NIEK TAX

METHODS FOR LARGE SCALE LEARNING-TO-RANK

[18th March 2014 at 20:22 – classicthesis version 0.1]

METHODS FOR LARGE SCALE LEARNING-TO-RANK

A study concerning parallelization of Learning-to-Rank algorithms using Hadoop

NIEK TAX BSC.
Research Chair Databases
Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)
University of Twente

UNIVERSITY OF TWENTE.

Februari 2014 – version 0.1

Ohana means family.
Family means nobody gets left behind, or forgotten.

— Lilo & Stitch

ABSTRACT

Short summary of the contents in English...

SAMENVATTING

Een korte samenvatting in het Nederlands...

God gave you a gift of 86 400 seconds today.
Have you used one to say thank you?

- William Arthur Ward

ACKNOWLEDGMENTS

Many thanks go to Dr. ir. Djoerd Hiemstra of the University of Twente and to Sander Bockting Msc. of Avanade Netherlands BV for their great supervision throughout the project.

In addition I would like to thank all other graduate interns at Avanade as well as all the Avanade employees for the great talks at the coffee machine or elsewhere. In particular I owe many thanks to Michel Barneveld for giving me insight in potential applications of Learning-to-Rank within Avanade.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

ADMM Alternating Direction Method of Multipliers

AP Average Precision

CART Classification and Regression Trees

CC Cooperative Coevolution

CUDA Computing Unified Device Architecture

DCG Discounted Cumulative Gain

DSN Deep Stacking Network

EA Evolutionary Algorithm

ERR Expected Reciprocal Rank

ET Extremely Randomised Trees

FPGA Field-Programmable Gate Array

GA Genetic Algorithm

GBDT Gradient Boosted Decision Tree

GPGPU General-Purpose computing on Graphical Processing Units

GPU Graphical Processing Unit

IDF Inverse Document Frequency

IP Immune Programming

MAP Mean Average Precision

MSE Mean Squared Error

NDCG Normalized Discounted Cumulative Gain

RLS Regularised Least-Squares

SGD Stochastic Gradient Descent

SIMD Single Instruction Multiple Data

SVM Support Vector Machine

TF Term Frequency

TF-IDF Term Frequency - Inverse Document Frequency

TREC Text REtrieval Conference

URL Uniform Resource Locator

1 INTRODUCTION

1.1 MOTIVATION AND PROBLEM STATEMENT

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 for 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. One can find useful applications of ranking in many application domains outside web search as well. For example, it plays a vital role in amongst others: automatic document summarisation, machine translation, drug discovery and in determining the ideal order of maintenance operations [70]. In addition, the ranking task has been argued to be more fitting to recommender systems than regression-based rating prediction on a continuous scale [3, 61].

Research in the field of ranking models has for a long time been based on manually designed ranking functions and has been an active area of research. Luhn [59] was the first to propose a model that assigned relevance scores to documents given a query back in 1957. The increasing amounts of potential training data have recently made it possible to leverage machine learning methods to obtain more effective models. Learning-to-Rank is the relatively new research area that covers the use of machine learning models for the ranking task.

In recent years several Learning-to-Rank benchmark datasets have been proposed that enable comparison of the performance of different Learning-to-Rank methods. Well-known benchmark datasets include the *Yahoo! Learning to Rank Challenge* dataset [20], the Yandex Internet Mathematics competition[1], and the LETOR dataset [66] that was published by Microsoft Research. One of the concluding observations of the *Yahoo! Learning to Rank Challenge* was that almost all work in the Learning-to-Rank field focuses on ranking accuracy, while efficiency and scalability of Learning-to-Rank algorithms is still an underexposed research area that is likely to become more important in the near future as training sets are becoming larger and larger [21]. Liu [56] confirms the observation that efficiency and scalability of Learning-to-Rank methods has so far been an overlooked research area in his influential book on Learning-to-Rank.

1 http://imat-relpred.yandex.ru/en/

Some research has been done in the area of parallel or distributed machine learning [26, 19]. However, almost none of these studies include the Learning-to-Rank sub-field of machine learning. The field of efficient Learning-to-Rank has been getting some attention lately [6, 7, 17, 74, 72], since Liu [56] first stated its growing importance back in 2007. Only several of these studies [74, 72] have explored the possibilities of efficient Learning-to-Rank through the use of parallel programming paradigms.

MapReduce [31] is a parallel programming framework that is inspired by the *Map* and *Reduce* functions commonly used in functional programming. Since Google developed the MapReduce parallel programming framework back in 2004 it has grown to be the industry standard model for parallel programming. Lin [55] observed that algorithms that are of iterative nature, which most Learning-to-Rank algorithms are, are not amenable to the MapReduce framework. Lin argued that as a solution to the non-amenability of iterative algorithms to the MapReduce framework, iterative algorithms can often be replaced with non-iterative alternatives or can still be optimized in such a way that its performance in a MapReduce setting is good enough.

The appearance of benchmark datasets gave insight in the performance of different Learning-to-Rank approaches, which resulted in increasing popularity of those methods that showed to perform well on one or more benchmark datasets. Up to now it remains unknown whether popular existing Learning-to-Rank methods scale well when they are used in a parallel manner using the MapReduce framework. This thesis aims to be an exploratory start in this little researched area of parallel Learning-to-Rank. A more extensive overview of my research goals and questions are described in section 1.2.

1.2 RESEARCH GOALS

The objective of this thesis is to explore the speed-up in execution time of Learning-to-Rank algorithms through parallelisation using the MapReduce framework. This work focuses on those Learning-to-Rank algorithms that have shown leading performance on relevant benchmark datasets. This thesis addresses the following research questions:

RQ1 What are the best performing Learning-to-Rank algorithms in terms of accuracy on relevant benchmark datasets?

RQ2 What is the speed-up of those Learning-to-Rank algorithms when executed using the MapReduce framework?

Where the definition of *relative speed-up* is used for speed-up [75]:

$$S_N = \frac{\text{execution time using one core}}{\text{execution time using } N \text{ cores}}$$

RQ3 Can those Learning-to-Rank algorithms be adjusted in such a way that the parallel execution speed-up increases without decreasing accuracy?

1.3 APPROACH

A literature study will be performed to get insight in relevant existing techniques for large scale Learning-to-Rank. The literature study will be performed by using the following query:

("learning to rank" OR *"learning-to-rank"* OR *"machine learned ranking")* AND *("parallel"* OR *"distributed")*

and the following bibliographic databases:

- Scopus
- Web of Science
- Google Scholar

The query incorporates different ways of writing of Learning-to-Rank, with and without hyphens, and the synonymous term *machine learned ranking* to increase search recall, i.e. to make sure that no relevant studies are missed. For the same reason the terms *parallel* and *distributed* are included in the search query. Even though *parallel* and *distributed* are not always synonymous is all definitions, we are interested in both approaches in non-sequential data processing.

A one-level forward and backward reference search is used to find relevant papers missed so far. To handle the large volume of studies involved in the backward and forward reference search, relevance of the studies will be evaluated solely on the title of the study.

To answer the first research question, I will implement Learning-to-Rank methods in the MapReduce framework and measure runtime as a factor of the number of cluster nodes used to complete the computation.

I will use the HDInsight cloud-based MapReduce implementation from Microsoft to implement the Learning-to-Rank algorithms. HDInsight is based on the popular open source MapReduce implementation Hadoop[2]. The algorithms that we include in the measurements

2 http://hadoop.apache.org/

will be determined based on experimental results on the *Yahoo! Learning to Rank Challenge* [20], the Yandex Internet Mathematics competition[3], the LETOR [66] dataset and the LETOR successors MSLR-WEB10k and MSLR-WEB30k.

1.4 THESIS OVERVIEW

CHAPTER II: BACKGROUND introduces the reader to the basic principles and recent work in the fields of Learning-to-Rank.

CHAPTER III: RELATED WORK concisely describes existing work in the field of parallel Learning-to-Rank.

CHAPTER IV: BENCHMARK RESULTS sketches the accuracy of existing Learning-to-Rank methods on several benchmark datasets and describes the selection of Learning-to-Rank methods for the parallelisation experiments.

CHAPTER V: SELECTED LEARNING-TO-RANK METHODS describes the algorithms and details of the Learning-to-Rank methods selected in Chapter IV.

CHAPTER VI: IMPLEMENTATION describes implementation details of the Learning-to-Rank algorithms in the Hadoop framework.

CHAPTER VII: RESULTS & DISCUSSION presents and discusses speed-up results for the implemented Learning-to-Rank methods.

CHAPTER VIII: CONCLUSION summarizes the results and answers our research questions based on the results. The limitations of our research as well as future research directions in the field are mentioned here.

3 http://imat-relpred.yandex.ru/en/

2 TECHNICAL BACKGROUND

This chapter is written with the goal to provide the reader with a subtle introduction in the Learning-to-Rank field. Models and theories explained in this chapter can be regarded as prior knowledge needed to understand the subsequent chapters of this thesis.

2.1 A BASIC INTRODUCTION TO LEARNING-TO-RANK

Different definitions of Learning-to-Rank exist. In general, all ranking methods that use machine learning technologies to solve the problem of ranking are called Learning-to-Rank methods. Figure 1 describes the general process of machine learning. It shows training elements from an input space that are mapped to an output space using a model such that the difference between the actual labels of the training elements and the labels predicted with with the model are minimal in terms of a loss function.

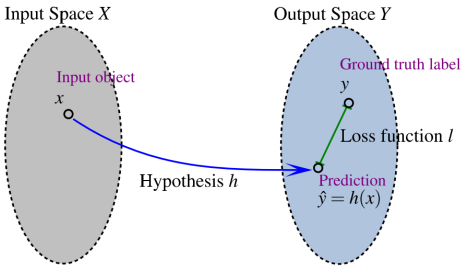

Figure 1: Machine learning framework for Learning-to-Rank, obtained from Liu [56]

Liu [56] proposes a more narrow definition and only considers ranking methods to be a Learning-to-Rank method when it is *feature based* and uses *discriminative training*, which are itself defined as follows:

FEATURE BASED means that all documents under investigation are represented by feature vectors. Those feature vectors reflect the relevance of the documents to the query, or the importance of the document in itself.

