Learning to Rank for Information Retrieval (LR4IR 2007)

Thorsten Joachims
Cornell University
tj@cs.cornell.edu

Hang Li
Microsoft Research Asia
hangli@microsoft.com

Tie-Yan Liu
Microsoft Research Asia
tyliu@microsoft.com

ChengXiang Zhai
University of Illinois at Urbana-Champaign
czhai@cs.uiuc.edu

1 Introduction

The task of "learning to rank" has emerged as an active and growing area of research both in information retrieval and machine learning. The goal is to design and apply methods to automatically learn a function from training data, such that the function can sort objects (e.g., documents) according to their degrees of relevance, preference, or importance as defined in a specific application.

The relevance of this task for IR is without question, because many IR problems are by nature ranking problems. Improved algorithms for learning ranking functions promise improved retrieval quality and less of a need for manual parameter adaptation. In this way, many IR technologies can be potentially enhanced by using learning to rank techniques.

A workshop on “Learning to Rank for Information Retrieval (LR4IR 2007)” was held in conjunction with the 30th Annual International ACM SIGIR Conference (SIGIR 2007), in Amsterdam, on July 27, 2007. The main purpose of this workshop was to bring together IR researchers and ML researchers working on or interested in the learning to rank technologies, and let them share their latest research results, express their opinions on the related issues, and discuss future directions.

A program committee was created, consisting of 25 prominent researchers in machine learning and information retrieval, and from both academia and industry. The call for papers attracted 14 submissions. 8 submissions were accepted by the program committee for presentations at the workshop, based on the novelty, significance, relevance, clarity, and technical soundness of the papers. Besides, two invited talks and one panel discussion were also organized at the workshop. A benchmark data set was made available for experiments before the workshop. There were about 100 registrations to the workshop, making the workshop the largest one at SIGIR’07 in terms of number of participants.

All the information on the workshop is available at the web site: http://research.microsoft.com/users/LR4IR-2007/ .

2 Technical Program

2.1 Invited Talk

Two distinguished researchers were invited to give keynote speeches: Bruce Croft from University of Massachusetts and Chris Burges from Microsoft Research.

Bruce Croft delivered his invited talk on "Learning about Ranking and Retrieval Models." Bruce first reviewed how retrieval models in IR have developed. He then pointed out that "learning to rank" approaches have a number of potential advantages for retrieval model construction, but there are also important issues that must be addressed. He then described how these new approaches fit into the development of retrieval models, and gave an overview of what might be addressed in future research.

Chris Burges's invited talk was with the title "Learning to Rank for Web Search: Some New Directions." Chris first pointed out that one aspect of the ranking problem in web search that attracted the interest of machine learning researchers is that the objective functions to optimize are either flat or discontinuous everywhere, and thus are difficult to handle by using the optimization techniques such as gradient descent. He then introduced their recent work on tackling the challenge, including LambdaRank, SPSA, MART, and XRank.

2.2 Paper Presentation

There were 8 papers presented at the workshop.

The paper by Tie-Yan Liu, Jun Xu, Tao Qin, Wenying Xiong, and Hang Li, entitled "LETOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval," was presented by Jun Xu. In the talk, Jun introduced the benchmark data set, LETOR, developed for research on learning to rank for information retrieval. The data set was derived from the existing data sets in OHSUMED and TREC. The details of LETOR were introduced, including queries, relevance judgments, and features.

The paper by Ronan Cummins and Colm O'Riordan, entitled "An Axiomatic Study of Learned Term Weighting Schemes" was presented by Ronan Cummins. Ronan first introduced a novel term-weighting scheme that employs an evolutionary learning approach and explained the relationships between the new scheme and existing term weighting schemes. Then, he described the effectiveness of the proposed scheme by giving some empirical study results.

Michael Taylor gave a talk on their paper "SoftRank: Optimising Non-Smooth Rank Metrics" by Michael Taylor, John Guiver, Stephen Robertson, and Tom Minka. Michael started his presentation by pointing out that most IR applications use evaluation metrics that are innately non-smooth, in contrast to the fact that machine learning algorithms usually require the smoothness of loss functions. They proposed, therefore, utilizing a new smooth utility function called soft NDCG, and employing a new algorithm to perform the optimization task called Soft Rank.

The paper entitled “Learning to Rank with Pairwise Regularized Least-Squares” by Tapio Pahikkala, Evgeni Tsivtsivadze, Antti Airola, Jorma Boberg, and Tapio Salakoski was presented by Tapio. They proposed a new preference learning algorithm based on regularized least squares within the framework of kernel methods. Tapio introduced in details about the algorithm, referred to as RankRLS, by showing its loss function, optimization techniques, and efficient implementations.

Guihong Cao presented the paper entitled “Learning to Rank Documents for Ad-Hoc Retrieval with Regularized Models” co-authored with Jian-Yun Nie, Luo Si, and Jing Bai. Guihong first pointed out the importance of accurate estimation of the weights of mixture models in language modeling for information retrieval. He then introduced two methods they have developed recently. One is based on Deterministic Annealing EM (DAEM), and the other L2-regularized log-linear model (RLM).

In his presentation of the paper “Learning to Rank for Information Retrieval Using Genetic Programming” by Jen-Yuan Yeh, Jung-Yi Lin, Hao-Ren Ke, and Wei-Pang Yang, Jen-Yuan proposed a new learning to rank method, called RankGP. RankGP employs genetic programming to learn a ranking function which utilizes various types of evidences in IR as features, including content features, structure features, and query-independent features.

