On the Theory of Weak Supervision for Information Retrieval

Hamed Zamani
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
Amherst, MA 01003
zamani@cs.umass.edu

W. Bruce Croft
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
Amherst, MA 01003
croft@cs.umass.edu

ABSTRACT

Neural network approaches have recently shown to be effective in several information retrieval (IR) tasks. However, neural approaches often require large volumes of training data to perform effectively, which is not always available. To mitigate the shortage of labeled data, training neural IR models with weak supervision has been recently proposed and received considerable attention in the literature. In weak supervision, an existing model automatically generates labels for a large set of unlabeled data, and a machine learning model is further trained on the generated "weak" data. Surprisingly, it has been shown in prior art that the trained neural model can outperform the weak labeler by a significant margin. Although these obtained improvements have been intuitively justified in previous work, the literature still lacks theoretical justification for the observed empirical findings. In this paper, we provide a theoretical insight into weak supervision for information retrieval, focusing on learning to rank. We model the weak supervision signal as a noisy channel that introduces noise to the correct ranking. Based on the risk minimization framework, we prove that given some sufficient constraints on the loss function, weak supervision is equivalent to supervised learning under uniform noise. We also find an upper bound for the empirical risk of weak supervision in case of non-uniform noise. Following the recent work on using multiple weak supervision signals to learn more accurate models, we find an information theoretic lower bound on the number of weak supervision signals required to guarantee an upper bound for the pairwise error probability. We empirically verify a set of presented theoretical findings, using synthetic and real weak supervision data.

ACM Reference Format:
Hamed Zamani and W. Bruce Croft. 2018. On the Theory of Weak Supervision for Information Retrieval. In *2018 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '18), September 14–17, 2018, Tianjin, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3234944.3234968

1 INTRODUCTION

Neural network models have recently shown promising results in a number of information retrieval (IR) tasks, including ad-hoc retrieval [15, 30], passage retrieval [6], and context-aware ranking [31]. Neural approaches often require a large volume of training data to perform effectively. Although large-scale relevance signals, e.g., clickthrough data, are available for a few IR tasks, e.g., web search, this data is not available for many real-world problems and domains. Moreover, academia and smaller companies also suffer from lack of access to large-scale labeled data or implicit user feedback. This is critical for fields, such as information retrieval, that have been developed based on extensive and accurate evaluations. The aforementioned limitations call for developing effective learning approaches to mitigate the shortage of training data. In this line of research, weak supervision has been proposed to train neural models for information retrieval tasks, such as learning to rank documents in the context of ad-hoc retrieval [12] and learning relevance-based word embedding [32]. The substantial improvements achieved by weakly supervised IR models have recently attracted attention of the IR community [11, 23, 29, 33]. Although the obtained improvements have been intuitively well justified in previous work [12, 32], to the best of our knowledge, no theoretical justification has been proposed to support the empirical findings from weak supervision in information retrieval.

To close the gap between theory and practice, this paper theoretically studies weak supervision in information retrieval to better understand how and why this learning strategy works. We build our theory upon the risk minimization framework, and model weak supervision as a noisy channel that introduces some noise to the true ranking labels. We further define *symmetric ranking loss functions*. Our major theoretical findings are summarized below:

- Assuming the noise distribution in the weak supervision noisy channel being uniform, we prove that risk minimization for symmetric ranking loss functions is noise tolerant. Informally, under the uniformity assumption, the globally optimum ranking model for weakly supervised data is also globally optimum for the true labeled data.
- For non-uniform noise distribution in the weak supervision noisy channel, we prove that if the risk function for the globally optimum ranking model on the true data is equal to zero, then again it is also the globally optimum ranking model on the weak data.
- For non-uniform noise distribution in the weak supervision noisy channel, we find an upper bound for the risk function for an optimum model trained on weak data based on the risk value of the optimum model trained on the true data. The upper bound is inversely correlated with the maximum error probability of the weak signal.
- We find a sufficient (but not necessary) constraint for pairwise loss functions to be symmetric ranking loss. We study a set of well-known pairwise loss functions and show that hinge loss and mean absolute error (i.e., L_1 loss) if used in pairwise setting satisfy

ICTIR '18, September 14–17, 2018, Tianjin, China

ACM ISBN 978-1-4503-5656-5/18/09…$15.00
https://doi.org/10.1145/3234944.3234968

the proposed sufficient constraint. On contrary, cross entropy and mean square error do not satisfy the proposed constraint.
- Following the recent empirical findings by Zamani et al. [33] on employing multiple weak supervision signals for learning more accurate models, we theoretically investigate the task of learning from multiple weak signals. Our theoretical study results in an information theoretic lower bound for the number of independent weak supervision signals required to guarantee the pairwise error being bounded by an arbitrary parameter ϵ.

The aforementioned theoretical findings provide insights into how and why training models on weakly supervised data can perform well and even outperform the weak labeler. They also introduce some guidelines on what loss functions and how many weak supervision signals to use, while training on weakly supervised data. We finally empirically study a set of our theoretical findings using synthetic and real weak supervision data. Our theoretical analysis can be easily generalized to different types of noisy labels, such as those collected via crowdsourcing.

2 RELATED WORK

Limited training data has been a perennial problem in information retrieval, and many machine learning-related domains [34]. This has motivated researchers to explore building models using *pseudo-labels*. For example, pseudo-relevance feedback (PRF) [2, 8] assumes that the top retrieved documents in response to a given query are relevant to the query. Although this assumption does not necessarily hold, PRF has been proven to be effective in many retrieval settings. Building pseudo-collections and simulated queries for various IR tasks, could be considered as another set of approaches that tackle this issue [1, 3].

As widely known, deep neural networks often require large volumes of training data. Recently, training neural IR models based on pseudo-labels (i.e., weak supervision) has shown to produce successful results [12, 32]. Dehghani et al. [12] proposed training a neural ranking model for the ad-hoc retrieval task based on the labels generated by an existing retrieval model, such as BM25. Zamani and Croft [32] argued that the objective functions of the general-purpose word embedding models, such as word2vec [22], are not necessarily equivalent to the objective that we seek in information retrieval. They proposed training of relevance-based word embeddings based on relevance models [20] as the weak label. Following these studies, the idea of training neural IR models with weak supervision has been further employed in [23, 29, 33].

