.369	2140.0	4987	
PISA: DAA T				
(2a) BM25	0.187	8.3	739	
(2b) BM25-T5	0.276	11.9	1150	
(2c) DeepImpact	0.326	19.4	1564	
(2d) uniCOIL-T5	0.352	36.9	1358	
(2e) uniCOIL-TILDE	0.350	28.4	2108	
(2f) SPLADEv2	0.369	220.3	4326	
JASS: SAA T				
<i>Exact</i>				
(3a) BM25	0.187	15.8	1156	
(3b) BM25-T5	0.277	50.2	1452	
(3c) DeepImpact	0.326	39.3	2039	
(3d) uniCOIL-T5	0.352	147.2	1310	
(3e) uniCOIL-TILDE	0.350	83.5	1976	
(3f) SPLADEv2	0.369	314.0	3813	
<i>Approximate</i>				
(4a) BM25	0.186	12.4	1156	
(4b) BM25-T5	0.275	10.1	1452	
(4c) DeepImpact	0.319	12.6	2039	
(4d) uniCOIL-T5	0.338	14.9	1310	
(4e) uniCOIL-TILDE	0.338	15.4	1976	
(4f) SPLADEv2	0.324	15.3	3813	

Table 1: Experimental results on the development queries of the MS MARCO passage ranking test collection.

BoW BM25 in row (1a) and the most effective model, SPLADEv2 in row (1f) is a whopping $\sim 50\times$ slowdown. With PISA, the slowdown is “only” $\sim 25\times$, but nevertheless still quite dramatic. Even for models that are slightly less effective than SPLADEv2, for example, the uniCOIL models in rows (d) and (e), we observe substantially worse query evaluation performance, in both Lucene and PISA.

Second, it is clear that the query evaluation performance of Anserini (Lucene) is far behind that of PISA, both in absolute terms as well as relative degradation with the models that we have explored. Of course, this is not a fair comparison because Lucene is production-ready search library that is already widely deployed, whereas PISA is a research system. The performance deficiencies of Lucene have been discussed elsewhere [13] so we don’t beleaguer the point further, so for the remainder of this paper we focus our comparisons on PISA.

To further examine different query evaluation algorithms, we applied WAND and BMW to SPLADEv2 in PISA in a side experiment and found that these algorithms resulted in *slower* processing than an *exhaustive ranked disjunction*, with mean latency values of 619ms, 681ms, and 553ms, respectively. In our experiments, MaxScore vastly outperforms the WAND-based approaches; this result

Method	V	Terms in Documents		Terms in Queries	
		Total	Unique	Total	Unique
BM25	2660824	39.8	30.1	5.9	5.8
BM25-T5	3929111	224.7	51.1	5.9	5.8
DeepImpact	3514102	4010.0	71.1	4.2	4.2
uniCOIL-T5	27678	5032.3	66.4	686.3	6.6
uniCOIL-TILDE	27646	8260.8	107.6	661.1	6.5
SPLADEv2	28131	10794.8	229.4	2037.8	25.0

Table 2: Term statistics of documents and queries for the different treatments of the MS MARCO passage corpus.

is due to MaxScore avoiding expensive sorting operations during query processing [32]. Essentially, WAND and BMW are “working hard” to compute which documents can be skipped, but the work is essentially wasted because in reality, few documents can actually be skipped if exact query evaluation is desired. This, once again, shows that procrastination pays.

Third, the JASS results from Table 1 are largely consistent with what we already know about SAA T approaches from the literature. Exact query evaluation (i.e., exhaustively traversing all postings) with JASS is slower than PISA, but achieves comparable effectiveness. For the learned sparse retrieval models, although we do observe performance degradation with JASS as well, the slowdowns are slightly less than what we observed with PISA; “only” around $20\times$ when comparing BoW BM25 (3a) with SPLADEv2 (3f).

The approximate query evaluation results with JASS are also consistent with previous findings. For “traditional” BM25 weighting, in rows (4a) and (4b), JASS is able to speed up query evaluation at the cost of small decreases in effectiveness. However, as the “model sophistication” increases, the heuristics provided by the JASS authors appear to result in an increased loss in effectiveness. Looking at rows (4d) to (4f), while JASS is much faster than PISA, it comes at a drop in effectiveness. Although the loss is less than 1% for BM25, and 4% on DeepImpact and UniCOIL, on SPLADEv2, the loss is over 12%. Table 1 only captures two operating points for JASS, but we explore additional configurations in Section 4.3 that yield different effectiveness/efficiency tradeoffs.

