Efficiency/Effectiveness Trade-offs in Learning to Rank

Tutorial @ ECML PKDD 2018
http://learningtorank.isti.cnr.it/

Claudio Lucchese
Ca’ Foscari University of Venice
Venice, Italy

Franco Maria Nardini
HPC Lab, ISTI-CNR
Pisa, Italy

The Ranking Problem

Ranking is at the core of several IR Tasks:

- Document Ranking in Web Search
- Ads Ranking in Web Advertising
- Query suggestion & completion
- Product Recommendation
- Song Recommendation
- …

The Ranking Problem

Definition:

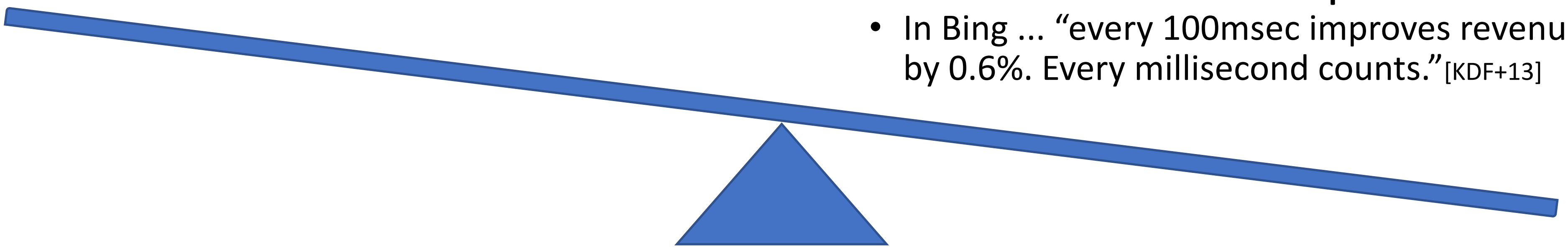
> Given a query q and a set of objects/documents D, to rank D so as to maximize users' satisfaction Q.

Goal #1: Effectiveness

- Maximize Q !
  - but how to measure Q?

Goal #2: Efficiency

- Make sure the ranking process is feasible and not too expensive
  - In Bing ... "every 100msec improves revenue by 0.6%. Every millisecond counts."[KDF+13]



[KDF+13] Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013, August). **Online controlled experiments at large scale**. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1168-1176). ACM.

Agenda

1. **Introduction to Learning to Rank** (LtR)
   - Background, algorithms, sources of cost in LtR, multi-stage ranking
2. **Dealing with the Efficiency/Effectiveness trade-off**
   - Feature Selection, Enhanced Learning, Approximate scoring, Fast Scoring
3. **Hands-on I**
   - Software, data and publicly available tools
   - Traversing Regression Forests, SoA tools and analysis
4. **Hands-on II**
   - Training models, Pruning strategies, Efficient scoring

At the end of the day you'll be able to train a high quality ranking model, and to exploit SoA tools and techniques to ***reduce its computational cost up to 18x !***

Document Representations and Ranking

Document Representations

A document is a multi-set of words

A document may have fields, it can be split into zones, it can be enriched with external text data (e.g., anchors)

Additional information may be useful, such as In-Links, Out-Links, PageRank, # clicks, social links, etc.

Hundred signals in public LtR Datasets

Ranking Functions

Term-weighting [SJ72]
Vector Space Model [SB88]

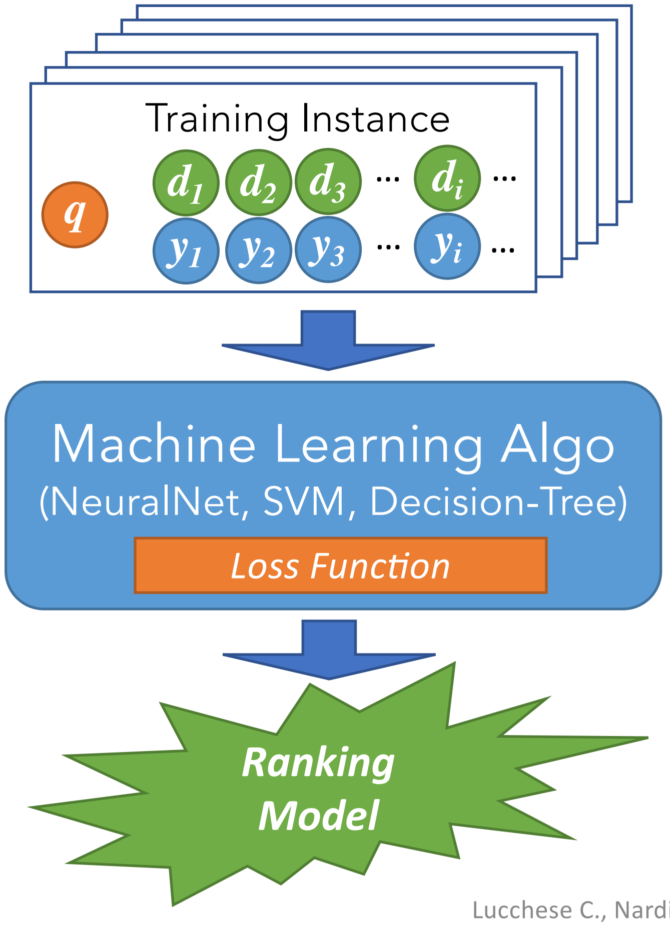
BM25 [JWR00], BM25f [RZT04]
Language Modeling [PC98]

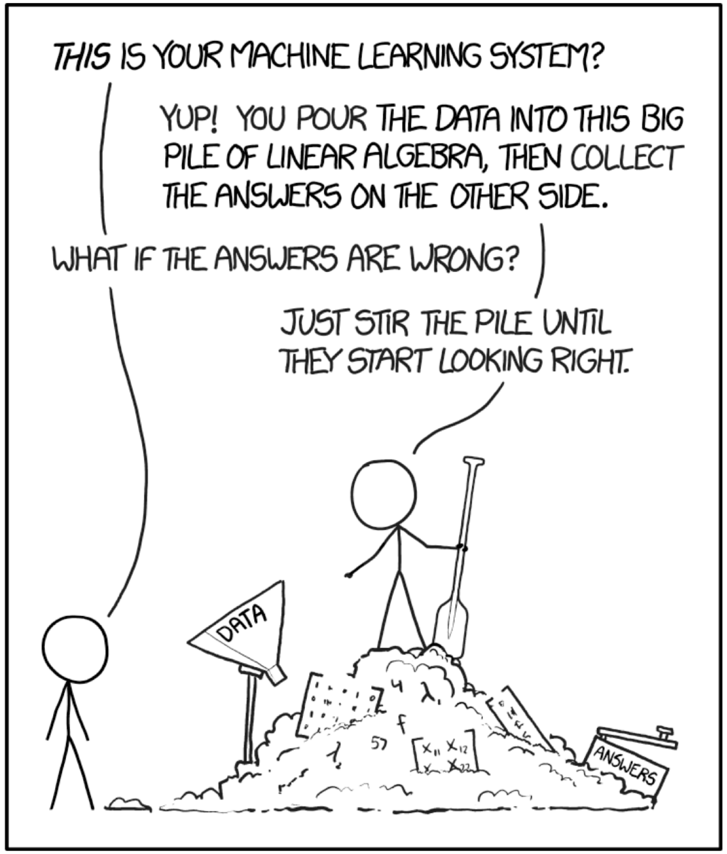
Linear Combination of features [MC07]

How to combine hundreds of signals?

