Learning to Rank for Information Retrieval (LR4IR 2008)

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

Learning to rank is a machine learning technique for constructing a ranking function from training data and it can be applied into a variety of tasks in information retrieval. Learning to rank has been receiving keen and growing interest in information retrieval and machine learning recently. For example,

1) Many papers on learning to rank and related topics have been published at the major machine learning and information retrieval conferences recently.
    a. ICML 2007 and ICML2008
    b. SIGIR2007 and SIGIR2008
2) Workshops have been organized with learning to rank as the main theme.
    a. NIPS 2005 workshop on learning to rank
    b. ICML 2006 workshop on learning in structured output space
    c. SIGIR 2007 workshop on learning to rank for information retrieval (LR4IR 2007).
3) Benchmark dataset LETOR has been released. Until now, there have been over 500 downloads of the dataset.

Following the big success of the first workshop, we organized the second workshop of the series entitled "Learning to Rank for Information Retrieval", in conjunction with the 31st Annual International ACM SIGIR Conference (SIGIR 2008), in Singapore, on July 24, 2008. The main purpose of this workshop was, again, 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.

The call for papers attracted 14 submissions. A program committee consisting of 24 members reviewed all the submissions. 8 papers were selected for presentations at the workshop. Besides, two invited talks and one opinion and commentary session were also organized. About 50 people attended the workshop. Detailed information on the workshop is available at http://research.microsoft.com/users/LR4IR-2008/.

2 Technical Program

2.1 Invited Talks

Two distinguished researchers were invited to give keynote speeches: Stephen Robertson from Microsoft Research Cambridge and Hongyuan Zha from Georgia Tech.

Stephen Robertson gave a talk entitled "On the Optimization of Evaluation Metrics". In the talk, he questioned the correctness of a commonly made assumption in learning to rank, that is, if we want to optimize a particular evaluation metric at a search system, we should try to optimize exactly the same metric on the training set. Stephen explained in details in the talk why he thinks the assumption may be inappropriate. (http://research.microsoft.com/users/LR4IR-2008/keynote.pdf)

The talk by Hongyuan Zha was about "A Structured Learning Framework for Learning to Rank in Web Search". Hongyuan first gave a survey on learning to rank. He then proposed a new method for learning DCG, a popular evaluation measure in ranking. In the work he considered an evaluation metric as a form of utility function which reflects the degree of satisfaction of the users. He argued that the evaluation metrics themselves should be learned from judgment data and user interaction data; and learning to rank algorithms should seek to optimize the evaluation metrics through structured learning methodology. (http://research.microsoft.com/users/LR4IR-2008/lr4ir.pdf)

2.2 Paper Presentations

The 8 papers presented at the workshop are as follows.

Leonardo Rigutini, Tiziano Papini, Marco Maggini, and Franco Scarselli presented a paper entitled "SortNet: Learning To Rank By a Neural-Based Sorting Algorithm". The authors proposed a new adaptive algorithm for learning to rank, called SortNet. The method orders objects using neural network model as a comparator. In the method, the training set provides examples of the desired ordering between pairs of items and it is constructed by an iterative procedure which, at each iteration, adds the most informative training examples. Moreover, the comparator adopts a connectionist architecture that is particularly suited for implementing a preference function.

The paper by Zhaohui Zheng, Hongyuan Zha, and Gordon Sun, entitled "Query-Level Learning to Rank Using Isotonic Regression", was presented by Hongyuan Zha. The authors proposed a new optimization method for learning to rank. They first pointed out that existing methods tend to treat labeled data equally across queries. To cope with the problem, they proposed a method that takes into account the entire data within a query as a group at each iteration of training (optimization). The optimization problem then is tackled using functional iterative methods where the update at each iteration is computed by solving an isotonic regression issue.

The paper by Vitor R. Carvalho, Jonathan L. Elsas, William W. Cohen, and Jaime G. Carbonell was about "A Meta-Learning Approach for Robust Rank Learning". The author pointed out that one disadvantage of pairwise ranking learning is that an incorrect relevance judgment on a single document can lead to a large number of mislabeled document pairs. The authors studied the effects of outlying pairs in pairwise ranking learning and introduced a new learning algorithm capable of suppressing these undesirable effects. It turns out that the proposed algorithm works as a second optimization step in which any linear baseline ranker can be used as input.

Onno Zoeter, Michael Taylor, Ed Snelson, John Guiver, Nick Craswell, and Martin Szummer made a presentation entitled "A Decision Theoretic Framework for Ranking using Implicit Feedback". This paper proposes a new decision theoretic ranking system that uses both explicit and implicit feedbacks in training. The system employs a model that predicts, given all available data at query time, different interactions a person might have with search results. A utility function is defined to provide a real valued score to the user's session. The optimal ranking is the list of documents that maximizes the utility for a user session.

