

# Learning More From Less

Towards Strengthening Weak Supervision for Ad-Hoc Retrieval

Dany Haddad\*

danyhaddad@utexas.edu

The University of Texas at Austin

## ABSTRACT

The limited availability of ground truth relevance labels has been a major impediment to the application of supervised methods to ad-hoc retrieval. As a result, unsupervised scoring methods, such as BM25, remain strong competitors to deep learning techniques which have brought on dramatic improvements in other domains, such as computer vision and natural language processing. Recent works have shown that it is possible to take advantage of the performance of these unsupervised methods to generate training data for learning-to-rank models. The key limitation to this line of work is the size of the training set required to surpass the performance of the original unsupervised method, which can be as large as  $10^{13}$  training examples. Building on these insights, we propose two methods to reduce the amount of training data required. The first method takes inspiration from crowdsourcing, and leverages multiple unsupervised rankers to generate soft, or noise-aware, training labels. The second identifies harmful, or mislabeled, training examples and removes them from the training set. We show that our methods allow us to surpass the performance of the unsupervised baseline with far fewer training examples than previous works.

## CCS CONCEPTS

• Information systems → Retrieval models and ranking.

## KEYWORDS

Information retrieval, Noisy Labels, Weak Supervision, Neural Network, Deep Learning

### ACM Reference Format:

Dany Haddad and Joydeep Ghosh. 2019. Learning More From Less: Towards Strengthening Weak Supervision for Ad-Hoc Retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France*. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3331184.3331272>

## 1 INTRODUCTION

Classical ad-hoc retrieval methods have relied primarily on unsupervised signals such as BM25, TF-IDF, and PageRank as inputs

\*Work done while interning at CognitiveScale.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '19, July 21–25, 2019, Paris, France

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6172-9/19/07...\$15.00

<https://doi.org/10.1145/3331184.3331272>

Joydeep Ghosh

ghosh@ece.utexas.edu

The University of Texas at Austin

to learning-to-rank (LEToR) models. Supervision for these models is often supplied in the form of click-stream logs or hand-curated rankings, both of which come with their issues and limitations. First, both sources are typically limited in availability and are often proprietary company resources. Second, click-stream data is typically biased towards the first few elements in the ranking presented to the user [2] and are noisy in general. Finally, such logs are only available after the fact, leading to a cold start problem. These issues motivate the search for an alternate source of “ground truth” ranked lists to train our LEToR model on.

In [7], Dehghani et al. show that the output of an unsupervised document retrieval method can be used to train a supervised ranking model that outperforms the original unsupervised ranker. More recently, [13] proposed a hierarchical interaction based model that is trained on a similarly generated training set. These works have shown the potential of leveraging unsupervised methods as sources of weak supervision for the retrieval task. However, they require training on as many as  $10^{13}$  training examples to surpass the performance of the unsupervised baseline [7, 13].

In this work, we substantially reduce this number by making more effective use of the generated training data. We present two methods that make improvements in this direction, and beat the unsupervised method using fewer than 10% of the training rankings compared to previous techniques.

In the first method, we take a crowdsourcing approach and collect the output of multiple unsupervised retrieval models. Following [19], we learn a joint distribution over the outputs of said retrieval models and generate a new training set of soft labels. We call this the *noise-aware* model. The *noise-aware* model does not require access to any gold labels<sup>1</sup>.

Our second method builds on the idea of dataset debugging and identifies training examples with the most harmful influence [10] (the labels most likely to be incorrect) and drops them from the training set. We call this the *influence-aware* model.

## 2 RELATED WORK

Much of the prior work in handling noisy datasets has been in the context of a classifier from noisy labels. In the binary classification context, noise is seen as a class-conditional probability that an observed label is the opposite of the true label [8, 14]. In the ranking context, we typically expect that models trained using pairwise or listwise loss functions will far outperform pointwise approaches [11]. Since the label of a pair is determined by the ordering of the documents within the pair, it is not immediately obvious how the class-conditional flip probabilities translate to this formulation. The relationship to listwise objectives is not straightforward either.