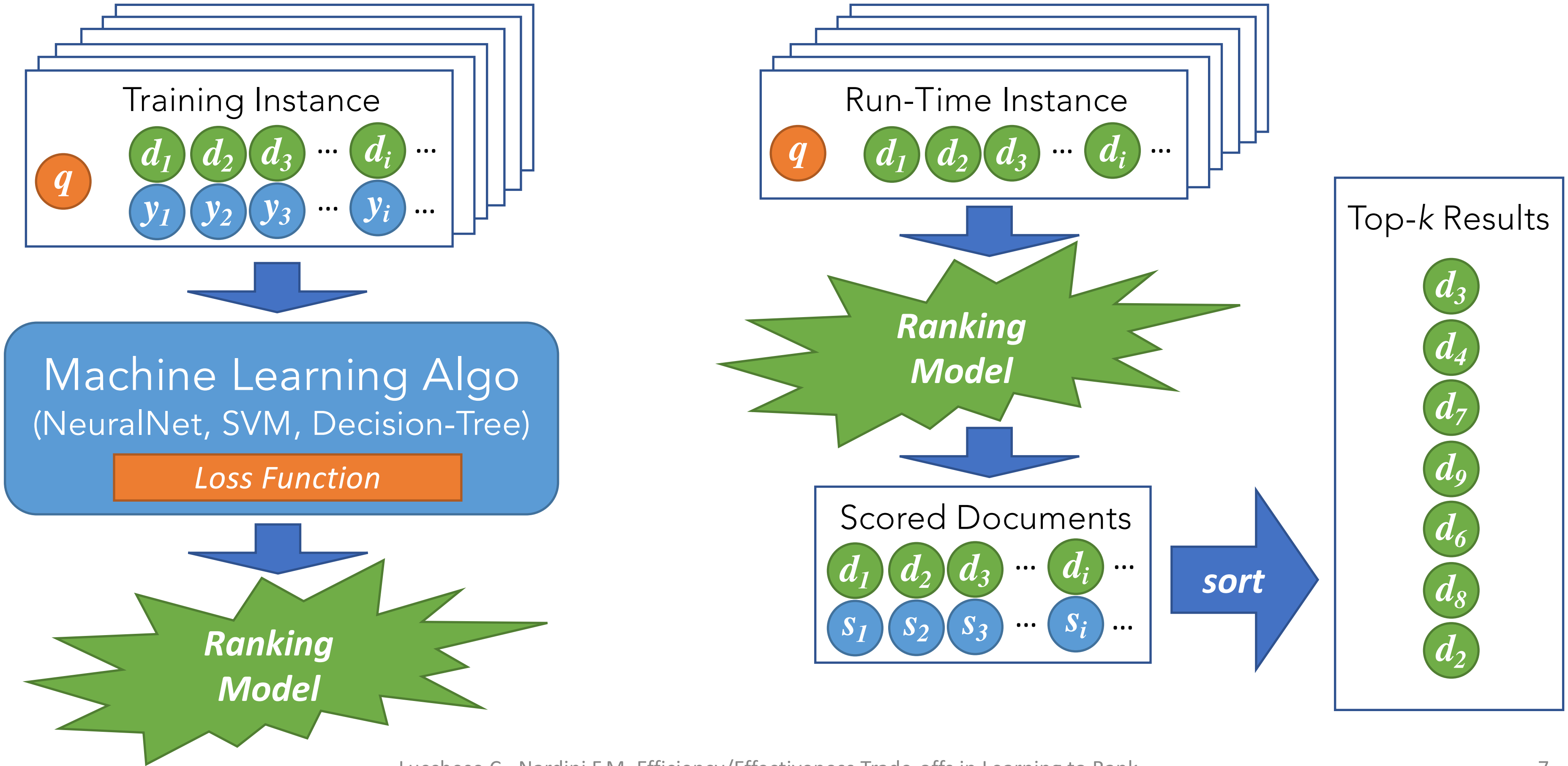
[SJ72] Karen Sparck Jones. **A statistical interpretation of term specificity and its application in retrieval**. Journal of documentation, 28(1):11–21, 1972.
[SB88] Gerard Salton and Christopher Buckley. **Term-weighting approaches in automatic text retrieval**. Information processing & management, 24(5):513–523, 1988.
[JWR00] K Sparck Jones, Steve Walker, and Stephen E. Robertson**. A probabilistic model of information retrieval: development and comparative experiments**. Information processing & management, 36(6):809–840, 2000
[RZT04] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. **Simple bm25 extension to multiple weighted fields**. In Proceedings of the thirteenth ACM international conference on Information and knowledge management, pages 42–49. ACM, 2004.
[PC98] Jay M Ponte and W Bruce Croft. **A language modeling approach to information retrieval**. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 275–281. ACM, 1998.
[MC07] Donald Metzler and W Bruce Croft**. Linear feature-based models for information retrieval**. Information Retrieval, 10(3):257–274, 2007.

Ranking as a Supervised Learning Task





Ranking as a Supervised Learning Task



Query/Document Representation

q d

Useful signals

- Link Analysis [H+00]
- Term proximity [RS03]
- Query classification [BSD10]
- Query intent mining [JLN16, LOP+13]
- Finding entities documents [MW08] and in queries [BOM15]
- Document recency [DZK+10]
- Distributed representations of words and their compositionality [MSC+13]
- Convolutional neural networks [SHG+14]
- ….

Relevance Labels Generation

y

- ***Explicit Feedback***
  - Thousands of Search Quality Raters
  - Absolute vs. Relative Judgments [CBCD08]
- ***Implicit Feedback***
  - clicks/query chains [JGP+05, Joa02, RJ05]
  - Unbiased learning-to-rank [JSS17]
- ***Minimize annotation cost***
  - Active Learning [LCZ+10]
  - Deep versus Shallow labelling [YR09]

Evaluation Measures for Ranking

Rank	Top 10 Retrieved Documents	Binary Relevance Labels		Graded Relevance Labels	
1	d_3	y_3	✓	y_3	☆ ☆ ☆ ☆
2	d_4	y_4	✗	y_4	✗
3	d_7	y_7	✓	y_7	☆
4	d_9	y_9	✗	y_9	✗
5	d_6	y_6	✗	y_6	✗
6	d_8	y_8	✗	y_8	✗
7	d_2	y_2	✓	y_2	☆ ☆ ☆
8	d_5	y_5	✗	y_5	✗
9	d_1	y_1	✗	y_1	✗
10	d_{10}	y_{10}	✗	y_{10}	✗

Precision @10

$$P@10 = \frac{3}{10}$$

Account for labels:

$$Q@10 = 4 + 1 + 3$$

Account for labels and ranks:

$$Q@10 = \frac{4}{1} + \frac{1}{3} + \frac{3}{7}$$

Evaluation Measures for Ranking

Many are in the form: $$Q@k = \sum_{\texttt{ranks}\ r=1\ldots k} \mathsf{Gain}(d^r) \cdot \mathsf{Discount}(r)$$

- (N)DCG [JK00]: $\mathsf{Gain}(d) = 2^y - 1 \quad \mathsf{Discount}(r) = \frac{1}{\log(r+1)}$
- RBP [MZ08]: $\mathsf{Gain}(d) = \mathbb{I}(y) \quad \mathsf{Discount}(r) = (1-p)p^{r-1}$
- ERR [CMZG09]: $\mathsf{Gain}(d) = R_i \prod_{j=1}^{i-1}(1 - R_j)\ with\ R_i = (2^y - 1)/2^{y_{max}} \quad \mathsf{Discount}(r) = 1/r$

Do they match User satisfaction ?

- ERR correlates better with user satisfaction (clicks and editorials) [CMZG09]
- Results Interleaving to compare two rankings [CJRY12]
  - "major revisions of the web search rankers [Bing] ... The differences between these rankers involve changes of over *half a percentage point*, in absolute terms, of NDCG"

[JK00] Kalervo J arvelin and Jaana Kekalainen. **IR evaluation methods for retrieving highly relevant documents**. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pages 41–48. ACM, 2000.

[MZ08] Alistair Moffat and Justin Zobel. **Rank-biased precision for measurement of retrieval effectiveness**. ACM Transactions on Information Systems (TOIS), 27(1):2, 2008.

[CMZG09] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. **Expected reciprocal rank for graded relevance**. In Proceedings of the 18th ACM conference on Information and knowledge management, pages 621–630. ACM, 2009.

[CJRY12] Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. **Large-scale validation and analysis of interleaved search evaluation**. ACM Transactions on Information Systems (TOIS), 30(1):6, 2012.

Is it an easy or difficult task?