In the realm of machine learning, learning from noisy data is a challenging and at the same time an important task. For example, training models on crowdsourced data, which is often noisy, is relatively well studied in the literature [18]. Ghosh et al. [14] theoretically studied the robustness of binary classification model to random noise. It has been recently generalized to multi-class classification [13]. In this area, Bekker and Goldberger [4] proposed a model for learning from noisy data by learning the noise distribution in addition to optimizing the neural network parameters. Dehghani et al. [10] suggested to weight noisy instances using a limited amount of labeled data. Recently, Rolnick et al. [27] showed that deep neural networks can be robust to multiple noise patterns in computer vision tasks. Although, learning from implicit feedback can be also considered as noisy data, it has been shown that implicit feedback provided by users, e.g., clickthrough data, is a strong signal for search engines [16]. The main challenge in learning from implicit feedback data is how to learn unbiased model [17], due to different biases in user behaviours, e.g., positional bias. Moreover, boosting [28] has been extensively used in the machine learning community to learn multiple weak learners from supervised data. In addition, co-training [5], generalized expectation [21], and other semi-supervised learning approaches are examples of popular techniques for dealing with limited data in machine learning. These lines of research often require a small set of labeled data and are out of the scope of this paper.

3 BACKGROUND

3.1 Learning to Rank Formulation

Suppose $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space for a ranking problem. Each element of $\mathcal{X}$, denoted as $\mathbf{x}$, is a list of n feature vectors corresponding to n objects that should be ranked, i.e., $\mathbf{x} = [x_1, x_2, \cdots, x_n]$. Each element of $\mathcal{Y}$, denoted as $\mathbf{y}$, is also a list of n labels corresponding to the objects appeared in $\mathbf{x}$, i.e., $\mathbf{y} = [y_1, y_2, \cdots, y_n]$. Let $P(X, Y)$ be an unknown joint distribution, where random variables X and Y respectively take $\mathbf{x}$ and $\mathbf{y}$ as their values.

Assume that $\mathcal{M}$ is a ranking model that takes $\mathbf{x}$ as input and generates a rank list from the objects appearing in $\mathbf{x}$. Without loss of generality, we assume that $\mathcal{M}$ produces a score for each object in $\mathbf{x}$ and the objects are then sorted based on their scores in descending order. Therefore, $\mathcal{M}(\mathbf{x})$ can be written as $[\mathcal{M}(x_1), \mathcal{M}(x_2), \cdots, \mathcal{M}(x_n)]$.

In typical statistical learning to rank problems, we are given a training set $\mathcal{T} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \cdots, (\mathbf{x}_m, \mathbf{y}_m)\}$ with m elements, each representing a rank list for n objects. The training instances are assumed to be drawn *iid* according to an unknown distribution over $\mathcal{X} \times \mathcal{Y}$. In document ranking, e.g., ad-hoc retrieval, $\mathbf{x}_i$ is equal to $\{(q_i, d_{i1}), (q_i, d_{i2}), \cdots, (q_i, d_{in})\}$, where q_i denotes the i^{th} query in the training set and d_{ij} denotes the j^{th} candidate document that should be ranked in response to the query q_i. $\mathbf{y}_i$ is also a list of n labels representing the relevance judgments for the corresponding query-document pairs.

3.2 Risk Minimization Framework

In this subsection, we briefly explain the risk minimization framework in statistical learning. The risk function for a given ranking model $\mathcal{M}$ and loss function $\mathcal{L}$ is defined as follows:

$$R_{\mathcal{L}}(\mathcal{M}) = \mathbb{E}_P[\mathcal{L}(\mathcal{M}(\mathbf{x}), \mathbf{y})] = \int \int \mathcal{L}(\mathcal{M}(\mathbf{x}), \mathbf{y}) P(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\mathbf{y} \quad (1)$$

where $\mathbb{E}$ denotes expectation and its subscript is a distribution (or a random variable) with respect to which the expectation is taken. As mentioned earlier in Section 3.1, P denotes an unknown true distribution over the $\mathcal{X} \times \mathcal{Y}$ space. $\mathcal{L}(\cdot, \cdot)$ is a loss function that computes the difference between its inputs which are two lists with the size of n.

Given the training data $\mathcal{T}$, the empirical risk is defined as follows:

$$\widehat{R}_{\mathcal{L}}(\mathcal{M}) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \in \mathcal{T}}[\mathcal{L}(\mathcal{M}(\mathbf{x}), \mathbf{y})] = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(\mathcal{M}(\mathbf{x}_i), \mathbf{y}_i) \quad (2)$$

Under the risk minimization framework, the objective is to learn a ranking model $\mathcal{M}$ that is a global minimizer of the risk function $\widehat{R}_{\mathcal{L}}$, which depends on the ranking loss function $\mathcal{L}$.[1]

3.3 Information Theory Preliminaries

In this subsection, we define a number of information theory concepts and lemmas used througout this paper.

Definition 1. [Total Variation Distance] The total variation distance between two probability measures P and Q with the same σ-algebra $\mathcal{F}$ is defined as:

$$\delta_{TV}(P, Q) = \sup_{A \in \mathcal{F}} \{P(A) - Q(A)\} \tag{3}$$

Informally, total variation distance is the largest possible difference between the probabilities that the two probability distributions can assign to the same measurable set.

Definition 2. [Kullback-Leibler Divergence] The Kullback-Leibler divergence (KL divergence), between two probability measures P and Q on a set $\mathcal{X}$ is defined as:

$$D(P||Q) = \int_{\mathcal{X}} dP \log \frac{dP}{dQ} \tag{4}$$

If P and Q are distributions of continuous random variables, then KL divergence is computed as:

$$\int_{-\infty}^{\infty} f_P(x) \log \frac{f_P(x)}{f_Q(x)} \, \mathrm{d}x \tag{5}$$

where f_P and f_Q denote the probability density functions of P and Q, respectively.

The KL divergence for discrete random variables is also computed as:

$$\sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \tag{6}$$

where p and q are two probability mass functions.

Lemma 1. [Chain Rule of KL Divergence] Let $X_1, X_2, \cdots, X_k$ be k discrete random variables, and let P and Q be two joint distributions of these random variables. Then, the KL divergence of the joint distribution is computed as:

$$D(P(x_1, x_2, \cdots, x_k)||Q(x_1, x_2, \cdots, x_k)) = \sum_{i=1}^{k} D(P(x_i|x_1, x_2, \cdots, x_{i-1})||Q(x_i|x_1, x_2, \cdots, x_{i-1})) \tag{7}$$

Remark 1. Let $X_1, X_2, \cdots, X_k$ denote k independent discrete random variables. Then, the KL divergence of two joint distributions P and Q is computed as:

$$D(P(x_1, x_2, \cdots, x_k)||Q(x_1, x_2, \cdots, x_k)) = \sum_{i=1}^{k} D(P(x_i)||Q(x_i)) \tag{8}$$

Lemma 2. [Pinsker's Inequality] For any two probability measures P and Q,

$$\delta_{TV}(P, Q) \leq \sqrt{\frac{1}{2} D(P||Q)} \tag{9}$$

[1]Since ranking metrics are often non-differentiable, a surrogate loss function is often used. We can assume that $\mathcal{L}$ is a surrogate ranking loss.

where $D(\cdot||\cdot)$ denotes the KL divergence between the two given probability measures.