Finally, in terms of index sizes, there’s a general trend of larger index sizes as effectiveness increases; the underlying reasons will become obvious in the next section. However, for all models and systems, the space requirements are quite modest on the whole. In the context of modern servers, where terabytes of disk storage are common, these differences are negligible.

4.2 Wacky Weights

What might be the cause of the behavior described in the previous section? Table 2 begins to answer some of these questions. The two main sections in the table show descriptive statistics of the documents and queries. An explanation about the total vs. unique terms: the total number of terms includes duplicated terms in our pseudo-documents, as described in Section 3.2. Thus, total terms is best understood as the sum of all the weights assigned to all unique terms, either in the document or the query. The second column, |V|, denotes the vocabulary size, or the total number of unique terms

in the collection; this is also the number of dimensions that are in the representation vectors of the documents and queries.

A few additional caveats are necessary for properly understanding these figures. All statistics are computed based on our replication of the retrieval models, and thus may differ from figures reported in the authors’ original papers due to differences in corpus preparation. In all cases, counts are computed by simply splitting text on whitespace. For BM25 and BM25-T5, these are *prior* to tokenization, stopword removal, etc. For uniCOIL and SPLADEv2, both document and queries are pretokenized. Thus, there are qualitative differences, especially since uniCOIL and SPLADEv2 tokens are subwords derived from BERT, e.g., “androgen receptor” is broken into “and” and “#rogen receptor”. Thus, for this reason, uniCOIL and SPLADEv2 have much smaller vocabulary sizes.

Nevertheless, these statistics go a long way to explaining many of the observations from Table 1. It is clear that learned sparse retrieval models achieve higher effectiveness based on document expansion, and in some models, query expansion. Document expansion obviously yields longer documents (and bigger indexes), and together with query expansion, it is no mystery why query evaluation efficiency degrades. These models, however, begin to give us a more nuanced look. For example, comparing DeepImpact and uniCOIL-T5, we see that the latter performs query expansion, whereas the former doesn’t. This is likely the biggest source of query latency differences, row (2c) vs. row (2d) in Table 1. Comparing the T5 and TILDE variants of uniCOIL, we observe that they achieve comparable effectiveness, but T5 is more “compact” in performing less document expansion. However, counter-intuitively, uniCOIL-TILDE, row (2e), is actually faster than uniCOIL-T5, row (2d). We don’t presently have a good explanation for this. Finally, we note that SPLADEv2 likely derives its superior effectiveness from performing more expansion. Compared to the uniCOIL models, the SPLADEv2 queries contain around 4× more unique terms and the documents contain 2 – 4× more unique terms.

The wackiness of term weights assigned by learned sparse retrieval models is evident based on manual examination of model output. Building on the tokenization example above, for the query “androgen receptor define”, the full SPLADEv2 query includes 25 unique tokens, representing an increase of 21 tokens beyond those in the original query, “and #rogen receptor define”. For these tokens, the model assigns the following weights: (“and”, 225), (“#rogen”, 251), (“receptor”, 242), and (“define”, 59). These weights seem reasonable, although the large weight on “and” is a bit surprising, given that the subword is conflated with a stopword. Many expansion terms that SPLADEv2 adds to the query do make sense, for example, (“hormone”, 179), (“definition”, 162), and (“meaning”, 99). However, puzzling is the fact that the model also adds many stopwords to the query, including (“is”, 70), (“the”, 56), (“for”, 46), and (“are”, 32). Most non-sensical is the fact that “,” (yes, the comma) is added as an expansion term, with a relatively large weight of 68 (srsly, wtf?). Note that as far as we are aware, SPLADEv2 was trained without any special corpus processing, and punctuation are indeed part of BERT’s vocabulary. In some cases, punctuation are semantically meaningful (e.g., apostrophes in names like O’Toole), but it is difficult to see the role that the comma might play in the context of this query. While it is difficult to argue with the overall

