Information Processing and Management 51 (2015) 757–772

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

# A cross-benchmark comparison of 87 learning to rank methods

Niek Tax[a,c,1,*], Sander Bockting[a], Djoerd Hiemstra[b]

[a] *Avanade Netherlands B.V., Versterkerstraat 6, 1322AP Almere, The Netherlands*
[b] *University of Twente, P.O. Box 217, 7500AE Enschede, The Netherlands*
[c] *Eindhoven University of Technology, P.O. Box 513, 5600MB Eindhoven, The Netherlands*

ARTICLE INFO

*Article history:*
Received 1 October 2014
Revised 14 May 2015
Accepted 7 July 2015
Available online 22 August 2015

*Keywords:*
Learning to rank
Information retrieval
Evaluation metric

ABSTRACT

Learning to rank is an increasingly important scientific field that comprises the use of machine learning for the ranking task. New learning to rank methods are generally evaluated on benchmark test collections. However, comparison of learning to rank methods based on evaluation results is hindered by the absence of a standard set of evaluation benchmark collections. In this paper we propose a way to compare learning to rank methods based on a sparse set of evaluation results on a set of benchmark datasets. Our comparison methodology consists of two components: (1) Normalized Winning Number, which gives insight in the ranking accuracy of the learning to rank method, and (2) Ideal Winning Number, which gives insight in the degree of certainty concerning its ranking accuracy. Evaluation results of 87 learning to rank methods on 20 well-known benchmark datasets are collected through a structured literature search. ListNet, SmoothRank, FenchelRank, FSMRank, LRUF and LARF are Pareto optimal learning to rank methods in the Normalized Winning Number and Ideal Winning Number dimensions, listed in increasing order of Normalized Winning Number and decreasing order of Ideal Winning Number.

## 1. Introduction

Ranking is a core problem in the field of information retrieval. The ranking task in information retrieval entails the ranking of candidate documents according to their relevance to a given query. Ranking has become a vital part of web search, where commercial search engines help users find their need in the extremely large collection of the World Wide Web. Among useful applications of machine learning based ranking outside web search are automatic text summarization, machine translation, drug discovery and determining the ideal order of maintenance operations (Rudin, 2009). In addition, McNee, Riedl, and Konstan (2006) found the ranking task to be a better fit for recommender systems than the regression task (continuous scale predictions), which is currently still frequently used within such systems.

Research in the field of ranking models has long been based on manually designed ranking functions, such as the well-known BM25 model (Robertson & Walker, 1994). Increased amounts of potential training data have recently made it possible to leverage machine learning methods to obtain more effective ranking models. Learning to rank is the relatively new research area that covers the use of machine learning models for the ranking task.

[*] Corresponding author at: Eindhoven University of Technology, Department of Mathematics and Computer Science, P.O. Box 513, 5600MB Eindhoven, The Netherlands. Tel.: +31 634085760.

*E-mail addresses:* n.tax@tue.nl (N. Tax), sander.bockting@avanade.com (S. Bockting), d.hiemstra@utwente.nl (D. Hiemstra).

[1] Author is affliated with Eindhoven University of Technology, but this paper was written during his stay at Avanade Netherlands B.V.

http://dx.doi.org/10.1016/j.ipm.2015.07.002
0306-4573/© 2015 Elsevier Ltd. All rights reserved.

In recent years, several learning to rank benchmark datasets have been proposed with the aim of enabling comparison of learning to rank methods in terms of ranking accuracy. Well-known benchmark datasets in the learning to rank field include the *Yahoo! Learning to Rank Challenge* datasets (Chapelle & Chang, 2011), the *Yandex Internet Mathematics 2009* contest,[2] the LETOR datasets (Qin, Liu, Xu, & Li, 2010), and the MSLR (Microsoft Learning to Rank) datasets.[3] There exists no agreement among authors in the learning to rank field on the benchmark collection(s) to use to evaluate a new model. Comparing ranking accuracy of learning to rank methods is largely hindered by this lack of a standard way of benchmarking.

Gomes, Oliveira, Almeida, and Gonçalves (2013) analyzed the ranking accuracy of a set of models on both LETOR 3.0 and 4.0. Busa-Fekete, Kégl, Éltető, and Szarvas (2013) compared the accuracy of a small set of models over the LETOR 4.0 datasets, both MSLR datasets, both the Yahoo! Learning to Rank Challenge datasets and one of the datasets from LETOR 3.0. Both studies did not aim to be complete in benchmark datasets and learning to rank methods included in their comparisons. To our knowledge, no structured meta-analysis on ranking accuracy has been conducted where evaluation results on several benchmark collections are taken into account. In this paper we will perform a meta-analysis with the aim of comparing the ranking accuracy of learning to rank methods. The paper will describe two stages in the meta-analysis process: (1) collection of evaluation results, and (2) comparison of learning to rank methods.