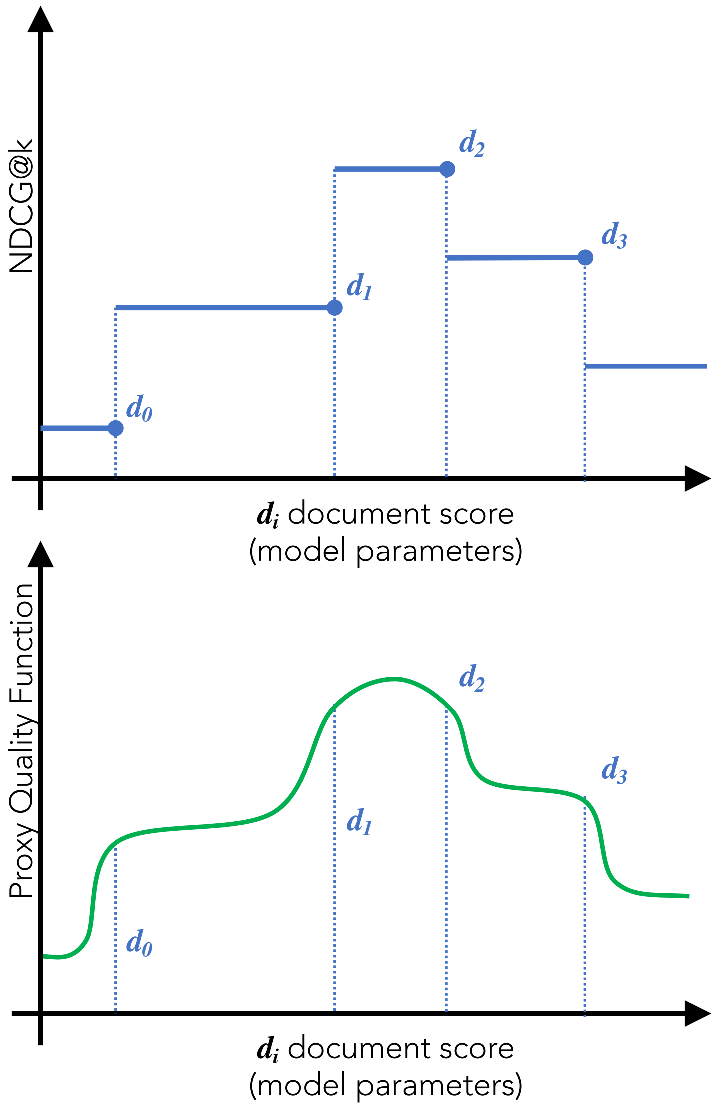
Gradient descent cannot be applied directly

Rank-based measures (NDCG, ERR, MAP, …) depend on ***documents sorted order***

- ***gradient is either 0*** (sorted order did not change) or ***undefined*** (discontinuity)

Solution: we need a proxy Loss function

- it should be ***differentiable***
- and with a ***similar behavior of the original cost function***



Point-Wise Algorithms

<u>Each document is considered independently</u> from the others

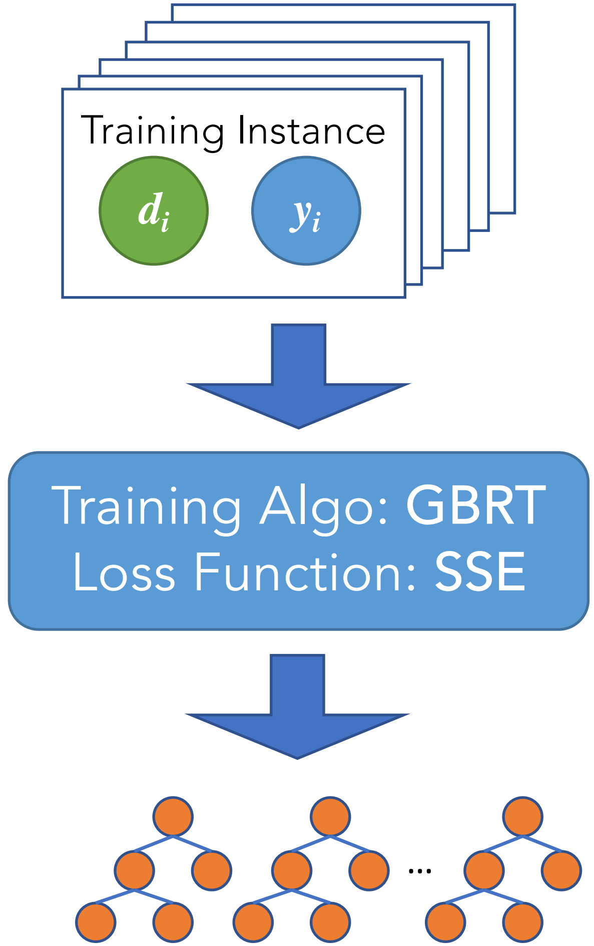
- No information about other candidates for the same query is used at training time

A different cost-function is optimized

- Several approaches: Regression, Multi-Class Classification, Ordinal regression, … [Liu11]

Among Regression-Based:
Gradient Boosting Regression Trees [Fri01]

- ***Sum of Squared Errors (SSE)*** is minimized



[Liu11] Tie-Yan Liu. **Learning to rank for information retrieval**, 2011. Springer.
[Fri01] Jerome H Friedman. **Greedy function approximation: a gradient boosting machine**. Annals of statistics, pages 1189–1232, 2001.

Gradient Boosting Regression Trees

Iterative algorithm: $F(d) = \sum_i f_i(d)$

Weak Learner

Each f_i is regarded as a step in the best optimization direction, i.e., a ***steepest descent step***:

by line-search

negative gradient

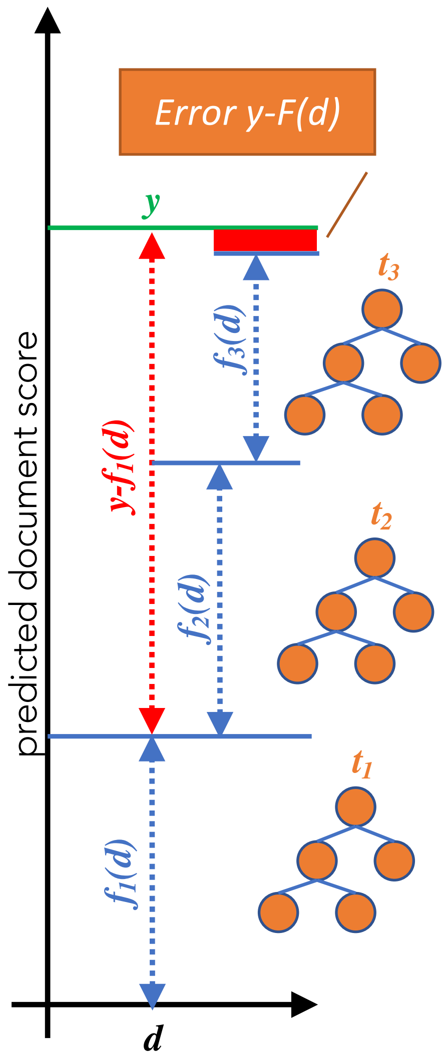
$$f_i(d) = -\rho_i \; g_i(d) \qquad -g_i(d) = -\left[\frac{\partial L(y, f(d))}{\partial f(d)}\right]_{f=\sum_{j<i} f_j}$$

Given $L = SSE/2$:

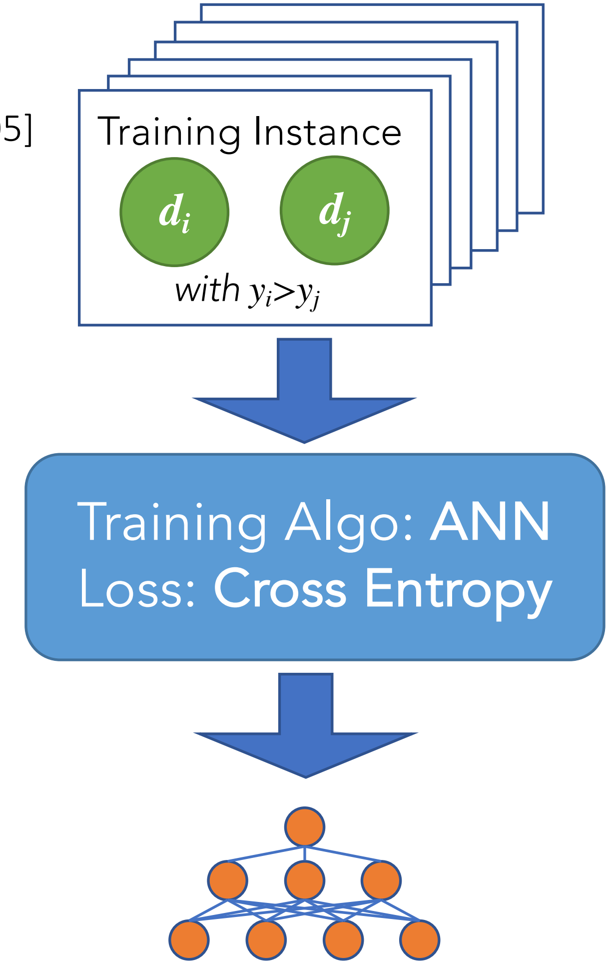
pseudo-response

$$-\frac{\partial[\frac{1}{2}SSE(y, f(d))]}{\partial f(d)} = -\frac{\partial[\frac{1}{2}\sum(y - f(d))^2]}{\partial f(d)} = y - f(d)$$