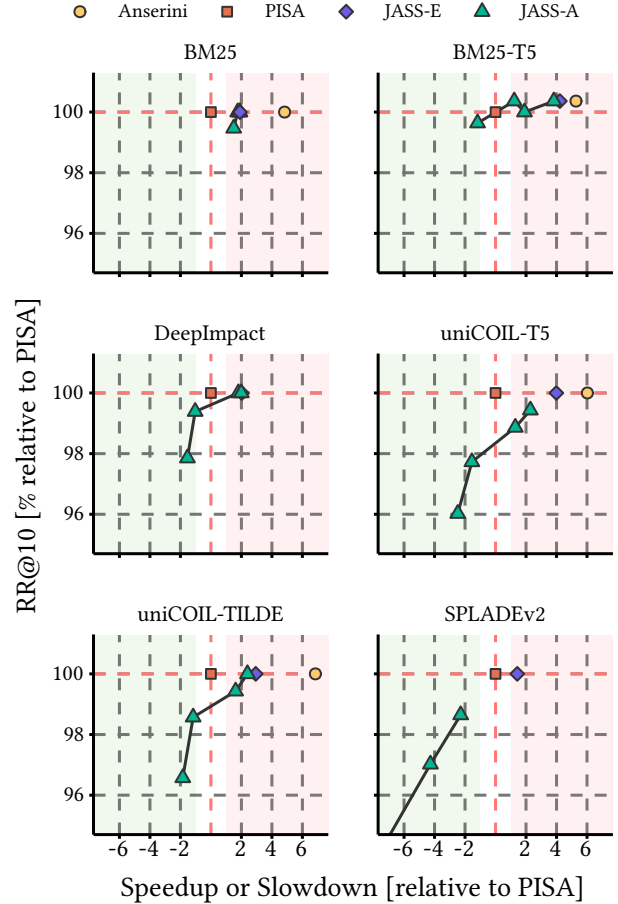


Figure 1: Relative comparisons between each system configuration for each retrieval model. The x -axis reports the relative speedup (negative values) or slowdown (positive values) with respect to PISA; the y -axis reports the relative effectiveness as a percentage of mean RR@10 achieved by PISA.

effectiveness results (the high mean RR@10), we nevertheless find these weights “wacky”.

4.3 Effectiveness/Efficiency Tradeoffs

One main advantage of SAAT approaches, and JASS specifically, is the ability to trade effectiveness for efficiency by processing varying amounts of postings. Since the query evaluation algorithm considers segments of postings in decreasing order of importance, it is possible to quit at any time, yielding approximate results. These tradeoffs are shown in Figure 1, with a plot for each model.

In each plot, we have normalized both the effectiveness and efficiency of the PISA configuration to one, represented as a red square at (1, 100), and thus both effectiveness (in percent) and latency speedup/slowdown are in relative terms (negative x represents speedup, positive x represents slowdown). Lucene is represented as a yellow circle; it is just as effective as PISA but substantially slower. JASS with exhaustive query evaluation (JASS-E) is shown as a blue diamond; in general, its effectiveness is comparable to PISA and

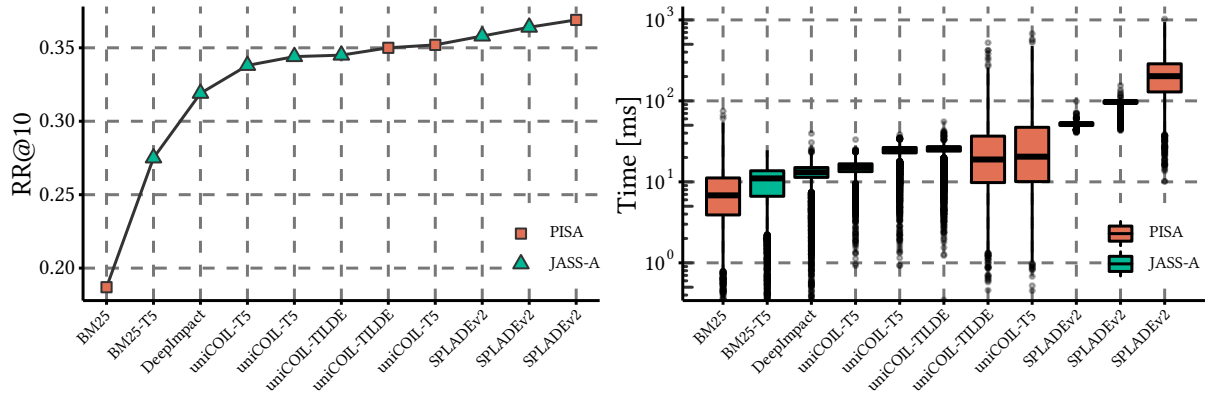


Figure 2: Effectiveness (left) and distribution of query latency (right) for all configurations along the Pareto-optimal frontier from Figure 3. DAAT MaxScore in PISA exhibits much greater variability compared to the SAAT approach in JASS.