用无监督模型的结果作为有监督模型的数据标签，通常需要大量的训练数据才能使有监督模型的性能超过无监督模型。

<sup>1</sup>To differentiate them from labels originating from weak supervision sources, we refer to relevance scores assigned by a human as “gold” labels

In [5] and [6], Dehghani et al. introduce two semi-supervised student-teacher models where the teacher weights the contribution of each sample in the student model's training loss based on its confidence in the quality of the label. They train the teacher on a small subset of gold labels and use the model's output as confidence weights for the student model. [5] shows that using this approach, they can beat the unsupervised ranker using ~75% of the data required when training directly on the noisy data. They train a cluster of 50 gaussian processes to form the teacher annotations which are used to generate soft labels to fine-tune the student model.

In [19], Ratner et al. transform a set of weak supervision sources, that may disagree with each other, into soft labels used to train a discriminative model. They show experimentally that this approach outperforms the naïve majority voting strategy for generating the target labels. This inspires our *noise-aware* approach.

In [10], Koh et al. apply classical results from regression analysis to approximate the change in loss at a test point caused by removing a specific point from the training set. They show experimentally that their method approximates this change in loss well, even for highly non-linear models, such as GoogLeNet. They also apply their method to prioritize training examples to check for labeling errors. Our *influence-aware* approach uses influence functions [10] to identify mislabeled training examples.

### 3 PROPOSED METHODS

#### 3.1 Model Architecture

In this work, we only explore pairwise loss functions since they typically lead to better performing models than the pointwise approach. Listwise approaches, although typically the most effective, tend to have high training and inference time computational complexity due to their inherently permutation based formulations [11].

We consider a slight variant of the *Rank* model proposed in [7] as our baseline model. We represent the tokens in the  $i^{th}$  query as  $t_i^q$  and the tokens in the  $i^{th}$  document as  $t_i^d$ . We embed these tokens in a low dimensional space with a mapping  $E : \mathcal{V} \mapsto \mathbb{R}^l$  where  $\mathcal{V}$  is the vocabulary and  $l$  is the embedding dimension. We also learn token dependent weights  $W : \mathcal{V} \mapsto \mathbb{R}$ . Our final representation for a query  $q$  is a weighted sum of the word embeddings:  $v_q = \sum_{t_j \in t^q} \tilde{W}_q(t_j^q)E(t_j^q)$  where  $\tilde{W}_q$  indicates that the weights are normalized to sum to 1 across tokens in the query  $q$  using a softmax operation. The vector representation for documents is defined similarly.

In addition, we take the difference and elementwise products of the document and query vectors and concatenate them into a single vector  $v_{q,d} = [v_q, v_d, v_q - v_d, v_q \odot v_d]$ . We compute the relevance score of a document,  $d$ , to a query,  $q$  by passing  $v_{q,d}$  through a feed-forward network with *ReLU* activations and scalar output. We use a *tanh* at the output of the *rank* model and use the raw logit scores otherwise. We represent the output of our model parameterized by  $\theta$  as  $f(x; \theta)$ .

Our training set  $\mathcal{Z}$  is a set of tuples  $z = (q, d_1, d_2, s_{q,d_1}, s_{q,d_2})$  where  $s_{q,d_i}$  is the relevance score of  $d_i$  to  $q$  given by the unsupervised ranker. The pairwise objective function we minimize is given by:

$$\mathcal{L}(\mathcal{Z}; \theta) = \sum_{z \in \mathcal{Z}} L(f(v_{q,d_1}; \theta) - f(v_{q,d_2}; \theta), rel_{q,(d_1, d_2)}) \quad (1)$$

$$L_{ce}(x, y) = y \cdot \log(\sigma(x)) + (1 - y) \cdot \log(1 - \sigma(x)) \quad (2)$$

$$L_{hinge}(x, y) = \max\{0, \epsilon - \text{sign}(y) \cdot x\} \quad (3)$$

Where  $rel_{q,(d_1, d_2)} \in [0, 1]$  gives the relative relevance of  $d_1$  and  $d_2$  to  $q$ .  $L$  is either  $L_{ce}$  or  $L_{hinge}$  for cross-entropy or hinge loss, respectively. The key difference between the *rank* and *noise-aware* models is how  $rel_{q,(d_1, d_2)}$  is determined. As in [7], we train the *rank* model by minimizing the max-margin loss and compute  $rel_{q,(d_1, d_2)}$  as  $\text{sign}(s_{q,d_1} - s_{q,d_2})$ .