Alexandre Klementiev, Dan Roth, and Kevin Small's paper was about "A Framework for Unsupervised Rank Aggregation". The authors pointed out that the current rank aggregation algorithms generally require either domain knowledge or supervised ranked data; both of which are expensive to acquire. To address the limitations, they proposed a mathematical and algorithmic framework for learning to aggregate (partial) rankings in an unsupervised setting, and to instantiate it for the cases of combining permutations and combining top-k lists. They also derived an unsupervised learning algorithm for rank aggregation (ULARA), which approximates the behavior of this framework by directly optimizing the weighted Borda count.

"Machine Learned Sentence Selection Strategies for Query-Biased Summarization" was a paper by Donald Metzler and Tapas Kanungo. The paper investigated how learning to rank approaches can be applied to the task of selecting relevant sentences for abstract generation in search. The authors analyzed and evaluated several learning to rank approaches, such as ranking support vector machines (SVMs), support vector regression (SVR), and gradient boosted decision trees (GBDTs). The results showed that the effectiveness of the machine learning approaches varies across collections with different characteristics. The results also showed that GBDTs provide a robust and powerful framework for the sentence selection task and significantly outperform SVR and ranking SVMs on several data sets.

The paper by Tom Minka and Stephen Robertson "Selection bias in the LETOR datasets" was presented by Stephen Robertson. The authors conducted an examination on the LETOR datasets and showed that the document selection in LETOR has (for each of the three corpora) a particular bias or skewness. This may have unexpected effects influencing any learning-to-rank exercise conducted on these datasets. The problems may be resolvable by modifying the datasets, indicated by the authors.

Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li gave a talk on their paper entitled "How to Make LETOR More Useful and Reliable". The paper discussed three ways to further improve LETOR to make it more useful and reliable. First, adding more information to the features, to enable the reproduction or optimization of the features. Second, employing a new document sampling strategy to reduce the bias in the previous LETOR datasets. Third, adding more queries to the datasets in LETOR, and enlarging the datasets.

2.3 Opinion and Commentary Session

There was a special 'opinion and commentary' session organized, as conclusion of the workshop. Participants were invited to give five-minute presentations to introduce their opinions on learning to rank for IR. After that, free discussions were made between the speakers and the audience. Donald Metzler, Jun Xu, ChengXiang Zhai talked about their views on learning to rank. Heated discussions were made after each of the presentations.

Donald talked about his thoughts on the important factors for learning to rank. He started the talk by casting a question to the audience, which is based on an imaginary competition on improving search relevance. "One team tries new features while the other new models. Who will win?" The answers from the audience were the same: the first team. His presentation and the related discussions from the audience were focusing on the relations between features and models. Here are the main points discussed. (1) Features are what IR community studied. Models trained with machine learning are somewhat new in learning to rank. Models and features are interdependent. For example, if we use a linear model then query dependent features will make no sense. For both feature and model research, there is no low hanging fruit any more. (2) Can we develop a learning method that can generalize well across different features? (3) Research on automatic learning of features (e.g., Donald' works MRF) is also important for IR. Employing learning methods which can learn both model and feature at the same time would also be an interesting direction. (4) Data and evaluation are also important for learning to rank, in addition to feature and model.

Jun introduced his view on “semi supervised learning”. Here are the major points in his presentation. (1) In learning to rank we should leverage IR knowledge as much as possible. This might be particularly true for semi-supervised learning. For example, BM25 is a powerful model created with “unsupervised” learning and thus a semi-supervised learning method in learning to rank should leverage this fact. (2) While similarity between instances is useful information for semi-supervised learning for classification, this does not seem to be true for ranking. Therefore, a semi-supervised learning method for ranking might not benefit from the use of such information. (3) There are several types of unlabeled data: queries with both labeled and unlabeled documents, and queries with unlabeled documents. We need to consider the uses of both of them in semi-supervised learning for ranking.

ChengXiang talked about the relationships between learning to rank and other learning methods for IR. Bruce Croft gave several keywords to learning to rank in his invited talk at the previous workshop: discriminative training, feature based, and data driven. Is this enough? The risk minimization framework which Zhai et al introduced should also be regarded as a kind of learning to rank method. Discussions were made among the audience. It seems that combining the discriminative and generative approaches would also be an important direction to explore. For example, in machine learning, the relations between generative and discriminative models in classification were studied. It is still not clear what the relations between the generative models and discriminative models are in learning to rank.

3 Conclusion

It was a very successful workshop in the sense that many participants actively took part in the discussions, particularly in the opinion and commentary session. There were many novel and inspiring ideas exchanged among the participants. The feedbacks on the workshop from the participants were also very positive. We hope that we can re-run the workshop next year and make it a platform for commutations among related researchers.

4 Acknowledgements

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