Report on LEARNER 2017: 1st International Workshop on LEARning Next gEneration Rankers

Nicola Ferro
University of 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, Italy
maistro@dei.unipd.it

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

Abstract

The LEARNER workshop was co-located with the "third ACM International Conference on the Theory of Information Retrieval", (ICTIR 2017). The goal of the workshop was to foster investigation on novel Learning-to-Rank algorithms, on their evaluation, on dataset creation and curation, and on domain specific applications of Learning-to-Rank. The half-day workshop hosted eight paper presentations and two invited talks: the first by Craig Macdonald on "Hypothesis Testing for Risk-Sensitive Evaluation and Learning to Rank in Web Search" and the second by Djoerd Hiemstra entitled "Ranking Learning-to-Rank Methods". Detailed information is available at http://learner2017.dei.unipd.it/.

1 Motivations and Goals

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 LEARNER [5] was to investigate how to improve ranking, in particular LtR, by bringing in new perspectives which have not been explored or fully addressed yet by our community after the 2011 Yahoo Learning to Rank Challenge [2]. The workshop solicited submission of works covering the following topics:

- Next Generation LtR Algorithms: unsupervised approaches to LtR, incremental LtR, feature engineering for ranking, deep neural networks for ranking, etc.;
- Evaluation of LtR Algorithms: accounting for user behaviour and perceived quality, reproducibility of LtR experiments, etc.
- Datasets: creation and curation of datasets, contributing novel datasets to the community, etc.
- Applications: LtR beyond documents, keyword-based access to structured data, multimedia, graphs, etc.

The LEARNER workshop covered half-day and it was structured as follows. There were four *full presentations* of reviewed and accepted papers, additional four *short presentations*, invited by the workshop chairs, reporting on preliminary results and two *keynote talks*. The topics discussed by the above presentations covered on-line learning, query-based specialization, axiomatic approaches, entity-enhanced rankers, dataset creation and tools for algorithm evaluation. The proceeding of the workshop are published in CEUR-WS [4]. Below we shortly report on the works presented at the workshop.

2 Keynotes

"Hypothesis Testing for Risk-Sensitive Evaluation and Learning to Rank in Web Search" by Craig Macdonald

When a user is unsatisfied with the quality of results of a web search engine, they may switch to another, leading to a loss of ad revenue to the engine. Use of a robust retrieval approach is therefore essential, to that the experience of the users of the search engine is not damaged by poorly-performing queries. For this reason, there has been growing interest in measuring robustness using a new class of risk-sensitive evaluation measures, which assess the extent to which a system exhibits risk, i.e. performs worse than a given baseline system on a set of queries. In this talk, the author presented his recent advances in two families of risk-sensitive evaluation measures both based upon hypothesis testing, and their integration into a state-of-the-art LtR algorithm, to create effective yet robust retrieval models. Experiments using 10,000 topics from the MSLR LtR dataset from the Bing search engine demonstrate that the proposed t-test and Chi-square based objective functions reduce the number of poorly performing queries exhibited by a state-of-the-art learning to rank algorithm.

“Ranking Learning-to-Rank Methods” by Djoerd Hiemstra [8]

Like most information retrieval methods, LtR methods are evaluated on benchmark datasets, such as the many datasets provided by Microsoft and the datasets provided by Yahoo and Yandex. In this talk, the author proposed a way to compare LtR methods based on a sparse set of evaluation results on many benchmark datasets. The comparison methodology consists of two components: (1) the Normalized Winning Number, a measure that gives insight in the ranking accuracy of the LtR method, and (2) the Ideal Winning Number, which gives insight in the degree of certainty concerning the ranking accuracy. Evaluation results of 87 LtR methods on 20 well-known benchmark datasets were collected. It was reported on the best performing methods by Normalized Winning Number and Ideal Winning Number and the author suggested what methods need more research to make the analysis more robust.

3 Paper Presentations

Online Learning of a Ranking Formula for Revenue and Advertiser ROI Optimization by Or Levi [10]

The presentation discussed LtR for ads ranking. First, the author addressed the task of ranking ads on the search results page for revenue optimization. While most works address this challenge by improving *Click Through Rate (CTR)* estimation, he considered the case when CTR estimation is somehow ineffective, and this is compensated by applying a larger weight to the bid factor. Second, the author aimed at improving advertiser *Return On Investment (ROI)*, while keeping a similar level of revenues for ads ranking on the home page feed. To this end, he introduced into the standard ranking formula a factor that favors ads with higher click-out rate and serves as an effective tie-breaker in cases of two competing ads with relatively similar revenue expectations. To optimize the ranking formula, for each case, he proposed an online learning procedure in a multi-armed bandit setting. Empirical evaluation demonstrates significant improvements over the existing ranking in production.