4 LEARNING TO RANK FROM WEAK SIGNALS

In this section, we present a set of theoretical results on learning to rank from weakly supervised data. In the following, we first formulate the problem, and then define symmetric ranking loss functions. We further study the problem under uniformity and non-uniformity assumptions. Finally, we study a set of pairwise loss functions to determine which ones are symmetric ranking losses.

4.1 Problem Statement

Weak supervision is a learning strategy that does not require labeled training data. Let M be a weak supervision signal that automatically produces a label for any given input. This gives us a set of weakly labeled training data $\widehat{\mathcal{T}} = \{(\mathbf{x}_1, \widehat{\mathbf{y}}_1), (\mathbf{x}_2, \widehat{\mathbf{y}}_2), \cdots, (\mathbf{x}_m, \widehat{\mathbf{y}}_m)\}$, where $\widehat{\mathbf{y}}_i = [M(x_{i1}), M(x_{i2}), \cdots, M(x_{in})]$. The input feature vectors $\mathbf{x}$ in weakly labeled data are the same as the ones in $\mathcal{T}$ described in Section 3.1.

The weakly supervised labels are generated automatically and thus are not accurate. In the following subsections, we theoretically study how to effectively train a ranking model on such noisy labels. Without loss of generality, we can look at weak supervision as a noisy channel that applies some noise on the actual true labels. Therefore, $\widehat{\mathbf{y}} = M(\mathbf{x})$ is modeled as a noisy channel that introduces noise on the true label $\mathbf{y}$.

Let us first define noise tolerance in ranking based on the risk minimization framework as follows:

Definition 3. [Noise Tolerance in Learning to Rank] Under the ranking loss function $\mathcal{L}$, risk minimization is noise tolerant if

$$Pr[\mathcal{M}(\mathbf{x}) \overset{\text{rank}}{=} \mathbf{y}] = Pr[\widehat{\mathcal{M}}(\mathbf{x}) \overset{\text{rank}}{=} \mathbf{y}] \qquad \forall (\mathbf{x}, \mathbf{y}) \in \mathcal{T} \tag{10}$$

where $\cdot \overset{\text{rank}}{=} \cdot$ denotes that the two given score lists are equal in terms of ranking (i.e., the corresponding objects are in the same order when sorted by their scores). $\mathcal{M}$ and $\widehat{\mathcal{M}}$ respectively denote the ranking models trained on the true data $\mathcal{T}$ and the weakly supervised data $\widehat{\mathcal{T}}$.

4.2 Symmetric Ranking Loss

Motivated by the symmetry condition defined for binary classification [14], in this subsection, we define symmetric loss function, which is heavily used in the paper.

Definition 4. [Symmetric Ranking Loss Function] A ranking loss function $\mathcal{L}$ is symmetric, if it satisfies the following constraint:

$$\sum_{\mathbf{y} \in \mathcal{Y}} \mathcal{L}(\mathcal{M}(\mathbf{x}), \mathbf{y}) = c \quad \forall \mathbf{x}, \forall \mathcal{M} \tag{11}$$

where c is a constant number.

Note that the above definition assumes that the output space $\mathcal{Y}$ is finite and discrete, which is a reasonable assumption for a ranking task, where the order of objects matters and the number of items is finite (i.e., equal to n in our setting). In case of binary relevance judgments, the output space $\mathcal{Y}$ is $\{0, 1\}^n$, thus $|\mathcal{Y}| = 2^n$.

Theorem 1. In case of binary relevance judgments (labels), any ranking loss function $\mathcal{L}$ based on a pairwise classification loss $\mathcal{L}_{pair}$ is symmetric, if the following condition, called the sufficient symmetric pairwise loss (SSPL) constraint, holds:

$$\mathcal{L}_{pair}(\mathcal{M}(x)-\mathcal{M}(x'),-1)+\mathcal{L}_{pair}(\mathcal{M}(x)-\mathcal{M}(x'),1)=c \quad \forall x,x',\forall \mathcal{M} \tag{12}$$

where c is a constant number.[2]

Proof. If the ranking loss function $\mathcal{L}$ is computed based on pairwise misclassifications, then without loss of generality, we have:

$$\begin{aligned}\sum_{\mathbf{y}\in\mathcal{Y}}\mathcal{L}(\mathcal{M}(\mathbf{x}),\mathbf{y}) &= \frac{1}{Z}\sum_{\mathbf{y}\in\mathcal{Y}}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathcal{L}_{pair}(\mathcal{M}(x_i)-\mathcal{M}(x_j),y_i-y_j)\\ &= \frac{1}{Z}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\sum_{\mathbf{y}\in\mathcal{Y}}\mathcal{L}_{pair}(\mathcal{M}(x_i)-\mathcal{M}(x_j),y_i-y_j)\end{aligned} \tag{13}$$

where Z is a constant normalization factor. Since the labels are binary, thus $\mathcal{Y}=\{0,1\}^n$. The terms inside the summations in Equation (13) only depend on the i^{th} and the j^{th} element of $\mathbf{y}$. Therefore, we can rewrite the above equations as below:

$$\begin{aligned}&= \frac{2^{n-2}}{Z}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\sum_{(y_i,y_j)\in\{0,1\}^2}\mathcal{L}_{pair}(\mathcal{M}(x_i)-\mathcal{M}(x_j),y_i-y_j)\\ &= \frac{2^{n-2}}{Z}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathcal{L}_{pair}(\mathcal{M}(x_i)-\mathcal{M}(x_j),-1)-\mathcal{L}_{pair}(\mathcal{M}(x_i)-\mathcal{M}(x_j),1)\end{aligned} \tag{14}$$

Note that the pairwise loss for two objects with the same labels is assumed to be zero. Given the SSPL condition mentioned in Equation (12), we rewrite the above equation as:

$$\Rightarrow \sum_{\mathbf{y}\in\mathcal{Y}}\mathcal{L}(\mathcal{M}(\mathbf{x}),\mathbf{y})=\frac{2^{n-2}}{Z}\sum_{i=1}^{n}\sum_{j=i+1}^{n}c=c' \tag{15}$$

which shows that $\mathcal{L}$ is a symmetric ranking function, and completes the proof. □

In Section 4.5, we study a number of well-known pairwise loss functions and discuss whether they satisfy the SSPL constraint. At this step, it is sufficient to know that there exist some loss functions that satisfy the SSPL constraint, and thus the following theoretical findings are useful in practice.