Gradient g_i is approximated by a Regression Tree t_i



Pair-wise Algorithms: RankNet[BSR+05]



Documents are considered in pairs

Estimated probability that ***d_i is better than d_j*** is:

$$P_{ij} = \frac{e^{o_{ij}}}{1 + e^{o_{ij}}}$$

$o_{ij} = F(d_i) - F(d_j)$

Let Q_{ij} be the true probability, the ***Cross Entropy Loss*** is:

$$C_{ij} = -Q_{ij} \log P_{ij} - (1 - Q_{ij}) \log(1 - P_{ij})$$

We consider ***only pairs where d_i is better than d_j, ie., $y_i > y_j$:***

If $o_{ij} \rightarrow +\infty$
(i.e., correctly ordered)
$C_{ij} \rightarrow 0$

$$C_{ij} = \log(1 + e^{-o_{ij}})$$

If $o_{ij} \rightarrow -\infty$
(i.e., mis-ordered)
$C_{ij} \rightarrow +\infty$

This is ***differentiable***: used to train a ***Neural Network with back-propagation***.

Other approaches: Ranking-SVM[Joa02], RankBoost[FISS03], …

[BSR+05] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. **Learning to rank using gradient descent**. In Proceedings of the 22nd international conference on Machine learning, pages 89–96. ACM, 2005.
[Joa02] Thorsten Joachims. **Optimizing search engines using clickthrough data**. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 133–142. ACM, 2002.
[FISS03] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. **An efficient boosting algorithm for combining preferences**. Journal of machine learning research, 4(Nov):933–969, 2003.

Pair-wise Algorithms

RankNet** performs* ***better *than other pairwise algorithms*

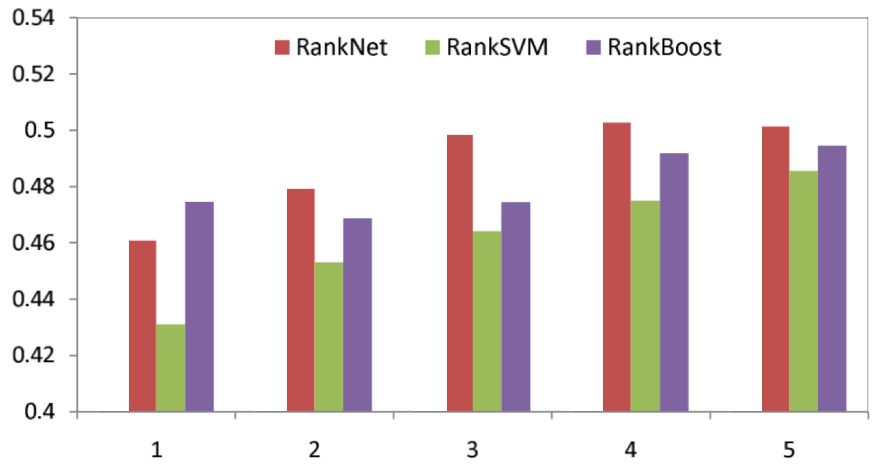


Figure 1. Ranking accuracies in terms of NDCG@n on TREC

RankNet** cost is* ***not *nicely correlated with NDCG quality*

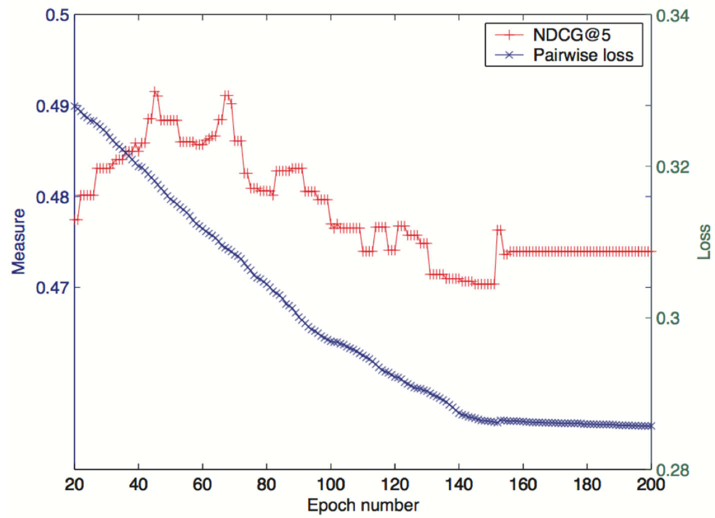


Figure 4. Pairwise loss v.s. NDCG@5 in RankNet

[CQL+07] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li**. Learning to rank: from pairwise approach to listwise approach**. In Proceedings of the 24th international conference on Machine learning, pages 129–136. ACM, 2007.

List-wise Algorithms: LambdaMart[Bur10]

Recall: ***GBRT requires a gradient g_i for every d_i***

First: ***estimate the gradient comparing to d_j***, with $y_i > y_j$:

derivative of the negative RankNet cost

Δ Quality after swapping d_i with d_j

$$\lambda_{ij} = \frac{1}{1 + e^{o_{ij}}} \; |\Delta NDCG| = -\lambda_{ji}$$

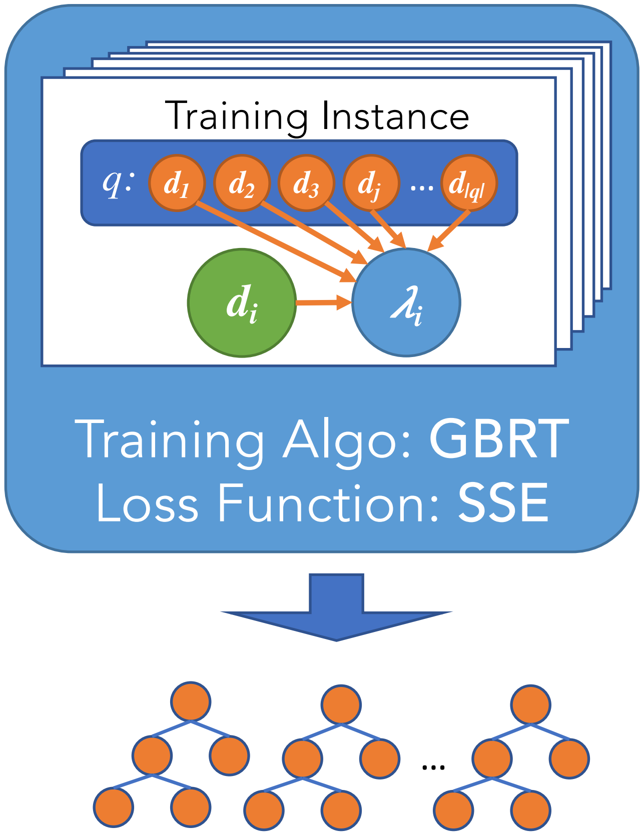
If $o_{ij} \rightarrow +\infty$ (i.e., correctly ordered) $\lambda_{ij} \rightarrow 0$

If $o_{ij} \rightarrow -\infty$ (i.e., mis-ordered) $\lambda_{ij} \rightarrow |\Delta\ NDCG|$

Top documents are more relevant !

Then: ***estimate the gradient comparing to every other d_j for q***

$$g_i = \lambda_i = \sum_{y_i > y_j} \lambda_{ij} - \sum_{y_i < y_j} \lambda_{ij}$$



[Bur10] Christopher J.C. Burges. **From ranknet to lambdarank to lambdamart: An overview**. Technical Report MSR-TR-2010-82, June 2010.