DISCRIMINATIVE TRAINING means that the learning process can be well described by the four components of discriminative learning. That is, a Learning-to-Rank method has its own *input space*, *output space*, *hypothesis space*, and *loss function*, like the machine learning process described by Figure 1. *Input space*, *out-*

put space, *hypothesis space*, and *loss function* are hereby defined as follows:

INPUT SPACE contains the objects under investigation. Usually objects are represented by feature vectors, extracted according to different applications.

OUTPUT SPACE contains the learning target with respect to the input objects.

HYPOTHESIS SPACE defines the class of functions mapping the input space to the output space. The functions operate on the feature vectors of the input object, and make predictions according to the format of the output space.

LOSS FUNCTION in order to learn the optimal hypothesis, a training set is usually used, which contains a number of objects and their ground truth labels, sampled from the product of the input and output spaces.

Figure 2 shows how the machine learning process as described in Figure 1 typically takes place in a ranking scenario. A set of queries q_i with $n > i > 1$, the documents associated with these queries which are represented by feature vector x_i, and the relevant judgements of those documents y_i are used together to train a model h. Model h can after training be used to predict a ranking of the documents y_i, such the difference between the document rankings predicted by h and the actual optimal rankings based on y_i is minimal in terms of a loss function.

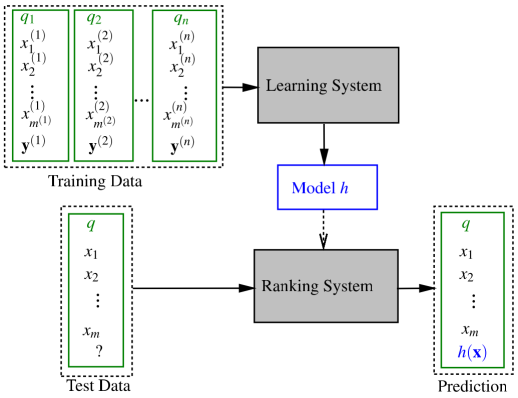

Figure 2: A typical Learning-to-Rank setting, obtained from Liu [56]

The predictions and the loss function might either be defined for:

1. the relevance of a single document
2. the classification of the most relevant document out of a document-pair
3. the ranking of documents directly

These three approaches are in literature respectively called the pointwise approach, the pairwise approach and the listwise approach. These three approaches to Learning-to-Rank will be described in more detail in section 2.3.

2.2 HOW TO EVALUATE A RANKING

Evaluation metrics have long been studied in the field of information retrieval. First in the form of evaluation of unranked retrieval sets and later, when the information retrieval field started focussing more on ranked retrieval, in the form of ranked retrieval evaluation. In this section several frequently used evaluation metrics for ranked results will be described.

No single evaluation metric that we are going to describe is indisputably better or worse than any of the other metrics. Different benchmarking settings have used different evaluation metrics. Metrics introduced in this section will be used to express the evaluation results in chapter 4 of this thesis.

2.2.1 *Normalized Discounted Cumulative Gain*

Cumulative gain, or its predecessor discounted cumulative gain and normalized discounted cumulative gain, is one of the most widely used measures for effectiveness of ranking methods.

2.2.1.1 *Discounted Cumulative Gain*

There are two definitions of Discounted Cumulative Gain (DCG) used in practice. DCG at position p was originally defined by Järvelin and Kekäläinen [49] as

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i - 1}{\log_2(i+1)}$$

with rel_i the graded relevance of the result at position i. The idea is that highly relevant documents that appear lower in a search result should be penalized (discounted). This discounting is done by reducing the graded relevance logarithmically proportional to the position of the result.

Burges et al. [13] proposed an alternative definition of DCG that puts stronger emphasis on document relevance

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

2.2.1.2 *Normalized Discounted Cumulative Gain*

nDCG normalizes the DCG metric to a value in the [0,1] interval by dividing by the DCG value of the optimal rank. This optimal rank is obtained by sorting documents on relevance for a given query. The definition of nDCG can be written down mathematically as

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

Table 1 shows an example calculation for nDCG for both the Järvelin and Kekäläinen [49] and Burges et al. [13] version of DCG.

	Rank										
	1	2	3	4	5	6	7	8	9	10	Sum
rel(i)	10	7	6	8	9	5	1	3	2	4	
$\frac{2^{rel_i}-1}{\log_2(i+1)}$	512	40.4	16	55.1	99.0	5.7	0.3	1.3	0.6	2.3	732.7
$\frac{rel_i}{\log_2(i+1)}$	10	4.42	3	3.45	3.48	1.78	0.33	0.95	0.6	1.16	29.17
optimal rank	10	9	8	7	6	5	4	3	2	1	
$\frac{2^{rel_i}-1}{\log_2(i+1)}$	512	161.5	64	27.6	12.4	5.7	2.7	1.3	0.6	0.2	788.0
$\frac{rel_i}{\log_2(i+1)}$	10	5.68	4	3.01	2.32	1.78	1.33	0.95	0.6	0.29	29.96

Järvelin and Kekäläinen [49] nDCG = $\frac{29.17}{29.96} = 0.97$

Burges [13] et al. nDCG = $\frac{732.7}{788.0} = 0.93$

Table 1: Example calculation for nDCG

2.2.2 *Expected Reciprocal Rank*

Expected Reciprocal Rank (ERR) [22] was designed based on the observation that nDCG is based on the false assumption that the usefulness of a document at rank i is independent of the usefulness of the documents at rank less than i. ERR is based on the reasoning that users are likely to stop exploring the result list once they have found a document that satisfied their information need. The ERR metric is defined as the expected reciprocal length of time that the user will take to find a relevant document. ERR is formally defined as

$$ERR = \sum_{r=1}^{n} \frac{1}{r} \prod_{i=1}^{r-1} (1 - R_i) R_i$$

where the product sequence part of the formula represents the chance that the user will stop at position r.
An algorithm to compute ERR is shown in Listing 1. The algorithm

	Rank										Sum
	1	2	3	4	5	6	7	8	9	10	
r_i	1	0	0	0	1	1	0	1	0	0	
P@i	1				0.4	0.5		0.5			2.4
									/#Relevant docs	=	7
									AP@10	=	0.34

Table 2: Average Precision example calculation. The number of relevant documents R is assumed to be seven.

requires relevance grades g_i, $1 \leqslant i \leqslant n$ and mapping function R that maps relevance grades to probability of relevance.

Listing 1: Algorithm to compute the ERR metric, obtained from [22]

```
p <- 1, ERR <- 0
for r=1 to n do
        R <- R(g[r])
        ERR <- ERR + p * R/r
        p <- p * (1-R)
end for
return ERR
```

2.2.3 *Mean Average Precision*

Average Precision (AP) [103] is an often used binary relevance judgement based metric that can be seen as a trade-off between precision and recall that is defined as

$$AP = \frac{\sum_{k=1}^{n} Precision(k) * relevance(k)}{\text{number of relevant docs}}$$

With k being the positions in the result set between 1 and n. Table 2 provides an example calculation of average precision where de documents at positions 1, 5, 6 and 8 in the ranking are relevant. MAP is the average AP for a set of queries.

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}$$

2.3 APPROACHES TO LEARNING-TO-RANK

2.3.1 *Pointwise Approach*

The pointwise approach can be seen as the most straightforward way of using machine learning for ranking. Pointwise Learning-to-Rank

methods directly apply machine learning methods to the ranking problem by observing each document in isolation. They can be subdivided in the following approaches:

1. regression-based, which estimate the relevance of a considered document using a regression model.
2. classification-based, which classify the relevance category of the document using a classification model.
3. ordinal regression-based, which classify the relevance category of the document using a classification model in such a way that the order of relevance categories is taken into account.

Well-known algorithms that belong to the pointwise approach include McRank [54] and PRank [28].

2.3.2 *Pairwise Approach*

Pointwise Learning-to-Rank methods have the drawback that they optimise real valued expected relevance, while evaluation metrics like nDCG and ERR are only impacted by a change in expected relevance when that change impacts a pairwise preference. The pairwise approach solves this drawback of the pointwise approach by regarding ranking as pairwise classification.

Aggregating a set of predicted pairwise preferences into the corresponding optimal rank is shown to be a NP-Hard problem [34]. An often used solution to this problem is to upper bound the number of classification mistakes by an easy to optimise function [8].

Well-known pairwise Learning-to-Rank algorithms include FRank [79], GBRank [102], LambdaRank [15], RankBoost [35], RankNet [13], Ranking SVM [45, 50], and SortNet [68].

2.3.3 *Listwise Approach*

Listwise ranking optimises the actual evaluation metric. The learner learns to predict an actual ranking itself without using an intermediate step like in pointwise or pairwise Learning-to-Rank. The main challenge in this approach is that most evaluation metrics are not differentiable. MAP, ERR and nDCG are non-differentiable, non-convex and discontinuous functions, what makes them very hard to optimize.

Although the properties of MAP, ERR and nDCG are not ideal for direct optimisation, some listwise approaches do focus on direct metric optimisation [99, 78, 23]. Most listwise approaches work around

optimisation of the non-differentiable, non-convex and discontinuous metrics by optimising surrogate cost functions that mimic the behaviour of MAP, ERR or nDCG, but have nicer properties for optimisation.

Well-known algorithms that belong to the listwise approach include AdaRank [90], BoltzRank [82], ListMLE [89], ListNet [18], RankCosine [67], SmoothRank [23], SoftRank [78], SVM^{map} [99].

3

RELATED WORK

3.1 LITERATURE STUDY CHARACTERISTICS

The literature research is performed by using the bibliographic databases Scopus and Web of Science with the following search query:

("learning to rank" OR *"learning-to-rank"* OR *"machine learned ranking")* AND *("parallel"* OR *"distributed")*

An abstract-based manual filtering step is applied where those results are filtered that use the terms *parallel* or *distributed* in context to *learning to rank*, *learning-to-rank* or *machine learned ranking*. As a last step I will filter out studies based on the whole document that only focus on efficient query evaluation and not on parallel or distributed learning of ranking functions, as those studies are likely to meet listed search terms.

On Scopus, the defined search query resulted in 65 documents. Only 14 of those documents used *large scale*, *parallel* or *distributed* terms in context to the *learning to rank*, *learning-to-rank* or *machine learned ranking*. 10 out of those 14 documents focussed on parallel or distributed learning of ranking functions.