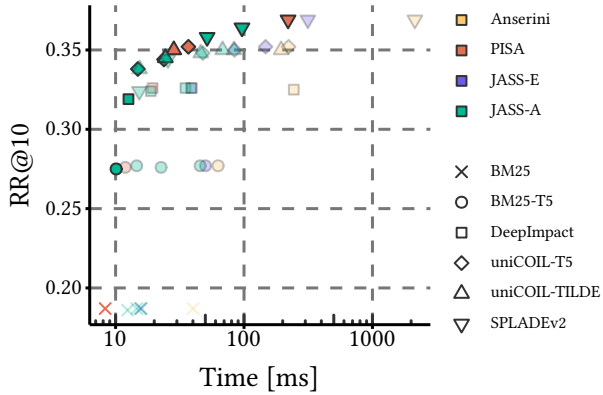


Figure 3: Efficiency (mean query latency) vs. effectiveness (mean RR@10) for all configurations; those along the Pareto-optimal frontier are highlighted. Note that *every* retrieval model is Pareto-optimal under at least one configuration for either PISA or JASS. Lucene does not appear on the Pareto frontier.

Lucene, while being slower than PISA but faster than Lucene. The green triangles capture the tradeoff curve of JASS with approximate query evaluation using $\rho \in \{1, 2, 5, 10\}$ million postings; in the case of SPLADEv2, some points are outside the plot area.

By altering ρ we can trade effectiveness for efficiency to differing degrees, beyond the two operating points reported in Table 1. For example, considering the most effective model, SPLADEv2, the loss in mean RR@10 is less than 2% with a mean query latency of 96ms (half the time of PISA) with $\rho = 10m$. If even more effectiveness loss is tolerable, using $\rho = 5m$ can yield a mean query latency of 52ms (quarter of the time of PISA) with a 3% loss in mean RR@10. On the other hand, PISA is a better choice if no effectiveness loss is acceptable, regardless of the retrieval model.

In Figure 3 we have combined all six subplots in Figure 1 into a single plot. Here, we report absolute measures of effectiveness (mean RR@10) as well as efficiency (mean query latency). This

plot captures all combinations of retrieval models (as shapes) and systems (as colors). Combinations that are on the Pareto-optimal frontier are highlighted.

In this context, the Pareto-optimal frontier represents the best tradeoff that can be obtained for all explored combinations of retrieval models \times systems. If a point lies on the frontier, it means that no other setting is able to achieve *both* higher effectiveness and lower mean query latency. There is no principled way to identify “better” or “worse” configurations along the frontier, as the selection of the operating point depends on the application scenario. Points along the frontier represent that best of “what’s possible” for the system designer to choose from.

Quite amazingly, we note that *all* models lie on some part of the frontier. This means that there is no single model that dominates all others. Furthermore, PISA and JASS-A (approximate query evaluation) share points along the frontier. This means that depending on the desired tradeoffs between effectiveness and efficiency, the optimal choice varies, both in terms of the retrieval model as well as the query evaluation approach!

Finally, Figure 2 provides more details on the configurations along the Pareto-optimal frontier. In the left plot, we have arranged the configurations along the x axis to clearly show each model/system combination, with the corresponding mean RR@10 shown along the y-axis. Here, we can clearly see that each retrieval model represents an optimal choice given some desired level of effectiveness. In the right plot, we show the distribution of query latency as standard Tukey box-and-whiskers plots; it is important to note that the selection of the Pareto-optimal configurations was performed using *mean* query latency, and means might differ from medians due to outliers.

Confirming prior work [5], we observe that while the DAAT query evaluation can outperform SAAT query evaluation in terms of mean latency, better performance comes at the expense of less predictable latency, i.e., tail latency. On the other hand, SAAT query evaluation, by design, yields much more predictable latency, as it enforces a strict limit on the total amount of allocated computation on a per query basis. It is also worth noting the lack of variance in the latency of JASS with the SPLADEv2 model. Since SPLADEv2

aggressively expands queries (see Table 2), JASS almost always processes ρ postings, resulting in highly predictable performance. In contrast, for some of the other models, queries may have fewer than ρ postings to process, resulting in the “lower tails” (i.e., queries that are much faster), for example, in JASS with BM25-T5.