Query-level Ranker Specialization by Rolf Jagerman, Harrie Oosterhuis, and Maarten de Rijke [9]

Traditional LtR models optimize a single ranking function for all available queries. This assumes that all queries come from a homogeneous source. Instead, it seems reasonable to assume that queries originate from heterogeneous sources, where certain queries may require documents to be ranked differently. The authors introduced the Specialized Ranker Model which assigns queries to different rankers that become specialized on a subset of the available queries. The authors provided a theoretical foundation for this model starting from the listwise Plackett-Luce ranking model and derived a computationally feasible expectation-maximization procedure to infer the model’s parameters. Furthermore they conducted experiments with a noisy oracle to model the risk/reward tradeoff that exists when deciding which specialized ranker to use for unseen queries.

Discovery and Promotion of Subtopic Level High Quality Domains for Programming Queries in Web Search by Arpita Das, Saurabh Shrivastava, and Manoj Chinnakotla [3]

With the advancement of technology in modern era, a significant segment of the Web serves to satisfy the programming related information need of the users. User satisfaction in this segment not only depends on the relevance of the retrieved pages, but also on the domains that these pages belong to. The authors aimed at discovering sub-topic level associations of the domains and queries. They proposed a supervised deep neural network based approach using the click-through data of a commercial Web search engine to discover and promote the domains which provide high quality and expert level content for a query intent. Experiments show that their domain specific ranker performs significantly well, both qualitatively as well as quantitatively, on real-world coding query sets when compared with standard web ranking baseline.

A Software Library for Conducting Large Scale Experiments on Learning to Rank Algorithms by Nicola Ferro, Paolo Picello, and Gianmaria Silvello [6]

This presentation proposed an efficient application for driving large scale experiments on LtR algorithms. The authors designed a software library that exploits caching mechanisms and efficient data structures to make the execution of massive experiments on LtR algorithms as fast as possible in order to try as many combinations of components as possible. The proposed software has been tested on different algorithms as well as on different implementations of the same algorithm in different libraries. This software is highly configurable and extensible in order to enable the seamless addition of new features, algorithms, and libraries.

4 Short Presentations

The Impact of Negative Samples on Learning to Rank by Claudio Lucchese, Franco Maria Nardini, Raffaele Perego, and Salvatore Trani [11]

The effectiveness of LtR ranking 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 presentation the authors reviewed the state-of-the-art proposals and they reported the results of a preliminary investigation of a new sampling strategy aimed at reducing the number of not relevant query-document pairs.

Multileaving for Online Evaluation of Rankers by Brian Brost [1]

In online LtR one of the main challenge is how to define a tradeoff between exploring new, potentially superior rankers, and exploiting preexisting knowledge of what rankers have performed well in the past. Multileaving methods offer an attractive approach to this problem since they can efficiently use online feedback to simultaneously evaluate a potentially arbitrary number of rankers.

In this talk the author discussed some of the main challenges in multileaving, and promising areas for future research.

When Learning to Rank Meets Axiomatic Thinking by Hui Fang and Chengxiang Zhai

Axiomatic thinking has been successfully applied to analyze and improve retrieval models and evaluation metrics. The basic idea of axiomatic thinking is to leverage formalized constraints to guide the search of the optimal solutions for a given problem. In this presentation, the authors talked about their vision on applying axiomatic thinking to the problem of LtR.

Learning to Rank Target Types for Entity-Bearing Queries by Darío Garigliotti, and Krisztian Balog [7]

Detecting the target types of entity-bearing queries can help to improve retrieval performances as well as the overall search experience. The authors proposed an LtR approach, with a rich variety of features, for automatically identifying the target types of a query with respect to a type taxonomy. Using a purpose-built test collection, they showed that their method outperforms existing ones by a remarkable margin.

5 Discussion and Conclusion

The workshop enjoyed an audience of about 30 participants, who actively participated to the discussions fostered by the paper presentations. On the basis of the presentations and their discussions, we highlight three main research challenges in the area of LtR.