The defined search query resulted in 16 documents on Web of Science. Four of those documents were part of the 10 relevant documents found using Scopus, leaving 12 new potentially relevant documents to consider. Four of those 12 documents used *large scale*, *parallel* or *distributed* terms in context to the *learning to rank*, *learning-to-rank* or *machine learned ranking*, none of them focused on parallel or distributed learning of ranking functions.

On Google Scholar, the defined search query resulted in 3300 documents. Because it infeasible to evaluate all 3300 studies we focus on the first 300 search results.

Backward reference search resulted in 10 studies regarded as potentially relevant based on the title, of which four were actually relevant and included in the literature description. Forward reference search resulted in 10 potentially relevant titles, of which seven studies turned out to be relevant.

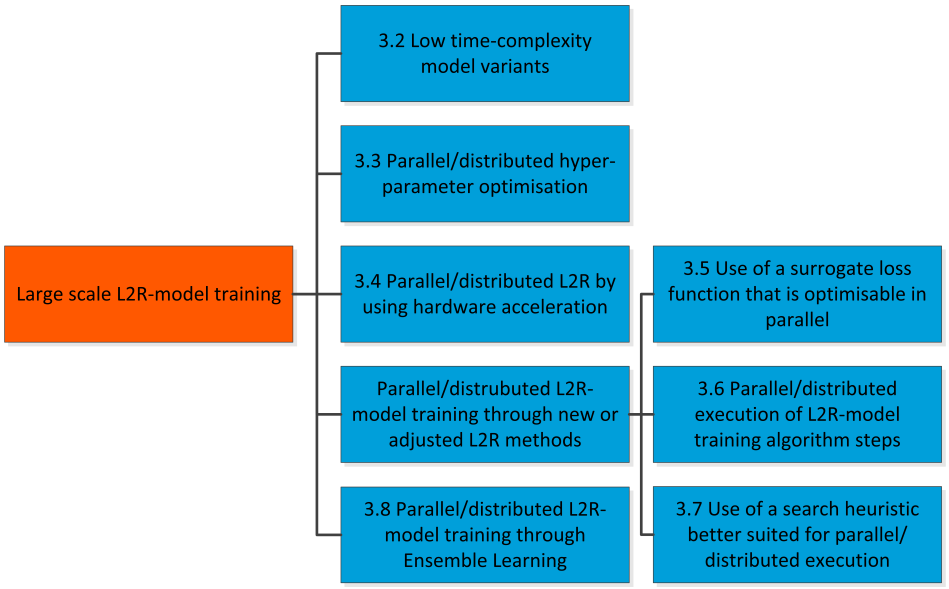

Figure 3: Categories in research on large scale training of Learning-to-Rank models

Research in scaling up the training phase of Learning-to-Rank models can be categorised according to the approach in scaling up. Figure 3 illustrates the categories of scalable training approaches in Learning-to-Rank. The numbers in Figure 3 correspond to the sections that describe the related work belonging to these categories.

3.2 LOW COMPUTATIONAL COMPLEXITY LEARNING-TO-RANK

One approach for handling large volumes of training data for Learning-to-Rank is through design of low time complexity Learning-to-Rank methods. Pahikkala et al. [63] described a pairwise Regularised Least-Squares (RLS) type of ranking function, RankRLS, with low time complexity. Airola et al. [4] further improved the training time complexity of RankRLS to $\mathcal{O}(tms)$, where t is the number of needed iterations, m the number of training documents and s the number of features. The RankRLS ranking function showed ranking performance similar to RankSVM [46, 50] on the BioInfer corpus [65], a corpus for information extraction in the biomedical domain.

Airola et al. [5] and Lee and Lin [53] both described lower time complexity methods to train a linear kernel ranking Support Vector Machine (SVM) [46, 50]. Lee and Lin [53] observed that linear kernel RankSVMs are inferior in accuracy compared to nonlinear kernel RankSVMs and Gradient Boosted Decision Tree (GBDT)s and are mainly useful to quickly produce a baseline model. Details of the lower time complexity version of the linear kernel RankSVM will not be discussed as it is shown to be an inferior Learning-to-Rank method

in terms of accuracy.

Learning-to-Rank methods that are specifically designed for their low computational complexity, like RankRLS and the linear kernel RankSVM methods described in this section, are generally not among the top achieving models in terms of accuracy. From results on benchmarks and competitions it can be observed that the best generalisation accuracy are often more complex ones. This makes low time complexity models as a solution for large scale Learning-to-Rank less applicable and increases the relevance of the search for efficient training of more complex Learning-to-Rank models.

3.3 DISTRIBUTED HYPERPARAMETER TUNING OF LEARNING-TO-RANK MODELS

Hyperparameter optimisation is the task of selecting the combination of hyperparameters such that the Learning-to-Rank model shows optimal generalisation accuracy. Ganjisaffar et al. [41, 39] observed that long training times are often a result of hyperparameter optimisation, because it results in training multiple Learning-to-Rank models. Grid search is the *de facto* standard of hyperparameter optimisation and is simply an exhaustive search through a manually specified subset of hyperparameter combinations. The authors show how to perform parallel grid search on MapReduce clusters, which is easy because grid search is an embarrassingly parallel method as hyperparameter combinations are mutually independent. They apply their grid search on MapReduce approach in a Learning-to-Rank setting to train a LambdaMART [88] ranking model, which uses the Gradient Boosting [38] ensemble method combined with regression tree weak learners. Experiments showed that the solution scales linearly in number of hyperparameter combinations. However, the risk of overfitting grows overwhelmingly as the number of hyperparameter combinations grow, even when validation sets grow large.

Burges et al. [16] described their Yahoo! Learning to Rank Challenge submission that was built by performing an extensive hyperparameter search on a 122-node MPI cluster, running Microsoft HPC Server 2008. The hyperparameter optimisation was performed on a linear combination ensemble of eight LambdaMART models, two LambdaRank models and two MART models using a logistic regression cost. This submission achieved the highest ERR score of all Yahoo! Learning to Rank Challenge submissions.

Notice that methods described in this section train multiple Learning-to-Rank models at the same time to find the optimal set of paramet-

ers for a model, but that the Learning-to-Rank models itself are still trained sequentially. In the next sections we will present literature focusing on training Learning-to-Rank models in such a way that steps in this training process can be executed simultaneously.

3.4 HARDWARE ACCELERATED LEARNING-TO-RANK

Hardware accelerators are special purpose processors designed to speed up compute-intensive tasks. An Field-Programmable Gate Array (FPGA) and a Graphical Processing Unit (GPU) are two different types of hardware that can achieve better performance on some tasks though parallel computing. In general, FPGAs provide better performance while GPUs tend to be easier to program [24]. Some research has been done in parallelising Learning-to-Rank methods using hardware accelerators.

3.4.1 *FPGA-based parallel Learning-to-Rank*

Yan et al. [93, 94, 95, 96] described the development and incremental improvement of a Single Instruction Multiple Data (SIMD) architecture FPGA designed to run, the Neural Network based LambdaRank Learning-to-Rank algorithm. This architecture achieved a 29.3X speed-up compared to the software implementation, when evaluated on data from a commercial search engine. The exploration of FPGA for Learning-to-Rank showed additional benefits other than the speed-up originally aimed for. In their latest publication [96] the FPGA based LambdaRank implementation showed it could achieve up to 19.52X power efficiency and 7.17X price efficiency for query processing compared to Intel Xeon servers currently used at the commercial search engine.

Xu et al. [91, 92] designed an FPGA-based accelerator to reduce the training time of the RankBoost algorithm [35], a pairwise ranking function based on Freund and Schapire's AdaBoost ensemble learning method [36]. Xu et al. [92] state that RankBoost is a Learning-to-Rank method that is not widely used in practice because of its long training time. Experiments on MSN search engine data showed the implementation on a FPGA with SIMD architecture to be 170.6x faster than the original software implementation [91]. In a second experiment in which a much more powerful FPGA accelerator board was used, the speed-up even increased to 1800x compared to the original software implementation [92].

3.4.2 *GPGPU for parallel Learning-to-Rank*

Wang et al. [83] experimented with a General-Purpose computing on Graphical Processing Units (GPGPU) approach for parallelising RankBoost. Nvidia Computing Unified Device Architecture (CUDA) and ATI Stream are the two main GPGPU computing platform and are released by the two main GPU vendors Nvidia and AMD. Experiments show a 22.9x speed-up on Nvidia CUDA and a 9.2x speed-up on ATI Stream.

De Sousa et al. [30] proposed a GPGPU approach to improve both training time and query evaluation through GPU use. An association rule based Learning-to-Rank approach, proposed by Veloso et al. [81], has been implemented using the GPU in such a way that the set of rules van be computed simultaneously for each document. A speed-up of 127X in query processing time is reported based on evaluation on the LETOR dataset. The speed-up achieved at learning the ranking function was unfortunately not stated.

3.5 PARALLEL EXECUTION OF LEARNING-TO-RANK ALGORITHM STEPS

Some research focused on parallelising the steps Learning-to-Rank algorithms that can be characterised as strong learners.

Tyree et al. [80] described a way of parallelising GBDT models for Learning-to-Rank where the boosting step is still executed sequentially, but instead the construction of the regression trees themselves is parallelised. The parallel decision tree building is based on Ben-Haim and Yom-Tov's work on parallel construction of decision trees for classification [9], which are build layer-by-layer. The calculations needed for building each new layer in the tree are divided among the nodes, using a master-worker paradigm. The data is partitioned and the data parts are divided between the workers, who compress their share into histograms and send these to the master. The master uses those histograms to approximate the split and build the next layer. The master then communicates this new layer to the workers who can use this new layer to compute new histograms. This process is repeated until the tree depth limit is reached. The tree construction algorithm parallelised with this master-worker approach is the well-known Classification and Regression Trees (CART) [11] algorithm. Speed-up experiments on the LETOR and the Yahoo! Learning to Rank challenge data sets were performed. This parallel CART-tree building approach showed speed-up of up to 42x on shared memory machines and up to 25x on distributed memory machines.

3.5.1 *Parallel ListNet using Spark*

Shukla et al. [72] explored the parallelisation of the well-known ListNet Learning-to-Rank method using Spark. Spark is a parallel computing model that is designed for cyclic data flows which makes it more suitable for iterative algorithms. Spark is incorporated into Hadoop since Hadoop 2.0. The Spark implementation of ListNet showed near linear training time reduction.

