# Effect of Dynamic Pruning Safety on Learning to Rank Effectiveness

Craig Macdonald[1], Nicola Tonellotto[2], Iadh Ounis[1]
[1] University of Glasgow, Glasgow, G12 8QQ, UK
[2] National Research Council of Italy, Via G. Moruzzi 1, 56124 Pisa, Italy
{craig.macdonald, iadh.ounis}@glasgow.ac.uk[1], {nicola.tonellotto}@isti.cnr.it[2]

## ABSTRACT

A dynamic pruning strategy, such as WAND, enhances retrieval efficiency without degrading effectiveness to a given rank $K$, known as safe-to-rank-$K$. However, it is also possible for WAND to obtain more efficient but unsafe retrieval without actually significantly degrading effectiveness. On the other hand, in a modern search engine setting, dynamic pruning strategies can be used to efficiently obtain the set of documents to be re-ranked by the application of a learned model in a learning to rank setting. No work has examined the impact of safeness on the effectiveness of the learned model. In this work, we investigate the impact of WAND safeness through experiments using 150 TREC Web track topics. We find that unsafe WAND is biased towards documents with lower docids, thereby impacting effectiveness.

**Categories & Subject Descriptors:** H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

**General Terms:** Performance, Experimentation

**Keywords:** Dynamic Pruning, Learning to Rank

## 1. INTRODUCTION

A search engine deploying learning to rank techniques reranks the top $K$ documents retrieved by a standard weighting model, known as the sample [3], as shown in Figure 1. To improve the efficiency of such a deployment, a dynamic pruning strategy such as WAND [1] could easily be used, which omits the scoring of documents that cannot reach the top $K$ retrieved set. In doing so, WAND is *safe-to-rank-K*, which we denote as *safe* for short.

WAND follows a Document-at-a-time retrieval strategy (DAAT), whereby the posting lists for all constituent terms of a query are processed in parallel, allowing immediate decisions as to whether a document has scored high enough to make the current top $K$ retrieved set. In particular, WAND repeatedly calculates a *pivot term*, by comparing the upper bounds $\sigma(t)$ of each query term $t$ to the current score of the $K$-th ranked document, known as the threshold $\tau$. The next document containing the pivot term is called the *pivot document*, which will be the next document to be fully scored. If the scored document exceeds the threshold $\tau$, then the current $K$-th ranked document is expelled from the retrieved set, the new document inserted, and $\tau$ updated. As the scoring for a query continues, the threshold $\tau$ rises, such that more documents can be omitted from scoring.

However, Broder et al. [1] showed that WAND can be made more efficient by relaxing the safeness guarantee. This is achieved by artificially increasing $\tau$ by a factor $F \geq 1$. $F = 1$ guarantees safe retrieval, while for $F > 1$, increased efficiency can be achieved without much degradation in effectiveness. However, to the best of our knowledge, no previous work in the literature has addressed how such unsafe document rankings affect retrieval performance within a modern learning to rank setting. This paper provides a first study into the effect of safeness within a learning to rank setting, while providing explanations for the observed inherent bias in unsafe pruning that can improve effectiveness in some settings. Indeed, in contrast to a safe setting, unsafe WAND is dependent on the ordering of the collection, suggesting that further research into addressing the bias is needed.

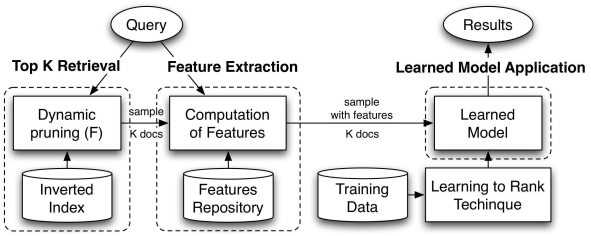


**Figure 1: Retrieval phases of a search engine.**

## 2. DATA & METHODS

Our experiments use the ClueWeb09 (cat. B) collection, which comprises 50 million English Web documents, and is aimed to represent the first tier index of a commercial search engine. We use the 150 corresponding topics and relevance assessments from the TREC Web tracks 2009-2011.