List-wise Algorithms: some results

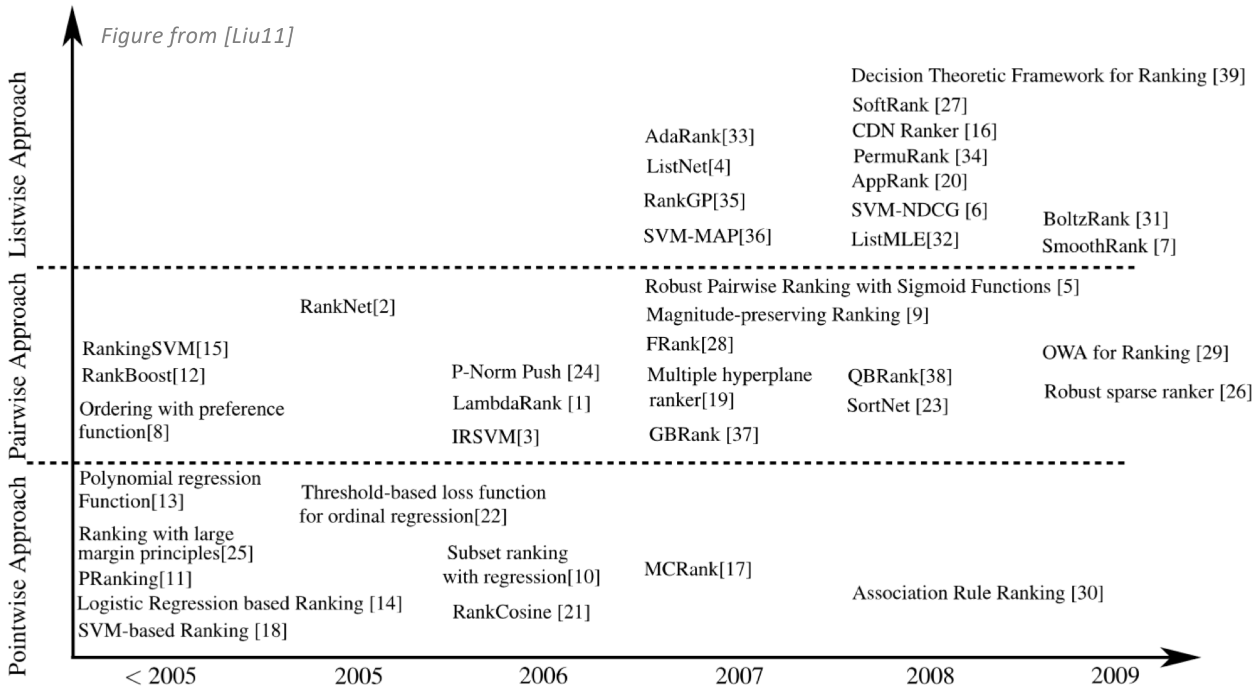
- NDCG@10 on public LtR Datasets

Algorithm	**MSN10K**	**Y!S1**	**Y!S2**	**Istella-S**
RankingSVM	0.4012	0.7238	0.7306	N/A
GBRT	0.4602	**0.7555**	**0.7620**	0.7313
LambdaMART	**0.4618**	0.7529	0.7531	**0.7537**

Other approaches: ListNet/ListMLE[CQL+07], Approximate Rank[QLL10], SVM AP[YFRJ07], RankGP[YLKY07], others ...

[CQL+ 07] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. **Learning to rank: from pairwise approach to listwise approach**. In Proceedings of the 24th international conference on Machine learning, pages 129–136. ACM, 2007.
[QLL10] Tao Qin, Tie-Yan Liu, and Hang Li. **A general approximation framework for direct optimization of information retrieval measures**. Information retrieval, 13(4):375–397, 2010.
[YFRJ08] Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims. **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, pages 271– 278. ACM, 2007.
[YLKY07] Jen-Yuan Yeh, Jung-Yi Lin, Hao-Ren Ke, and Wei-Pang Yang. **Learning to rank for information retrieval using genetic programming**. In Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval (LR4IR 2007), 2007.

Learning to Rank Algorithms



- New approaches to **optimize IR measures:**
  - DirectRank[XLL+08], LambdaMart[Bur10], BLMart[GCL11], SSLambdaMART[SY11], CoList[GY14], LogisticRank[YHT+16], …
    See [Liu11][TBH15].
- **Deep Learning** to improve query-document matching:
  - Conv.DNN[SM15], DSSM[HHG+13], Dual-Embedding[MNCC16], Local and Distributed repr.[MDC17], Weak Supervision[DZS+17], Neural Click Model[BMdRS16], …
- **On-line learning**:
  - Multi-armed bandits [RKJ08], Dueling bandits [YJ09], K-armed dueling bandits[YBKJ12], online learning[HSWdR13][HWdR13], …

[Liu11] Tie-Yan Liu. **Learning to rank for information retrieval**, 2011. Springer.
[TBH15] Niek Tax, Sander Bockting, and Djoerd Hiemstra. **A cross-benchmark comparison of 87 learning to rank methods**. Information processing & management, 51(6):757–772, 2015.

In this tutorial we focus on GBRTs

Ads Click Prediction: GBDT as a ***feature extractor***, then LogReg [HPJ+14]

Ads Click Prediction: refine/***boost NN*** output [LDG+17]

Product Ranking: 100 GBDTs with pairwise ranking [SCP16]

Document Ranking: GBDT named LogisticRank [YHT+16]

Ranking, forecasting & recommendations: ***Oblivious GBRT***

[HPJ+14] Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, et al. **Practical lessons from predicting clicks on ads at facebook**. In Proceedings of the Eighth International Workshop on Data Mining for Online Advertising, pages 1–9. ACM, 2014.
[LDG+17] Xiaoliang Ling, Weiwei Deng, Chen Gu, Hucheng Zhou, Cui Li, and Feng Sun. **Model ensemble for click prediction in bing search ads**. In Proceedings of the 26th International Conference on World Wide Web Companion, pages 689–698. International World Wide Web Conferences Steering Committee, 2017.
[SCP16] Daria Sorokina and Erick Cantu ́-Paz. **Amazon search: The joy of ranking products**. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pages 459–460. ACM, 2016.
[YHT+16] Dawei Yin, Yuening Hu, Jiliang Tang, Tim Daly, Mianwei Zhou, Hua Ouyang, Jianhui Chen, Changsung Kang, Hongbo Deng, Chikashi Nobata, et al. **Ranking relevance in yahoo search**. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 323–332. ACM, 2016.

In this tutorial we focus on GBRTs

- Successful in several ***Data Challenges***:
  - Winner of the ***Yahoo! LtR Challenge***: combination of 12 ranking models, 8 of which were Lambda-MART models, each having up to 3,000 trees [CC11]
  - According to the 2015 statistics, GBRTs were adopted by the majority of the winning solutions among the ***Kaggle*** competitions, even more than the popular deep networks, and all the top-10 teams qualified in the ***KDDCup 2015*** used GBRT-based algorithms [CG16]

- New interesting ***open-source implementations***:
  - XGBoost, LightGBM by ***Microsoft***, CatBoost by ***Yandex***

- ***Pluggable*** within ***Apache Lucene/Solr***

[CC11] Olivier Chapelle and Yi Chang. **Yahoo! learning to rank challenge overview**. In Proceedings of the Learning to Rank Challenge, pages 1–24, 2011.
[CG16] Tianqi Chen and Carlos Guestrin. **Xgboost: A scalable tree boosting system**. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16, pages 785–794, New York, NY, USA, 2016. ACM.

Single-Stage Ranking

Requires to apply the learnt *model* to *every matching document*, and to generate the required *features*.

Not feasible!

We have at least ***3 efficiency vs. effectiveness trade-offs***.

Single-Stage Ranking

① ***Feature Computation Trade-off***

- Computationally ***Expensive*** & ***highly discriminative*** features vs. computationally ***Cheap*** & ***slightly discriminative*** features

Two-Stage Ranking

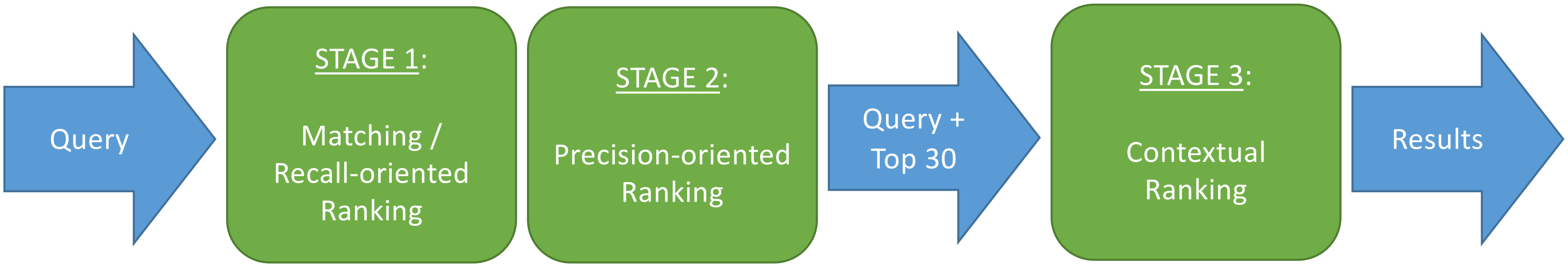
Expensive features are computed only for the *top-K candidate documents* passing the first stage.
How to chose K?