4.3 Weak Supervision as Uniform Noisy Channel

In this subsection, we generalize Ghosh et al.'s findings [14] on binary classification to ranking and we assume that independent from the input $\mathbf{x}$, the noisy channel applies a *uniform* noise on the true label and produces the weak label. Although, this is a strong assumption that does not hold in many real-world situations, it gives insights into understanding learning from weak supervision, and is a first step towards more complex situations (e.g., for the non-uniformity assumption).

[2]We assume that the pairwise loss for two objects with the same label is zero (or constant), otherwise the condition on this theorem should be modified.

Theorem 2. In learning to rank from weak supervision, where the weak signal is drawn from the true label with a uniform noise, let $\mathcal{L}$ be a symmetric ranking loss function (see Definition 4). Then, $\mathcal{L}$ is noise tolerant (see Definition 3) under the noise probability $\rho < \frac{2^n-1}{2^n}$ (meaning that the weak supervision signal performs better than random).

Proof. We prove this theorem based on the risk minimization framework. To do so, we show that any ranking model $\widehat{\mathcal{M}}^*$ that is a global minimizer for the empirical risk function on the weak data is also a global minimizer of the true empirical risk.

The empirical risk function for a ranking model $\mathcal{M}$ over the weak data $\widehat{\mathcal{T}}$ is defined as:

$$\begin{aligned}\widehat{R}'_{\mathcal{L}}(\mathcal{M}) &= \mathbb{E}_{(\mathbf{x},\widehat{\mathbf{y}})\in\widehat{\mathcal{T}}}[\mathcal{L}(\mathcal{M}(\mathbf{x}),\widehat{\mathbf{y}})]\\ &= \mathbb{E}_{\mathbf{x}}\,\mathbb{E}_{\mathbf{y}|\mathbf{x}}\,\mathbb{E}_{\widehat{\mathbf{y}}|\mathbf{x},\mathbf{y}}[\mathcal{L}(\mathcal{M}(\mathbf{x}),\widehat{\mathbf{y}})]\end{aligned} \tag{16}$$

Given the uniform noise assumption, the expected value of the loss for the weak supervision label is equal to the loss for the true label with the probability of $1-\rho$ and the probability of $\frac{\rho}{2^n-1}$ for any other labels with binary judgments. Hence, the empirical risk $\widehat{R}'_{\mathcal{L}}(\mathcal{M})$ is equal to:

$$\mathbb{E}_{\mathbf{x}}\,\mathbb{E}_{\mathbf{y}|\mathbf{x}}\left[(1-\rho)\mathcal{L}(\mathcal{M}(\mathbf{x}),\mathbf{y})+\frac{\rho}{2^n-1}\sum_{\mathbf{y}'\in\mathcal{Y}\setminus\{\mathbf{y}\}}\mathcal{L}(\mathcal{M}(\mathbf{x}),\mathbf{y}')\right] \tag{17}$$

Since $\mathbb{E}_{\mathbf{x}}\,\mathbb{E}_{\mathbf{y}|\mathbf{x}}=\mathbb{E}_{\mathbf{x},\mathbf{y}}$ and $\mathcal{L}$ is a symmetric ranking loss, given the definition of empirical risk for true labels, we have:

$$\begin{aligned}\widehat{R}'_{\mathcal{L}}(\mathcal{M}) &= (1-\rho)\widehat{R}_{\mathcal{L}}(\mathcal{M})+\frac{\rho}{2^n-1}(c-\widehat{R}_{\mathcal{L}}(\mathcal{M}))\\ &= \frac{c\rho}{2^n-1}+(1-\frac{2^n\rho}{2^n-1})\widehat{R}_{\mathcal{L}}(\mathcal{M})\end{aligned} \tag{18}$$

Now assume that $\widehat{\mathcal{M}}^*$ is a global minimizer for the risk function $\widehat{R}'_{\mathcal{L}}$, thus for any ranking model $\mathcal{M}$, we have:

$$\widehat{R}'_{\mathcal{L}}(\widehat{\mathcal{M}}^*)-\widehat{R}'_{\mathcal{L}}(\mathcal{M})\leq 0 \tag{19}$$

On the other hand, from Equation (18) we have:

$$\widehat{R}'_{\mathcal{L}}(\widehat{\mathcal{M}}^*)-\widehat{R}'_{\mathcal{L}}(\mathcal{M})=(1-\frac{2^n\rho}{2^n-1})(\widehat{R}_{\mathcal{L}}(\widehat{\mathcal{M}}^*)-\widehat{R}_{\mathcal{L}}(\mathcal{M})) \tag{20}$$

According to Equations (19) and (20) and because $\rho<\frac{2^n-1}{2^n}$, then $\widehat{R}_{\mathcal{L}}(\widehat{\mathcal{M}}^*)-\widehat{R}_{\mathcal{L}}(\mathcal{M})\leq 0$ and thus $\widehat{\mathcal{M}}^*$ is also a global minimizer for the true empirical risk function $\widehat{R}_{\mathcal{L}}$. Therefore, risk minimization under symmetric ranking losses is noise tolerant, which completes the proof. □

This theorem shows that symmetric ranking losses are robust to uniform noise used to generate weak supervision labels. Note that the only condition is ρ being less than $\frac{2^n-1}{2^n}$. This means that the weak signal should be better than random, which is not a restrictive condition. It is also interesting that this finding is independent of the size of training data.

4.4 Weak Supervision as Non-uniform Noisy Channel

In the last subsection, we assume that the error probability of weak labeler is the same for all training instances. This means that the quality of weak supervision signal is the same for all queries, which is not a true assumption in practice, i.e., some queries are more difficult and some are easier to respond. In this subsection, we relax this assumption and find an upper bound for the empirical risk function.

Theorem 3. Let $\mathcal{L}$ be a symmetric ranking loss function (see Definition 4). For each pair $(\mathbf{x}, \widehat{\mathbf{y}}) \in \widehat{\mathcal{T}}$, assume that the weak label $\widehat{\mathbf{y}}$ is equal to the true label with a probability of $\rho_{\mathbf{x}}$ which depends on the input $\mathbf{x}$. Then, the empirical risk function $\widehat{R}_{\mathcal{L}}(\widehat{\mathcal{M}}^*)$ is upper bounded by $\widehat{R}_{\mathcal{L}}(\mathcal{M}^*)/(1-\frac{2^n\rho_{max}}{2^n-1})$, where $\widehat{\mathcal{R}}_{\mathcal{L}}$ is empirical risk on the true data, ρ_{max} is the maximum error probability, and $\widehat{\mathcal{M}}^*$ and $\mathcal{M}^*$ are the global minimizers of the weak risk $\widehat{R}'_{\mathcal{L}}$ and the true risk $\widehat{R}_{\mathcal{L}}$, respectively.