## 2. Collecting evaluation results

We collect evaluation results on the datasets of benchmark collections through a structured literature search. Table 1 presents an overview of the benchmark collections included in the meta-analysis. Note that all these datasets offer feature set representations of the to-be-ranked documents instead of the documents themselves. Therefore, any difference in ranking performance is due to the ranking algorithm and not the features used.

For the LETOR collections, the evaluation results of the baseline models will be used from LETOR 2.0,[4] 3.0[5] and 4.0[6] as listed on the LETOR website.

LETOR 1.0 and 3.0, Yahoo! Learning to Rank Challenge, WCL2R and AOL have accompanying papers that were released with the collection. Authors publishing evaluation results on these benchmark collections are requested to cite these papers. We collect evaluation measurements of learning to rank methods on these benchmark collections through forward literature search. Table 2 presents an overview of the results of this forward literature search performed using Google Scholar.

The LETOR 4.0, MSLR-web10/30k and Yandex Internet Mathematics Competition 2009 benchmark collections are not accompanied by a paper. To collect evaluation results for learning to rank methods on these benchmarks, a Google Scholar search is performed on the name of the benchmark. Table 3 shows the results of this literature search.

### 2.1. Literature selection

Table A.5 in the appendix gives an overview of the learning to rank methods for which evaluation results were found through the described procedure. Occurrences of L2, L3 and L4 in Table A.5 imply that these algorithms are evaluated as official LETOR 2.0, 3.0 and 4.0 baselines respectively.

Some studies with evaluation results found through the literature search procedure were not usable for the meta-analysis. The following enumeration enumerates those properties that made one or more studies unusable for the meta-analysis. The references between brackets are the studies to which these properties apply.

1. A different evaluation methodology was used in the study compared to what was used in other studies using the same benchmark (Geng, Qin, Liu, Cheng, & Li, 2011; Lin, Yeh, & Liu, 2012).
2. The study focuses on a different learning to rank task (e.g. rank aggregation or transfer ranking) (Ah-Pine, 2008; Argentini, 2012; Chen et al., 2010; Dammak, Kammoun, & Ben Hamadou, 2011; De, 2013; De & Diaz, 2011, 2012; De, Diaz, & Raghavan, 2010, 2012; Derhami, Khodadadian, Ghasemzadeh, & Zareh Bidoki, 2013; Desarkar, Joshi, & Sarkar, 2011; Duh & Kirchhoff, 2011; Hoi & Jin, 2008; Lin, Yu, & Chen, 2011; Miao & Tang, 2013; Pan, Lai, Liu, Tang, & Yan, 2013; Qin, Geng, & Liu, 2010; Volkovs & Zemel, 2012, 2013; Wang, Tang et al., 2009).
3. The study used an altered version of a benchmark that contained additional features (Bidoki & Thom, 2009; Ding, Qin, & Zhang, 2010).
4. The study provides no exact data of the evaluation results (e.g. results are only in graphical form) (Adams & Zemel, 2011; Agarwal & Collins, 2010; Alejo, Fernández-Luna, Huete, & Pérez-Vázquez, 2010; Benbouzid, Busa-Fekete, & Kégl, 2012; Chang & Zheng, 2009; Chen, Weinberger, Chapelle, Kedem, & Xu, 2012; Ciaramita, Murdock, & Plachouras, 2008; Geng, Yang, Xu, & Hua, 2012; He, Ma, & Niub, 2010; Huang & Frey, 2008; Karimzadehgan, Li, Zhang, & Mao, 2011; Kuo, Cheng, & Wang, 2009; Li, Wang, Ni, Huang, & Xie, 2008; Ni, Huang, & Xie, 2008; Pan, Luo, Tang, & Huang, 2011; Petterson, Yu, Mcauley, & Caetano, 2009; Qin, Liu, Zhang, Wang, Xiong et al., 2008; Sculley, 2009; Shivaswamy &

[2] http://imat2009.yandex.ru/en.

[3] http://research.microsoft.com/en-us/projects/mslr/.

[4] http://research.microsoft.com/en-us/um/beijing/projects/letor/letor2.0/baseline.aspx.

[5] http://research.microsoft.com/en-us/um/beijing/projects/letor/letor3baseline.aspx.

[6] http://research.microsoft.com/en-us/um/beijing/projects/letor/letor4baseline.aspx.