3.6 PARALLELISABLE SEARCH HEURISTICS FOR LISTWISE RANKING

Direct minimisation of ranking metrics is a hard problem due to the non-continuous, non-differentiable and non-convex nature of the nDCG, ERR and MAP evaluation metrics. This optimisation problem is generally addressed either by replacing the ranking metric with a convex surrogate, or by heuristic optimisation methods such as Simulated Annealing or a Evolutionary Algorithm (EA). One EA heuristic optimisation method that is successfully used in direct rank evaluation functions optimisation is the Genetic Algorithm (GA) [98]. GAs are search heuristic functions that mimic the process of natural selection, consisting of mutation and cross-over steps [47]. The following subsection describe related work that uses search heuristics for parallel/distributed training.

3.6.1 *Immune Programming*

Wang et al. [86] proposed a Immune Programming (IP) solution to direct ranking metric optimisation and observed that all EAs are generally easy to implement such that it can be carried out on different computers distributively. However, no statements on the possible speed-up of a distributed implementation of the IP solution has been made and no speed-up experiments have been conducted.

3.6.2 *CCRank*

Wang et al. [84, 85] proposed a parallel evolutionary algorithm based on Cooperative Coevolution (CC), a type of evolutionary algorithm. The CC algorithm is capable of directly optimizing non-differentiable functions, as nDCG, in contrary to many optimization algorithms. the divide-and-conquer nature of the CC algorithm enables parallelisation. CCRank showed an increase in both accuracy and efficiency on the LETOR 4.0 benchmark dataset compared to its baselines. However, the increased efficiency was achieved through speed-up and not scale-up. Two reasons have been identified for not achieving linear

scale-up with CCRank: 1) parallel execution is suspended after each generation to perform combination in order to produce the candidate solution, 2) Combination has to wait until all parallel tasks have finished, which may spend different running time.

3.6.3 *nDCG-Annealing*

Karimzadeghan et al. [51] proposed a method using Simulated Annealing along with the Simplex method for its parameter search. This method directly optimises the often non-differentiable Learning-to-Rank evaluation metrics like nDCG and MAP. The authors successfully parallelised their method in the MapReduce paradigm using Hadoop. The approach showed to be effective on both the LETOR 3.0 dataset and their own dataset with contextual advertising data. Unfortunately their work does not directly report on the speed-up obtained by parallelising with Hadoop, but it is mentioned that further work needs to be done to effectively leverage parallel execution.

3.7 PARALELLY OPTIMISABLE SURROGATE LOSS FUNCTIONS

3.7.1 *Alternating Direction Method of Multipliers*

Duh et al. [33] proposed the use of Alternating Direction Method of Multipliers (ADMM) for the Learning-to-Rank task. ADMM is a general optimization method that solves problems of the form

$$\begin{aligned} &\text{minimize } f(x) + g(x) \\ &\text{subject to } Ax + Bz = c \end{aligned}$$

by updating x and z in an alternating fashion. It holds the nice properties that it can be executed in parallel and that it allows for incremental model updates without full retraining. Duh et al. [33] showed how to use ADMM to train a RankSVM [46, 50] model in parallel. Experiments showed the ADMM implementation to achieve a 13.1x training time speed-up on 72 workers, compared to training on a single worker.

Another ADMM Learning-to-Rank based approach was proposed by Boyd et al. [10]. They implemented an ADMM based Learning-to-Rank method in Pregel [60], a parallel computation framework for graph computations. No experimental results on parallelisation speed-up were reported on this Pregel based approach.

3.7.2 *Bregman Divergences and Monotone Retargeting*

Acharyya et al. [2, 1] proposed a Learning-to-Rank method searches for an order preserving transformation (*monotone retargeting*) of the target scores that is easier for a regressor to fit. This approach is based on the observation that it is not necessary to fit scores exactly, since the evaluation is dependent on the order and not on the pointwise predictions themselves.

Bregman divergences are a family of distance-like functions that do not satisfy symmetry nor the triangle inequality. A well-known member of the class of Bregman divergences is Kullback-Leibler divergence, also known as *information gain*.

Acharyya et al. [2, 1] defined a parallel algorithm that optimises a Bregman divergence function as surrogate of nDCG that is claimed to be well suited for implementation of a GPGPU. No experiments on speed-up have been performed.

3.7.3 *Parallel robust Learning-to-Rank*

Robust learning [48] is defined as the task to lean a model in the presence of outliers. Yun et al. described a [100] robust Learning-to-Rank model called RoBiRank that has the additional advantage that it is executable in parallel. RoBiRank uses parameter expansion to linearise a surrogate loss function, after which the elements of the linear combination can be divided over the available nodes. The Parameter expansion trick was proposed by Gopal and Yang [44], who originally proposed it for multinomial logistic regression. Unfortunately, no speed-up experiments were mentioned for the RoBiRank method, since Yun et al. focussed their research more on robust ranking than on parallel ranking. The only reference to performance of RoBiRank in terms of speed is the statement that its training time on a computing cluster is comparable to the more efficient implementation by Lee and Lin [53] of a linear kernel RankSVM [46, 50] described in section 3.2.

3.7.4 *Distributed Stochastic Gradient Descent*

Long et al. [58] described special case of their pairwise cross-domain factor Learning-to-Rank method using distributed optimization of Stochastic Gradient Descent (SGD) based on Hadoop MapReduce. Parallelisation of the SGD optimization algorithm was performed using the MapReduce based method described by Zinkevich et al. [104] has been used. Real world data from Yahoo! has been used to show that the model is effective. Unfortunately the speed-up obtained by train-

ing their model in parallel is not reported.

3.8 ENSEMBLE LEARNING FOR PARALLEL LEARNING-TO-RANK

Schapire proved in 1990 that a strong model can be generated by combining weak models through a procedure called boosting [71]. The invention of the boosting method resulted in an increasing focus in machine learning research on methods to combine multiple weak models, which are called ensemble learning methods. Well-known ensemble methods include gradient boosting [37], bagging [12], AdaBoost [36] and stacked generalisation (often called stacking) [87].

Ensemble methods can be used to parallelise learning methods by training the weak models in the ensemble on different nodes. Parallelisation of the training phase of Learning-to-Rank models is often achieved through ensemble learning, mainly boosting, and combined with decision trees, jointly called Gradient Boosted Decision Tree (GBDT). GBDTs are shown to be able to achieve good accuracy in a Learning-to-Rank setting when used in a pairwise [102] or listwise [25] setting.

3.8.1 *Gradient Boosting*

Gradient Boosting [37] is an ensemble learning method in which multiple weak learners are iteratively added together into one ensemble model in such a way that new models focus more on those data instances that were misclassified before.

Kocsis et al. [52] proposed a way to train multiple weak learners in parallel by extending those models that are likely to yield a good model when combined through boosting. Authors showed through theoretical analysis that the proposed algorithm asymptotically achieve equal performance to regular gradient boosting. GBDT models could be trained in parallel using their BoostingTree algorithm.

Ye et al. [97] described how to implement the stochastic GBDT model in a parallel manner using both MPI and Hadoop. Stochastic GBDT differs from regular GBDT models by using stochastic gradient boosting instead of regular gradient boosting. Stochastic gradient boosting is a minor alteration to regular gradient boosting that, at each iteration of the algorithm, trains a base learner on a sub-sample of the training set that is drawn at random without replacement [38]. Experiments showed the Hadoop implementation to resulted in too expensive communication cost to be useful. Authors believed that these high communication costs were a result of the communication intensive im-

plementation that was not well suited for the MapReduce paradigm. The MPI approach proved to be successful and obtained near linear speed-ups.

Svore and Burges [77, 76] designed two approaches to train the LambdaMART [88] model in a distributed manner. Their first approach is applicable in case the whole training set fits in main memory on every node, in that case the tree split computations of the trees in the LambdaMART tree-ensemble are split instead of the data. The second approach distributes the training data and corresponding training computations and is therefore scalable to high amounts of training data. Both approaches achieves a six times speed-up over LambdaMART on 32 nodes compared to a single node.

3.8.2 *Boosting wrapped in Bagging*

Some approaches combine both boosting and bagging. Bagging [12], also called bootstrap aggregating, is a ensemble learning method in which m training sets $D_1..D_m$ are constructed from the training set D by uniformly sampling data items from D with replacement. The bagging method is parallelisable by training each D_i with $i \in \{1..m\}$ on a different node.

Yandex researchers Pavlov et al. [64] were the first to proposed a boosting-wrapped-in-bagging approach, which they called BagBoo. The boosting step itself is not parallelisable, but the authors state that learning a short boosted sequence on a single node is still a do-able task.

Ganjisaffar et al. [40, 39] proposed a pairwise boosting-wrapped-in-bagging model called BL-MART, contrasting the pointwise BagBoo model. BL-MART adds a bagging step to the LambdaMART [88] ranking model, that uses the Gradient Boosting [38] ensemble method combined with regression tree weak learners. In contract to BagBoo, BL-MART is limited in the number of trees. An experiment on the TD2004 dataset resulted in 4500 trees using BL-MART while 1.1 million trees were created with the BagBoo model.

3.8.3 *Stacked generalisation*

A Deep Stacking Network (DSN) is a processing architecture developed from the field of *Deep Learning*, that uses the Mean Squared Error (MSE) loss function that is easily optimisable. DSN is based on the stacked generalization (stacking) ensemble learning method [87], an approach where several learning models are stacked on top of each other in such a way that the output of a model is used as one of the in-

put features of models higher up in the stack. Each layer in the stack learns has the task of learning the weights of the inputs of that layer. The close-form constraints between input and output weights allow the input weight matrices to be estimated in a parallel manner. Deng et al. [32]

3.8.4 *Randomisation*

Most ensemble learning based Learning-to-Rank methods are based on boosting. Although boosting have shown to be an effective method in Learning-to-Rank, it has the drawback that the models in the ensemble are not completely mutually independent, which makes the process harder to parallelise. Geurts and Louppe [43] proposed a tree based ensemble method in which the trees are built using random subsets of the available features. The authors called this method Extremely Randomised Trees (ET) and originally proposed this method for classification and regression settings instead of ranking settings [42]. The ET algorithm is similar to the well-known Random Forest algorithm, but with two differences: 1) in ET each tree is build on the complete dataset instead of random subsets of the data and 2) ET sets the discretisation thresholds that define the splits in the tree randomly, instead of based on the sample. The randomisations in the process make the models in the ensemble mutually independent, which makes it trivial to parallelise by training each tree on a different node. In contrary to what one might think, ETs show very good performance on benchmark data with a 10th place in the Yahoo! Learning to Rank Challenge.