Proof. Similar to Equation (17), the empirical risk function for any ranking model $\mathcal{M}$ over the weakly labeled data $\widehat{\mathcal{T}}$ is defined as:

$$\begin{aligned}\widehat{R}'_{\mathcal{L}}(\mathcal{M}) &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[(1-\rho_{\mathbf{x}})\mathcal{L}(\mathcal{M}(\mathbf{x}),\mathbf{y}) + \frac{\rho_{\mathbf{x}}}{2^n-1}\sum_{\mathbf{y}'\in\mathcal{Y}\setminus\{\mathbf{y}\}}\mathcal{L}(\mathcal{M}(\mathbf{x}),\mathbf{y}')\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[(1-\rho_{\mathbf{x}})\mathcal{L}(\mathcal{M}(\mathbf{x}),\mathbf{y}) + \frac{\rho_{\mathbf{x}}}{2^n-1}(c-\mathcal{L}(\mathcal{M}(\mathbf{x}),\mathbf{y}))\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\frac{c\rho_{\mathbf{x}}}{2^n-1}\right] + \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[(1-\frac{2^n\rho_{\mathbf{x}}}{2^n-1})\mathcal{L}(\mathcal{M}(\mathbf{x}),\mathbf{y})\right] \quad (21)\end{aligned}$$

Note that the term $\mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\frac{c\rho_{\mathbf{x}}}{2^n-1}\right]$ is independent of the ranking model. Let $\widehat{\mathcal{M}}^*$ and $\mathcal{M}^*$ be the global minimizers of the empirical risk functions $\widehat{R}'_{\mathcal{L}}$ and $\widehat{R}_{\mathcal{L}}$, respectively. Therefore, we have:

$$\widehat{R}'_{\mathcal{L}}(\widehat{\mathcal{M}}^*) - \widehat{R}'_{\mathcal{L}}(\mathcal{M}^*) \le 0 \quad (22)$$

According to Equation (21), we can rewrite the above inequality as:

$$\begin{aligned}&\mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\left(1-\frac{2^n\rho_{\mathbf{x}}}{2^n-1}\right)\left(\mathcal{L}(\widehat{\mathcal{M}}^*(\mathbf{x}),\mathbf{y}) - \mathcal{L}(\mathcal{M}^*(\mathbf{x}),\mathbf{y})\right)\right] \le 0 \\ &\Rightarrow \min_{\rho_{\mathbf{x}}}\left\{1-\frac{2^n\rho_{\mathbf{x}}}{2^n-1}\right\}\mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\left(\mathcal{L}(\widehat{\mathcal{M}}^*(\mathbf{x}),\mathbf{y}) - \mathcal{L}(\mathcal{M}^*(\mathbf{x}),\mathbf{y})\right)\right] \le 0 \\ &\Rightarrow (1-\frac{2^n\rho_{max}}{2^n-1})\mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathcal{L}(\widehat{\mathcal{M}}^*(\mathbf{x}),\mathbf{y})] \le \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathcal{L}(\mathcal{M}^*(\mathbf{x}),\mathbf{y})] \\ &\Rightarrow \widehat{R}_{\mathcal{L}}(\widehat{\mathcal{M}}^*) \le \frac{1}{1-\frac{2^n\rho_{max}}{2^n-1}}\widehat{R}_{\mathcal{L}}(\mathcal{M}^*) \quad (23)\end{aligned}$$

Therefore, the true empirical risk for $\widehat{\mathcal{M}}^*$ is upper bounded by $\widehat{R}_{\mathcal{L}}(\mathcal{M}^*)/(1-\frac{2^n\rho_{max}}{2^n-1})$. This completes the proof. □

Theorem 3 shows that the ratio of empirical risk for the global minimizer of the weak risk to to the one for the global minimizer of the true risk is upper bounded by $\frac{1}{1-\frac{2^n\rho_{max}}{2^n-1}}$.

Remark 2. Theorem 3 shows that if the minimum empirical risk on the true labeled data is equal to 0, then the model $\widehat{\mathcal{M}}^*$ is the global minimizer of the empirical risk on the true labeled data. Therefore, if $\widehat{R}_{\mathcal{L}}(\mathcal{M}^*) = 0$, any symmetric ranking loss $\mathcal{L}$ is robust to non-uniform noise.

4.5 A Study of Pairwise Loss Functions

In this subsection, we study a number of pairwise loss functions to identify the ones that satisfy the SSPL constraint introduced by Theorem 1. Without loss of generality, assume that $\mathcal{M}(x) \in [0,1] : \forall x$ which can be obtained via a sigmoid function. With some relaxation of notation throughout this section, for a pair of objects (o, o') with feature vectors (x, x'), let $s_{o\ge o'} = \mathcal{M}(x) - \mathcal{M}(x')$ denote the score of o being ranked higher than o' by the ranking model $\mathcal{M}$. Therefore, $s_{o\ge o'} = -s_{o<o'} \in [-1,1]$. Let $y_{o\ge o'} \in \{-1,1\}$ be a pairwise label indicating whether o should be ranked higher than o' or not.

Lemma 3. In pairwise learning to rank if $\mathcal{M}(x) \in [0,1] : \forall x$, hinge loss and mean absolute error (L_1 loss) satisfy the SSPL constraint. On contrary, cross entropy loss and mean square error (L_2 loss) do not satisfy the SSPL constraint.

Proof. In the following, we study the loss functions mentioned in the lemma one by one.