We index this collection using the Terrier information retrieval platform[1], with stemming and stopword removal. Following the three phase architecture of Figure 1, the top $K = 1000$ documents are ranked by WAND using the DPH Divergence from Randomness weighting model. We use a total of 33 standard query-dependent features (e.g. term weighting models, proximity features) and query-independent document features (e.g. link analysis, URL length, content quality). To re-rank the documents in the sample, we use the LambdaMART learning to rank technique [2, 4], which represents a state-of-the-art learning to rank technique, as per its recent performance in the 2011 Yahoo! learning to rank challenge. In particular, the 150 TREC topics are randomly split into three sets, namely training, validation and test. In the following, we experiment with the effectiveness of samples and LambdaMART models for various $F$ values, while comparing and contrasting their retrieval effectiveness, in terms of NDCG@20 and relevant documents retrieved.

[1] `http://terrier.org`

Copyright is held by the author/owner(s).
*SIGIR'12,* August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

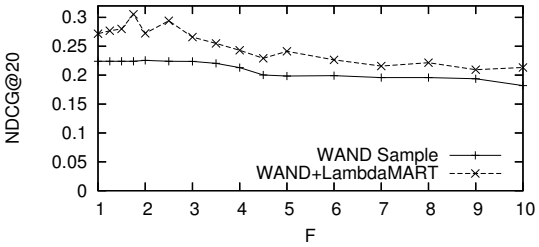


**Figure 2: NDCG@20 for the WAND sample, and LambdaMART applied on the WAND sample.**

| | $F = 1$ | $F = 1.75$ |
|---|---|---|
| LambdaMART NDCG@20 | 0.2718 | 0.3055* |
| Sample NDCG@20 | 0.2242 | 0.2242 |
| Relevant Retrieved in Sample | 1931 | 1849 |
| Mean docid of Relevant Retrieved | 26.7M | 25.6M |
| | $F = 1$ (A) → $F = 1.75$ (B) | $F = 1.75$ (A) → $F = 1$ (B) |
| Mean docid of docs present in sample A not present in sample B | 35.8M | 11.2M |

**Table 1: Analysis of safe and unsafe samples and learned models, as well as comparative statistics.**

## 3. RESULTS & ANALYSIS

Figure 2 shows the NDCG@20 effectiveness of both the WAND sample, and LambdaMART applied on the sample document rankings from WAND, as $F$ is varied. Note that as the learned model obtained by LambdaMART may be sensitive to a given $F$ setting, a different model is learned for each $F$ value. From Figure 2, we observe that the effectiveness to rank 20 of the WAND sample is unchanged for $1 \leq F \leq 3$, mirroring the original observations of Broder et al. [1]. However, the LambdaMART performance is much less stable for different $F$ values - indeed, the overall NDCG@20 trend is downward for larger $F$. This is explained in that the learned model ranks documents from deep in the sample, and hence is affected by degradations in the number of relevant documents retrieved in unsafe samples. Indeed, on analysing $F = 1$, we find that of the top 20 documents ranked by LambdaMART, some were found as deep as rank 935 in the input sample ranking, while the mean rank in the sample of LambdaMART's top 20 documents was 89.

However, for some small $F$ values in Figure 2 ($1 < F < 2$), a learned model obtained from an unsafe sample could improve over the NDCG@20 of the learned model obtained from the safe $F = 1$ sample. To analyse this unexpected characteristic, in Table 1 we compare and contrast four settings: $F = 1$ and $F = 1.75$, with and without the application of LambdaMART. Indeed, $F = 1.75$ is an interesting setting as while it is unsafe, it does not degrade NDCG@20 of the sample ranking obtained from WAND, but significantly improves the effectiveness of LambdaMART, according to a paired t-test ($p < 0.01$). Moreover, $F = 1.75$ is an efficient setting (we find that it reduces the mean response time of 1000 queries from a query log by 19% compared to $F = 1$, while larger $F$ values do not cause further time reductions).