4 BENCHMARK RESULTS

This chapter describes benchmark characteristics like collection size and features and gives an overview of the performance of different Learning-to-Rank methods. Performance differences for a given Learning-to-Rank method are explained in terms of differences in benchmark characteristics. This chapter is concluded with a selection of Learning-to-Rank methods to include in the experiments.

Benchmarking datasets enable fair comparisons of ranking models over a fixed set of documents and queries. The approach of using fixed benchmarking datasets to compare the performance of Information Retrieval systems under equal circumstances has been the standard evaluation methodology in the Information Retrieval field since the release of the Cranfield collection [27].

4.1 YAHOO! LEARNING TO RANK CHALLENGE

Yahoo's observation that all existing benchmark datasets were too small to draw reliable conclusions prompted Yahoo to release two internal datasets from Yahoo! search. Datasets used at commercial search engines are many times larger than available benchmark datasets. Yahoo! published a subset of their own commercial training data and launched a Learning-to-Rank competition based on this data. The Yahoo! Learning to Rank Challenge [20] is a public Learning-to-Rank competition which took place from March to May 2010, with the goal to promote the datasets and encourage the research community to develop new Learning-to-Rank algorithms.

The Yahoo! Learning to Rank Challenge consists of two tracks that uses the two datasets respectively: a standard Learning-to-Rank track and a transfer learning track where the goal was to learn a specialized ranking function that can be used for a small country by leveraging a larger training set of another country. For this experiment, I will only look at the standard Learning-to-Rank dataset, because transfer learning is a separate research area that is not included in this thesis.

Both nDCG and ERR are measured as performance metrics, but the final standings of the challenge were based on the ERR values. Model validation on the Learning-to-Rank methods participating in the challenge is performed using a train/validation/test-set split following the characteristics shown in Table 3. Competitors could train on the

	Train	Validation	Test
# of queries	19,994	2,994	6,983
# of documents	473,134	71,083	165,660
# of features	519	519	519

Table 3: Yahoo! Learning to Rank Challenge dataset characteristics, as described in the challenge overview paper [20]

training set and get immediate feedback on their performance on the validation set. The test set performance is used to create the final standings and is only measured after the competition has ended to avoid overfitting on the test set. The large number of documents, queries and features compared to other benchmark datasets makes the Yahoo! Learning to Rank Challenge dataset interesting. Yahoo did not provide detailed feature descriptions to prevent competitors to get detailed insight in the characteristics of the Yahoo data collection and features used at Yahoo. Instead high level descriptions of feature categories are provided. The following categories of features are described in the challenge overview paper [20]:

WEB GRAPH Quality and popularity metrics of web documents, e.g. PageRank [62].

DOCUMENT STATISTICS Basic document statistics such as the number of words and url characteristics.

DOCUMENT CLASSIFIER Results of various classifiers on the documents. These classifiers amongst others include: spam, adult, language, main topic, and quality classifiers.

QUERY Basic query statistics, such as the number of terms, query frequency, and click-through rate.

TEXT MATCH Textual similarity metrics between query and document. Includes Term Frequency - Inverse Document Frequency (TF-IDF), BM25 [69] and other metrics for different sections of the document.

TOPICAL MATCHING These features go beyond similarity at word level and compute similarity on topic level. For example by classifying both the document and the query in a large topical taxonomy.

CLICK Click-based user feedback.

EXTERNAL REFERENCES Document meta-information such as Delicious[1] tags

1 https://delicious.com/

	Authors	ERR
1	Burges et al. (Microsoft Research)	0.46861
2	Gottschalk (Activision Blizzard) & Vogel (Data Mining Solutions)	0.46786
3	Parakhin (Microsoft)	0.46695
4	Pavlov & Brunk (Yandex Labs)	0.46678
5	Sorokina (Yandex Labs)	0.46616

Table 4: Final standings of the Yahoo! Learning to Rank Challenge, as presented in the challenge overview paper [20]

TIME Document age and historical in- and outlink data that might help for time sensitive queries.

4.1.1 *Results*

1055 teams send in at least one submission to the Yahoo! Learning to Rank challenge. The top eight participants of the Yahoo! Learning to Rank challenge all used decision trees combined with ensemble methods. The mainly used ensemble method within these top performers is boosting. The combination of boosting with decision tree learners is often called Gradient Boosted Decision Tree (GBDT). Figure 4 shows the top five participants in the Yahoo! Learning to Rank Challenge in terms of ERR score. Burges [16] created a linear combination ensemble of eight LambdaMART [14], two LambdaRank and two Logistic Regression models. Gottschalk and Vogel used a combination of RandomForest models and GBDT models. Pavlov and Brunk used a regression based model using the BagBoo [64] ensemble technique, which combines bagging and boosting. Sorokina used a similar combination of bagging and boosting that is called Additive Groves [73].

The challenge overview paper [20] states as one of the lessons of the challenge that the simple baseline GBDT model performed very well with a small performance gap to the complex ensemble submissions at the top of the table.

Although the winning GBDT models performs very well in terms of nDCG their high complexity makes them unsuitable to use in production. The winning models take weeks to train and are very slow during query evaluation. An exception in training time is the BagBoo [64] method used by Pavlov & Brunk. The bagging component of the BagBoo method enables it to achieve high scalability through parallelism. Pavlov et al. [64] managed to train half a million trees on 200 nodes in 2 hours.

4.2 YANDEX INTERNET MATHEMATICS COMPETITION

4.2.1 *Results*

The Yandex Internet Mathematics competition was won by Pavlov et al. [64] with the BagBoo method.

4.3 LETOR

The LETOR benchmark set was first released by Microsoft Research Asia in April 2007 [57] to solve the absence of a experimental platform for Learning-to-Rank at that time. LETOR has been updated several times since: LETOR 2.0 was released at the end of 2007, LETOR 3.0 in December 2008 and LETOR 4.0 in July 2009.

4.3.1 *LETOR 3.0*

The LETOR 3.0 benchmark collection [66] as released in 2007 contained two data sets: the OHSUMED collection and the .gov collection. The OHSUMED collection is a subset of the medical publication database MEDLINE and contains medical publications from 270 journals that were published between 1987 and 1991. The .gov collection is a web crawl obtained in January 2002, which was used for the Text REtrieval Conference (TREC) web track in 2003 and 2004. Tables 10 and 9 in Appendix A provide the descriptions of the features of the .gov and the OHSUMED collections respectively.

Different query sets exists for the .gov corpus. Those query sets are categorized in the following categories [29]:

TOPIC DISTILLATION (TD) these queries involve finding relevant pages given a broad query. E.g., the query 'cotton industry' which is entered with the aim to find pages that provide information about the cotton industry.

NAMED PAGE FINDING (NP) in these queries the user asks for one specific page by name. E.g., the user queries for 'ADA Enforcement' to find the page http://www.usdoj.gov/crt/ada/enforce.htm.

HOME PAGE FINDING (HP) these queries are similar to NP queries in that the user is looking for one specific page, but now this specific page is a homepage. E.g., the user queries for 'Tennessee Valley Authority' to find the homepage http://www.tva.gov.

With query sets available from both the TREC 2003 and 2004 conferences there are six query sets in total: TD2003, TD2004, NP2003,

NP2004, HP2003 and HP2004.

The evaluation metrics used are nDCG and MAP. The *winning number* metric is defined as the number of other algorithms that it can beat over all of the seven datasets (six .gov sets + OHSUMED). The LETOR organisation implemented and evaluated several well-known Learning-to-Rank algorithms themselves and in addition gathers and publications and results of new algorithms evaluated on the LETOR benchmark. The baseline Learning-to-Rank algorithms evaluated by the organisation are:

POINTWISE

- Linear regression

PAIRWISE

- RankSVM [46, 50]
- RankBoost [35]
- FRank [79]

LISTWISE

- ListNet [18]
- AdaRank-MAP [90]
- AdaRank-nDCG [90]
- SVM^{MAP} [99]

4.3.2 *Results*

The LETOR paper [66] describes the performance of the LETOR baseline methods. Figures 4 and 5 show ListNet to be the best performing baseline in terms of winning number on both nDCG and MAP. The LETOR website lists[2] a few algorithms that have since been evaluated on the LETOR benchmark collection.

I will describe the performance of the algorithms that were evaluated by the LETOR team on LETOR 3.0 after publication of the LETOR paper by Qin et al. [66], as listed on the LETOR website[2] by comparing their performance with the ListNet baseline. We will consider those methods to be better than ListNet when they beat the ListNet baseline in at least four of the seven datasets in nDCG value. Note that this does not necessarily imply that these methods would have scored a higher nDCG winning number than ListNet. Table 5 shows the ListNet performance on the seven LETOR datasets in terms of nDCG@10 and MAP.