- **Hinge loss**: Hinge loss, also known as the max-margin loss, is defined as follows:
$$\max\{0, 1 - y_{o\ge o'} s_{o\ge o'}\} \quad (24)$$
where $y_{o\ge o'} \in \{-1,1\}$. Given the above definition, we have:
$$\begin{aligned}&\mathcal{L}_{hinge}(s_{o\ge o'}, -1) + \mathcal{L}_{hinge}(s_{o\ge o'}, 1) \\ &= \max\{0, 1 - s_{o\ge o'}\} + \max\{0, 1 + s_{o\ge o'}\} \quad (25)\end{aligned}$$
Since $s_{o\ge o'} \in [-1,1]$, the above equation is equal to 2, and thus hinge loss satisfies the SSPL constraint.
- **Mean absolute error or L_1 loss**: Given the definition of L_1 loss, we have:
$$\begin{aligned}&\mathcal{L}_{MAE}(s_{o\ge o'}, -1) + \mathcal{L}_{MAE}(s_{o\ge o'}, 1) \\ &= |1 - s_{o\ge o'}| + |-1 - s_{o<o'}| + |-1 - s_{o\ge o'}| + |1 - s_{o<o'}| \quad (26)\end{aligned}$$
Since $s_{o\ge o'} = -s_{o<o'} \in [-1,1]$, then the above equation is equal to 4, and thus mean absolute error satisfies the SSPL constraint.
- **Cross entropy loss**: Given the definition of cross entropy loss, we have:
$$\begin{aligned}&\mathcal{L}_{ce}(s_{o\ge o'}, -1) + \mathcal{L}_{ce}(s_{o\ge o'}, 1) \\ &= \log p_{o\ge o'} + \log(1 - p_{o\ge o'}) \\ &= \log p_{o\ge o'}(1 - p_{o\ge o'}) \quad (27)\end{aligned}$$
Note that we should use the pairwise probability $p_{o\ge o'}$ for the cross entropy loss. The above equation is not equal to a constant and is not bounded. Therefore, the cross entropy loss function does not satisfy the SSPL constraint.
- **Mean square error or L_2 loss**: Given the definition of L_2 loss, we have:
$$\begin{aligned}&\mathcal{L}_{MSE}(s_{o\ge o'}, -1) + \mathcal{L}_{MSE}(s_{o\ge o'}, 1) \\ &= 2(1 - s_{o\ge o'})^2 + 2(1 + s_{o\ge o'})^2 \quad (28)\end{aligned}$$
Thus, the L_2 loss function does not satisfy the SSPL constraint. However, $\mathcal{L}_{MSE}(s_{o\ge o'}, -1) + \mathcal{L}_{MSE}(s_{o\ge o'}, 1)$ is bounded by $[4, 8]$.

□

5 LEARNING FROM MULTIPLE WEAK SIGNALS

As pointed out in Section 4.4, the expected risk upper bound for the global minimizer of the weak risk under the non-uniformity noise assumption is inversely correlated with the maximum error probability of the weak labeler. Recently, Zamani et al. [33] proposed to employ multiple weak supervision signals to improve the accuracy of weakly supervised models. In this section, we theoretically show how to guarantee a maximum arbitrary error rate to tighten the upper bound found in Section 4.4 using multiple weak supervision signals.

Theorem 4. For any object pair (o, o'), at least $\frac{\rho \ln 2}{2}\left(\frac{1-2\epsilon}{1-2\rho}\right)^2$ independent pairwise weak supervision signals, each with a pairwise noise probability $\rho < \frac{1}{2}$ are required to guarantee the pairwise error probability of less than or equal to an arbitrary $\epsilon < \frac{1}{2}$.

Proof. We prove this theorem based on binary hypothesis testing. For any given object pair (o, o') with feature vectors (x, x'), we define two hypotheses:

- **Hypothesis 1 (H_1)**: o should be ranked higher than or equal to o' (i.e., $o \geq o'$).
- **Hypothesis 2 (H_2)**: o' should be ranked higher than o (i.e., $o < o'$).

The probability mass functions for the two probability distributions P_1 and P_2 respectively corresponding to H_1 and H_2 are as follows:

$$\begin{cases} P_1(o \geq o') = 1-\rho \\ P_1(o < o') = \rho \end{cases} \qquad \begin{cases} P_2(o \geq o') = \rho \\ P_2(o < o') = 1-\rho \end{cases} \tag{29}$$

Therefore, the probability of error P_e for identifying the correct pairwise label in binary hypothesis testing is lower bounded as follows:

$$P_e \geq \frac{1}{2}\left[1 - \delta_{TV}(P_1^{(k)}, P_2^{(k)})\right] \tag{30}$$

where k is the number of weak supervision signals. Given the Pinsker's inequality (see Lemma 2), we rewrite the above inequality as below:

$$P_e \geq \frac{1}{2}\left[1 - \sqrt{\frac{2}{\ln 2} D(P_1^{(k)} || P_2^{(k)})}\right] \tag{31}$$

where $D(\cdot||\cdot)$ denotes the KL divergence between two probability distributions (see Section 3.3). Since the weak supervision signals are assumed to be independent, we use Lemma 1 to rewrite the above inequality:

$$\begin{aligned} P_e &\geq \frac{1}{2}\left[1 - \sqrt{\frac{2k}{\ln 2} D(P_1||P_2)}\right] \\ &= \frac{1}{2}\left[1 - \sqrt{\frac{2k}{\ln 2}(1-2\rho)\ln(\frac{1-\rho}{\rho})}\right] \end{aligned} \tag{32}$$

Given the inequality $1 + x \leq e^x : \forall x$, we have $\ln(\frac{1-\rho}{\rho}) = \ln(1 + \frac{1-2\rho}{\rho}) \leq \frac{1-2\rho}{\rho}$. Hence, if we want the probability of error being upper bounded by any arbitrary ϵ, we have:

$$\begin{aligned} \epsilon \geq P_e &\geq \frac{1}{2}\left[1 - \sqrt{\frac{2k}{\ln 2}\frac{(1-2\rho)^2}{\rho}}\right] \\ \Rightarrow k &\geq \frac{\rho \ln 2}{2}\left(\frac{1-2\epsilon}{1-2\rho}\right)^2 \end{aligned} \tag{33}$$

This completes the proof. □

Remark 3. Theorem 4 assumes that the error probability of all independent weak labelers are equal to ρ. If they have different error probabilities, the theorem should be slightly modified. In this case, let P_{i1} and P_{i2} respectively denote the probability distribution of the i^{th} weak signal for H_1 and H_2. Therefore, from Equation (32), we have:

$$\epsilon \geq P_e \geq \frac{1}{2}\left[1 - \sqrt{\frac{2}{\ln 2}\sum_{i=1}^{k} D(P_{i1}||P_{i2})}\right] \tag{34}$$

Thus, the minimum k that requires to satisfy the above inequality is the answer.

Remark 4. Theorem 4 assumes that the noise distribution in weak supervision signals are independent. If they are not independent, the divergence term in Equation (32) should be computed using the chain rule mentioned in Lemma 1.

6 EXPERIMENTS

In this section, we first provide the results on a synthetic noisy data, and then experiment with real weak supervision data for the ad-hoc retrieval task.

6.1 Evaluation on Synthetic Data

In this subsection, we empirically verify a set of our theoretical findings in Section 4. To do so, we create a synthetic data based on the MQ2008 dataset, which is a part of the LETOR 4.0 dataset.[3] Each training and test instance in this dataset contains 46 features, extracted via various retrieval techniques. In our experiments, we performed 5-fold cross-validation based on the queries. We trained a fully-connected feed-forward neural network on the training data. The network consists of two hidden layers with 500 and 100 neurons. Relu was used as the non-linear activation function in hidden layers, and sigmoid was used as the output activation. We trained the model using a pairwise setting; any two documents with different labels with respect to each query were considered as a pairwise training instance. The model was trained for one epoch using the Adam optimizer [19], and the learning rate was selected from $\{0.0001, 0.0005, 0.001\}$ based on the performance on the validation set. The batch size was set to 128. We evaluated the performance of the models in terms of normalized discounted cumulative gain for the top 20 documents (NDCG@20).