Despite the results in [21] showing that the max-margin loss exhibits stronger empirical risk guarantees for ranking tasks using noisy training data, we minimize the cross-entropy loss in each of our proposed models for the following reasons: in the case of the *noise-aware* model, each of our soft training labels are a distribution over  $\{0, 1\}$ , so we seek to learn a calibrated model rather than one which maximizes the margin (as would be achieved using a hinge loss objective). For the *influence-aware* model, we minimize the cross-entropy rather than the hinge loss since the method of influence functions relies on having a twice differentiable objective.

#### 3.2 Noise-aware model

In this approach,  $rel_{q,(d_i, d_j)} \in [0, 1]$  are soft relevance labels. For each of the queries in the training set, we rank the top documents by relevance using  $k$  unsupervised rankers. Considering ordered pairs of these documents, each ranker gives a value of 1 if it agrees with the order, -1 if it disagrees and 0 if neither document appears in the top 10 positions of the ranking. We collect these values into a matrix  $\Lambda \in \{-1, 0, 1\}^{m \times k}$  for  $m$  document pairs. The joint distribution over these pairwise preferences and the true pairwise orderings  $y$  is given by:

$$P_w(\Lambda, y) = \frac{1}{Z(w)} \exp\left(\sum_i^m w^T \phi(\Lambda_i, y_i)\right) \quad (4)$$

Where  $w$  is a vector of learned parameters and  $Z(w)$  is the partition function. A natural choice for  $\phi$  is to model the accuracy of each individual ranker in addition to the pairwise correlations between each of the rankers. So for the  $i^{th}$  document pair, we have the following expression for  $\phi_i := \phi(\Lambda_i, y_i)$ :

$$\phi_i = [\{\Lambda_{ij} = y_i\}_{1 \leq j \leq k} || \{\Lambda_{ij} = \Lambda_{il} \neq 0\}_{j \neq l}]$$

Since the true relevance preferences are unknown, we treat them as latent. We learn the parameters for this model without any gold relevance labels  $y$  by maximizing the marginal likelihood (as in [19]) given by:

$$\max_w \log \sum_y P_w(\Lambda, y) \quad (5)$$

We use the Snorkel library<sup>2</sup> to optimize equation 5 by stochastic gradient descent, where we perform Gibbs sampling to estimate the gradient at each step. Once we have determined the parameters of the model, we can evaluate the posterior probabilities  $P_w(y_i | \Lambda_i)$  which we use as our soft training labels.

<sup>2</sup><https://github.com/HazyResearch/snorkel>

### 3.3 Influence Aware Model

In this approach, we identify training examples that hurt the generalization performance of the trained model. We expect that many of these will be incorrectly labeled, and that our model will perform better if we drop them from the training set. The influence of removing a training example  $z_i = (x_i, y_i)$  on the trained model's loss at a test point  $z_{test}$  is computed as [10]:

$$\Delta L(z_{test}; \theta) \approx \mathcal{I}_{drop}(z_i, z_{test}) \quad (6)$$

$$= \frac{1}{n} \nabla_\theta L(z_{test}; \theta)^T H_\theta^{-1} \nabla_\theta L(z_i; \theta) \quad (7)$$

where  $H_\theta$  is the Hessian of the objective function. If  $\mathcal{I}_{drop}(z_i, z_{test})$  is negative, then  $z_i$  is a harmful training example for  $z_{test}$  since its inclusion in the training set causes an increase in the loss at that point. Summing this value over the entire test set gives us  $\mathcal{I}_{drop}(z_i)$ . We compute  $\mathcal{I}_{drop}(z_i)$  for each training example  $z_i$ , expecting it to represent  $z_i$ 's impact on the model's performance at test time. In our setup, we know that some of our training examples are mislabeled; we expect that these points will have a large negative value for  $\mathcal{I}_{drop}$ . Of course, for a fair evaluation, the  $z_{test}$  points are taken from the development set used for hyperparameter tuning (see section 4).