The first is related to the evaluation and consolidation of current results. The invited talk by Djoerd Hiemstra [8] reported on a comparison, based on literature review, among several LtR algorithms, while the presentation by Brian Brost [1] covered the topic of online evaluation, the invited talk by Craig Macdonald discussed risk-sensitive evaluation measures, and finally Ferro et al. [6] proposed a novel software library for large scale LtR experiments. Indeed, it is not clear whether or not there is a reference algorithm for LtR to which any new proposed approach should be compared to, and choosing the appropriate evaluation measure for such comparison is not trivial. In this regard, as discussed by Claudio Lucchese et al. [11], also the creation of LtR datasets should deserve further attention.

The second challenge arises from the application of LtR to specific fields such as Web advertisement, as discussed by Or Levi [10], or from the specialization of ranking models to different kinds of queries, as discussed by Rolf Jagerman et al. [9], by Dario Gigliotti et al. [7] or by Arpita Das et al. [3]. Specialization of rankers leads to improved ranking quality at the cost of a more complex query analysis. Moreover, some of the insights found in specific domains may apply to other domains or to Web search in general.

The third challenge regards the application of novel methodologies to provide a significant shift beyond the current state of the art. Hui Zang proposed axiomatic thinking as a novel approach. The benefits in the LtR task are not yet determined.

After the workshop, feedback from the attendees was collected by means of an online form. Among the most important research challenges ahead in the LtR area, participants highlighted deep learning, learning from synthetic data and weakly supervised training, online optimization and evaluation, and learning to rank directly from user signals and interactions. Furthermore, the participants suggested more industry involvement for future edition of the workshop. Indeed annotated offline LtR datasets are limited and industries have insights into many LtR problems and on much more interesting types of data. Opening discussion towards the problem of data in our community should also be included.

We can conclude that the workshop was appreciated, topics discussed were found interesting, and that a longer workshop, allowing a longer open discussion, would meet the favour of the attendees.

Acknowledgements

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] Brian Brost. Multileaving for Online Evaluation of Rankers. In Ferro et al. [4].

[2] Olivier Chapelle, Yi Chang, and Tie-Yan Liu. Future directions in learning to rank. In *Proceedings of the 2010 International Conference on Yahoo! Learning to Rank Challenge - Volume 14*, Proceedings of Machine Learning Research, Vol. 14, pages 91–100. PMLR, 2011. URL http://dl.acm.org/citation.cfm?id=3045754.3045764.

[3] Arpita Das, Saurabh Shrivastava, Manoj Chinnakotla, Prateek Agrawal, and Sandeep Sahoo. Discovery and Promotion of Subtopic Level High Quality Domains for Programming Queries in Web Search. In Ferro et al. [4].

[4] N. Ferro, C. Lucchese, M. Maistro, and R. Perego, editors. *1st International Workshop on LEARning Next gEneration Rankers (LEARNER 2017)*, 2017. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073.

[5] Nicola Ferro, Claudio Lucchese, Maria Maistro, and Raffaele Perego. LEARning Next gEneration Rankers (LEARNER 2017). In J. Kamps, E. Kanoulas, M. de Rijke, H. Fang, and E. Yilmaz, editors, *Proc. 3rd ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR 2017)*, pages 331–332. ACM Press, New York, USA, 2017. doi: 10.1145/3121050.3121110. URL http://doi.acm.org/10.1145/3121050.3121110.

[6] Nicola Ferro, Paolo Picello, and Gianmaria Silvello. A Software Library for Conducting Large Scale Experiments on Learning to Rank Algorithms. In Ferro et al. [4].

[7] Darío Garigliotti and Krisztian Balog. Learning to Rank Target Types for Entity-Bearing Queries. In Ferro et al. [4].

[8] Djoerd Hiemstra, Niek Tax, and Sander Bockting. Ranking Learning-to-Rank Methods. In Ferro et al. [4].

[9] Rolf Jagerman, Harrie Oosterhuis, and Maarten de Rijke. Qery-level Ranker Specialization. In Ferro et al. [4].

[10] Or Levi. Online Learning of a Ranking Formula for Revenue and Advertiser ROI Optimization. In Ferro et al. [4].

[11] Claudio Lucchese, Franco Maria Nardini, Raffaele Perego, and Salvatore Trani. The Impact of Negative Samples on Learning to Rank. In Ferro et al. [4].