Uniform noise. In the first set of experiments, we applied a uniform noise on the training data. The pairwise noise probability was swept from 0.0 to 0.45. Note that the pairwise noise should be less than 0.5 which means that the weak labeler should perform

[3] https://www.microsoft.com/en-us/research/project/letor-learning-rank-information-retrieval/

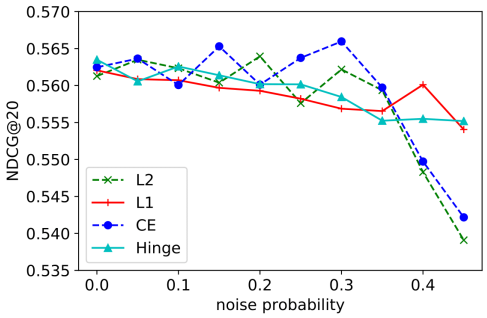

Figure 1: The retrieval performance on MQ2008 with respect to the uniform noise probability ($\rho < 0.5$).

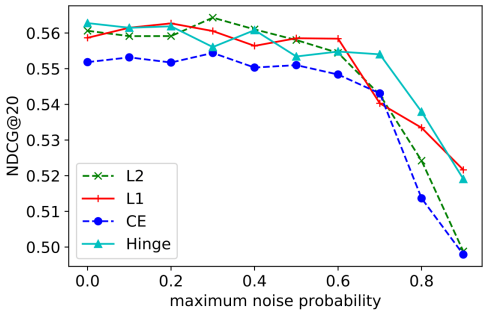

Figure 2: The retrieval performance on MQ2008 with respect to the non-uniform maximum noise probability.

better than random. We evaluated models with the same neural architecture, but different pairwise loss functions. The results are plotted in Figure 1. As depicted, performance of the models based on the L_2 loss and cross entropy (CE) significantly drop when the noise probability increases. However, the models with the L_1 loss and the hinge loss are robust to uniform noise. This observation empirically validates our theoretical findings on the robustness of symmetric ranking losses to uniform noise.

Non-uniform noise. In the next set of experiments, we applied non-uniform noise on the training data. In other words, the probability of noise varies across queries. We swept the maximum noise probability from 0.0 to 0.9. The results are plotted in Figure 2. According to this plot, the performance of all the models significantly drop when the maximum noise probability passes the 0.7 threshold. Figure 2 shows that the models with symmetric ranking losses (i.e., L_1 and hinge loss) perform substantially better than those based on the L_2 loss and cross entropy, when the maximum noise probability is high (i.e., > 0.7).

6.2 Evaluation on Weak Supervision Data

In this subsection, we focus on a real weak supervision setting for ad-hoc retrieval. To do so, we trained a fully-connected feed-forward pairwise model, exactly the same as Rank Model introduced in [12]. Network parameters were optimized using the Adam optimizer [19]. In this experiment, the learning rate and the batch size were selected from $\{5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}\}$ and $\{32, 64, 128\}$, respectively. The hidden layer sizes were selected from $\{100, 300, 500\}$. We initialized the word embedding matrix by pre-trained GloVe [25] vectors learned from Wikipedia dump 2014 plus Gigawords 5.[4] The embedding dimensionality was set to 300. All retrieval experiments were carried out using the Galago search engine [9].[5] We performed 2-fold cross-validation over the queries in each collection for hyper-parameter tuning.

We collected our training queries from AOL query logs [24]. We only used the query strings, and no session and click information was obtained from the query logs. We filtered out the navigational queries containing URL substrings, i.e., "http", "www.", ".com", ".net", ".org", ".edu". All non-alphanumeric characters were removed from the queries. As a sanity check, we made sure that no queries from the training set appear in our evaluation query sets. Applying all of these constraints leads to over 6 million unique queries as our training query set. We used query likelihood [26] with Dirichlet prior smoothing [35] as the weak supervision signal. In more detail, for each training query, we retrieved 100 documents from the target evaluation collection using the query likelihood model and created our pairwise training instances based on the query likelihood scores.

We evaluate our models using the following two TREC collections: The first collection, Robust, consists of thousands of news articles and is considered as homogeneous collections. Robust was previously used in TREC 2004 Robust Track. The second collection, ClueWeb, is a challenging and large-scale web collection containing heterogeneous and noisy documents. ClueWeb (i.e., ClueWeb09-Category B) is a common web crawl collection that only contains English web pages. ClueWeb was previously used in TREC 2009-2012 Web Track. We cleaned the ClueWeb collection by filtering out the spam documents, using the Waterloo spam scorer[6] [7] with the threshold of 60%. Stopwords were removed from all collections. For Robust, TREC topics 301-450 & 601-700, and for ClueWeb, topics 1-200 were used for the experiments.

Results. The results reported in Table 1 show that all weakly supervised models outperform the query likelihood (QL) model, which is also the weak labeler. The results demonstrate that the models with L_1 and hinge loss functions significantly outperform the models with L_2 loss or cross entropy as the loss function. The statistical differences are computed using the two-tailed paired t-test at 95% confidence interval ($p\_value < 0.05$). Recall that the L_1 loss and the hinge loss satisfy the SSPL constraint, while the L_2 loss and the cross entropy loss function do not. This is an empirical validation on real weak supervision data for our theory presented in Section 4.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we provided a theoretical study on learning to rank from inaccurate relevance signals, motivated by the recent advancements in developing weakly supervised models for information retrieval tasks. We looked at weak supervision as a noisy channel

[4] https://nlp.stanford.edu/projects/glove/
[5] https://www.lemurproject.org/galago.php
[6] http://plg.uwaterloo.ca/~gvcormac/clueweb09spam/

Table 1: Retrieval performance of weakly supervised neural ranking models (NRM) with different loss functions. The highest value per column is marked in bold, and the superscripts 0/1/2 denote statistically significant improvements compared to QL/NRM-CE/NRM-L2, respectively.