We address the computational constraints of computing (7) by treating our trained model as a logistic regression on the bottleneck features. We freeze all model parameters except the last layer of the feed-forward network and compute the gradient with respect to these parameters only. This gradients can be computed in closed form in an easily parallelizable way, allowing us to avoid techniques that rely on autodifferentiation operations [16]. We compute  $H_\theta^{-1} \nabla_\theta L(z_{test}; \theta)$  for every  $z_{test}$  using the method of conjugate gradients following [20]. We also add a small damping term to the diagonal of the Hessian to ensure that it is positive definite [12].

## 4 DATA PREPROCESSING AND MODEL TRAINING

We evaluate the application of our methods to ad-hoc retrieval on the Robust04 corpus with the associated test queries and relevance labels. As in [7], our training data comes from the AOL query logs [15] on which we perform similar preprocessing. We use the Indri<sup>3</sup> search engine to conduct indexing and retrieval and use the default parameters for the query likelihood (QL) retrieval model [18] which we use as the weak supervision source. We fetch only the top 10 documents from each ranking in comparison to previous works which trained on as many as the top 1000 documents for each query. To compensate for this difference, we randomly sample  $n_{neg}$  additional documents from the rest of the corpus for each of these 10 documents. We train our model on a random subset of 100k rankings generated by this process. This is fewer than 10% the number of rankings used in previous works [7, 13], each of which contains far fewer document pairs.

<sup>3</sup><https://www.lemurproject.org/indri.php>

Table 1: Results comparison with smoothing.

	Rank Model	Noise-Aware	Influence-Aware	QL
NDCG@10	0.3881	† 0.3952	† <b>0.4008</b>	0.3843
Prec@10	0.3535	† 0.3621	† <b>0.3657</b>	0.3515
MAP	0.2675	† 0.2774	† <b>0.2792</b>	0.2676

For the word embedding representations,  $W$ , we use the 840B.300d GloVe [17] pretrained word embedding set<sup>4</sup>. The feed-forward network hidden layer sizes are chosen from {512, 256, 128, 64} with up to 5 layers. We use the first 50 queries in the Robust04 dataset as our development set for hyperparameter selection, computation of  $\mathcal{I}_{drop}$  and early stopping. The remaining 200 queries are used for evaluation.

During inference, we rank documents by the output of the feed-forward network. Since it is not feasible to rank all the documents in the corpus, we fetch the top 100 documents using the QL retrieval model and then rerank using the trained model's scores.

## 4.1 Model Specific Details

For the *noise-aware* model, we generate separate rankings for each query using the following retrieval methods: Okapi BM25, TF-IDF, QL, QL+RM3 [1] using Indri with the default parameters.

For the *influence-aware* model, we train the model once on the full dataset and then compute  $\mathcal{I}_{drop}(z_i)$  for each training point dropping all training examples with a negative value for  $\mathcal{I}_{drop}(z_i)$  which we find to typically be around half of the original training set. We then retrain the model on this subset.

Interestingly, we find that using a smaller margin,  $\epsilon$ , in the training loss of the *rank* model leads to improved performance. Using a smaller margin incurs 0 loss for a smaller difference in the model's relative preference between the two documents. Intuitively, this allows for less overfitting to the noisy data. We use a margin of 0.1 chosen by cross-validation.

The *noise-aware* and *influence-aware* models train end-to-end in around 12 and 15 hours respectively on a single NVIDIA Titan Xp.

## 5 EXPERIMENTAL RESULTS

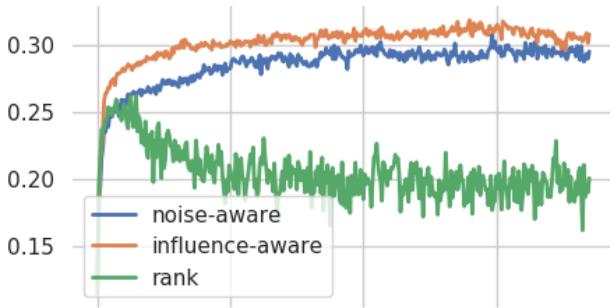
We compare our two methods against two baselines, the unsupervised ranker (QL) and the *rank* model. Compared to the other unsupervised rankers (see section 4.1) used as input to the *noise-aware* model, the QL ranker performs the best on all metrics. Training the *rank* model on the results of the majority vote of the set of unsupervised rankers used for the *noise-aware* model performed very similarly to the *rank* model, so we only report results of the *rank* model. We also compare the results after smoothing with the normalized QL document scores by linear interpolation.

