LEARning Next gEneration Rankers (LEARNER 2017)

Nicola Ferro
University of Padua, Padua, Italy
ferro@dei.unipd.it

Claudio Lucchese
Ca' Foscari University of Venice and ISTI-CNR, Pisa, Italy
c.lucchese@isti.cnr.it

Maria Maistro
University of Padua, Padua, Italy
maistro@dei.unipd.it

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

ABSTRACT

The aim of LEARNER@ICTIR2017 is to investigate new solutions for LtR. In details, we identify some research areas related to LtR which are of actual interest and which have not been fully explored yet. We solicit the submission of position papers on novel LtR algorithms, on evaluation of LtR algorithms, on dataset creation and curation, and on domain specific applications of LtR.

LEARNER@ICTIR2017 will be a gathering of academic people interested in IR, ML and related application areas. We believe that the proposed workshop is relevant to ICTIR since we look for novel contributions to LtR focused on foundational and conceptual aspects, which need to be properly framed and modeled.

KEYWORDS

learning to rank; user behaviour; datasets; evaluation

1 WORKSHOP DESCRIPTION AND MOTIVATIONS

Ranking is forever at the core of *Information Retrieval (IR)* since it allows to sift out non relevant information and to select a list of items ordered by their estimated relevance to a given query. Documents, information needs, search tasks and interaction mechanisms between users and information systems are getting more and more complex and diversified, and this calls for more and more sophisticated techniques able to cope with this emerging complexity and the high expectations of users.

Learning to Rank (LtR), and *Machine Learning (ML)* in general, have proven to be very effective methodologies to address these issues, significantly improving over state-of-the-art traditional algorithms. Popular areas of investigation in LtR are related to efficiency, feature selection, supervised learning, but many new angles are still overlooked. The goal of this workshop is to investigate how to improve ranking, in particular LtR, by bringing in new perspectives which have not explored or fully addressed yet by our community after the 2011 Yahoo Learning to Rank Challenge [3].

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

For example, user behaviour in interacting with search results is a prominent research area in IR, but traditional offline LtR approaches limit themselves to consider, e.g., the number of clicks on a document or the dwell time, as features without embedding or exploiting user dynamics in the LtR algorithm itself. The user dynamic changes with queries of different types, with the length of the user session, with the difficulty and complexity of the search task, and this calls for novel evaluation measures and LtR algorithms [8]. On the other hand, online ranking algorithms iteratively exploit user interaction in A/B testing-like settings to choose among rankers and interleave their outputs, but they do not modify the ranking algorithms according to such user interaction. We would like to explore how to bridge the gap between these two approaches to LtR and create new solutions that leverage the best of the two worlds.

As another example, efficiency is a core theme in LtR and it is traditionally approached by studying how to optimize such algorithms from a time and/or space perspective [2, 12]. Nevertheless, the cost of applying the models generated by those algorithms in a real production environment has received limited attention. Ranking is typically organized in pipelined stages, where increasingly accurate and expensive models are used to rank documents coming from the previous stage and prune the ones that are not likely to appear in the top-K results. How to optimally design and train such a complex pipeline is still an open problem.

On the other side, learning with effectiveness centric techniques is getting more and more costly, especially when large amounts of data need to be processed. Indeed, to sustain a reasonable ranking speed, additional hardware and energy costs are required [17]. Therefore, a tradeoff between the effectiveness and quality of the ranking and the ranking speed represents a challenging and compelling area of research. This is of particular importance when academic research and knowledge need to be transferred to business applications [16].

A possible approach to reduce annotation costs consists in limiting the size of the training dataset. LtR algorithms tend to perform the training phase on the whole datasets to leverage the benefit of using a large amount of annotated data. However, when the whole dataset is used, even low quality data are included and this affects the global effectiveness [13]. New research directions deal with the proposal of novel techniques to select training data by preserving and/or increasing the effectiveness and efficiency of rankers. For instance, a compression-based selection process can be used to create small and yet highly informative and effective initial sets that can later be labeled and used to bootstrap a LtR system [14].

Another key area of investigation to bridge between offline and online approaches is personalized LtR algorithms able to account for user profile, interests, current user context, preferences in diversity of results, relevance feedback and more. When rankers are trained on batch data they can not adapt to user preferences and satisfy changing user needs [4]. On the other side, utilizing interaction data to improve search algorithms is challenging. Recently [9] analyzed the problem of bias in click data that affects LtR algorithms when used as training signal. For example, using clicks and no clicks as relevant or non relevant judgements leads to biased rankers, since the presentation position has a strong influence on where users click. Moreover, when it comes to personal content search, there are no multiple observations across different users for the same query-document pair [1]. Indeed, user queries are personal, therefore they might not generalize well across different users, and documents (e.g. emails or private files) are not shared between different users.

We also solicit the submission of papers which deal with hybrid approaches between LtR and deep learning. Recently, deep learning applied to IR has received an increasing attention from the research community, as witnessed by the tutorial on deep learning for IR [11] and the neu-IR workshop [5] proposed at SIGIR 2016. Neural networks were proposed to improve multiple aspects of IR, for instance to learn user interests and preferences [15], to model the novelty of a document for diversification [18], and in combination with traditional models as BM25 to define new ranking functions [6].

