The Impact of Negative Samples on Learning to Rank

Claudio Lucchese
Ca' Foscari University of Venice, Italy
claudio.lucchese@unive.it

Franco Maria Nardini
ISTI-CNR, Pisa, Italy
f.nardini@isti.cnr.it

Raffaele Perego
ISTI-CNR, Pisa, Italy
r.perego@isti.cnr.it

Salvatore Trani
ISTI-CNR, Pisa, Italy
s.trani@isti.cnr.it

ABSTRACT

Learning-to-Rank (LtR) techniques leverage machine learning algorithms and large amounts of training data to induce high-quality ranking functions. Given a set of documents and a user query, these functions are able to predict a score for each of the documents that is in turn exploited to induce a relevance ranking. The effectiveness of these learned functions has been proved to be significantly affected by the data used to learn them. Several analysis and document selection strategies have been proposed in the past to deal with this aspect. In this paper we review the state-of-the-art proposals and we report the results of a preliminary investigation of a new sampling strategy aimed at reducing the number of not relevant query-document pairs, so to significantly decrease the training time of the learning algorithm and to increase the final effectiveness of the model by reducing noise and redundancy in the training set.

ACM Reference format:
Claudio Lucchese, Franco Maria Nardini, Raffaele Perego, and Salvatore Trani. 2017. The Impact of Negative Samples on Learning to Rank. In *Proceedings of the first international Workshop on LEARning Next gEneration Rankers, Amsterdam, The Netherlands, October 1, 2017 (LEARNER'17),* 2 pages.

1 INTRODUCTION

Ranking is one of the most important problems in information retrieval. Modern Web search engines exploit and combine hundreds of features modelling the relevance between the user query and the candidate documents. A specific area of machine learning, i.e., learning to rank, has been developed to deal with the ranking problem. Given a training set of feature vectors and relevance pairs, a learning to rank algorithm learns how to combine the query and the document features so to optimize a specific effectiveness ranking metric. In the last years important efforts have been spent on feature engineering/extraction and the development of sophisticated learning to rank algorithms while little research has been conducted on how to choose queries and documents for learning to rank datasets nor on the effect of these choices on the ability of learning to rank algorithms to learn effectively and efficiently. Yilmaz and Robertson attack the problem by observing that the number of judgments in the training set directly affects the quality of the learned system [5]. Given the expense of obtaining relevance judgments for constructing training data, the major problem is how to well distribute this judgment effort. Authors thus investigate the trade-off between the number of queries and the number of judgments per query when building training sets. In particular, they show that training sets with more queries but less judgments per query are more cost effective than training sets with less queries but more judgments per query.

Asham *et al.* propose several document selection methodologies, i.e., depth-k pooling, sampling (infAP, statAP), active-learning (MTC), and on-line heuristics (hedge) [1]. Authors prove that some of the proposed methods, i.e., infAP, statAP and depth pooling, are better than others (hedge and the LETOR method) for building efficient and effective learning to rank collections. Authors also propose a comparison with the document selection methodology used to create the LETOR datasets. The proposed study also deals with both i) the proportion of relevant documents to non-relevant documents and ii) the similarity between relevant and non-relevant documents in the datasets. Results confirm that both characteristics highly affect the quality of the learning-to-rank collections, with the latter having more impact. As a side result, authors also observed that some learning to rank algorithms, RankNet and LambdaRank, are more robust to document selection methodologies than other, i.e., Regression, RankBoost and Ranking SVM.

In a later contribution, Kanoulas *et al.* propose a large-scale study on the effect of the distribution of labels across the different grades of relevance in the training set on the performance of trained ranking functions [3]. Authors propose a methodology that allows to generate a large number of training datasets with different label distributions. The datasets are then employed by three learning to rank algorithms to fit a ranking model. Authors investigate the effect of these distributions on the accuracy of the obtained ranking functions to characterize how training sets should be constructed. Authors conclude that the relevance grade distribution in the training set is an important factor for the effectiveness of learning to rank algorithms. They provide qualitative advise about the construction of learning to rank datasets: i) distributions with a balance between the number of documents in the extreme grades should be favoured as the middle relevance grades play less important role than the extreme ones.

In this paper, we investigate a new technique to sample documents in order to improve both efficiency and effectiveness of learning to rank models. Indeed, the improved efficiency is a consequence of a reduced size of the sampled dataset, with the ranking algorithm that have to learn from a lower number of query-document pairs. On the other hand, an effective sampling technique may lead to improve the effectiveness of the resulting model by filtering out noise and reducing the redundancy of the query-document pairs.

LEARNER'17, October 1, 2017, Amsterdam, The Netherlands

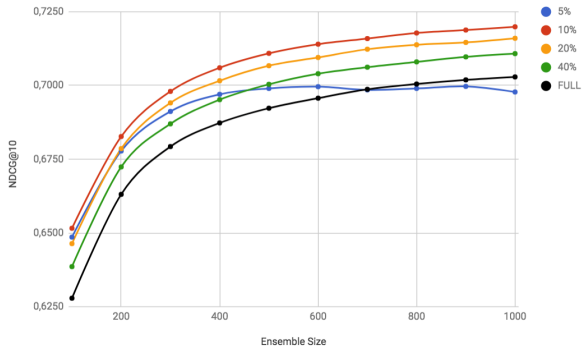


Figure 1: Effectiveness of λ-MART models trained on full dataset vs sampled ones.