The results in tables 1 and 2 show that the *noise-aware* and *influence-aware* models perform similarly, with both outperforming the unsupervised baseline. Bold items are the largest in their row and daggers indicate statistically significant improvements over the *rank* model at a level of 0.05 using Bonferroni correction. Figure 1 shows that the *rank* model quickly starts to overfit. This does

<sup>4</sup><https://nlp.stanford.edu/projects/glove/>

**Table 2: Results comparison without smoothing.**

	Rank Model	Noise-Aware	Influence-Aware
NDCG@10	0.2610	† 0.2886	† <b>0.2966</b>
Prec@10	0.2399	† <b>0.2773</b>	† 0.2742
MAP	0.1566	† 0.1831	† <b>0.1839</b>

**Figure 1: Test NDCG@10 during training**

not contradict the results in [7] since in our setup we train on far fewer pairs of documents for each query, so each relevance label error has much greater impact. For each query, our distribution over documents is uniform outside the results from the weak supervision source, so we expect to perform worse than if we had a more faithful relevance distribution. Our proposed approaches use an improved estimate of the relevance distribution at the most important positions in the ranking, allowing them to perform well.

We now present two representative training examples showing how our methods overcome the limitations of the *rank* model.

**Example 5.1.** The method in section 3.2 used to create labels for the *noise-aware* model gives the following training example an unconfident label (~0.5) rather than a relevance label of 1 or 0: ( $q = \text{"town of davie post office"}$ ,  $(d_1=\text{FBIS3-25584}, d_2=\text{FT933-13328})$ ) where  $d_1$  is ranked above  $d_2$ . Both of these documents are about people named “Davie” rather than about a town or a post office, so it is reasonable to avoid specifying a hard label indicating which one is explicitly more relevant.

**Example 5.2.** One of the most harmful training points as determined by the method described in section 3.3 is the pair ( $q = \text{"pictures of easter mice"}$ ,  $(d_1=\text{FT932-15650}, d_2=\text{LA041590-0059})$ ) where  $d_1$  is ranked above  $d_2$ .  $d_1$  discusses the computer input device and  $d_2$  is about pictures that are reminiscent of the holiday. The incorrect relevance label explains why the method identifies this as a harmful training example.

## 6 CONCLUSIONS AND FUTURE WORK

We have presented two approaches to reduce the amount of weak data needed to surpass the performance of the unsupervised method that generates the training data. The *noise-aware* model does not require ground truth labels, but has an additional data dependency on multiple unsupervised rankers. The *influence-aware* model requires a small set of gold-labels in addition to a re-train of the model,

although empirically, only around half the dataset is used when training the second time around.

Interesting paths for future work involve learning a better joint distribution for training the *noise-aware* model or leveraging ideas from [22] to construct soft training labels rather than for the query performance prediction task. Similarly, we could apply ideas from unsupervised LETOR [4] to form better noise-aware labels. For the *influence-aware* model, we could use the softmax loss [3] rather than cross-entropy and instead compute set influence rather than the influence of a single training example [9].