Finally, reproducibility of the experimental results is an increasing concern in IR and in computer science, in general [7]. In particular, we think that the lack of large real-world information-rich datasets and the difficulty of generating them is hindering further developments of LtR in the academic community. We welcome the contribution of novel datasets to the community and the proposal of methodologies for evaluating the quality of the datasets currently available. Moreover, we encourage the submission of papers related to new approaches to evaluate bias and variance of LtR as well as to explain and interpret the outputs of LtR algorithms, also using visual analytics techniques [10].

2 SCOPE

A (not exhaustive) list of relevant topics for the LEARNER@ICTIR2017 workshop is reported below:

- Next Generation LtR Algorithms:
  - Unsupervised approaches to LtR, active learning for LtR, transfer learning for LtR;
  - Incremental LtR, online, or personalized LtR;
  - Embedding user behaviour and dynamic in LtR;
  - Cost-Aware LtR;
  - List-based approaches for result list diversification and/or clustering;
  - Bias/Variance and other theoretical characterizations or ranking models;
  - Feature engineering for ranking;
  - Deep neural networks for ranking;
  - Understanding and explaining complex LtR models, also via visual analytics solutions.
- Evaluation of LtR Algorithms:
  - Quality measures accounting for user behaviour and perceived quality;
  - Quality measures accounting for models failures, redundancy, robustness, sensitivity, etc.;
  - Evaluation of ranking efficiency vs. quality trade-off;
  - Visual analytics solutions for exploring and interpreting experimental data;
  - Reproducibility of LtR experiments.
- Datasets:
  - Measuring quality of training datasets: noise, contradictory examples, redundancy, difficulty of building a good model, features quality, coverage of application domain use cases;
  - Creation and curation of datasets: compression, negative sampling, aging, dimensionality reduction;
  - Contributing novel datasets to the community.
- Applications:
  - Application of LtR to verticals or to other domains (e.g., recommendation, news, product search, social media, job search, ...);
  - LtR beyond documents: keyword-based access to structured data, multimedia, graphs, etc.

Acknowledgments. This work was partially supported by the EC H2020 Program INFRAIA-1-2014-2015 SoBigData: Social Mining & Big Data Ecosystem (654024), and SID16 Ferro, PRAT 2016: Improving Information Retrieval Effectiveness via Markovian User Models.

REFERENCES

[1] M. Bendersky, X. Wang, D. Metzler, and M. Najork. Learning from User Interactions in Personal Search via Attribute Parameterization. In WSDM, pages 791–799, ACM, 2017.
[2] B. B. Cambazoglu, and R. Baeza-Yates. Scalability and Efficiency Challenges in Large-Scale Web Search Engines. In SIGIR, pages 1223–1226, ACM, 2016.
[3] O. Chapelle, Y. Chang, and T. Liu. Future Directions in Learning to Rank. Yahoo! Learning to Rank Challenge, pages 91–100, PMLR, 2011.
[4] S. Chaudhuri, and A. T. Tewari. Online Learning to Rank with Feedback at the Top. In AISTAT, pages 277–285, PMLR, 2016.
[5] N. Craswell, W. B. Croft, J. Guo, B. Mitra, and M. de Rijke. Neu-IR: The SIGIR 2016 Workshop on Neural Information Retrieval. In SIGIR, pages 1245–1246, ACM, 2016.
[6] M. Dehghani, H. Zamani, A. Severyn, J. Kamps, and W. B. Croft. Neural Ranking Models with Weak Supervision. In SIGIR, ACM, 2017.
[7] N. Ferro. Reproducibility Challenges in Information Retrieval Evaluation. In Data and Information Quality, 2(8):1–4, ACM, 2017.
[8] N. Ferro, C. Luchesse, M. Maistro, and R. Perego. On Including the User Dynamic in Learning to Rank. In SIGIR, ACM, 2017.
[9] T. Joachims, A. Swaminathan, and T. Schnabel. Unbiased Learning-to-Rank with Biased Feedback. In WSDM, pages 781–789, ACM, 2017.
[10] J. Krause, A. Perer, and K. Ng. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In CHI, pages 5686–5697, ACM, 2016.
[11] H. Li, and Z. Lu. Deep Learning for Information Retrieval. In SIGIR, pages 1203–1206, ACM, 2016.
[12] C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, F. Silvestri, and S. Trani. Post-Learning Optimization of Tree Ensembles for Efficient Ranking. In SIGIR, pages 949–952, ACM, 2016.
[13] Q. Ma, B. He, Ben and J. Xu. Direct Measurement of Training Query Quality for Learning to Rank. In SAC, pages 1035–1040, ACM, 2016.
[14] R. M. Silva, G. C. M. Gomes, M. S. Alvim, and M. A. Gonçalves. Compression-Based Selective Sampling for Learning to Rank. In CIKM, pages 247–256, ACM, 2016.
[15] Y. Song, A. M. Elkahky, and X. He. Multi-Rate Deep Learning for Temporal Recommendation. In SIGIR, pages 909–912, ACM, 2016.
[16] M. Tsagkias, and W. Weerkamp. Building a Self-Learning Search Engine: From Research to Business. In SIGIR, pages 523–524, ACM, 2016.
[17] L. Wang, J. Lin, D. Metzler, and J. Han. Learning to Efficiently Rank on Big Data. In WWW, pages 209–210, ACM, 2014.
[18] L. Xia, J. Xu, Y. Lan, J. Guo, and X. Cheng. Modeling Document Novelty with Neural Tensor Network for Search Result Diversification. In SIGIR, pages 395–404, ACM, 2016.