②***Number of Matching Candidates Trade-off*** :

- a ***Large set*** of candidates is ***Expensive*** and produces ***high-quality*** results vs.
  a ***Small set*** of candidates is ***Cheap*** and produces ***low-quality*** results
  - 1000 documents [DBC13] (Gov2, ClueWeb09-B collections)
  - 1500-2000 documents [MSO13] (ClueWeb09-B)
  - "hundreds of thousands" (over "hundreds of machines") [YHT+16a]

[DBC13] Van Dang, Michael Bendersky, and W Bruce Croft. **Two-stage learning to rank for information retrieval**. In Advances in Information Retrieval, pages 423–434. Springer, 2013.
[MSO13] Craig Macdonald, Rodrygo LT Santos, and Iadh Ounis. **The whens and hows of learning to rank for web search**. Information Retrieval, 16(5):584–628, 2013.
[YHT+16] Dawei Yin, Yuening Hu, Jiliang Tang, Tim Daly, Mianwei Zhou, Hua Ouyang, Jianhui Chen, Changsung Kang, Hongbo Deng, Chikashi Nobata, et al. **Ranking relevance in yahoo search**. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 323–332. ACM, 2016.

Multi-Stage Ranking



- 3 stages [YHT+16]: ***Contextual features*** are considered in the 3rd stage
  - Contextual => ***about the current result set***
  - Rank based on specific features, Mean, Variance, Standardized features (see also [LNO+15a]), topic model similarity
  - First two stages are executed at each serving node
- *N* stages [CGBC17]: Which ***model*** in each stage? Which ***features***? How many ***documents***?
  - About ***200 configurations tested,*** best results with N=3 stages, 2500 and 700 docs between stages
- Predict the best k for STAGE 1 [CCL16], and the best processing pipeline [MCB+18]
- A proper methodology/algorithm for choosing the best configuration is still missing.

[YHT+16] Dawei Yin, Yuening Hu, Jiliang Tang et al. **Ranking relevance in yahoo search**. In Proceedings of the 22nd ACM SIGKDD. ACM, 2016.
[CGBC17] Ruey-Cheng Chen, Luke Gallagher, Roi Blanco, and J. Shane Culpepper. **Efficient cost-aware cascade ranking in multi-stage retrieval**. In Proceedings of ACM SIGIR ACM, 2017.
[MCB+18] Mackenzie, J., Culpepper, J. S., Blanco, R., et al. **Query Driven Algorithm Selection in Early Stage Retrieval**. In Proceedings of WSDM. ACM, 2018.
[CCL16] Culpepper, J. S., Clarke, C. L., & Lin, J. **Dynamic cutoff prediction in multi-stage retrieval systems**. In Proceedings of the 21st Australasian Document Computing Symposium. ACM, 2016.

Multi-Stage Ranking

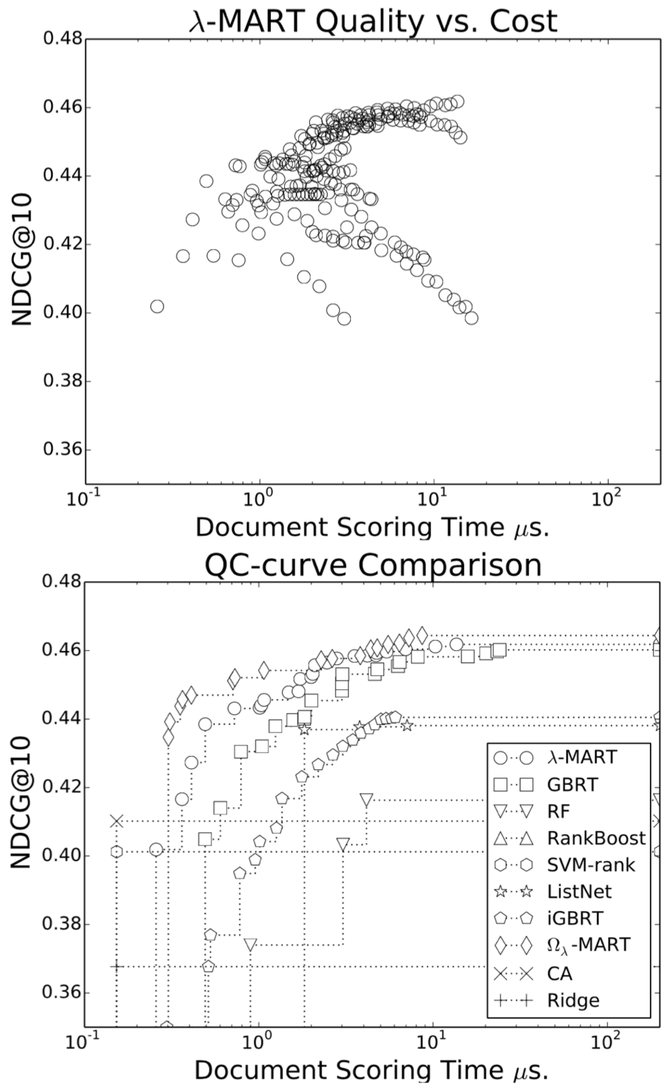
③ ***Model Complexity Trade-off*** :

- ***Complex*** & ***Slow*** ***high-quality*** vs. ***Simple*** & ***Fast*** ***low-quality*** models:
  - ***Complex*** as: Random Forest, GBRT, Initialized GBRT, Lambda-MART,
  - ***Simple*** as: Coordinate Ascent, Ridge Regression, SVM-Rank, RankBoost
  - ***In-between*** as: Oblivious Lambda-Mart, ListNet

Model Complexity Trade-off

- Comparison on varying training parameters [CLN+16]:
  - #trees, #leaves, learning rate, etc.
- ***Complex models*** achieve significantly ***higher quality***
- Best model depends on ***time budget***

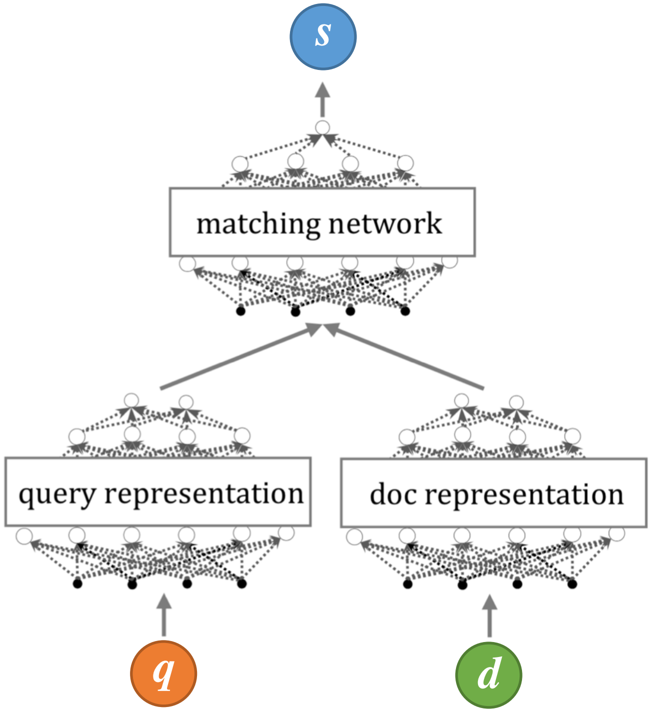
- ***Today is about Model Complexity Trade-off!***



[CLN+16] Gabriele Capannini, Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, and Nicola Tonellotto. **Quality versus efficiency in document scoring with learning-to-rank models**. Information Processing & Management, 2016.

…some recent advances…

Learning to Rank with Deep Neural Networks



- Issues: ***Multiple Fields / Multi-instance Fields***
- Proposed solution:
  - ***Instance-level DNN***:
    - Layers: 3-gram hashing, embedding, 1D convolution, pooling, dense
    - ***Per-field tunable***
  - Multi-instance aggregation by averaging
  - Multi-field aggregation concatenation
- Pair-wise training, ***non-public Bing data***

Model	NDCG@1	NDCG@10
BM25F	0.4431	0.6020
LTR	0.4888	0.6341
NRM-F	**0.4906***	**0.6380***

[AM+18] Zamani, H., Mitra, B., Song, X., Craswell, N., & Tiwary, S. (2018, February). **Neural ranking models with multiple document fields**. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (pp. 700-708). ACM.