## REFERENCES

- [1] Nasreen Abdul-Jaleel, James Allan, Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Donald Metzler, Mark D. Smucker, Trevor Strohman, Howard Turtle, and Courtney Wade. 2004. Umass at trec 2004: Notebook. *academia.edu* (2004).
- [2] Qingyao Ai, Keping Bi, Cheng Luo, Jiafeng Guo, and W Bruce Croft. 2018. Unbiased Learning to Rank with Unbiased Propensity Estimation. In *The 41st International ACM SIGIR Conference*. ACM Press, New York, New York, USA, 385–394. <https://doi.org/10.1145/3209978.3209986>
- [3] Ricardo Baeza-Yates, Berthier de Araújo Neto Ribeiro, et al. 2007. Learning to rank with nonsmooth cost functions. *NIPS* (2007).
- [4] Avradeep Bhowmik and Joydeep Ghosh. 2017. LETOR Methods for Unsupervised Rank Aggregation. In *The 26th International Conference*. ACM Press, New York, New York, USA, 1331–1340. <https://doi.org/10.1145/3038912.3052689>
- [5] Mostafa Dehghani, Arash Mehrjou, Stephan Gouws, Jaap Kamps, and Bernhard Schölkopf. 2017. Fidelity-Weighted Learning. *arXiv.org* (Nov. 2017). arXiv:cs.LG/1711.02799v2
- [6] Mostafa Dehghani, Aliaksei Severyn, Sascha Rothe, and Jaap Kamps. 2017. Learning to Learn from Weak Supervision by Full Supervision. *arXiv.org* (Nov. 2017), 1–8. arXiv:1711.11383
- [7] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural Ranking Models with Weak Supervision. In *The 40th International ACM SIGIR Conference*. ACM Press, New York, New York, USA, 65–74. <https://doi.org/10.1145/3077136.3080832>
- [8] Xinxin Jiang, Shirui Pan, Guodong Long, Fei Xiong, Jing Jiang, and Chengqi Zhang. 2017. Cost-sensitive learning with noisy labels. *JMLR* (2017).
- [9] Rajiv Khanna, Been Kim, Joydeep Ghosh, and Oluwasanmi Koyejo. 2018. Interpreting Black Box Predictions using Fisher Kernels. *arXiv.org* (Oct. 2018). arXiv:cs.LG/1810.10118v1
- [10] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. *arXiv.org* (March 2017), 1–11. arXiv:1703.04730
- [11] Tie-Yan Liu. 2009. Learning to Rank for Information Retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331. <https://doi.org/10.1561/1500000016>
- [12] James Martens. 2010. Deep learning via Hessian-free optimization. (2010).
- [13] Yifan Nie, Alessandro Sordoni, and Jian-Yun Nie. 2018. Multi-level Abstraction Convolutional Model with Weak Supervision for Information Retrieval. In *The 41st International ACM SIGIR Conference*. ACM Press, New York, New York, USA, 985–988. <https://doi.org/10.1145/3209978.3210123>
- [14] Curtis G Northcutt, Tailin Wu, and Isaac L Chuang. 2017. Learning with Confident Examples: Rank Pruning for Robust Classification with Noisy Labels. *arXiv.org* (May 2017). arXiv:1705.01936
- [15] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A Picture of Search. *InfoScale* (2006), 1–es. <https://doi.org/10.1145/1146847.1146848>
- [16] Barak Pearlmutter. 1994. Fast exact multiplication by the Hessian. *MIT Press* 6, 1 (Jan. 1994), 147–160. <https://doi.org/10.1162/neco.1994.6.1.147>
- [17] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2015. GloVe: Global Vectors for Word Representation.
- [18] Jay M Ponte and W Bruce Croft. 1998. A Language Modeling Approach to Information Retrieval. *SIGIR* (1998), 275–281. <https://doi.org/10.1145/290941.291008>
- [19] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel. *Proceedings of the VLDB Endowment* 11, 3 (Nov. 2017), 269–282. <https://doi.org/10.14778/3157794.3157797>
- [20] Jonathan R Shewchuk. 1994. An introduction to the conjugate gradient method without the agonizing pain. (1994).
- [21] Hamed Zamani and W Bruce Croft. 2018. *On the Theory of Weak Supervision for Information Retrieval*. ACM, New York, New York, USA. <https://doi.org/10.1145/3234944.3234968>
- [22] Hamed Zamani, W Bruce Croft, and J Shane Culpepper. 2018. Neural Query Performance Prediction using Weak Supervision from Multiple Signals. In *The 41st International ACM SIGIR Conference*. ACM Press, New York, New York, USA, 105–114. <https://doi.org/10.1145/3209978.3210041>