Managing Tail Latencies in Large Scale IR Systems

Joel Mackenzie
RMIT University
Melbourne, Australia
joel.mackenzie@rmit.edu.au

KEYWORDS

Tail Latency; Efficiency; Scalability

ACM Reference format:
Joel Mackenzie. 2017. Managing Tail Latencies in Large Scale IR Systems. In *Proceedings of SIGIR '17, Shinjuku, Tokyo, Japan, August 07-11, 2017,* 1 pages.
DOI: http://dx.doi.org/10.1145/3077136.3084152

ABSTRACT

With the growing popularity of the world-wide-web and the increasing accessibility of smart devices, data is being generated at a faster rate than ever before. This presents scalability challenges to web-scale search systems – how can we efficiently index, store and retrieve such a vast amount of data? A large amount of prior research has attempted to address many facets of this question, with the invention of a range of efficient index storage and retrieval frameworks that are able to efficiently answer most queries. However, the current literature generally focuses on improving the mean or median query processing time in a given system. In the proposed PhD project, we focus on improving the efficiency of high percentile *tail latencies* in large scale IR systems while minimising end-to-end effectiveness loss.

Although there is a wealth of prior research involving improving the efficiency of large scale IR systems, the most relevant prior work involves predicting long-running queries and processing them in various ways to avoid large query processing times. Prediction is often done through pre-trained models based on both static and dynamic features from queries and documents. Many different approaches to reducing the processing time of long running queries have been proposed, including parallelising queries that are predicted to run slowly [5, 6], scheduling queries based on their predicted run time [7], and selecting or modifying the query processing algorithm depending on the load of the system [1, 11].

Considering the specific focus on tail latencies in large-scale IR systems, the proposed research aims to: (i) study what causes large tail latencies to occur in large-scale web search systems, (ii) propose a framework to mitigate tail latencies in multi-stage retrieval systems through the prediction of a vast range of query-specific efficiency parameters, (iii) experiment with mixed-mode query semantics to provide efficient and effective querying to reduce tail latencies, and (iv) propose a time-bounded solution for Document-at-a-Time (DaaT) query processing which is suitable for current web search systems.

SIGIR '17, Shinjuku, Tokyo, Japan
 978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3084152

As a preliminary study, Crane et al. [3] compared some state-of-the-art query processing strategies across many modern collections. They found that although modern DaaT dynamic pruning strategies are very efficient for ranked disjunctive processing, they have a much larger variance in processing times than *Score-at-a-Time* (SaaT) strategies which have a similar efficiency profile regardless of query length or the size of the required result set. Furthermore, Mackenzie et al. [8] explored the efficiency trade-offs for paragraph retrieval in a multi-stage question answering system. They found that DaaT dynamic pruning strategies could efficiently retrieve the top-1,000 candidate paragraphs for very long queries.

Extending on prior work [3, 4, 7], Mackenzie et al. [9] showed how a range of per-query efficiency settings can be accurately predicted such that 99.99 percent of queries are serviced in less than 200 ms without noticeable effectiveness loss. In addition, a reference list framework [2, 4, 10] was used for training models such that no relevance judgements or annotations were required. Future work will focus on improving the candidate generation stage in large-scale multi-stage retrieval systems. This will include further exploration of index layouts, traversal strategies [3], and query rewriting, with the aim of improving early stage efficiency to reduce the system tail latency, while potentially improving end-to-end effectiveness.

REFERENCES

[1] D. Broccolo, C. Macdonald, S. Orlando, I. Ounis, R. Perego, F. Silvestri, and N. Tonellotto. 2013. Load-sensitive selective pruning for distributed search. In *Proc. CIKM*. 379–388.
[2] C. L. A. Clarke, J. S. Culpepper, and A. Moffat. 2016. Assessing efficiency—effectiveness tradeoffs in multi-stage retrieval systems without using relevance judgments. *Information Retrieval* 19, 4 (2016), 351–377.
[3] M. Crane, J. S. Culpepper, J. Lin, J. Mackenzie, and A. Trotman. 2017. A comparison of Document-at-a-Time and Score-at-a-Time query evaluation. In *Proc. WSDM*. 201–210.
[4] J. S. Culpepper, C. L. A. Clarke, and J. Lin. 2016. Dynamic cutoff prediction in multi-stage retrieval systems. In *Proc. ADCS*. 17–24.
[5] S-W. Hwang, S. Kim, Y. He, S. Elnikety, and S. Choi. 2016. Prediction and predictability for search query acceleration. *ACM Trans. Web* 10, 3 (Aug. 2016), 19:1–19:28.
[6] M. Jeon, S. Kim, S-W. Hwang, Y. He, S. Elnikety, A. L. Cox, and S. Rixner. 2014. Predictive parallelization: taming tail latencies in web search. In *Proc. SIGIR*. 253–262.
[7] C. Macdonald, N. Tonellotto, and I. Ounis. 2012. Learning to predict response times for online query scheduling. In *Proc. SIGIR*. 621–630.
[8] J. Mackenzie, R-C. Chen, and J. S. Culpepper. 2016. RMIT at the TREC 2016 LiveQA Track. In *Proc. TREC-25*.
[9] J. Mackenzie, J. S. Culpepper, R. Blanco, M. Crane, C. L. A. Clarke, and J. Lin. 2017. Efficient and Effective Tail Latency Minimization in Multi-Stage Retrieval Systems. (2017). arXiv:1704.03970 [cs.IR]
[10] L. Tan and C. L. A. Clarke. 2015. A Family of Rank Similarity Measures Based on Maximized Effectiveness Difference. *TKDE* 27, 11 (2015), 2865–2877.
[11] N. Tonellotto, C. Macdonald, and I. Ounis. 2013. Efficient and effective retrieval using selective pruning. In *Proc. WSDM*. 63–72.