Filip Radlinski gave a presentation on his paper entitled “Addressing Malicious Noise in Clickthrough Data”. Filip first made an analysis on click-spam from a utility standpoint, and investigated whether personalizing web search results can reduce or eliminate the financial incentives of spammers. He then gave a formalization of click-spam and introduced investigations on the model.

The paper “Efficient Query Delegation by Detecting Redundant Retrieval Strategies” by Christian Scheel, Nicolas Neubauer, Andreas Lommatzsch, Klaus Obermayer, and Sahin Albayrak was presented by Christian Scheel. Their work was about data fusion in information retrieval, particularly the weighting strategy for combining ranking functions (strategies). They proposed using similarity measures between ranking functions, such as correlation measures, as the weighing strategy and demonstrated the effectiveness of the proposal through experimental results.

2.3 Panel Discussion

A panel discussion was also held, with contribution from four panelists, Thorsten Joachims from Cornell, Steven Robertson from Microsoft Research, Hugo Zaragoza from Yahoo Research, and ChengXiang Zhai from UIUC. During the first part, the panelists outlined their perspectives on learning to rank for information retrieval.

Thorsten expressed his belief in machine learning as a fundamental model for IR. On an abstract level, supervised machine learning aims to model the relationship between an input x (e.g., query and information need of a user) and an output y (e.g., relevant information, for example in the form of a ranking) over a distribution P(x,y) and a given performance measure. This setup obviously fits many information retrieval problems, and much of what is known in ML could be transferred to IR. If it is possible to extend ML algorithms and theory from the “simple” x and y typically studied in ML to the complex x and y in IR, this could provide a very fruitful paradigm for studying the complexity of retrieval problems theoretically, and for deriving new retrieval algorithms. This would make ML not only a tool for IR as it is today, but a fundamental model of IR. As a second point, Thorsten raised the question of what training data to use when learning or tuning retrieval systems. A consequent

machine learning approach implies that the learning algorithms needs access to training data from P(x,y), namely the distribution of users their information needs. The most direct access to such data is through implicit feedback from user interactions. He argues that such data should be used for training retrieval systems to make sure that they actually fit the needs of the user population, which may be quite different from what is conjectured by expert relevance judges.

Stephen talked about the necessity of conducting research on the optimization measures in learning to rank for IR. He first questioned the correctness of the machine learning principle when applied to IR, that in training one should optimize the objective function which is exactly used in prediction (testing). He then gave two reasons: (1) for ease of manipulation it is better to utilize smooth and convex functions as objective functions, while in IR applications objective functions are usually neither smooth nor convex; (2) the objective functions in IR may be highly granular (e.g., P@5), and thus in training one may need define more sensitive objective functions to optimize in order to make better generalization. He concluded, therefore, that we may need to treat the choice of optimization measure in training as a separate issue from that in prediction.

Hugo argued that in many cases we need machine learning to improve IR; however, machine learning is still not powerful enough to meet the request. He proposed that we should move learning to rank beyond its current formulation. The learning to rank problem is difficult, because it imposes a special structure on the output. He pointed out that there exist many other forms of structure that we need to exploit in order to solve the IR challenges. He then took entity ranking as an example. Entities are inter-related in different ways, and this yields an input space with a rich structure, which current machine learning techniques do not take into account.

ChengXiang pointed out that we should not think that machine learning is omnipotent for IR, although we are excited about the huge potential of applying machine learning techniques to IR. He expressed his concern that simply casting IR as a supervised learning problem without looking into modeling relevance deeply (e.g., through semantic analysis) would not go very far. He challenged machine learning researchers to "rediscover" an effective weighting formula such as BM25 through pure learning without relying prior knowledge about some known forms of formulas such TF normalization and IDF. He pointed out the need for integrating conditional models with generative language models. ChengXiang also indicated that the current formulation of learning to rank only reflects a simplified view of ranking in IR, and an interesting challenge would be to apply machine learning techniques to solve more complex ranking problems such as the subtopic retrieval task, which goes beyond "independent relevance" and requires considering redundancy among the documents while ranking them.

The second part of the panel consisted of questions by the audience to the panelists and free discussions between the panelists and the audience. Issues such as (1) how to make a balance between leveraging the power of machine learning techniques and enhancing the understanding of IR problems, (2) how to apply machine learning techniques to subareas of IR, (3) what the challenges are when applying learning to rank to IR, etc were discussed.

2.4 Shared Benchmark Data

A data set for learning to rank on information retrieval was released prior to the workshop. The data set, created based on OHSUMED and TREC data, contains features and relevance judgments for

training and evaluation of learning to rank methods. It is available at (http://research.microsoft.com/users/tyliu/LETOR/).

3 Conclusion

Learning to rank for information retrieval is an emerging area in IR. This is the first workshop devoted to the theme at SIGIR. The workshop attracted a broad range of interest. The feedbacks to the workshop from the participants and related people were very positive. We sincerely hope that this workshop helps attract more and more people to work on the many challenging yet exciting issues in the area of learning to rank.

A workshop with the same title and scope is currently planned for the SIGIR next year. A related special issue on "learning to rank for information retrieval" is also under consideration for the journal of Information Retrieval.

4 Acknowledgements

We extend our sincere gratitude to the SIGIR workshop committee, the LR4IR program committee, the invited speakers, panelists, authors, presenters, and all the participants of the workshop.