5 CONCLUSIONS

Retrieval models based on sparse learned representations are relatively new innovations. Most work has focused on evaluating model effectiveness, but here we build on previous studies to examine effectiveness/efficiency tradeoffs. We find that, indeed, term weights generated by such models have substantive effects on the behavior of query processing algorithms, and that they differentially affect DAAT vs. SAAT. The net effect is to make SAAT algorithms more attractive in general.

In future work, it would be interesting to include the recently proposed *anytime* DAAT algorithms to the comparison [27], as they provide similar work-limiting (and, in turn, tail-latency minimizing) guarantees as SAAT retrieval. Furthermore, expanding our experimental comparisons to different collections would help us understand corpus effects. Finally, an obvious improvement for retrieval models based on learned sparse representations would be to incorporate efficiency considerations into the model training objective, much like the “learning to efficiently rank” thread of work from a decade ago [42, 43]. How to formulate an appropriate loss, however, is not obvious. Nevertheless, there are many exciting directions we are interested in pursuing in the near future.

REFERENCES

- [1] V. N. Anh, O. de Kretser, and A. Moffat. Vector-space ranking with effective early termination. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pages 35–42, 2001.
- [2] Y. Bai, X. Li, G. Wang, C. Zhang, L. Shang, J. Xu, Z. Wang, F. Wang, and Q. Liu. SparTerm: Learning term-based sparse representation for fast text retrieval. *arXiv:2010.00768*, 2020.
- [3] 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.
- [4] A. Z. Broder, D. Carmel, M. Herscovici, A. Soffer, and J. Zien. Efficient query evaluation using a two-level retrieval process. In *Proceedings of the Twelfth International Conference on Information & Knowledge Management (CIKM 2003)*, pages 426–434, 2003.
- [5] M. Crane, J. S. Culpepper, J. Lin, J. Mackenzie, and A. Trotman. A comparison of document-at-a-time and score-at-a-time query evaluation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM 2017)*, pages 201–210, 2017.
- [6] Z. Dai and J. Callan. Context-aware sentence/passage term importance estimation for first stage retrieval. *arXiv:1910.10687*, 2019.
- [7] J. Dean and L. A. Barroso. The tail at scale. *Communications of the ACM*, 56(2):74–80, 2013.
- [8] L. Dhulipala, I. Kabiljo, B. Karrer, G. Ottaviano, S. Pupyrev, and A. Shalita. Compressing graphs and indexes with recursive graph bisection. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*, pages 1535–1544, 2016.
- [9] S. Ding and T. Suel. Faster top- k document retrieval using block-max indexes. In *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*, pages 993–1002, 2011.
- [10] T. Formal, C. Lassance, B. Piwowarski, and S. Clinchant. SPLADE v2: Sparse lexical and expansion model for information retrieval. *arXiv:2109.10086*, 2021.
- [11] T. Formal, B. Piwowarski, and S. Clinchant. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2288–2292, 2021.
- [12] L. Gao, Z. Dai, T. Chen, Z. Fan, B. V. Durme, and J. Callan. Complementing lexical retrieval with semantic residual embedding. In *Proceedings of the 43rd European Conference on Information Retrieval (ECIR 2021), Part I*, pages 146–160, 2021.
- [13] A. Grand, R. Muir, J. Ferenczi, and J. Lin. From MaxScore to Block-Max WAND: The story of how Lucene significantly improved query evaluation performance. In *Proceedings of the 42nd European Conference on Information Retrieval, Part II (ECIR 2020)*, pages 20–27, 2020.
- [14] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.
- [15] O. Khattab, M. Hammoud, and T. Elsayed. Finding the best of both worlds: Faster and more robust top- k document retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, pages 1031–1040, 2020.
- [16] D. Lemire and L. Boytsov. Decoding billions of integers per second through vectorization. *Software: Practice and Experience*, 41(1):1–29, 2015.
- [17] J. Lin. A proposed conceptual framework for a representational approach to information retrieval. *arXiv:2110.01529*, 2021.
- [18] J. Lin and X. Ma. A few brief notes on DeepImpact, COIL, and a conceptual framework for information retrieval techniques. *arXiv:2106.14807*, 2021.
- [19] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, and R. Nogueira. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362, 2021.
- [20] J. Lin, X. Ma, J. Mackenzie, and A. Mallia. On the separation of logical and physical ranking models for text retrieval applications. In *Proceedings of the 2nd International Conference on Design of Experimental Search & Information Retrieval Systems (DESIRE 2021): CEUR Workshop Proceedings Vol-2950*, pages 176–178, 2021.
- [21] J. Lin, J. Mackenzie, C. Kamphuis, C. Macdonald, A. Mallia, M. Siedlaczek, A. Trotman, and A. de Vries. Supporting interoperability between open-source search engines with the common index file format. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, pages 2149–2152, 2020.
- [22] J. Lin, R. Nogueira, and A. Yates. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Morgan & Claypool Publishers, 2021.
- [23] J. Lin and A. Trotman. Anytime ranking for impact-ordered indexes. In *Proceedings of the ACM International Conference on the Theory of Information Retrieval (ICTIR 2015)*, pages 301–304, 2015.
- [24] J. Mackenzie, Z. Dai, L. Gallagher, and J. Callan. Efficiency implications of term weighting for passage retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, pages 1821–1824, 2020.
- [25] J. Mackenzie and A. Moffat. Examining the additivity of top- k query processing innovations. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM 2020)*, pages 1085–1094, 2020.
- [26] J. Mackenzie, M. Petri, and A. Moffat. Faster index reordering with bipartite graph partitioning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 1910–1914, 2021.
- [27] J. Mackenzie, M. Petri, and A. Moffat. Anytime ranking on document-ordered indexes. *ACM Transactions on Information Systems*, 40(1):13.1–13.32, Jan. 2022. To appear.
- [28] Y. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020.
- [29] A. Mallia, O. Khattab, T. Suel, and N. Tonello. Learning passage impacts for inverted indexes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 1723–1727, 2021.
- [30] A. Mallia, G. Ottaviano, E. Porciani, N. Tonello, and R. Venturini. Faster Block-Max WAND with variable-sized blocks. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, pages 625–634, 2017.
- [31] A. Mallia, M. Siedlaczek, J. Mackenzie, and T. Suel. PISA: Performant indexes and search for academia. In *Proceedings of the Open-Source IR Replicability Challenge (OSIRRC 2019): CEUR Workshop Proceedings Vol-2409*, pages 50–56, 2019.
- [32] A. Mallia, M. Siedlaczek, and T. Suel. An experimental study of index compression and DAAT query processing methods. In *Proceedings of the 41st European Conference on Information Retrieval (ECIR 2019)*, pages 353–368, 2019.
- [33] R. Nogueira and J. Lin. From doc2query to docTTTTTquery, 2019.
- [34] R. Nogueira, W. Yang, J. Lin, and K. Cho. Document expansion by query prediction. *arXiv:1904.08375*, 2019.
- [35] M. Petri, J. S. Culpepper, and A. Moffat. Exploring the magic of WAND. In *Proceedings of the 18th Australasian Document Computing Symposium (ADCS 2013)*, pages 58–65, 2013.
- [36] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text

- transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [37] S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.
 - [38] 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.
 - [39] A. Trotman and M. Crane. Micro- and macro-optimizations of SAAT search. *Software: Practice and Experience*, 49(5):942–950, 2019.
 - [40] A. Trotman and K. Lilly. Elias revisited: Group Elias SIMD coding. In *Proceedings of the 23rd Australasian Document Computing Symposium (ADCS 2018)*, pages 4.1–4.8, 2018.
 - [41] H. R. Turtle and J. Flood. Query evaluation: Strategies and optimizations. *Information Processing & Management*, 31(6):831–850, 1995.
 - [42] L. Wang, J. Lin, and D. Metzler. Learning to efficiently rank. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*, pages 138–145, 2010.
 - [43] L. Wang, J. Lin, and D. Metzler. A cascade ranking model for efficient ranked retrieval. In *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*, pages 105–114, 2011.
 - [44] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. N. Bennett, J. Ahmed, and A. Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*, 2021.
 - [45] P. Yang, H. Fang, and J. Lin. Anserini: Reproducible ranking baselines using Lucene. *Journal of Data and Information Quality*, 10(4):Article 16, 2018.
 - [46] S. Zhuang and G. Zuccon. Fast passage re-ranking with contextualized exact term matching and efficient passage expansion. *arXiv:2108.08513*, 2021.
 - [47] S. Zhuang and G. Zuccon. TILDE: Term independent likelihood moDEL for passage re-ranking. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 1483–1492, 2021.