WORKSHOP REPORT

Learning to Rank for Information Retrieval (LR4IR 2009)

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 Overview

As an interdisciplinary field between information retrieval and machine learning, learning to rank is concerned with automatically constructing a ranking model using training data. Learning to rank technologies have been successfully applied to many tasks in information retrieval such as search and collaborative filtering, and have been attracting more and more attention recently.

At SIGIR 2007 and SIGIR 2008, we have successfully organized two workshops on learning to rank for information retrieval with very good attendance. The reports of those two workshops can be found at http://www.sigir.org/forum/2007D/2007d_sigirforum_joachims.pdf
http://www.sigir.org/forum/2008D/sigirwksp/2008d_sigirforum_li.pdf

From the experiences of running those two workshops, we have found that there is a community emerging, consisting of people from both academia and industry and including both researchers and practitioners. They have rich experiences of IRand machine learning, and are also deeply interested in the learning to rank technologies. We have, therefore, organized a workshop on the same theme again, in conjunction with SIGIR 2009. The main purpose remains to bring together IR researchers and ML researchers working on or interested in the technologies, and enable them to share their latest research results, to express their opinions on the related issues, and to discuss future directions.

The call for papers attracted 12 submissions. A program committee consisting of 24 members reviewed all the submissions. 7 papers were selected for presentations at the workshop. Besides, two invited talks and one opinion session were also organized. About 30 people attended the workshop. Detailed information on the workshop is available at http://research.microsoft.com/en-us/um/beijing/events/lr4ir-2009/ .

The feedback on the workshop from the participants was very positive. All the people who took part in a survey after the workshop gave rates of 4-5 for the question of how much did you enjoy the workshop. Many of them said that they really liked the invited talks and the opinion session, and they were really impressed by the quality of the papers presented at the workshop. One participant commented: “The organizers did a fantastic job! This was the highlight of my SIGIR experience”. It was a very successful workshop also in the sense that many participants actively took part in the discussions, particularly in the opinion session.

2 Technical Program

2.1 Invited Talks

Two distinguished researchers were invited to give keynote speeches: Paul B. Kantor from Rutgers University and Olivier Chapelle from Yahoo Research.

In his invited speech, Paul Kantor talked about the problem of “Learning to Rank for Diversity”. He started the talk by pointing out that ranking of search results is aimed at maximizing both user service and server revenue, which might be either consonant or competing goals. He mentioned that there are two aspects to address queries whose ambiguity poses a challenge in ranking: identifying the salient distinct meanings in the returned data and presenting them in the most effective way. He introduced their ongoing work on random projections and clustering algorithms for identifying the meanings of queries. He concluded the talk by raising the questions of whether it is possible to learn to rank for diversity and whether it is possible to present aspects in a non-linear way (beyond linear reading). Slides of Paul’s talk are available at http://research.microsoft.com/en-us/um/beijing/events/lr4ir-2009/Kantor%20PPApproach%20-LR4IR%202009.pdf

Olivier Chapelle shared his thoughts on “Direct Optimization for Web Search Ranking”. Olivier indicated that most ranking algorithms, such as pairwise ranking, are based on the optimization of standard loss functions, but the quality measure to test web page rankers is often different. He then introduced two algorithms they developed, which aim at directly optimizing one of the popular measures, Normalized Discounted Cumulative Gain. One of the methods uses a continuous approximation of NDCG, and the other is based on the framework of structured output learning. Finally he shared his perspectives on learning to rank, such as the differences among the pointwise, pairwise, and listwise approaches, future research topics for learning to rank. Slides of Olivier’s talk are available at http://research.microsoft.com/en-us/um/beijing/events/lr4ir-2009/Olivier-LR4IR%202009.pdf

2.2 Paper Presentations

The 7 papers presented at the workshop are as follows.

“Efficient and Accurate Local Learning for Ranking” by Somnath Banerjee, Avinava Dubey, Jinesh Machchhar, Soumen Chakrabarti. In this paper the authors propose a local learning to rank algorithm based on a new similarity measure between queries. First, the authors represent (relevant and irrelevant) document vectors for the query as a point cloud. Second, the authors define a similarity between the shapes of two point clouds, based on principal component analysis (PCA). Their local learning algorithm then clusters queries at training time, using the PCA-based query similarity measure. During test time, they locate the nearest training cluster, and use in ranking the model trained for that cluster. The test time is small, training time is reasonable, and the accuracy in ranking beats several local learning approaches, as tested on the LETOR dataset.

“Learning to Rank QA Data” by Suzan Verberne, Hans van Halteren, Daphne Theijssen, Stephan Raaijmakers, and Lou Boves. In this work, the authors evaluate a number of machine learning techniques for ranking of answers in why-type question answering. They use a set of 37 linguistically motivated features that characterize questions and answers and experiment with a number of learning techniques in various settings. The purpose of the experiments is to assess how different machine learning techniques can cope with the highly imbalanced binary relevance data. The authors find that with all machine learning techniques, it is possible to obtain an MRR score that is significantly above the TF-IDF baseline of 0.25 and not significantly lower than the best score of 0.35. Regression techniques seem to be the best option for the learning problem.

“Ranking Experts with Discriminative Probabilistic Models” by Yi Fang, Luo Si, and Aditya P. Mathur. In this paper, the authors investigate how to merge and weight heterogeneous knowledge sources in the task of expert finding. A relevance-based supervised learning framework is presented to learn the combination weights from training data. Beyond just learning a fixed combination strategy for all the queries and experts, the authors propose a series of probabilistic models which have increasing capability to associate the combination weights with specific experts and queries. In the last (and also the most sophisticated) proposed model, the combination weights depend on both expert classes and query topics, and these classes and topics are derived from expert and query features. Compared with expert and query independent combination methods, the proposed combination strategy can better adjust to different types of experts and queries. In consequence, the model yields much flexibility of combining data sources when dealing with a broad range of expertise areas and a large variation in experts. Empirical studies on a real world faculty expertise testbed demonstrate the effectiveness and robustness of the proposed learning based models.