Method	Robust			ClueWeb		
	MAP	P@20	NDCG@20	MAP	P@20	NDCG@20
QL	0.2499	0.3556	0.4143	0.1044	0.3139	0.2294
NRM-CE	0.2743^{0}	0.3682^{0}	0.4272^{0}	0.1233^{0}	0.3286^{0}	0.2308^{0}
NRM-L2	0.2765^{0}	0.3696^{0}	0.4290^{0}	0.1214^{0}	0.3271^{0}	0.2315^{0}
NRM-L1	$\mathbf{0.2831}^{012}$	$\mathbf{0.3769}^{012}$	$\mathbf{0.4333}^{012}$	0.1321^{012}	$\mathbf{0.3368}^{012}$	0.2386^{012}
NRM-Hinge	0.2815^{012}	0.3752^{012}	0.4327^{012}	$\mathbf{0.1329}^{012}$	0.3351^{012}	$\mathbf{0.2392}^{012}$

that introduces some noise on the true labels. We defined symmetric ranking loss functions, and further proved that learning to rank models with symmetric loss functions are noise tolerant under uniform noise. We also found an upper bound for the risk obtained by the global minimizer of weakly supervised data, based on the risk for the true global minimizer. We also proposed a sufficient constraint for pairwise loss functions to be symmetric ranking loss. Motivated by the recent work on learning from multiple weak supervision signals, we also found a lower bound for the number of weak supervision signals required to guarantee any arbitrary maximum noise probability. We also empirically validated a set of our theoretical discoveries using synthetic and real weak supervision data. Our theoretical findings not only justify the recent empirical results obtained by the weakly supervised IR models, but also provide guidelines on how to train effective models with weak supervision.

In this paper, we only focus on studying the effectiveness of weakly supervised models. Since weak supervision requires large volumes of training data, we intend to theoretically study the efficiency of weakly supervised models in terms of training time. Reducing their training time is an interesting direction.

Acknowledgements. This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] N. Asadi, D. Metzler, T. Elsayed, and J. Lin. 2011. Pseudo Test Collections for Learning Web Search Ranking Functions. In *SIGIR '11*. 1073–1082.
[2] R. Attar and A. S. Fraenkel. 1977. Local Feedback in Full-Text Retrieval Systems. *J. ACM* 24, 3 (1977), 397–417.
[3] L. Azzopardi, M. de Rijke, and K. Balog. 2007. Building Simulated Queries for Known-item Topics: An Analysis Using Six European Languages. In *SIGIR '07*. 455–462.
[4] Alan Joseph Bekker and Jacob Goldberger. 2016. Training deep neural-networks based on unreliable labels. In *ICASSP '16*. 2682–2686.
[5] Avrim Blum and Tom Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-training. In *COLT' 98*. 92–100.
[6] Daniel Cohen and W. Bruce Croft. 2018. A Hybrid Embedding Approach to Noisy Answer Passage Retrieval. In *ECIR '18*. 127–140.
[7] Gordon V. Cormack, Mark D. Smucker, and Charles L. Clarke. 2011. Efficient and Effective Spam Filtering and Re-ranking for Large Web Datasets. *Inf. Retr.* 14, 5 (Oct. 2011), 441–465.
[8] W. B. Croft and D. J. Harper. 1979. Using Probabilistic Models of Document Retrieval Without Relevance Information. *J. Doc.* 35, 4 (1979), 285–295.
[9] W. Bruce Croft, Donald Metzler, and Trevor Strohman. 2009. *Search Engines: Information Retrieval in Practice* (1st ed.). Addison-Wesley Publishing Company.
[10] Mostafa Dehghani, Arash Mehrjou, Stephan Gouws, Jaap Kamps, and Bernhard Schölkopf. 2018. Fidelity-Weighted Learning. In *ICLR '18*.
[11] M. Dehghani, A. Severyn, S. Rothe, and J. Kamps. 2017. Avoiding Your Teacher's Mistakes: Training Neural Networks with Controlled Weak Supervision. *CoRR* abs/1711.00313 (2017).
[12] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural Ranking Models with Weak Supervision. In *SIGIR '17*. 65–74.
[13] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. 2017. Robust Loss Functions under Label Noise for Deep Neural Networks. In *AAAI '18*. 1919–1925.
[14] Aritra Ghosh, Naresh Manwani, and P. S. Sastry. 2015. Making risk minimization tolerant to label noise. *Neurocomputing* 160 (2015), 93–107.
[15] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. In *CIKM '16*. 55–64.
[16] Thorsten Joachims. 2002. Optimizing Search Engines Using Clickthrough Data. In *KDD '02*. 133–142.
[17] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-Rank with Biased Feedback. In *WSDM '17*. 781–789.
[18] Ashish Khetan, Zachary C. Lipton, and Anima Anandkumar. 2018. Learning From Noisy Singly-labeled Data. In *ICLR '18*.
[19] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR '15*.
[20] Victor Lavrenko and W. Bruce Croft. 2001. Relevance Based Language Models. In *SIGIR '01*. 120–127.
[21] Gideon S. Mann and Andrew McCallum. 2010. Generalized Expectation Criteria for Semi-Supervised Learning with Weakly Labeled Data. *J. Mach. Learn. Res.* 11 (2010), 955–984.
[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS '13*. 3111–3119.
[23] Yifan Nie, Alessandro Sordoni, and Jian-Yun Nie. 2018. Multi-level Abstraction Convolutional Model with Weak Supervision for Information Retrieval. In *SIGIR '18*. 985–988.
[24] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A Picture of Search. In *InfoScale '06*.
[25] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP '14*. 1532–1543.
[26] Jay M. Ponte and W. Bruce Croft. 1998. A Language Modeling Approach to Information Retrieval. In *SIGIR '98*. 275–281.
[27] David Rolnick, Andreas Veit, Serge J. Belongie, and Nir Shavit. 2017. Deep Learning is Robust to Massive Label Noise. *CoRR* abs/1705.10694 (2017).
[28] Robert E. Schapire and Yoav Freund. 2012. *Boosting: Foundations and Algorithms*. The MIT Press.
[29] Nikos Voskarides, Edgar Meij, Ridho Reinanda, Abhinav Khaitan, Miles Osborne, Giorgio Stefanoni, Prabhanjan Kambadur, and Maarten de Rijke. 2018. Weakly-supervised Contextualization of Knowledge Graph Facts. In *SIGIR '18*. 765–774.
[30] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. In *SIGIR '17*. 55–64.
[31] Hamed Zamani, Michael Bendersky, Xuanhui Wang, and Mingyang Zhang. 2017. Situational Context for Ranking in Personal Search. In *WWW '17*. 1531–1540.
[32] Hamed Zamani and W. Bruce Croft. 2017. Relevance-based Word Embedding. In *SIGIR '17*. 505–514.
[33] Hamed Zamani, W. Bruce Croft, and J. Shane Culpepper. 2018. Neural Query Performance Prediction Using Weak Supervision from Multiple Signals. In *SIGIR '18*. 105–114.
[34] Hamed Zamani, Mostafa Dehghani, Fernando Diaz, Hang Li, and Nick Craswell. 2018. SIGIR 2018 Workshop on Learning from Limited or Noisy Data for Information Retrieval. In *SIGIR '18*. 1439–1440.
[35] Chengxiang Zhai and John Lafferty. 2001. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *SIGIR '01*. 334–342.