Speeding-up Document Scoring with Tree Ensembles using CPU SIMD Extensions

Claudio Lucchese[1,3], Franco Maria Nardini[1,3], Salvatore Orlando[2], Raffaele Perego[1,3], Nicola Tonellotto[1,3], and Rossano Venturini[4,3]

[1] ISTI-CNR, Pisa, [2] University Ca' Foscari of Venice,
[3] Istella Srl, [4] University of Pisa.

Abstract. Scoring documents with *learning-to-rank* (LtR) models based on large ensembles of regression trees is currently deemed one of the best solutions to effectively rank query results to be returned by large scale Information Retrieval systems. This extended abstract shortly summarizes the work in [4] proposing V-QuickScorer (vQS), an algorithm which exploits SIMD vector extensions on modern CPUs to perform the traversal of the ensamble in parallel by evaluating multiple documents simultaneously. We summarize the results of a comprehensive evaluation of vQS against state-of-the-art scoring algorithms showing that vQS outperforms competitors with speed-ups up to a factor of 2.4x.

Additive ensembles of regression trees, such as GBRT [2] and λ-MART [5], are nowadays considered among the most advanced LtR models for ranking documents in IR systems, although these require very efficient scoring algorithms for processing queries by strict time budgets [1]. The state-of-the-art algorithm for efficient scoring via additive ensemble of regression trees is QuickScorer (QS) [3]. In this extended abstract we shortly summarize the work in [4] where we introduce vQS, a parallelized version of QS that exploits the SIMD capabilities of mainstream CPUs. Streaming SIMD Extensions (SSE) and Advanced Vector Extensions (AVX) are sets of instructions exploiting wide registers of 128 and 256 bits that allow parallel operations to be performed on simple data types, e.g., a 128 bit containing four single precision or two double precision floats. Using SIMD capabilities of mainstream CPUs, namely SSE 4.2 and AVX 2, vQS can process up to 8 documents in parallel. Results of a comprehensive evaluation of vQS on public datasets against state-of-the-art scoring algorithms show that vQS outperforms competitors with speed-ups up to a factor of 2.4x.

References

1. G. Capannini, C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, and N. Tonellotto. Quality versus efficiency in document scoring with learning-to-rank models. *Information Processing & Management*, 2016.
2. J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.
3. C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. Quickscorer: A fast algorithm to rank documents with additive ensembles of regression trees. In *Proc. ACM SIGIR*, pages 73–82. ACM, 2015.
4. C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. Exploiting cpu simd extensions to speed-up document scoring with tree ensembles. In *Proc. ACM SIGIR 2016*. ACM, 2016.
5. Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 2010.