We present preliminary experimental results proving the benefits of the new proposed sampling methodology.

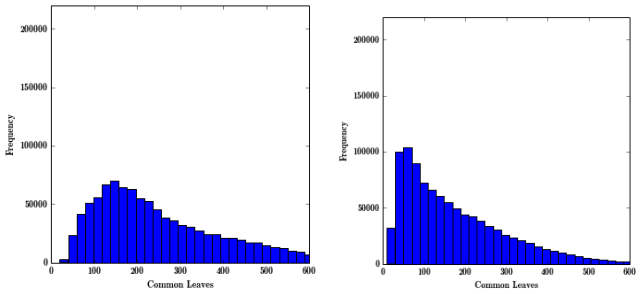
2 METHODOLOGY AND DISCUSSION

We propose a sampling methodology that acts in two steps: i) we train a ranking model on the full dataset and we use it to rank the query-document pairs of the same dataset; ii) we select a given fraction of the top-ranked negative samples, i.e., label = 0, together with all the positive ones, i.e., label> 0. The rationale is that the top-ranked negative samples by the learned model are the most likely to be *mis-ranked*. Consequently, we reduce the dataset to such positive and negative examples to let the learning algorithm focuses on the most critical and discriminating examples. Moreover, by sampling only negative documents we indirectly reduce the unbalance between positive and negative classes and reduce the redundancy in the negative class.

We evaluate the effectiveness of our proposed technique on the Istella[1] [2] publicly available dataset. It is composed of 33,018 queries and each of the 10,454,629 query-document pairs is represented by means of 220 features. Training and validation data were used to train a reference algorithm. i.e.. λ-MART. a *list-wise* algorithm that is capable of using *NDCG* in its loss function, resulting in a predictor of the ranking [4]. The λ-MART algorithm was fine-tuned by sweeping its parameters to maximize *NDCG*@10. The best performance were obtained with a learning rate equal to 0.05 and using 64 leaves.

We built several sampled datasets by varying the fraction of negatives sampled to $\{5\%, 10\%, 20\%, 40\%\}$. We trained a λ-MART model on each of these sampled datasets and we evaluated their performance on the full test set. As shown in Figure 1, the best performing model is the one trained on the dataset where only the top 10% negatives were selected. This model significantly outperforms the model trained on the full training set by a large margin. Moreover, when the fraction of documents selected is lowered to 5%, we observe a drop in performance, while in the first iterations (up to 200 trees) results are in line with other models. This behavior suggests that the use of 5% of the negative instances does not fully represent the negative class of query-document pairs. As a consequence the resulting model is not able to correctly identify them and this harm the ranking accuracy.

[1]http://blog.istella.it/istella-learning-to-rank-dataset/



(a) Model trained on Full Dataset (b) Model trained on 10% Sample

Figure 2: Similarity distribution of negative documents.

The proposed sampling strategy has been showed to provide a significant boost to the effectiveness of the ranking model. To understand the reasons that lead to this gain, we investigate the similarity distribution between pairs of documents in the negative class. Figure 2 reports such instance analysis, where the similarity between a pair of documents is computed by counting the number of exit leaves in common in scoring the two documents using the given ensemble-based ranking model. The x-axis corresponds to the number of leaves in common (1,000 is the max for two documents exiting in the same leaves for all the trees), while the y-axis corresponds to the number of pairs with the given similarity. The negative examples in the full dataset were analyzed with the models trained on the full dataset and on the 10% sample. According to the model trained on the sampled dataset, the document pairs distribution is significantly skewed towards dissimilar pairs, i.e., with a few common exit leaves. This not only signifies that the initial dataset was redundant in the negative class, but it also implies that the model trained on the sampled dataset is much more effective in discriminating among the negative examples and thus resulting in an improved ranking accuracy.

The preliminary experimental evaluation confirms the validity of the proposed sampling technique. As future work we intend to investigate in depth for a robust and systematic sampling technique that is able to provide benefits independently from the class-distribution of the dataset and where the best sampling ratio is automatically chosen accordingly to the dataset properties. We believe that sampling strategies can both improve the quality of the training data generate and the effectiveness of the learning algorithms.

REFERENCES

[1] Javed A. Aslam, Evangelos Kanoulas, Virgil Pavlu, Stefan Savev, and Emine Yilmaz. 2009. Document Selection Methodologies for Efficient and Effective Learning-to-rank. In *Proc. ACM SIGIR'09*. 468–475.
[2] D. Dato, C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. 2016. Fast Ranking with Additive Ensembles of Oblivious and Non-Oblivious Regression Trees. *ACM Trans. Inf. Syst.* 35, 2, Article 15 (2016).
[3] Evangelos Kanoulas, Stefan Savev, Pavel Metrikov, Virgil Pavlu, and Javed Aslam. 2011. A Large-scale Study of the Effect of Training Set Characteristics over Learning-to-rank Algorithms. In *Proc. ACM SIGIR'11*. 1243–1244.
[4] Q. Wu, C.J.C. Burges, K.M. Svore, and J. Gao. 2010. Adapting boosting for information retrieval measures. *Information Retrieval* (2010).
[5] Emine Yilmaz and Stephen Robertson. 2009. Deep Versus Shallow Judgments in Learning to Rank. In *Proc. ACM SIGIR'09*. 662–663.