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

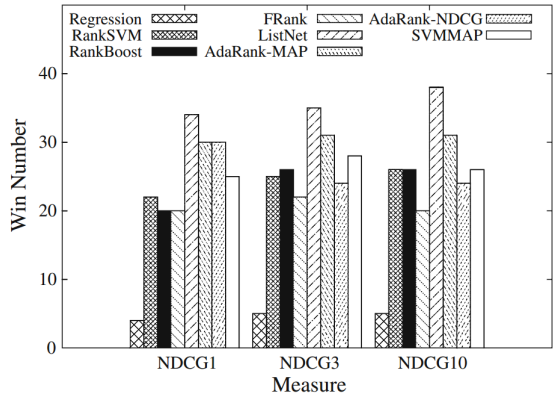

Figure 4: Comparison across the seven datasets in LETOR by nDCG, obtained from Qin et al. [66]

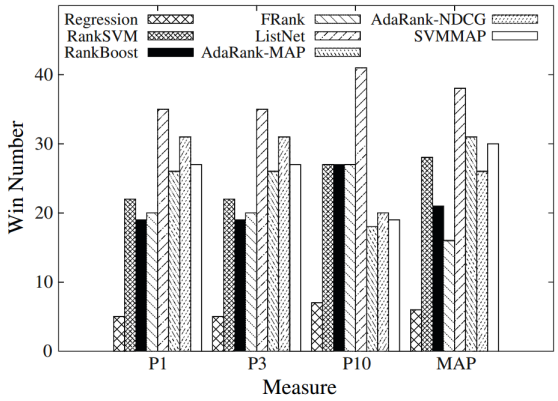

Figure 5: Comparison across the seven datasets in LETOR by MAP, obtained from Qin et al. [66]

Since LETOR is arguably the most well-known benchmark collection in the field, it is conceivable that creators of new Learning-to-Rank methods evaluate their new method on the LETOR collection to show how well their new method works. To create an overview of the performance of newer Learning-to-Rank methods on LETOR 3.0 we perform a forward reference search on the LETOR 3.0 paper [66]. It should be noted that results of evaluations individually performed by the creator of the evaluated Learning-to-Rank method can not be regarded as equally reliable as the evaluations performed by the LETOR team. TODO: 307 citations, x evaluated a new Learning-to-Rank method on LETOR 3.0, x were usable. The unusable studies did test a new Learning-to-Rank method on one or more datasets of the LETOR collection, but 1) did not use a ranking metric used in LETOR (nDCG or MAP) or 2) did not list evaluation results at all.

	nDCG@10	MAP
TD2003	0.348	0.2753
TD2004	0.317	0.2231
NP2003	0.801	0.6895
NP2004	0.812	0.6720
HP2003	0.837	0.7659
HP2004	0.784	0.6899
OHSUMED	0.441	0.4457

Table 5: Performance of ListNet on LETOR 3.0

	Regression + L2 regulatisation	RankSVM-Primal	RankSVM-Struct	SmoothRank	ListNet
TD2003	0.3297	0.3571	0.3467	0.3367	0.348
TD2004	0.2832	0.2913	0.3090	0.3343	0.317
NP2003	0.8025	0.7894	0.7955	0.7986	0.801
NP2004	0.8040	0.7950	0.7977	0.8075	0.812
HP2003	0.8216	0.8180	0.8162	0.8325	0.837
HP2004	0.7188	0.7720	0.7666	0.8221	0.784
OHSUMED	0.4436	0.4504	0.4523	0.4568	0.441
# winning datasets	2	2	1	3	-

Table 6: nDCG@10 comparison of algorithms recently evaluated on LETOR 3.0 with the ListNet baselines

Dataset	MQ2007	MQ2008
Queries	1692	784
Documents	69,622	15,211
Features	46	46

Table 7: Characteristics of the LETOR 4.0 collection

Model	Dataset	Mean nDCG	nDCG@10
RankSVM-Struct	MQ2007	0.4966	0.4439
	MQ2008	0.4832	0.2279
ListNet	MQ2007	0.4988	0.4440
	MQ2008	0.4914	0.2303
AdaRank-nDCG	MQ2007	0.4914	0.4369
	MQ2008	0.4950	0.2307
AdaRank-MAP	MQ2007	0.4891	0.4335
	MQ2008	0.4915	0.2288
RankBoost	MQ2007	0.5003	0.4464
	MQ2008	0.4850	0.2255

Table 8: Comparison of LETOR 4.0 baseline models

4.3.3 *LETOR 4.0*

The LETOR 4.0 benchmark collection[3] consists of the Gov-2 document collection and two query datasets from the Million Query Track at TREC 2007 (MQ2007) and TREC 2008 (MQ2008). LETOR 4.0 consists of a semi-supervised ranking task, a rank aggregation task and a listwise ranking task next to the supervised ranking task. Table **??** shows the collection characteristics in number of queries, documents and features for LETOR 4.0. Evaluation on this collection is performed using a five-fold cross-validation, where partitioning of the data into folds was performed beforehand by the creators of the MQ2007 and MQ2008 datasets. Documents in the dataset were

4.3.3.1 *Results*

RankSVM-Struct, ListNet, AdaRank-nDCG, AdaRank-MAP and RankBoost were used as baseline models on the LETOR 4.0 dataset and were implemented and evaluated by the publishers of LETOR 4.0. Table 8 shows the performance of those baseline models on the LETOR 4.0 benchmark collection.

3 http://http://research.microsoft.com/en-us/um/beijing/projects/letor/

BoostedTree model [52] showed an nDCG of 0.5071 on the MQ2007 dataset and thereby beat all baseline models.

4.4 MSLR-WEB10K/MSLR-WEB30K

Contrary to the Yahoo! Learning to Rank Challenge dataset, the MSLR-WEB30k provides detailed feature descriptions. The MSLR-WEB30k dataset however contains no proprietary features but only features that are commonly used in the research community.

4.4.1 *Results*

4.5 SELECTING LEARNING-TO-RANK METHODS

The accuracies of the Learning-to-Rank methods described in the preceding sections must only be compared within the benchmark and not between benchmarks for the following reasons:

1. Differences in feature sets between data sets detract from fair comparison
2. Although the nDCG definition is unambiguous, Busa-Fekete et al. [17] found that nDCG evaluation tools of benchmark data sets produced different scores

5

SELECTED LEARNING-TO-RANK METHODS

Algorithms and details of the well-performing Learning-to-Rank methods as selected in the afore-going part are presented and explained in this part.

6 IMPLEMENTATION

Describes implementation details of the Learning-to-Rank algorithm parallel implementations in HDInsight that are either Hadoop, Microsoft Azure, or HDInsight and are not part of the algorithm specification itself.

7

RESULTS & DISCUSSION

8 CONCLUSIONS

APPENDIX A: LETOR FEATURE SET

ID	Feature Description	ID	Feature Description
1	$\sum_{q_i \in q \cap d} c(q_i, d)$ of title	24	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot \log(\frac{\|C\|}{df(q_i)})$ of abstract
2	$\sum_{q_i \in q \cap d} \log(c(q_i, d) + 1)$ of title	25	$\sum_{q_i \in q \cap d} \log(\frac{c(q_i,d)}{\|d\|}) \cdot \frac{\|C\|}{c(q_i,C)} + 1$ of abstract
3	$\sum_{q_i \in q \cap d} \frac{c(q_i,d)}{\|d\|}$ of title	26	BM25 of abstract
4	$\sum_{q_i \in q \cap d} \log(\frac{c(q_i,d)}{\|d\|} + 1)$ of title	27	log(BM25) of abstract
5	$\sum_{q_i \in q \cap d} \log(\frac{\|C\|}{df(q_i)})$ of title	28	LMIR.DIR of abstract
6	$\sum_{q_i \in q \cap d} \log(\log(\frac{\|C\|}{df(q_i)}))$ of title	29	LMIR.JM of abstract
7	$\sum_{q_i \in q \cap d} \log(\frac{\|C\|}{c(q_i,C)} + 1)$ of title	30	LMIR.ABS of abstract
8	$\sum_{q_i \in q \cap d} \log(\frac{c(q_i,d)}{\|d\|} \cdot \frac{\|C\|}{df(q_i)} + 1)$ of title	31	$\sum_{q_i \in q \cap d} c(q_i, d)$ of title + abstract
9	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot \log(\frac{\|C\|}{df(q_i)})$ of title	32	$\sum_{q_i \in q \cap d} \log(c(q_i, d) + 1)$ of title + abstract
10	$\sum_{q_i \in q \cap d} \log(\frac{c(q_i,d)}{\|d\|}) \cdot \frac{\|C\|}{c(q_i,C)} + 1$ of title	33	$\sum_{q_i \in q \cap d} \frac{c(q_i,d)}{\|d\|}$ of title + abstract
11	BM25 of title	34	$\sum_{q_i \in q \cap d} \log(\frac{c(q_i,d)}{\|d\|} + 1)$ of title + abstract
12	log(BM25) of title	35	$\sum_{q_i \in q \cap d} \log(\frac{\|C\|}{df(q_i)})$ of title + abstract
13	LMIR.DIR of title	36	$\sum_{q_i \in q \cap d} \log(\log(\frac{\|C\|}{df(q_i)}))$ of title + abstract
14	LMIR.JM of title	37	$\sum_{q_i \in q \cap d} \log(\frac{\|C\|}{c(q_i,C)} + 1)$ of title + abstract
15	LMIR.ABS of title	38	$\sum_{q_i \in q \cap d} \log(\frac{c(q_i,d)}{\|d\|} \cdot \frac{\|C\|}{df(q_i)} + 1)$ of title + abstract
16	$\sum_{q_i \in q \cap d} c(q_i, d)$ of abstract	39	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot \log(\frac{\|C\|}{df(q_i)})$ of title + abstract
17	$\sum_{q_i \in q \cap d} \log(c(q_i, d) + 1)$ of abstract	40	$\sum_{q_i \in q \cap d} \log(\frac{c(q_i,d)}{\|d\|}) \cdot \frac{\|C\|}{c(q_i,C)} + 1$ of title + abstract
18	$\sum_{q_i \in q \cap d} \frac{c(q_i,d)}{\|d\|}$ of abstract	41	BM25 of title + abstract
19	$\sum_{q_i \in q \cap d} \log(\frac{c(q_i,d)}{\|d\|} + 1)$ of abstract	42	log(BM25) of title + abstract
20	$\sum_{q_i \in q \cap d} \log(\frac{\|C\|}{df(q_i)})$ of abstract	43	LMIR.DIR of title + abstract
21	$\sum_{q_i \in q \cap d} \log(\log(\frac{\|C\|}{df(q_i)}))$ of abstract	44	LMIR.JM of title + abstract
22	$\sum_{q_i \in q \cap d} \log(\frac{\|C\|}{c(q_i,C)} + 1)$ of abstract	45	LMIR.ABS of title + abstract
23	$\sum_{q_i \in q \cap d} \log(\frac{c(q_i,d)}{\|d\|} \cdot \frac{\|C\|}{df(q_i)} + 1)$ of abstract		

Table 9: Features of the LETOR OHSUMED dataset, obtained from Qin et al [66]