Are ANN faster than Regression Forests?

- ***Rationale***:
  - Trust accuracy of *LambdaMART*, use a “similar” ANN at run time
- ***Methodology***:
  - *Ranking Distillation* [TW18]
  - ***Train a ANN that approximates the output of LambdaMART***
    - … rather than the training labels!
    - Enrich the dataset with points around discontinuities, i.e., trees’ split points
    - Networks used: Fully connected 4 layers 2000x500x500x100 and 2 layers 500x100

Method	# Layers	MSN30k MAP	GOV2 MAP
Regression Forest	-	0.6004	0.2995
N_{approx}	4	0.5950	0.2995
N_{approx}	2	0.5955	0.3007
$N_{\text{relevance}}$	4	0.5639*	0.2531*

		1000 Trees		20,000 Trees	
Impl	Source	8 Leaves	64 Leaves	8 Leaves	64 Leaves
Generated C++ for Forest	If-Then-Else	8.2-10.3	55.9-55.1	709.0-772.2	4462.0-4809.0
Tensorflow-CPU 4-layer	N				51.1-53.3
Tensorflow-CPU 2-layer	N				5.83-7.04
PyTorch GPU 4-layer	N				0.976-1.01
PyTorch GPU 2-layer	N				0.323-0.335

[CF+18] Cohen, D., Foley, J., Zamani, H., Allan, J., & Croft, W. B. (2018, June). **Universal Approximation Functions for Fast Learning to Rank: Replacing Expensive Regression Forests with Simple Feed-Forward Networks**. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 1017-1020). ACM.

[TW18] Tang, J., & Wang, K. (2018, July). Ranking Distillation: Learning Compact Ranking Models With High Performance for Recommender System. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2289-2298). ACM.

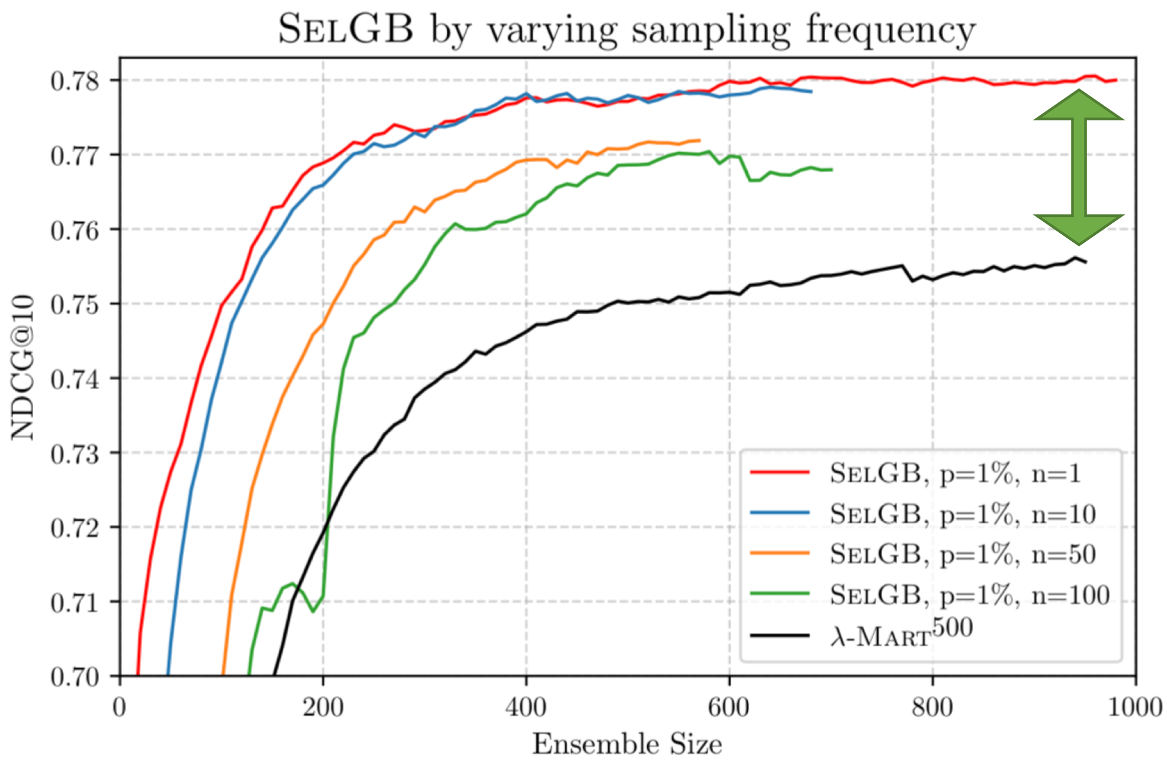
Dealing with large and unbalanced datasets

GBRT/LambdaMART is typically applied to (trained on) a *large* set of documents out of which *only a few are relevant*

Can we achieve faster and more effective training?

Selective Gradient Boosting

- uses a small percentage of ***non-relevant documents*** (e.g., 1%)
- chosen, at each iteration, ***among the top ranked***
- Achieves >3% NDCG relative improvement!



[LNP+18] Lucchese, C., Nardini, F. M., Perego, R., Orlando, S., & Trani, S. (2018, June). **Selective Gradient Boosting for Effective Learning to Rank.** In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 155-164). ACM.

Next …

Efficiency/Effectiveness trade-offs in:

- Feature Selection
- Enhanced Learning Algorithms
- Approximate scoring
- Fast Scoring

References

[BMdRS16] Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov.A neural click model for web search.In Proceedings of the 25th International Conference on World Wide Web, pages 531--541. International World Wide Web Conferences Steering Committee, 2016.

[BOM15] Roi Blanco, Giuseppe Ottaviano, and Edgar Meij.Fast and space-efficient entity linking for queries.In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pages 179--188. ACM, 2015.

[BSD10] Paul N Bennett, Krysta Svore, and Susan T Dumais.Classification-enhanced ranking.In Proceedings of the 19th international conference on World wide web, pages 111--120. ACM, 2010.

[BSR+05] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender.Learning to rank using gradient descent.In Proceedings of the 22nd international conference on Machine learning, pages 89--96. ACM, 2005.

[Bur10] Christopher J.C. Burges.From ranknet to lambdarank to lambdamart: An overview.Technical Report MSR-TR-2010-82, June 2010.

[CBCD08] Ben Carterette, Paul Bennett, David Chickering, and Susan Dumais.Here or there: Preference Judgments for Relevance. Advances in Information Retrieval, pages 16--27, 2008.

[CC11] Olivier Chapelle and Yi Chang.Yahoo! learning to rank challenge overview.In Proceedings of the Learning to Rank Challenge, pages 1--24, 2011.

[CCL11] Olivier Chapelle, Yi Chang, and T-Y Liu.Future directions in learning to rank.In Proceedings of the Learning to Rank Challenge, pages 91--100, 2011.

[CG16] Tianqi Chen and Carlos Guestrin.Xgboost: A scalable tree boosting system.In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 785--794, New York, NY, USA, 2016. ACM.

References

[CGBC17] Ruey-Cheng Chen, Luke Gallagher, Roi Blanco, and J. Shane Culpepper.Efficient cost-aware cascade ranking in multi-stage retrieval.In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17, pages 445--454, New York, NY, USA, 2017. ACM.

[CJRY12] Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue.Large-scale validation and analysis of interleaved search evaluation.ACM Transactions on Information Systems (TOIS), 30(1):6, 2012.

[CLN+16] Gabriele Capannini, Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, and Nicola Tonellotto.Quality versus efficiency in document scoring with learning-to-rank models.Inf. Process. Manage., 52(6):1161--1177, November 2016.

[CMZG09] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan.Expected reciprocal rank for graded relevance.In Proceedings of the 18th ACM conference on Information and knowledge management, pages 621--630. ACM, 2009.

[CQL+07] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li.Learning to rank: from pairwise approach to listwise approach.In Proceedings of the 24th international conference on Machine learning, pages 129--136. ACM, 2007.

[DBC13] Van Dang, Michael Bendersky, and W Bruce Croft.Two-stage learning to rank for information retrieval.In Advances in Information Retrieval, pages 423--434. Springer, 2013.

[DZK+10] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha.Time is of the essence: improving recency ranking using twitter data.In Proceedings of the 19th international conference on World wide web, pages 331--340. ACM, 2010.

[DZS+17] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft.Neural ranking models with weak supervision.In Proceedings of the 40th International {ACM {SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017, pages 65--74. {ACM, 2017.

