Scalability and Performance of Random Forest based Learning-to-Rank for Information Retrieval

Muhammad Ibrahim
Faculty of Information Technology, Monash University, Australia
313ibrahim@gmail.com

Abstract

For a query submitted by a user, the goal of an information retrieval system is to return a list of documents which are highly relevant with respect to that query. Traditionally different scoring methods, ranging from simple heuristic models to probabilistic models, have been used for this task. Recently researchers have started to use supervised machine learning techniques for solving this problem which is then called the learning-to-rank (LtR) problem. Many supervised learning methods have been tested so far with empirical success over conventional methods [4].

The random forest is a relatively simple but effective and efficient learning algorithm which aggregates the predictions of a large number of independent and variant base learners, namely decision trees. Its major benefits over other state-of-the-art methods include inherent parallelizability, ease of tuning and competitive performance. These benefits attract researchers across various disciplines where a random forest is a very popular choice. However, for LtR task, the random forest has not been thoroughly investigated.

In this research, we investigate the random forest based LtR algorithms. We aim at improving the efficiency, effectiveness, and understanding of these algorithms. With respect to the first goal, we employ undersampling techniques and leverage the inherent structure of a random forest to achieve better scalability, especially for highly imbalanced datasets [2]. We also reduce the correlation among the trees to reduce learning time and to improve performance [3]. With respect to the second goal, we investigate various objective functions ranging from completely randomized splitting criterion to so-called listwise splitting [1]. We also conduct a thorough study on random forest based pointwise algorithms. With respect to the third goal, we develop methods for estimating the bias and variance of rank-learning algorithms, and examine their empirical behavior against parameters of the learning algorithm.

The thesis is available at: `https://figshare.com/articles/Scalability_and_Performance_of_Random_Forest_based_Learning-to-Rank_for_Information_Retrieval/4407443`

References

[1] Muhammad Ibrahim and Mark Carman. *Comparing pointwise and listwise objective functions for random forest based learning-to-rank.* ACM Transactions on Information Systems (TOIS), 34(4): Article No.: 20, 2016.

[2] Muhammad Ibrahim and Mark Carman. *Undersampling techniques to re-balance training data for large scale learning-to-rank.* Information Retrieval Technology, pages 444-457. Springer, 2014.

[3] Muhammad Ibrahim and Mark Carman. *Improving scalability and performance of random forest based learning-to-rank algorithms by aggressive subsampling.* Proceedings of the 12th Australasian Data Mining Conference, pages 91-99. 2014.

[4] Muhammad Ibrahim and Manzur Murshed. *From tf-idf to learning-to-rank: An overview.* Handbook of Research on Innovations in Information Retrieval, Analysis, and Management, pages 62-109. IGI Global, USA, 2016.