Next, comparing the sample rankings obtained from WAND for $F = 1$ and $F = 1.75$, we note a decrease of 82 relevant documents retrieved across the 50 test queries. Moreover, we examined the docids (in the range 0..50M for ClueWeb09 cat. B) of the documents selected in the safe and unsafe samples. We found that, on average, the safe sample retrieved documents from later in the posting lists (i.e. higher docids) than the unsafe sample (mean docids: 24.3M vs 21.9M). This observation is mirrored in the documents retrieved in one sample and not the other: the mean docid of documents

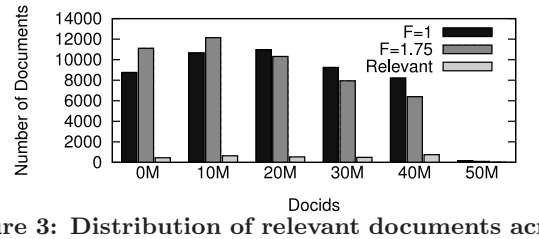


**Figure 3: Distribution of relevant documents across the docid range, and for the safe & unsafe samples.**

in the safe sample that are missing from the unsafe sample is 35M, while unsafe sample documents missing from safe have a mean docid of 11.2M. Finally, Figure 3 presents the distribution of relevant docids, as well as those retrieved in the safe and unsafe samples. This shows that while there is no docid bias for relevant documents, unsafe WAND is more biased towards low docids than safe WAND. Indeed, the mean docid of the relevant documents retrieved in the safe sample is higher than those found in the unsafe sample (26.7M vs 25.6M in Table 1), explaining the change in retrieval effectiveness. Overall, this shows that aggressive, unsafe pruning by WAND can change the selected documents in a biased manner that is not present in safe pruning.

This behaviour of WAND is explained as follows: by artificially increasing the threshold $\tau$ by the factor $F$, the threshold for unsafe WAND causes more documents to be prevented from entering the top $K$. Early in the traversal of the posting lists, when the $\tau$ is lower, documents can still enter into the retrieved set. However as $\tau$ gets higher, more pruning occurs, even for documents that would have made the retrieved set for $F = 1$. This explains unsafe WAND's comparative preference for lower docid documents.

## 4. CONCLUSIONS

We contrasted the effectiveness of safe vs. unsafe rankings from WAND, and its impact on the effectiveness of a learning to rank technique, using ClueWeb09 cat. B and 150 TREC Web track topics. We found that while unsafe retrieval effectiveness has little impact on the top ranked documents directly retrieved by WAND, it does impact deeper down, which can be to the detriment of a learned model applied on that sample. Some unsafe settings were even found to benefit the learned model. Through further analysis, we found that unsafe retrieval has an inherent bias towards documents with lower docids in the applied index ordering.

The observations in this paper can give rise to several further research lines. In particular, static collection orderings may be devised that counteract unsafe WAND's preference for lower docids. On the other hand, it may be possible to devise different manners of changing the threshold for increasing WAND's efficiency in a less biased manner.

## Acknowledgements

Craig Macdonald and Iadh Ounis acknowledge the support of EC-funded project SMART (FP7-287583).

## 5. REFERENCES

[1] A. Z. Broder, D. Carmel, M. Herscovici, A. Soffer, and J. Zien. Efficient query evaluation using a two-level retrieval process. In *Proc. of CIKM 2006*, 426–434.

[2] Y. Ganjisaffar, R. Caruana, and C. Lopes. Bagging gradient-boosted trees for high precision, low variance ranking models. In *Proc. of SIGIR 2011*, 85–94.

[3] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in IR*, 3(3):225–331, 2009.

[4] Q. Wu, C. J. C. Burges, K. M. Svore, and J. Gao. Ranking, boosting, and model adaptation. Technical Report MSR-TR-2008-109, Microsoft, 2008.