References

[FISS03] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer.An efficient boosting algorithm for combining preferences.Journal of machine learning research, 4(Nov):933--969, 2003.

[Fri01] Jerome H Friedman.Greedy function approximation: a gradient boosting machine.Annals of statistics, pages 1189--1232, 2001.

[GCL11] Yasser Ganjisaffar, Rich Caruana, and Cristina Videira Lopes.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, pages 85--94. ACM, 2011.

[GY14] Wei Gao and Pei Yang.Democracy is good for ranking: Towards multi-view rank learning and adaptation in web search.In Proceedings of the 7th ACM international conference on Web search and data mining, pages 63--72. ACM, 2014.

[H+00] Monika Rauch Henzinger et al.Link analysis in web information retrieval.IEEE Data Eng. Bull., 23(3):3--8, 2000.

[HHG+13] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck.Learning deep structured semantic models for web search using clickthrough data.In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, pages 2333--2338. ACM, 2013.

[HPJ+14] Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, et al.Practical lessons from predicting clicks on ads at facebook.In Proceedings of the Eighth International Workshop on Data Mining for Online Advertising, pages 1--9. ACM, 2014.

References

[HSWdR13] Katja Hofmann, Anne Schuth, Shimon Whiteson, and Maarten de Rijke.Reusing historical interaction data for faster online learning to rank for ir.In Proceedings of the sixth ACM international conference on Web search and data mining, pages 183--192. ACM, 2013.

[HWdR13] Katja Hofmann, Shimon Whiteson, and Maarten de Rijke.Balancing exploration and exploitation in listwise and pairwise online learning to rank for information retrieval.Information Retrieval, 16(1):63--90, 2013.

[JGP+05] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay.Accurately interpreting clickthrough data as implicit feedback.In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pages 154--161. Acm, 2005.

[JK00] Kalervo J{\"arvelin and Jaana Kekalainen.Ir evaluation methods for retrieving highly relevant documents.In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pages 41--48. ACM, 2000.

[JLN16] Di Jiang, Kenneth Wai-Ting Leung, and Wilfred Ng.Query intent mining with multiple dimensions of web search data.World Wide Web, 19(3):475--497, 2016.

[Joa02] Thorsten Joachims. Optimizing search engines using clickthrough data .In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 133--142. ACM, 2002.

[JSS17] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017.

[JWR00] K Sparck Jones, Steve Walker, and Stephen E. Robertson.A probabilistic model of information retrieval: development and comparative experiments: Part 2.Information processing & management, 36(6):809--840, 2000.

References

[KDF+13] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann.Online controlled experiments at large scale.In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1168--1176. ACM, 2013.

[LCZ+10] Bo Long, Olivier Chapelle, Ya Zhang, Yi Chang, Zhaohui Zheng, and Belle Tseng.Active learning for ranking through expected loss optimization.In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pages 267--274. ACM, 2010.

[LDG+17] Xiaoliang Ling, Weiwei Deng, Chen Gu, Hucheng Zhou, Cui Li, and Feng Sun.Model ensemble for click prediction in bing search ads.In Proceedings of the 26th International Conference on World Wide Web Companion, pages 689--698. International World Wide Web Conferences Steering Committee, 2017.

[Liu11] Tie-Yan Liu.Learning to rank for information retrieval, 2011.

[LNO+15] Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, and Nicola Tonellotto.Speeding up document ranking with rank-based features.In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 895--898. ACM, 2015.

[LOP+13] Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei.Discovering tasks from search engine query logs.ACM Transactions on Information Systems (TOIS), 31(3):14, 2013.

[MC07] Donald Metzler and W Bruce Croft.Linear feature-based models for information retrieval.Information Retrieval, 10(3):257--274, 2007.

[MDC17] Bhaskar Mitra, Fernando Diaz, and Nick Craswell.Learning to match using local and distributed representations of text for web search.In Proceedings of the 26th International Conference on World Wide Web, pages 1291--1299. International World Wide Web Conferences Steering Committee, 2017.

References

[MNCC16] Bhaskar Mitra, Eric Nalisnick, Nick Craswell, and Rich Caruana.A dual embedding space model for document ranking.arXiv preprint arXiv:1602.01137, 2016.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean.Distributed representations of words and phrases and their compositionality.In Advances in neural information processing systems, pages 3111--3119, 2013.

[MSO13] Craig Macdonald, Rodrygo LT Santos, and Iadh Ounis.The whens and hows of learning to rank for web search.Information Retrieval, 16(5):584--628, 2013.

[MW08] David Milne and Ian H Witten.Learning to link with wikipedia.In Proceedings of the 17th ACM conference on Information and knowledge management, pages 509--518. ACM, 2008.

[MZ08] Alistair Moffat and Justin Zobel.Rank-biased precision for measurement of retrieval effectiveness.ACM Transactions on Information Systems (TOIS), 27(1):2, 2008.

[PC98] Jay M Ponte and W Bruce Croft.A language modeling approach to information retrieval.In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 275--281. ACM, 1998.

[QLL10] Tao Qin, Tie-Yan Liu, and Hang Li.A general approximation framework for direct optimization of information retrieval measures.Information retrieval, 13(4):375--397, 2010.

[RJ05] Filip Radlinski and Thorsten Joachims.Query chains: learning to rank from implicit feedback.In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pages 239--248. ACM, 2005.

References

[RKJ08] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims.Learning diverse rankings with multi-armed bandits.In Proceedings of the 25th international conference on Machine learning, pages 784--791. ACM, 2008.

[RS03] Yves Rasolofo and Jacques Savoy.Term proximity scoring for keyword-based retrieval systems.Advances in information retrieval, pages 79--79, 2003.

[RZT04] Stephen Robertson, Hugo Zaragoza, and Michael Taylor.Simple bm25 extension to multiple weighted fields.In Proceedings of the thirteenth ACM international conference on Information and knowledge management, pages 42--49. ACM, 2004.

[SB88] Gerard Salton and Christopher Buckley.Term-weighting approaches in automatic text retrieval.Information processing & management, 24(5):513--523, 1988.

[SCP16] Daria Sorokina and Erick Cantu-Paz.Amazon search: The joy of ranking products.In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pages 459--460. ACM, 2016.

[SHG+14] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Gregoire Mesnil.Learning semantic representations using convolutional neural networks for web search.In Proceedings of the 23rd International Conference on World Wide Web, pages 373--374. ACM, 2014.

[SJ72] Karen Sparck Jones.A statistical interpretation of term specificity and its application in retrieval.Journal of documentation, 28(1):11--21, 1972.

[SM15] Aliaksei Severyn and Alessandro Moschitti.Learning to rank short text pairs with convolutional deep neural networks.In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 373--382. ACM, 2015.

[SY11] Martin Szummer and Emine Yilmaz.Semi-supervised learning to rank with preference regularization.In Proceedings of the 20th ACM international conference on Information and knowledge management, pages 269--278. ACM, 2011.

References

[TBH15] Niek Tax, Sander Bockting, and Djoerd Hiemstra.A cross-benchmark comparison of 87 learning to rank methods.Information processing & management, 51(6):757--772, 2015.

[XLL+08] Jun Xu, Tie-Yan Liu, Min Lu, Hang Li, and Wei-Ying Ma.Directly optimizing evaluation measures in learning to rank.In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 107--114. ACM, 2008.

[YBKJ12] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims.The k-armed dueling bandits problem.Journal of Computer and System Sciences, 78(5):1538--1556, 2012.

[YFRJ07] Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims.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, pages 271--278. ACM, 2007.

[YHT+16] Dawei Yin, Yuening Hu, Jiliang Tang, Tim Daly, Mianwei Zhou, Hua Ouyang, Jianhui Chen, Changsung Kang, Hongbo Deng, Chikashi Nobata, et al.Ranking relevance in yahoo search.In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 323--332. ACM, 2016.

[YJ09] Yisong Yue and Thorsten Joachims.Interactively optimizing information retrieval systems as a dueling bandits problem.In Proceedings of the 26th Annual International Conference on Machine Learning, pages 1201--1208. ACM, 2009.

[YLKY07] Jen-Yuan Yeh, Jung-Yi Lin, Hao-Ren Ke, and Wei-Pang Yang.Learning to rank for information retrieval using genetic programming.In Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval (LR4IR 2007), 2007.

[YR09] Emine Yilmaz and Stephen Robertson.Deep versus shallow judgments in learning to rank.In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pages 662--663. ACM, 2009.