“Is learning to rank effective for Web search” by Min Zhang, Da Kuang, Guichun Hua, Yiqun Liu, Shaoping Ma. This paper empirically studies the effectiveness of the state-of-art learning to rank algorithms, especially in Web search scenario. Besides LETOR, the benchmark data for learning to rank, a Web search data set is used, which is from a commercial search engine. Five approaches have been studied, including linear regression, RankBoost, ListNet, top k optimization of ListMLE, and SVM-MAP. Comparative study has been made among algorithms and across different datasets. Furthermore, the effects of learning to rank algorithms are compared with that of content-based and link-based ranking features. Essential differences have been observed and analyzed in the paper in terms of the effectiveness and stability of the algorithms and the feature selection.

“Priors in Web Search” by Michael Bendersky and Kenneth W. Church. Web search combines information obtained at query time with prior knowledge to form a posterior. This paper focuses on the prior, which is believed to be important, given the poverty of the query stimulus (many of the web queries are no more than a word or two). The authors propose a learning framework based on the Noisy Channel Model for combining prior evidence from multiple sources including both the authors' perspectives (e.g., PageRank - the principal eigenvector of the web graph) as well as the readers' perspectives (e.g., click logs and toolbar activity). The framework is general enough that it can be applied to both documents and queries. The authors show that even features that appear to depend on the combination of queries and documents and are often used for learning a ranking function (such as relevance judgments or retrieval scores) can be included in the prior model using multiple mechanisms of aggregation (e.g., moments or entropy). The prior model improves with both more features and more aggregates. The authors conduct an empirical evaluation of the proposed framework, demonstrating its benefits over a diverse set of learning tasks including: (1) query difficulty estimation, (2) click types prediction and (3) document ranking.

“Learning to Rank with Low Rank” by Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Yanjun Qi, Kunihiko Sadamasa, Olivier Chapelle, Kilian Weinberger. In this article the authors present Supervised Semantic Indexing (SSI) which defines a class of nonlinear (quadratic) models that are discriminatively trained to directly map from the word content in a query-document or document-document pair to a ranking score. Like Latent Semantic Indexing (LSI), the models take account of correlations between words (synonymy, polysemy). However, unlike LSI the models are trained with a supervised signal directly on the ranking task of interest. As the query and target texts are modeled separately, the proposed approach is easily generalized to different retrieval tasks, such as cross-language retrieval or online advertising placement. Dealing with models on all pairs of words features is computationally challenging. The authors further propose several improvements to the basic model for addressing this issue, including low rank (diagonal preserving) representations, correlated feature hashing (CFH). The authors provide an empirical study of all these methods on retrieval tasks based on Wikipedia documents as well as an Internet advertisement task. They obtain state-of-the-art performance while providing realistically scalable methods.

“On the Choice of Effectiveness Measures for Learning to Rank” by Emine Yilmaz, Stephen Robertson. Most current machine learning methods for building search engines are based on the assumption that there is a target evaluation metric that evaluates the quality of the search engine with respect to an end user and the engine should be trained to optimize for that metric. Treating the target evaluation metric as a given, many different approaches (e.g. LambdaRank, SoftRank, RankingSVM, etc.) have been proposed to develop methods for optimizing for retrieval metrics. Target metrics used in optimization act as bottlenecks that summarize the training data and it is known that some evaluation metrics are more informative than others. In this paper, the authors consider the effect of the target evaluation metric on learning to rank. In particular, they question the current assumption that retrieval systems should be designed to directly optimize for a metric that is assumed to evaluate user satisfaction. The authors show that even if user satisfaction can be measured by a metric X, optimizing the engine on a training set for a more informative metric Y may result in a better test performance according to X (as compared to optimizing the engine directly for X on the training set). The authors analyze the situations as to when there is a significant difference in the two cases in terms of the amount of available training data and the number of dimensions of the feature space.

2.3 Opinion Session

There was a special ‘opinion session’ organized at the workshop. Participants were asked to express their opinions on learning to rank for IR. After that, free discussions were made among the workshop participants.

Yisong Yue made a brief presentation on “Interactive Approaches to Learning to Rank”. He argued that learning to rank has gained significant attention in recent years, and as methods become more sophisticated, the lack of larger and more realistic datasets is likewise becoming a greater limiting factor. He then pointed out that one increasingly popular alternative is to leverage implicit feedback such as click-through data, but the challenge is that most existing methods use passively collected feedback and such data is inherently biased towards the incumbent retrieval function. He then introduced their recent work on learning interactively from users from observing implicit feedback. By controlling which results to show to users, interactive methods can elicit more meaningful (and less biased) feedback.

Tie-Yan Liu gave an introduction to LETOR 4.0, a new release of the benchmark datasets for learning to rank research. LETOR 4.0 data is derived from TREC million query track. It supports more settings of learning to rank, semi-supervised ranking, listwise ranking, rank aggregation, in addition to supervised ranking. There are in total 2,500 queries and thus the size of data has been significantly increased. LETOR web site: http://research.microsoft.com/~LETOR/.

Other issues such as future directions of learning to rank research, evaluation of learning to rank for IR, the use of click-through data in learning, and teaching of learning to rank were also discussed.

3 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.