ID	Feature Description	ID	Feature Description
1	Term Frequency (TF) of body	36	LMIR.JM of body
2	TF of anchor	37	LMIR.JM of anchor
3	TF of title	38	LMIR.JM of title
4	TF of Uniform Resource Locator (URL)	39	LMIR.JM of URL
5	TF of whole document	40	LMIR.JM of whole document
6	Inverse Document Frequency (IDF) of body	41	Sitemap based term propagation
7	IDF of anchor	42	Sitemap based score propagation
8	IDF of title	43	Hyperlink based score propagation: weighted
9	IDF of URL	44	Hyperlink based score propagation: weighted
10	IDF of whole document	45	Hyperlink based score propagation: uniform c
11	TF-IDF of body	46	Hyperlink based feature propagation: weighte
12	TF-IDF of anchor	47	Hyperlink based feature propagation: weighte
13	TF-IDF of title	48	Hyperlink based feature propagation: uniform
14	TF-IDF of URL	49	HITS authority
15	TF-IDF of whole document	50	HITS hub
16	Document length of body	51	PageRank
17	Document length of anchor	52	HostRank
18	Document length of title	53	Topical PageRank
19	Document length of URL	54	Topical HITS authority
20	Document length of whole document	55	Topical HITS hub
21	BM25 of body	56	In-link number
22	BM25 of anchor	57	Out-link number
23	BM25 of title	58	Number of slashes in URL
24	BM25 of URL	59	Length of URL
25	BM25 of whole document	60	Number of child page
26	LMIR.ABS [101] of body	61	BM25 of extracted title
27	LMIR.ABS of anchor	62	LMIR.ABS of extracted title
28	LMIR.ABS of title	63	LMIR.DIR of extracted title
29	LMIR.ABS of URL	64	LMIR.JM of extracted title
30	LMIR.ABS of whole document		
31	LMIR.DIR of body		
32	LMIR.DIR of anchor		
33	LMIR.DIR of title		
34	LMIR.DIR of URL		
35	LMIR.DIR of whole document		

Table 10: Features of the LETOR .GOV dataset, obtained from Qin et al [66]

BIBLIOGRAPHY

[1] Acharyya, S. *Learning to rank in supervised and unsupervised settings using convexity and monotonicity*. PhD thesis, The University of Texas, Austin, 2013.

[2] Acharyya, S., Koyejo, O., and Ghosh, J. Learning to rank with Bregman divergences and monotone retargeting. *arXiv preprint arXiv:1210.4851* (2012).

[3] Adomavicius, G., and Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on 17*, 6 (2005), 734–749.

[4] Airola, A., Pahikkala, T., and Salakoski, T. Large scale training methods for linear rankrls. In *Proceedings of the ECML/PKDD-10 Workshop on Preference Learning* (2010).

[5] Airola, A., Pahikkala, T., and Salakoski, T. Training linear ranking SVMs in linearithmic time using red-black trees. *Pattern Recognition Letters 32*, 9 (2011), 1328–1336.

[6] Asadi, N., and Lin, J. Training Efficient Tree-Based Models for Document Ranking. In *Proceedings of the 25th European Conference on Advances in Information Retrieval* (2013), vol. 7814, pp. 146–157.

[7] Asadi, N., Lin, J., and de Vries, A. Runtime Optimizations for Prediction with Tree-Based Models. *IEEE Transactions on Knowledge and Data Engineering PP*, 99 (2013), 1.

[8] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association 101*, 473 (2006), 138–156.

[9] Ben-Haim, Y., and Tom-Tov, E. A streaming parallel decision tree algorithm. *The Journal of Machine Learning Research 11* (2010), 849–872.

[10] Boyd, S., Cortes, C., Jiang, C., Mohri, M., Radovanovic, A., and Skaf, J. Large-scale Distributed Optimization for Improving Accuracy at the Top. In *NIPS Workshop on Optimization for Machine Learning* (2012).

[11] Breiman, J. H., Friedman, R. A., Stone, J. C., and Olshen, L. *Classification and regression trees*. Wadsworth International Group, 1984.

[12] Breiman, L. Bagging predictors. *Machine learning 24*, 2 (1996), 123–140.

[13] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning* (2005), pp. 89–96.

[14] Burges, C. J. C. From ranknet to lambdarank to lambdamart: An overview. *Learning 11* (2010), 23–581.

[15] Burges, C. J. C., Ragno, R., and Le, Q. V. Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems (NIPS)* (2006), vol. 6, pp. 193–200.

[16] Burges, C. J. C., Svore, K. M., Bennett, P. N., Pastusiak, A., and Wu, Q. Learning to Rank Using an Ensemble of Lambda-Gradient Models. *Journal of Machine Learning Research-Proceedings Track 14* (2011), 25–35.

[17] Busa-Fekete, R., Szarvas, G., Elteto, T., and Kégl, B. An apple-to-apple comparison of Learning-to-rank algorithms in terms of Normalized Discounted Cumulative Gain. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI)* (2012), vol. 242.

[18] Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning* (2007), pp. 129–136.

[19] Chang, E. Y., Zhu, K., Wang, H., Bai, H., Li, J., Qiu, Z., and Cui, H. PSVM: Parallelizing support vector machines on distributed computers. In *Advances in Neural Information Processing Systems (NIPS)* (2007), pp. 257–264.

[20] Chapelle, O., and Chang, Y. Yahoo! Learning to Rank Challenge Overview. *Journal of Machine Learning Research-Proceedings Track 14* (2011), 1–24.

[21] Chapelle, O., Chang, Y., and Liu, T.-Y. Future directions in learning to rank. *JMLR Workshop and Conference Proceedings 14* (2011), 91–100.

[22] Chapelle, O., Metlzer, D., Zhang, Y., and Grinspan, P. Expected reciprocal rank for graded relevance. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management - CIKM '09* (Nov. 2009), ACM Press, p. 621.

[23] Chapelle, O., and Wu, M. Gradient descent optimization of smoothed information retrieval metrics. *Information retrieval 13*, 3 (2010), 216–235.

[24] Che, S., Li, J., Sheaffer, J. W., Skadron, K., and Lach, J. Accelerating Compute-Intensive Applications with GPUs and FPGAs. In *Proceedings of the Symposium on Application Specific Processors* (2008), pp. 101–107.

[25] Chen, K., Lu, R., Wong, C. K., Sun, G., Heck, L., and Tseng, B. Trada: tree based ranking function adaptation. In *Proceedings of the 17th ACM conference on Information and knowledge management* (2008), pp. 1143–1152.

[26] Chu, C., Kim, S. K., Lin, Y.-A., Yu, Y., Bradski, G., Ng, A. Y., and Olukotun, K. Map-reduce for machine learning on multicore. *Advances in neural information processing systems 19* (2007), 281.

[27] Cleverdon, C. W., and Keen, M. Aslib Cranfield research project-Factors determining the performance of indexing systems; Volume 2, Test results. Tech. rep., Cranfield University, 1966.

[28] Crammer, K., and Singer, Y. Pranking with Ranking. In *Advances in Neural Information Processing Systems (NIPS)* (2001), pp. 641–647.

[29] Craswell, N., Hawking, D., Wilkinson, R., and Wu, M. Overview of the TREC 2003 Web Track. In *Proceedings of the 12th Text Retrieval Conference (TREC)* (2003), vol. 3, p. 12th.

[30] De Sousa, D. X., Rosa, T. C., Martins, W. S., Silva, R., and Gonçalves, M. A. Improving on-demand learning to rank through parallelism. In *Proceedings of the 13th International Conference on Web Information Systems Engineering (WISE)* (2012), pp. 526–537.

[31] Dean, J., and Ghemawat, S. MapReduce: simplified data processing on large clusters. *Communications of the ACM 51*, 1 (2004), 107–113.

[32] Deng, L., He, X., and Gao, J. Deep stacking networks for information retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2013), pp. 3153–3157.

[33] Duh, K., Suzuki, J., and Nagata, M. Distributed Learning-to-Rank on Streaming Data using Alternating Direction Method of Multipliers. In *Proceedings of the NIPS Big Learning Workshop* (2011).

[34] Feldman, V., Guruswami, V., Raghavendra, P., and Wu, Y. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing 41*, 6 (2012), 1558–1590.

[35] Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research 4* (Dec. 2003), 933–969.

[36] Freund, Y., and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences 55*, 1 (1997), 119–139.

[37] Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* (2001), 1189–1232.

[38] Friedman, J. H. Stochastic gradient boosting. *Computational Statistics & Data Analysis 38*, 4 (2002), 367–378.

[39] Ganjisaffar, Y. *Tree ensembles for learning to rank*. PhD thesis, University of California, Irvine, 2011.

[40] Ganjisaffar, Y., Caruana, R., and Lopes, C. V. Bagging gradient-boosted trees for high precision, low variance ranking models. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2011), pp. 85–94.

[41] Ganjisaffar, Y., Debeauvais, T., Javanmardi, S., Caruana, R., and Lopes, C. V. Distributed tuning of machine learning algorithms using MapReduce clusters. In *Proceedings of the Third Workshop on Large Scale Data Mining: Theory and Applications* (2011), p. 2.

[42] Geurts, P., Ernst, D., and Wehenkel, L. Extremely randomized trees. *Machine learning 63*, 1 (2006), 3–42.

[43] Geurts, P., and Louppe, G. Learning to rank with extremely randomized trees. In *JMLR: Workshop and Conference Proceedings* (2011), vol. 14.

[44] Gopal, S., and Yang, Y. Distributed training of Large-scale Logistic models. In *Proceedings of the 30th International Conference on Machine Learning (ICML)* (2013), pp. 289–297.

[45] Herbrich, R., Graepel, T., and Obermayer, K. Large margin rank boundaries for ordinal regression. *Advances in Neural Information Processing Systems* (1999), 115–132.

[46] Herbrich, R., Graepel, T., and Obermayer, K. Support vector learning for ordinal regression. In *Proceedings of the Ninth International Conference on Artificial Neural Networks* (1999), vol. 1, pp. 97–102 vol.1.

[47] Holland, J. H. *Hidden Order: How Adaptation Builds Complexity*. Ulam lecture series. Perseus Books, 1995.

[48] Huber, P. J. *Robust statistics*. John Wiley and Sons, New York, 1981.

[49] Järvelin, K., and Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS) 20*, 4 (2002), 422–446.

[50] Joachims, T. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2002), pp. 133–142.

[51] Karimzadehgan, M., Li, W., Zhang, R., and Mao, J. A stochastic learning-to-rank algorithm and its application to contextual advertising. In *Proceedings of the 20th International conference on World Wide web* (2011), pp. 377–386.

[52] Kocsis, L., György, A., and Bán, A. N. BoostingTree: parallel selection of weak learners in boosting, with application to ranking. *Machine learning 93*, 2-3 (2013), 293–320.

