

# Learning to Efficiently Rank on Big Data

Lidan Wang<sup>†</sup>, Jimmy Lin<sup>‡</sup>, Donald Metzler<sup>#</sup>, Jiawei Han<sup>†</sup>

<sup>†</sup>Univ. of Illinois at Urbana-Champaign, lidan, hanj@illinois.edu

<sup>‡</sup>Univ. of Maryland at College Park, jimmylin@umd.edu

<sup>#</sup>Google, Inc. metzler@gmail.com

Ranking in response to user queries is a central problem in information retrieval, data mining, and machine learning. In the era of “Big data”, traditional effectiveness-centric ranking techniques tend to get more and more costly (requiring additional hardware and energy costs) to sustain reasonable ranking speed on large data. The mentality of combating big data by throwing in more hardware/machines will quickly become highly expensive since data is growing at an extremely fast rate oblivious to any cost concerns from us. “Learning to efficiently rank” [5, 7, 6] offers a cost-effective solution to ranking on large data (e.g., billions of documents). That is, it addresses a critically important question – whether it is possible to improve ranking effectiveness on large data *without* incurring (too much) additional cost?

At the basic level, the “Learning to efficiently rank” framework explores speed-effectiveness tradeoffs in the context of building highly effective, *and* very fast ranking models for large-scale document collections. A key observation is that there is an inherent *tradeoff* between the speed and effectiveness of ranking models – models that are highly effective typically are slow (due to potentially using a larger number of costly features and complex model structures), and models that are fast typically are not as effective. We would like to design and learn new ranking models that can break through the tradeoff barrier and be *simultaneously* fast and effective when applied to large data (e.g., billions of documents). In view of the inherent tradeoffs, this is quite challenging, as most of the previous ranking models focus on a fixed operating point along the space of effectiveness vs speed.

Our tutorial contains four key sections. First, we will present the background information on ranking in response to user queries [2], and address factors that contribute to online model speed and effectiveness, and how such tradeoffs influence ranking system operational costs. Second, we introduce the audience to three important challenges in breaking through the tradeoff barrier: new models, metrics, and learning. We present systematic solutions to address the challenges. This leads to drastic new models and approaches for cost-effective ranking at scale, which will be discussed in depth in the tutorial. Furthermore, we will present new directions in this topic that have flourished since the inception of the basic framework. Examples include query-dependent learning to efficiently rank [3] and classification under time budgets [8]. Finally, viewing the framework from a learning perspective, it jointly optimizes multiple metrics (e.g., ranking speed and effectiveness). The general area of multi-objective learning to rank has drawn very strong interests in recent years, due to the multi-objective nature of many real-world ranking systems. We will broadly survey this general direction in the context of ranking, including learning to rank for freshness and relevance [1] and learning risk-sensitive ranking models [4].

## 1. References

- [1] N. Dai, M. Shokouhi, and B. D. Davison. Learning to rank for freshness and relevance. In *SIGIR*, 2011.
- [2] T. Liu. Learning to Rank for Information Retrieval. In *Foundations and Trends in Information Retrieval*, 2009.
- [3] N. Tonello, C. Macdonald, and I. Ounis. Efficient and effective retrieval using selective pruning. In *WSDM*, 2013.
- [4] L. Wang, P. Bennett, and K. Collins-Thompson. Robust ranking models via risk-sensitive optimization. In *SIGIR*, 2012.
- [5] L. Wang, J. Lin, and D. Metzler. Learning to efficiently rank. In *SIGIR*, 2010.
- [6] L. Wang, J. Lin, and D. Metzler. A cascade ranking model for efficient ranked retrieval. In *SIGIR*, 2011.
- [7] L. Wang, D. Metzler, and J. Lin. Ranking under temporal constraints. In *CIKM*, 2010.
- [8] Z. E. Xu, K. Q. Weinberger, and O. Chapelle. The greedy miser: Learning under test-time budgets. In *ICML*, 2012.