[53] Lee, C.-P., and Lin, C.-J. Large-scale linear RankSVM. *Neural Computation* (2014), 1–37.

[54] Li, P., Burges, C. J. C., Wu, Q., Platt, J. C., Koller, D., Singer, Y., and Roweis, S. McRank: Learning to Rank Using Multiple Classification and Gradient Boosting. In *Advances in Neural Information Processing Systems (NIPS)* (2007), vol. 7, pp. 845–852.

[55] Lin, J. Mapreduce is Good Enough? If All You Have is a Hammer, Throw Away Everything That's Not a Nail! *Big Data 1*, 1 (2013), 28–37.

[56] Liu, T.-Y. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval 3*, 3 (Mar. 2007), 225–331.

[57] Liu, T.-Y., Xu, J., Qin, T., Xiong, W., and Li, H. Letor: Benchmark dataset for research on learning to rank for information retrieval. In *Proceedings of the SIGIR 2007 workshop on learning to rank for information retrieval* (2007), pp. 3–10.

[58] Long, B., Chang, Y., Dong, A., and He, J. Pairwise cross-domain factor model for heterogeneous transfer ranking. In *Proceedings of the fifth ACM International Conference on Web Search and Data Mining* (2012), pp. 113–122.

[59] Luhn, H. P. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development 1*, 4 (1957), 309–317.

[60] Malewicz, G., Austern, M. H., Bik, A. J. C., Dehnert, J. C., Horn, I., Leiser, N., and Czajkowski, G. Pregel: a system for large-scale graph processing. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (2010), pp. 135–146.

[61] McNee, S. M., Riedl, J., and Konstan, J. A. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing systems* (2006), pp. 1097–1101.

[62] Page, L., Brin, S., Motwani, R., and Winograd, T. The PageRank citation ranking: Bringing order to the web.

[63] Pahikkala, T., Tsivtsivadze, E., Airola, A., Järvinen, J., and Boberg, J. An efficient algorithm for learning to rank from preference graphs. *Machine Learning 75*, 1 (2009), 129–165.

[64] Pavlov, D. Y., Gorodilov, A., and Brunk, C. A. BagBoo: a scalable hybrid bagging-the-boosting model. In *Proceedings of the 19th ACM International conference on Information and Knowledge Management* (2010), pp. 1897–1900.

[65] Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., and Salakoski, T. BioInfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics 8*, 1 (2007), 50.

[66] Qin, T., Liu, T.-Y., Xu, J., and Li, H. LETOR: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval 13*, 4 (2010), 346–374.

[67] Qin, T., Zhang, X.-D., Tsai, M.-F., Wang, D.-S., Liu, T.-Y., and Li, H. Query-level loss functions for information retrieval. *Information Processing & Management 44*, 2 (2008), 838–855.

[68] Rigutini, L., Papini, T., Maggini, M., and Scarselli, F. Sortnet: Learning to rank by a neural-based sorting algorithm. In *Proceedings of the SIGIR Workshop on Learning to Rank for Information Retrieval (LR4IR)* (2008), pp. 76–79.

[69] Robertson, S., and Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval 3*, 4 (Apr. 2009), 333–389.

[70] Rudin, C. The P-Norm Push: A simple convex ranking algorithm that concentrates at the top of the list. *The Journal of Machine Learning Research 10* (2009), 2233–2271.

[71] Schapire, R. E. The strength of weak learnability. *Machine learning 5*, 2 (1990), 197–227.

[72] Shukla, S., Lease, M., and Tewari, A. Parallelizing ListNet Training Using Spark. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2012), pp. 1127–1128.

[73] Sorokina, D., Caruana, R., and Riedewald, M. Additive groves of regression trees. In *Proceedings of the 18th European Conference on Machine Learning* (2007), pp. 323–334.

[74] Sousa, D., Rosa, T., Martins, W., Silva, R., and Gonçalves, M. Improving On-Demand Learning to Rank through Parallelism. In *Proceedings of the 13th International Conference on Web Information Systems Engineering* (2012), vol. 7651, pp. 526–537.

[75] Sun, X.-H., and Gustafson, J. L. Toward a better parallel performance metric. *Parallel Computing 17*, 10 (1991), 1093–1109.

[76] Svore, K. M., and Burges, C. J. Large-scale learning to rank using boosted decision trees. *Scaling Up Machine Learning: Parallel and Distributed Approaches* (2012).

[77] Svore, K. M., and Burges, C. J. C. Learning to Rank on a Cluster using Boosted Decision Trees. In *Proceedings of NIPS Learning to Rank on Cores, Clusters and Clouds Workshop* (2010), p. 16.

[78] Taylor, M., Guiver, J., Robertson, S., and Minka, T. Softrank: optimizing non-smooth rank metrics. In *Proceedings of the International Conference on Web Search and Data Mining (WSDM)* (2008), pp. 77–86.

[79] Tsai, M.-F., Liu, T.-Y., Qin, T., Chen, H.-H., and Ma, W.-Y. FRank: a ranking method with fidelity loss. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2007), pp. 383–390.

[80] Tyree, S., Weinberger, K. Q., Agrawal, K., and Paykin, J. Parallel boosted regression trees for web search ranking. In *Proceedings of the 20th International Conference on World Wide Web* (2011), pp. 387–396.

[81] Veloso, A. A., Almeida, H. M., Gonçalves, M. A., and Meira Jr, W. Learning to rank at query-time using association rules. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2008), pp. 267–274.

[82] Volkovs, M. N., and Zemel, R. S. Boltzrank: learning to maximize expected ranking gain. In *Proceedings of the 26th Annual International Conference on Machine Learning* (2009), pp. 1089–1096.

[83] Wang, B., Wu, T., Yan, F., Li, R., Xu, N., and Wang, Y. RankBoost Acceleration on both NVIDIA CUDA and ATI Stream platforms. In *Proceedings of the 15th International Conference on Parallel and Distributed Systems (ICPADS)* (2009), pp. 284–291.

[84] Wang, S., Gao, B. J., Wang, K., and Lauw, H. W. CCRank: Parallel Learning to Rank with Cooperative Coevolution. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence* (2011).

[85] Wang, S., Gao, B. J., Wang, K., and Lauw, H. W. Parallel learning to rank for information retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (2011), pp. 1083–1084.

[86] Wang, S., Ma, J., and Liu, J. Learning to rank using evolutionary computation: immune programming or genetic programming? In *Proceedings of the 18th ACM conference on Information and knowledge management* (2009), pp. 1879–1882.

[87] Wolpert, D. H. Stacked generalization. *Neural networks 5*, 2 (1992), 241–259.

[88] Wu, Q., Burges, C. J. C., Svore, K. M., and Gao, J. Ranking, boosting, and model adaptation. Tech. rep., Microsoft Research, 2008.

[89] Xia, F., Liu, T.-Y., Wang, J., Zhang, W., and Li, H. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th Annual International Conference on Machine Learning* (2008), pp. 1192–1199.

[90] Xu, J., and Li, H. Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2007), pp. 391–398.

[91] Xu, N.-Y., Cai, X.-F., Gao, R., Zhang, L., and Hsu, F.-H. FPGA-based Accelerator Design for RankBoost in Web Search Engines. In *Proceedings of the International Conference on Field-Programmable Technology (ICFPT)* (2007), pp. 33–40.

[92] Xu, N.-Y., Cai, X.-F., Gao, R., Zhang, L., and Hsu, F.-H. FPGA Acceleration of RankBoost in Web Search Engines. *ACM Transactions on Reconfigurable Technology and Systems (TRETS) 1*, 4 (2009), 19.

[93] Yan, J., Xu, N.-Y., Cai, X.-F., Gao, R., Wang, Y., Luo, R., and Hsu, F.-H. FPGA-based acceleration of neural network for ranking in web search engine with a streaming architecture. In *Pro-*

ceedings of the FPL International Conference on Field Programmable Logic and Applications (2009), pp. 662–665.

[94] Yan, J., Xu, N.-Y., Cai, X.-F., Gao, R., Wang, Y., Luo, R., and Hsu, F.-H. LambdaRank Acceleration for Relevance Ranking in Web Search Engines. In *Proceedings of the 18th Annual ACM/SIGDA International Symposium on Field Programmable Gate Arrays* (2010), p. 285.

[95] Yan, J., Xu, N.-Y., Cai, X.-F., Gao, R., Wang, Y., Luo, R., and Hsu, F.-H. An FPGA-based accelerator for LambdaRank in Web search engines. *ACM Transactions on Reconfigurable Technology and Systems (TRETS) 4*, 3 (2011), 25.

[96] Yan, J., Zhao, Z.-X., Xu, N.-Y., Jin, X., Zhang, L.-T., and Hsu, F.-H. Efficient Query Processing for Web Search Engine with FPGAs. In *Proceedings of the 20th IEEE International Symposium on Field-Programmable Custom Computing Machines* (2012), pp. 97–100.

[97] Ye, J., Chow, J.-H., Chen, J., and Zheng, Z. Stochastic gradient boosted distributed decision trees. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (2009), pp. 2061–2064.

[98] Yeh, J.-Y., Lin, J.-Y., Ke, H.-R., and Yang, W.-P. Learning to rank for information retrieval using genetic programming. In *Proceedings of SIGIR Workshop on Learning to Rank for Information Retrieval (LR4IR)* (2007).

[99] Yue, Y., Finley, T., Radlinski, F., and Joachims, T. A support vector method for optimizing average precision. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2007), pp. 271–278.

[100] Yun, H., Raman, P., and Vishwanathan, S. V. N. Ranking via Robust Binary Classification and Parallel Parameter Estimation in Large-Scale Data. *arXiv preprint arXiv:1402.2676* (2014).

[101] Zhai, C., and Lafferty, J. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2001), pp. 334–342.

[102] Zheng, Z., Chen, K., Sun, G., and Zha, H. A regression framework for learning ranking functions using relative relevance judgments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2007), pp. 287–294.

[103] Zhu, M. Recall, precision and average precision. Tech. rep., Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, 2004.

[104] Zinkevich, M., Weimer, M., Smola, A. J., and Li, L. Parallelized Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems (NIPS)* (2010), pp. 2595–2603.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

DECLARATION

Put your declaration here.

Enschede, Februari 2014

Niek Tax